SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 1–13

A REVIEW OF AN OPTIMAL DESIGN PROBLEM FOR A PLATE OF VARIABLE THICKNESS*

JULIO MUÑOZ† AND PABLO PEDREGAL‡

Abstract. We revisit a classic design problem for a plate of variable thickness under the model of Kirchhoff. Our main contribution has two goals. One is to provide a rather general existence result under a main assumption on the structure of the tensor of material constants. The other focuses on providing a minimal number of additional design variables for a relaxation of the problem when that assumption on the tensor of elastic constants does not hold. In both situations, the cost functional can be pretty general.

Key words. optimal design, direct method, existence, relaxation

AMS subject classifications. 49J25, 49J20, 35J50, 74P04

DOI. 10.1137/050639569

1. Introduction.

The problem of the optimal design of a plate of variable thickness under Kirchhoff's model can be stated as finding the optimal, symmetric profile

$$h : \Omega \subset \mathbf{R}^2 \to \mathbf{R},$$

where Ω is supposed to be the midplane with respect to which the plate is symmetric, so that it minimizes the value of the compliance functional

$$I(h) = \int_\Omega f(x)u(x)\,dx,$$

where f is the vertical load over the plate, and u is the vertical displacement in equilibrium which is obtained from the profile h by solving the equation of equilibrium

$$\sum_{i,j,k,l} \frac{\partial^2}{\partial x_i \partial x_j}\left(h^3(x)\, M_{ijkl} \frac{\partial^2 u(x)}{\partial x_k \partial x_l}\right) = f(x)$$

in Ω, supplemented with clamped boundary conditions around $\partial\Omega$ by demanding $u = \nabla u = 0$ over $\partial\Omega$. Here the fourth-order tensor M encloses the various material constants for the type of elastic material the plate is made of. In addition, there should be some other constraints on the admissible profiles so that the problem is meaningful. On the one hand, we assume that there is a minimum and a maximum height for h so that

$$0 < h_- \le h(x) \le h_+$$

*Received by the editors September 5, 2005; accepted for publication (in revised form) August 12, 2006; published electronically February 23, 2007. This work was supported by project MTM2004-07114 from Ministerio de Educación y Ciencia (Spain).

http://www.siam.org/journals/sicon/46-1/63956.html

†Departamento de Matemáticas, Facultad de Medio Ambiente, Universidad de Castilla-La Mancha, 071 Toledo, Spain (julio.munoz@uclm.es). The work of this author was supported by project PBC-05-010-1 from JCCM (Castilla-La Mancha).

‡Departamento de Matemáticas, ETSI Industriales, Universidad de Castilla-La Mancha, 13071 Ciudad Real, Spain (pablo.pedregal@uclm.es). The work of this author was supported by project PAI05-029 from JCCM (Castilla-La Mancha).

and h_-, h_+ are given parameters. On the other hand, we must limit the amount of material that can be used so that

$$\int_\Omega h(x)\,dx \le V\,|\Omega|$$

and $h_- < V < h_+$.

This problem has received some attention over the years in two different directions. First, it was noticed long ago that, at least in some situations, this problem is not well-posed in the sense that there might not exist optimal profiles (see [6], [7]). Today, this is a well-understood fact. It is typically associated with some lack of convexity, often taken in a suitable broad sense. This direction was further pursued and explored from the horizon of finding a minimal relaxation in the sense of using a minimal number of generalized design variables (see [5], [9]). Several later works emphasized this perspective and proved various types of results always trying to minimize in various ways the number of design variables needed to describe minimizing profiles. In many of these contributions, Young measures associated with minimizing profiles were used in one way or another (see [1], [3], [4], [11], [15]). Second, in some other situations, existence of optimal profiles has been shown despite the fact of the just-mentioned difficulties (see [14], [16], [17]), coming to a situation where it is not completely clear when, depending on the ingredients, one can trust existence results or else anticipate highly oscillating optimal profiles. Another point in many of these works is that the only cost functional considered is the compliance written before, along with some other variants of order zero (no derivatives of u).

The aim of our contribution here is twofold. First, we examine the structural ingredients of the problem that enable an existence result, and how existence of optimal profiles is compromised when such requirements are not fulfilled. As we will see, this is basically related to the structure of the tensor M of material constants so that the existence of optimal profiles for this problem depends (for many relevant cost functionals) on the elasticity properties of the material we use to manufacture the plate. Second, we would like to be able to examine more general cost functionals and not just the compliance. We will give results for much more general objective functionals in both existence as well as nonexistence cases.

Let $F(x,u,\lambda,\xi,h) : \Omega\times\mathbf{R}\times\mathbf{R}^2\times\mathbf{M}^{2\times 2}\times\mathbf{R}\to\mathbf{R}$ be a given integrand, continuous in the variables (u,λ,ξ,h) and measurable in x (here $\mathbf{M}^{2\times 2}$ is the space of the 2×2 real matrices). Define

$$I(h) = \int_\Omega F(x,u(x),\nabla u(x),\nabla^2 u(x),h(x))\,dx,$$

where u solves

$$\sum_{i,j,k,l}\frac{\partial^2}{\partial x_i\partial x_j}\left(h^3(x)M_{ijkl}\frac{\partial^2 u(x)}{\partial x_k\partial x_l}\right) = f(x) \text{ in } \Omega,$$

$$u(x) = \frac{\partial u(x)}{\partial n} = 0 \text{ on } \partial\Omega,$$

Specifically, we consider the following optimal design problem:

$$\begin{aligned} &\text{Minimize}\quad I(h)\\ &\text{subject to}\quad h_-\le h(x)\le h_+ \text{ in } \Omega,\quad \int_\Omega h(x)\,dx \le V\,|\Omega|\,. \end{aligned}$$

The main structural assumption to distinguish between existence and nonexistence of optimal solutions for this optimal design problem refers to the material tensor M. We will say that M is decomposable if

$$M = M_1 \otimes M_2,$$

where M_i are positive definite, second-order tensors (matrices). Notice how in this case the equilibrium equation basically reduces to the biharmonic operator. In this situation, we have a general existence result.

THEOREM 1. *Suppose that $M = M_1 \otimes M_2$, i.e., M is decomposable, and the integrand F in the cost functional I is such that the functions*

$$\xi \in \left\{ \mathbf{M}^{2\times 2} : \frac{c}{M_2 \cdot \xi} > 0 \right\} \mapsto F\left(x, u, \lambda, \xi, \frac{c}{(M_2 \cdot \xi)^{1/3}} \right)$$

and

$$(\xi, z) \in \left\{ \mathbf{M}^{2\times 2} : M_2 \cdot \xi = 0 \right\} \times \mathbf{R} \mapsto \min_{h \in [h_-, z] \cap Q} F(x, u, \lambda, \xi, h)$$

are convex for any constant c and fixed (x, u, λ). Then there are optimal profiles for the associated optimal design problem for the plate.

A corollary worth stating covers many situations of interest.

COROLLARY 2. *Suppose that the integrand F does not depend on ξ and h, and M is decomposable. Then for any such F (even nonconvex), the corresponding optimal design problem admits optimal solutions.*

Explicit cases like the compliance $F = f(x)u(x)$ are covered with this corollary. But also examples like $F = g(x)u(x)$, $F = |\nabla u(x)|^2$, etc., can be treated through this result as well.

When the tensor M is not decomposable, the situation is drastically different. In many cases, this fact is responsible for the lack of optimal solutions and the analysis is much more complex. See the references cited above. In the particular situation where we assume that the profile h is a function of x_1 alone, so that $h(x) = h(x_1)$, and the tensor M is that of an orthotropic material, a relaxed formulation of the problem can be pursued. It has been a principal goal over the years to find a minimal full relaxation of this problem, that is, one which requires the least number of additional design variables. For the compliance functional, the best result we know of has been obtained in [4] (also in [15] within a more general framework). Here, by revisiting some of our own old ideas [11], we are able to show that this same result holds true even for much more general functionals. Recall that M for orthotropic materials is defined in terms of two main material parameters: Young's modulus E, and Poisson's ratio r, so that the nonvanishing components of M are

$$M_{1111} = M_{2222} = \frac{2}{3}\frac{E}{1-r^2}, \quad M_{1122} = M_{2211} = \frac{2}{3}\frac{Er}{1-r^2},$$
$$M_{1212} = M_{1221} = M_{2112} = M_{2121} = \frac{1}{3}\frac{E}{1+r}.$$

THEOREM 3. *Let admissible profiles depend only on x_1, M corresponding to an orthotropic material, and let the integrand for the cost functional*

$$F(x, u, \lambda) : \Omega \times \mathbf{R} \times \mathbf{R}^2 \to \mathbf{R}$$

be measurable in $x \in \Omega$ and continuous (not necessarily convex) in the pairs (u, λ). Consider the optimal design problem

$$\text{Minimize in } (\theta, h): \quad J(\theta, h) = \int_\Omega F(x, u(x), \nabla u(x))\, dx$$

subject to

$$\theta \in [0,1], \quad h \in [h_-, h_+],$$
$$\int_\Omega \left[\theta(x) h_+ + (1-\theta(x)) h(x)\right] dx \leq V\,|\Omega|\,,$$

and where u solves

$$\sum_{i,j,k,l} \frac{\partial^2}{\partial x_i \partial x_j}\left(\overline{M}_{ijkl} \frac{\partial^2 u}{\partial x_k \partial x_l}\right) = f \text{ in } \Omega,$$
$$u = \frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega,$$

and the nonvanishing components of $\overline{M}$ depend on the design variables through formulae

$$\overline{M}_{1111} = \frac{2}{3} c(x) \frac{E}{1-r^2}, \quad \overline{M}_{2222} = \frac{2}{3} m(x) E + \frac{2}{3} c(x) \frac{E r^2}{1-r^2},$$
$$\overline{M}_{1122} = \overline{M}_{2211} = \frac{2}{3} c(x) \frac{E r}{1-r^2},$$
$$\overline{M}_{1212} = \overline{M}_{1221} = \overline{M}_{2112} = \overline{M}_{2121} = \frac{1}{3} m(x) \frac{E}{1+r},$$
$$m(x) = \theta(x) h_+^3 + (1-\theta(x)) h^3(x),$$
$$c(x)^{-1} = \theta(x) h_+^{-3} + (1-\theta(x)) h(x)^{-3}.$$

This problem is a full relaxation of the initial optimal design problem in the sense

$$\inf_h I(h) = \min_{(\theta,h)} J(\theta, h).$$

The relevance of this result is in the fact that only one additional design variable, θ (a certain weight), is required to produce a full relaxation of the problem, and this is so for a rather huge class of cost functionals and not only for the compliance. We will later provide further details as to how one is to interpret these pairs (θ, h) in terms of sequences of profiles for the original problem.

This work includes another three sections. The second one contains the full proof of Theorem 1 as well as some observations on some explicit examples. Section 3 covers a brief, elementary discussion on the structure of the material tensor M. Finally, the last section is devoted to the proof of Theorem 3. We will also dwell on the interpretation of the proposed relaxed formulation in terms of the ingredients of the original optimal design problem.

2. Existence results. It is our aim to study a type of design problem for plates whose state equation has the format

$$\operatorname{div}\left(\operatorname{div}\left(h^3(x)\,(M_1 \otimes M_2)\,\nabla^2 u\,(x)\right)\right) = f\,(x) \text{ in } \Omega. \tag{1}$$

It is assumed that $f \in L^2(\Omega)$ is the applied vertical load, Ω is a smooth bounded domain in $\mathbf{R}^2$ that represents the midplane of the plate, $h \in L^\infty(\Omega)$ is the design variable, and the tensors M_i are assumed to be positive and symmetric. To this equation we add the boundary conditions

$$u(x) = \frac{\partial u(x)}{\partial n} = 0 \text{ on } \partial\Omega \tag{2}$$

(clamped plate). We assume further natural constraints on the feasible designs by limiting the height of the thicknesses and the amount of material: the set $\mathcal{H}$ of admissible designs is defined as

$$\mathcal{H} = \left\{ h \in L^\infty(\Omega) : h(x) \in [h_-, h_+] \doteq Q \text{ a.e. } x \in \Omega,\ V(h) \doteq \int_\Omega h(x)dx \le V \right\} \tag{3}$$

(V, $0 < h_- < h_+$ are given positive constants). Associated with this state equation, we consider the general optimization problem

$$\min_{h\in\mathcal{H}} \left\{ L(h) \doteq \int_\Omega F\left(x, u, \nabla u(x), \nabla^2 u(x), h(x)\right) dx \right\}, \tag{4}$$

where u solves (1)–(2) and F is a given integrand such that

$$F : (x, u, \lambda, \xi, h) \in \Omega \times \mathbf{R} \times \mathbf{R}^2 \times \mathbf{M}^{2\times 2} \times Q \to \overline{\mathbf{R}} = \mathbf{R} \cup \{+\infty\}.$$

F is measurable in x and continuous in (u, λ, ξ, h).

Our goal is the optimization problem that consists of looking for an admissible h and the corresponding displacement u, the only weak solution of problem (1)–(2) in the Sobolev space $H_0^2(\Omega)$ (the subspace of $H^2(\Omega)$ under the constraints (2)), which minimizes the objective functional L defined in (4). We denote this problem by $(\mathcal{P})$.

For the study of the above minimization problem, we shall consider a new equivalent variational problem. The underlying idea is to use the differential expression (1) in order to define a new objective functional subject to a set of constraints which are easier to deal with. The construction of this equivalent problem is performed in an elementary way [13]:

1. We introduce an auxiliary function u_0: the solution of the elliptic problem

$$-\operatorname{div}(M_1 \nabla u_0) = f, \quad u_0 \in H_0^1(\Omega). \tag{5}$$

2. Equations (1) and (5) give

$$\operatorname{div}\operatorname{div}(h^3 (M_1 \otimes M_2) \nabla^2 u) + \operatorname{div}(M_1 \nabla u_0) = 0,$$

which is equivalent to writing

$$\operatorname{div}\left(M_1 \nabla (h^3 [\operatorname{div}(M_2 \nabla u)]) + M_1 \nabla u_0\right) = 0 \tag{6}$$

or even

$$\operatorname{div}\left(M_1 \nabla (h^3 \left[M_2 \cdot \nabla^2 u\right]) + M_1 \nabla u_0\right) = 0, \tag{7}$$

i.e.,

$$\operatorname{div}(M_1 \nabla v) = 0, \tag{8}$$

where

$$v = h^3 \operatorname{div}(M_2 \nabla u) + u_0 = h^3 M_2 \cdot \nabla^2 u + u_0. \tag{9}$$

3. The new optimization problem, denoted by $(\mathcal{EP})$, is described as follows. The new integrand for the cost functional is

$$\varphi(x,u,\lambda,\xi,v,z)=\min_{\widetilde{h}\in Q}\left\{F\left(x,u,\lambda,\xi,\widetilde{h}\right):v=\widetilde{h}^3\left(M_2\cdot\xi\right)+u_0\left(x\right),z\geq\widetilde{h}\right\},$$

understood as taking the value $+\infty$ whenever the set of admissible h's is empty, and the objective functional to be minimized in the variables (u,v,z) is

$$J(u,v,z)=\int_\Omega\varphi\left(x,u\left(x\right),\nabla u\left(x\right),\nabla^2u\left(x\right),v\left(x\right),z\left(x\right)\right)dx,$$

under the constraints

$$u\in H_0^2\left(\Omega\right),\ v\in H^1\left(\Omega\right),\ \operatorname{div}(M_1\nabla v)=0,\ z\in L^\infty\left(\Omega\right),\ \int_\Omega z(x)dx=V.$$

PROPOSITION 4. *The two optimization problems $(\mathcal{P})$ and $(\mathcal{EP})$ are equivalent in the following sense: for any admissible pair (h,u)[1] for $(\mathcal{P})$ there is a triplet (u,v,z) admissible for $(\mathcal{EP})$ such that*

$$L(h)\geq J(u,v,z).$$

Conversely, for any admissible triplet (u,v,z) for $(\mathcal{EP})$, there is an admissible pair (h,u) for $(\mathcal{P})$ and

$$L(h)=J(u,v,z).$$

In particular, if (u,v,z) is optimal for $(\mathcal{EP})$, then

$$h(x)=\left(\frac{v\left(x\right)-u_0\left(x\right)}{M_2\cdot\nabla^2u\left(x\right)}\right)^{\frac{1}{3}}$$

whenever $M_2\cdot\nabla^2u\left(x\right)\neq0$, and

$$h(x)=\arg\min_{\widetilde{h}\in Q}\left\{F\left(x,u(x),\nabla u\left(x\right),\nabla^2u\left(x\right),\widetilde{h}\right):z\left(x\right)\geq\widetilde{h}\right\}$$

otherwise, is an optimal profile for $(\mathcal{P})$.

Proof. The proof is almost straightforward. We include some details for the convenience of the reader.

Let (h,u) be admissible for $(\mathcal{P})$, so that problem (1)–(2) holds. We consider u_0 (solution of (5)), and

$$v\left(x\right)=h^3\left(x\right)\left(M_2\cdot\nabla^2u\left(x\right)\right)+u_0\left(x\right).$$

By following the construction explained above v solves (8) and the classical regularity results on elliptic systems ensure that v is in $H^2\left(\Omega\right)$. We select z verifying

$$z\left(x\right)\geq h\left(x\right),\quad z(x)\in Q,\quad\text{and}\quad\int_\Omega z(x)dx=V. \tag{10}$$

[1]Here (h,u) is said to be admissible in the sense that for any $h\in\mathcal{H}$ we find the only solution u of problem (1)–(2).

Then for any $x \in \Omega$,

$$\begin{aligned}
&\varphi\left(x, u\left(x\right), \nabla u\left(x\right), \nabla^2 u\left(x\right), v\left(x\right), z\left(x\right)\right) \\
&= \min_{\widetilde{h} \in Q} \left\{ F\left(x, u, \nabla u\left(x\right), \nabla^2 u\left(x\right), \widetilde{h}\right) : v\left(x\right) = \widetilde{h}^3 \left(M_2 \cdot \nabla^2 u\left(x\right)\right) + u_0\left(x\right), z\left(x\right) \geq \widetilde{h} \right\} \\
&\leq F\left(x, u, \nabla u\left(x\right), \nabla^2 u\left(x\right), h(x)\right).
\end{aligned}$$

It is clear that our triplet, (u, v, z), is admissible for $(\mathcal{EP})$ and $J(u, v, z) \leq L(h)$.

Let (u, v, z) now be admissible for $(\mathcal{EP})$. The multifunction H given by

$$\arg\min_{\widetilde{h} \in Q} \left\{ F\left(x, u(x), \nabla u\left(x\right), \nabla^2 u\left(x\right), \widetilde{h}\right) : v\left(x\right) = \widetilde{h}^3 \left(M_2 \cdot \nabla^2 u\left(x\right)\right) + u_0\left(x\right) z\left(x\right) \geq \widetilde{h} \right\}$$

is measurable and takes closed set values. Then H admits a measurable selection (see [10, Thm. 2.23]) and we can select a measurable function h such that $h(x) \in Q$, and for any $x \in \Omega$

$$\begin{aligned}
&\varphi\left(x, u\left(x\right), \nabla u\left(x\right), \nabla^2 u\left(x\right), v\left(x\right), z\left(x\right)\right) \\
&= \min_{\widetilde{h} \in Q} \left\{ F\left(x, u\left(x\right), \nabla u\left(x\right), \nabla^2 u\left(x\right), \widetilde{h}\right) : v\left(x\right) = \widetilde{h}^3 \left(M_2 \cdot \nabla^2 u\left(x\right)\right) + u_0\left(x\right), z\left(x\right) \geq \widetilde{h} \right\} \\
&= F\left(x, u\left(x\right), \nabla u\left(x\right), \nabla^2 u\left(x\right), h\left(x\right)\right).
\end{aligned}$$

Moreover, by definition we have $v\left(x\right) = h^3\left(x\right)\left(M_2 \cdot \nabla^2 u\left(x\right)\right) + u_0\left(x\right)$, $z\left(x\right) \geq h\left(x\right) \in Q$. This is enough to fulfill the state equation

$$\operatorname{div}\operatorname{div}(h^3 \left(M_1 \otimes M_2\right) \nabla^2 u) = f\left(x\right),$$

the bound on the volume

$$\int_\Omega h(x)dx \leq \int_\Omega z(x)dx = V,$$

and the equality $I(u, v, z) = L(h, u)$. □

We can now establish the existence of optimal solutions for $(\mathcal{EP})$.

THEOREM 5. *Assume that the two functions*

$$\xi \in \left\{ \mathbf{M}^{2\times 2} : \frac{c}{M_2 \cdot \xi} > 0 \right\} \mapsto F\left(x, u, \lambda, \xi, \frac{c}{\left(M_2 \cdot \xi\right)^{1/3}}\right)$$

and

$$(\xi, z) \in \left\{ \mathbf{M}^{2\times 2} : M_2 \cdot \xi = 0 \right\} \times \mathbf{R} \mapsto \min_{h \in [h_-, z] \cap Q} F(x, u, \lambda, \xi, h)$$

are convex for any constant c and fixed (x, u, λ). Problem $(\mathcal{EP})$,

$$\inf_{(u,v,z)} J(u, v, z) \doteq \int_\Omega \varphi\left(x, u\left(x\right), \nabla u\left(x\right), \nabla^2 u\left(x\right), v\left(x\right), z\left(x\right)\right) dx,$$

where

$$\varphi\left(x, u, \lambda, \xi, v, z\right) = \min_{h \in Q} \left\{ F\left(x, u, \lambda, \xi, h\right) : v = h^3 \left(M_2 \cdot \xi\right) + u_0\left(x\right), z \geq h \right\}$$

under the restrictions

$$u \in H_0^2(\Omega),\ v \in H^1(\Omega),\ \operatorname{div}(M_1 \nabla v) = 0,\ z \in L^\infty(\Omega),\ \int_\Omega z(x)dx = V,$$

has optimal solutions.

Proof. Let (u_j, v_j, z_j) be a minimizing sequence for $(\mathcal{EP})$. As we have seen in the proof of Proposition 4 we can build the corresponding sequence h_j such that

$$\operatorname{div}\operatorname{div}(h_j^3 (M_1 \otimes M_2) \nabla^2 u_j) = f(x),\ u_j \in H_0^2(\Omega).$$

Then we can ensure that u_j is uniformly bounded in $H^2(\Omega)$. This sequence converges to u weakly in $H_0^2(\Omega)$ and, consequently, u_j and ∇u_j converge strongly in $L^2(\Omega)$ to u and ∇u, respectively. On the other hand, by elliptic theory, $\operatorname{div}(M_1 \nabla v_j) = 0$ implies v_j converges almost everywhere to a function $v \in H^1(\Omega)$ verifying the same elliptic equation (see [16] for a very neat proof in the case of the Laplacian), so that v_j converges strongly to v. Finally, notice that because $\varphi(x, u, \lambda, \xi, v, z) = \varphi(x, u, \lambda, \xi, v, h_+)$ if $z_j \geq h_+$, we may assume without lost of generality that $h_j \leq z_j \leq h_+$. Then z_j converges to some z in $L^\infty(\Omega)$ weak-$\star$, and this limit must satisfy $\int_\Omega z(x)dx = V$.

On the basis of these remarks, it remains to prove that φ is jointly convex in (ξ, z) for fixed (x, u, λ, v). To do that, it is enlightening to rewrite φ as

$$\begin{cases} F\left(x, u, \lambda, \xi, \left(\frac{v-u_0(x)}{M_2 \cdot \xi}\right)^{1/3}\right), & M_2 \cdot \xi \neq 0,\ z \geq \left(\frac{v-u_0(x)}{M_2 \cdot \xi}\right)^{1/3} \in Q, \\ \min_{h \in [h_-, z] \cap Q} F(x, u, \lambda, \xi, h), & M_2 \cdot \xi = 0,\ v = u_0(x), \\ +\infty & \text{else} \end{cases}$$

and discuss the convexity by considering two main cases:

1. $v \neq u_0(x)$: in this case φ is given by

$$\begin{cases} F\left(x, u, \lambda, \xi, \left(\frac{v-u_0(x)}{M_2 \cdot \xi}\right)^{1/3}\right), & M_2 \cdot \xi \neq 0,\ z \geq \left(\frac{v-u_0(x)}{M_2 \cdot \xi}\right)^{1/3} \in Q, \\ +\infty & \text{else.} \end{cases}$$

   This is a convex function of (ξ, z), as the set where it is finite is convex, and, by hypothesis, the function on such a set is also convex. Checking this is elementary but a bit tedious.
2. $v = u_0(x)$: in this situation we have

$$\varphi(x, u, \lambda, \xi, v, z) = \begin{cases} \min_{h \in [h_-, z] \cap Q} F(x, u, \lambda, \xi, h), & M_2 \cdot \xi = 0, \\ +\infty & \text{else.} \end{cases}$$

   This is again convex by our main structural assumption on F. ∎

The proof of Theorem 1 is a direct consequence of Theorem 5 and Proposition 4.

The generality of the cost functional permits us to associate with the state equation a huge class of optimization problems. We give some examples of such densities: the compliance case $F = f(x)\, u(x)$ or some other typical densities like $F = g(x)u(x)$, $F = |\nabla u(x)|^2$, or the identification-type problem $F = |u(x) - u_d(x)|^2 + |\nabla u(x) - \nabla u_d(x)|^2$, where $u_d \in H^1(\Omega)$ is the observed deflection of the plate. Also, $F = F_1(x, u, \nabla u)$, where F_1 is continuous on the last two variables (but not necessarily

convex) or $F = F_2\left(x, u, \nabla u, \nabla^2 u\right)$, where F_2 is continuous on the last three variables and only convex in the $\nabla^2 u$ variable, are densities for which the existence is ensured. Even with $F = F_1\left(x, u, \nabla u\right) + F_3(h)$ or $F = F_2\left(x, u, \nabla u, \nabla^2 u\right) + F_3(h)$, where F_3 is any convex and nondecreasing function (F_1 and F_2 as above), the existence of optimal classical minimizers is guaranteed.

3. The structure of the tensor M. We have seen so far that the possibility of decomposing the tensor M as the tensor product of two matrices is the crucial ingredient for having existence of optimal profiles. This will be so for special types of materials. In many sources from mechanics this is assumed as part of the model. See [2], [8]. Indeed, the equilibrium equation for the plate is often taken as

$$D\Delta^2 u = f,$$

where the coefficient D is the so-called flexural rigidity or the bending stiffness given by

$$D = \frac{Eh^3}{12(1-r^2)},$$

where h is the (constant) thickness of the plate, and E and r are, as before, Young's modulus and Poisson's ratio. When h is nonconstant, then the equation of equilibrium must be written in the form

$$\overline{D}\Delta(h^3\Delta) = f.$$

This time

$$\overline{D} = \frac{E}{12(1-r^2)}.$$

Within this sort of model, the tensor M is clearly decomposable with M_1 and M_2 multiples of the identity. In these cases, we can apply Theorem 1 to ensure the existence of optimal profiles.

The case of orthotropic materials is, however, very different. In fact, an orthotropic tensor is not decomposable.

PROPOSITION 6. *An orthotropic tensor is never decomposable.*

The proof is elementary and well known to specialists. Indeed, by writing a fourth-order tensor as a 4×4 matrix in an organized way, we realize that the matrix corresponding to a orthotropic material is of the form

$$\frac{4}{9}\frac{E^4}{(1-r^2)(1+r)^2}\begin{pmatrix} 1 & 0 & 0 & r \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ r & 0 & 0 & 1 \end{pmatrix}.$$

If such an M were decomposable, this matrix would have to be a rank-one matrix, which is easily seen not to be the case.

4. Design with a nondecomposable tensor. We investigate in this section the same design problem for the plate under the assumption that the tensor of elastic constants is not decomposable so that Theorem 1 is not applicable. Indeed, it is well known, as indicated in the introduction, that in this situation nonexistence of an

optimal profile may result, as the creation of highly oscillatory stiffeners may favor the overall rigidity of the plate.

As before, the goal is to choose the half-thickness h and its corresponding deflection u, which minimizes an integral functional $L = L(h, u)$ given by

$$L(h,u) = \int_\Omega F(x, u(x), \nabla u(x))\, dx, \tag{11}$$

where F is assumed to be measurable in $x = (x_1, x_2)$ and continuous on the variables u and ∇u.

The class of materials is restricted by imposing an orthotropic condition, namely, the nonzero components of M_{ijkl} are

$$\begin{aligned} M_{1111} &= M_{2222} = \frac{2}{3}\frac{E}{1-r^2}, \\ M_{1122} &= M_{2211} = \frac{2}{3}\frac{Er}{1-r^2}, \\ M_{1212} &= M_{1221} = M_{2112} = M_{2121} = \frac{E}{3(1+r)}, \end{aligned}$$

where r and E stand for the Poisson ratio and the Young modulus, respectively. By our comments in the preceding section, this tensor is not decomposable.

We analyze this problem under the simplification that the thickness depends just on one variable $h(x) = h(x_1)$ for any x_1 in the interval

$$(a,b) \doteq \{x_1 : \text{ there exists } x_2 \in \mathbf{R} \text{ such that } (x_1, x_2) \in \Omega\}.$$

The design criterion is to minimize L among all the plates whose half-thickness h satisfies all the constraint indicated above. In other words, the aim is to solve

$$\min_{h \in \mathcal{H}} L, \tag{12}$$

where

$$\mathcal{H} = \left\{ h \in L^\infty(a,b) : h_- \le h(x_1) \le h_+ \text{ a.e. } x_1 \in (a,b),\ \int_\Omega h(x_1)\, dx_1 dx_2 \le V \right\}. \tag{13}$$

Here h_-, h_+, and V are as before.

As indicated before, it is widely recognized that the principle described in (12)–(13) may have no solution. At least, Theorem 1 cannot be applied. This fact suggests performing a relaxation of the design problem to understand the nature of minimizing sequences of profiles. This entails defining a new admissibility set $\overline{\mathcal{H}}$ containing $\mathcal{H}$, and an extension $\overline{L}$ of L such that

$$\inf_{\mathcal{H}} L = \min_{\overline{\mathcal{H}}} \overline{L}. \tag{14}$$

It is interesting to notice that by introducing the relaxation $\min_{\overline{\mathcal{H}}} \overline{L}$, we are considering a problem whose solutions provide information about minimizing sequences of (12). However, it is important to look for the (full) relaxation, which introduces a minimal number of additional design variables. Ideally, just one more variable would

be optimal. For the compliance functional, this was shown to be the case in [4] by making use of optimality conditions. We will prove that this is the case for many more cost functionals by revisiting some our our previous ideas on this problem [11].

The starting point for a relaxation is the lemma by Murat [12] and Tartar [18] related to H-convergence. It explains why the cubic-average and the harmonic cubic-average play an important role in the relaxation for this problem. This lemma is only valid under our assumption of profiles depending only on x_1. The reader can consult [5] for a detailed proof of this result.

LEMMA 7. *Let $\left\{M^{(r)}\right\}$ be a sequence of orthotropic tensors bounded uniformly by (d, D), i.e.,*

$$d\left|t\right|^2 \leq \sum_{i,j,k,l} M^{(r)}_{ijkl} t_{ij} t_{kl}, \quad \left|\sum_{ij} M^{(r)}_{ijkl} t_{ij}\right| \leq D\left|t\right| \text{ for every } k, l.$$

Suppose that $M^{(r)} = M^{(r)}(x_1)$ and

$$\begin{gathered}
\left(M^{(r)}_{1111}\right)^{-1} \overset{*}{\rightharpoonup} \left(M^{(\infty)}_{1111}\right)^{-1}, \\
\left(M^{(r)}_{1122}\right)\left(M^{(r)}_{1111}\right)^{-1} \overset{*}{\rightharpoonup} \left(M^{(\infty)}_{1122}\right)\left(M^{(\infty)}_{1111}\right)^{-1}, \\
\left(M^{(r)}_{2222}\right) - \left(M^{(r)}_{1122}\right)^2\left(M^{(r)}_{1111}\right)^{-1} \overset{*}{\rightharpoonup} \left(M^{(\infty)}_{2222}\right) - \left(M^{(\infty)}_{1122}\right)^2\left(M^{(\infty)}_{1111}\right)^{-1}, \\
M^{(r)}_{1212} \overset{*}{\rightharpoonup} M^{(\infty)}_{1212}.
\end{gathered}$$

If $u^{(r)}$ are the solutions of the equilibrium equation for the clamped plate with tensor $M^{(r)}$, and $u^{(\infty)}$ is the solution corresponding to $M^{(\infty)}$, then $u^{(r)} \rightharpoonup u^{(\infty)}$ in $H^2_0(\Omega)$.

Because of the structure of the components of an orthotropic tensor, it is elementary to check that for a given sequence of designs $\{h_j\}$, if we define (in a unique way) the pair (h, θ) by putting

$$\begin{aligned}
h_j^3 &\overset{*}{\rightharpoonup} \theta h_+^3 + (1-\theta)h^3, \\
h_j^{-3} &\overset{*}{\rightharpoonup} \theta h_+^{-3} + (1-\theta)h^{-3}
\end{aligned} \tag{15}$$

for $\theta \in [0,1]$, $h \in [h_-, h_+]$, and

$$\begin{aligned}
\overline{M}_{1111} &= \frac{2}{3} c \frac{E}{(1-r^2)}, \\
\overline{M}_{2222} &= \frac{2}{3} mE + \frac{2}{3} c \frac{Er^2}{1-r^2}, \\
\overline{M}_{1122} &= \overline{M}_{2211} = \frac{2}{3} c \frac{Er}{1-r^2}, \\
\overline{M}_{1212} &= \overline{M}_{1221} = \overline{M}_{2112} = \overline{M}_{2121} = \frac{1}{3} m \frac{E}{(1+r)},
\end{aligned} \tag{16}$$

where m and c denote the cubic average and the harmonic cubic-average of the pair (θ, h), respectively,

$$\begin{aligned}
m &= \theta h_+^3 + (1-\theta)h^3, \\
c &= \left(\theta h_+^{-3} + (1-\theta)h^{-3}\right)^{-1},
\end{aligned} \tag{17}$$

then

$$L(h_j, u_j) \to \overline{L}(\theta, h),$$

where

$$\overline{L}(\theta, h) = \int_\Omega F(x, \overline{u}, \nabla \overline{u})\, dx$$

and $\overline{u}$ is the solution of the plate equation corresponding to the tensor $\overline{M}$. Notice that weak convergence in $H_0^2(\Omega)$ implies strong convergence in $H^1(\Omega)$.

This discussion suggests defining a relaxation as an optimization problem for pairs (θ, h) in

$$\overline{\mathcal{H}} = \{(\theta, h) : 0 \le \theta \le 1, h_- \le h \le h_+\}$$

with cost

$$\overline{L}(\theta, h) = \int_\Omega F(x, \overline{u}, \nabla \overline{u})\, dx,$$

where as above $\overline{u}$ is the solution of the equilibrium plate problem for tensor $\overline{M}$ obtained from (θ, h) through the cubic-average and the harmonic cubic-average as in (16) and (17).

This would essentially be the proof of Theorem 3 except for the fact that the parameter V has not entered into our discussion. In fact, minimizing sequences of admissible designs must comply with

$$\int_\Omega h_j(x)\, dx \le V,$$

and we have not told how this parameter V enters into the relaxation. How is V to restrict further the pairs in $\overline{\mathcal{H}}$?

We observe that admissible pairs in $\overline{\mathcal{H}}$ come from the weak convergence of sequences $\left(h_j^3, h_j^{-3}\right)$. In order to relate h_j to $\left(h_j^3, h_j^{-3}\right)$, we will look for a function G so that

$$h = G(h^3, h^{-3}), \quad h \in [h_-, h_+],$$

and extend it by putting

$$G\left(\theta h_+^3 + (1-\theta)h^3, \theta h_+^{-3} + (1-\theta)h^{-3}\right) = \theta h_+ + (1-\theta)h.$$

If G so defined turns out to be convex, then by the weak convergences in (15),

(18)

$$\begin{aligned}
\lim_{j\to\infty} \int_\Omega h_j(x)\, dx &= \lim_{j\to\infty} \int_\Omega G\left(h_j^3(x), h_j^{-3}(x)\right) dx \\
&\ge \int_\Omega G\left(\theta(x)h_+^3 + (1-\theta(x))h(x)^3, \theta(x)h_+^{-3} + (1-\theta(x))h(x)^{-3}\right) dx \\
&= \int_\Omega \left[\theta(x)h_+ + (1-\theta(x))h(x)\right] dx,
\end{aligned}$$

so that we have

$$\int_\Omega \left[\theta(x)h_+ + (1-\theta(x))h(x)\right] dx \le V.$$

We then add this volume constraint to feasible pairs in $\overline{\mathcal{H}}$:

$$\overline{\mathcal{H}} = \left\{ (\theta, h) : 0 \le \theta \le 1, h_- \le h \le h_+, \int_\Omega [\theta(x)h_+ + (1-\theta(x))h(x)]\, dx \le V \right\}.$$

After the previous remarks, the full proof of Theorem 3 has been reduced to proving the convexity of the mapping G described above. This convexity property for G was proved in [11] (proof of Theorem 4.1). It is a nice, geometric argument, which we do not include here for the sake of brevity. It has nothing to do with the rest of the analysis in this work. One can also find in that paper how to recover admissible sequences of designs which are minimizing for the original problem from optimal pairs in $\overline{\mathcal{H}}$. This can be done in an elegant way by using Young measures associated with such sequences of designs.

REFERENCES

[1] N. Antonić and N. Balenović, *Optimal design for plates and relaxation*, in Proceedings of the 7th Annual International Conference in Operations Research (Rovinj, 1998), Math. Commun., 4 (1999), pp. 111–119.

[2] M. P. Bendsoe, *Optimization of Structural Topology, Shape and Material*, Springer, Berlin, 1995.

[3] E. Bonnetier and C. Conca, *Approximation of Young measures by functions and application to an optimal design problem for plates*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 399–442.

[4] E. Bonnetier and C. Conca, *Optimality conditions for a relaxed layout optimization problem*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 1005–1010.

[5] E. Bonnetier and M. Vogelius, *Relaxation of a compliance functional for a plate optimization problem*, in Application of Multiple Scaling in Mechanics, P. G. Ciarlet and E. Sánchez-Palencia, eds., Masson, Paris, 1987, pp. 31–53.

[6] K. T. Cheng and N. Olhoff, *An investigation concerning optimal design of solid elastic plates*, Internat. J. Solids Structures, 17 (1981), pp. 305–323.

[7] K. T. Cheng and N. Olhoff, *Regularized formulation for optimal design of axisymmetric plates*, Internat. J. Solids Structures, 18 (1982), pp. 153–169.

[8] D. Halim and S. O. Reza Moheimani, *An optimization approach to optimal placement of collocated piezoelectric actuators and sensors on a thin plate*, Mechatronics, 13 (2003), pp. 27–47.

[9] R. V. Kohn and M. Vogelius, *Thin plates with varying thickness, and their relation to structural optimization*, in Homogenization and Effective Moduli of Materials and Media, IMA Vol. Math. Appl. 1, J. Ericksen, D. Kinderlehrer, R. Kohn, and J. L. Lions, eds., Springer, New York, 1986, pp. 126–149.

[10] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems. Systems and Control: Foundations and Applications*, Birkhäuser, Boston, 1995.

[11] J. Muñoz and P. Pedregal, *On the relaxation of an optimal design problem for plates*, Asymptot. Anal., 16 (1996), pp. 125–140.

[12] F. Murat, *H-convergence*, Séminaire d'analyse fonctionelle et numérique de l'Université d'Alger, 1977.

[13] P. Pedregal, *On the generality of variational principles*, Milan J. Math., 71 (2003), pp. 319–356.

[14] P. Pedregal and A. Donoso, *Optimal design of a plate of variable thickness: A variational approach in dimension one*, Comput. Appl. Math., 22 (2003), pp. 75–89.

[15] T. Roubíček, *Maximum principle in optimal design of plates with stratified thickness*, Appl. Math. Optim., 51 (2005), pp. 183–200.

[16] J. Sprekels and D. Tiba, *A duality approach in the optimization of beams and plates*, SIAM J. Control Optim., 37 (1998), pp. 486–501.

[17] J. Sprekels and D. Tiba, *Optimization of clamped plates with discontinuous thickness*, Systems Control Lett., 48 (2003), pp. 289–295.

[18] L. Tartar, *Cours Peccot*, Collège de France, 1977.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 14–34

CONVERGENCE OF THE PRIMAL-DUAL ACTIVE SET STRATEGY FOR DIAGONALLY DOMINANT SYSTEMS*

KAZUFUMI ITO† AND KARL KUNISCH‡

Abstract. Sufficient conditions for global convergence of the primal-dual active set strategy for finite and infinite dimensional quadratic, as well as nonlinear optimization, problems with affine equality and inequality constraints are presented. These conditions involve diagonal dominance and cone preserving properties of the operator defining the cost functional. Globalization strategies are also provided, and specific sufficient conditions for the primal-dual active set step to have a descent property are given.

Key words. primal-dual active set strategy, diagonally dominant systems, bilateral constraints, globalization

AMS subject classifications. 49M29, 65J99, 65K10, 90C26, 90C46

DOI. 10.1137/050632713

1. Introduction. In this paper we discuss the primal-dual active set method for variational problems with simple constraints in function space or in $\mathbb{R}^n$. Let us consider the quadratic programming problem

$$\begin{cases} \min_{x\in Z} \frac{1}{2}\langle Ax, x\rangle_Z - \langle a, x\rangle_Z \\ \\ \text{subject to} \quad \phi \le x \le \psi, \end{cases} \tag{1.1}$$

where $a \in Z$, $A \in \mathcal{L}(Z)$ is a self-adjoint operator, and $Z = \mathbb{R}^n$ or $Z = L^2(\Omega)$, endowed with the usual Hilbert space structure and the natural ordering, where Ω is a domain in $\mathbb{R}^d$. We assume that (1.1) admits a unique solution denoted again by x. If x is a regular point [MZ] with respect to the constraints in (1.1), then there exists a Lagrange multiplier $\mu \in Z$ such that

$$\begin{aligned} &Ax + \mu = a, \\ &\mu = \max(0, \mu + c\,(x - \psi)) + \min(0, \mu + c\,(x - \phi)), \end{aligned} \tag{1.2}$$

where $c > 0$ is a fixed constant and max, min are interpreted pointwise a.e. in Ω if $Z = L^2(\Omega)$ and coordinatewise if $Z = \mathbb{R}^n$. The second equation in (1.2) constitutes the complementarity condition associated with the inequality constraint in (1.1) [IK1]. We note that in iterative methods such as sequential quadratic programming or second order augmented Lagrangian methods, quadratic optimization problems with linear constraints must be solved which take the form of (1.1). The primal-dual active set method that will be analyzed in this paper is an efficient technique for solving (1.2).

*Received by the editors May 31, 2005; accepted for publication (in revised form) August 11, 2006; published electronically February 23, 2007.
http://www.siam.org/journals/sicon/46-1/63271.html

†Center for Research in Scientific Computation, Department of Mathematics, North Carolina State University, Raleigh, NC 27695 (kito@unity.ncsu.edu). This author's research was partially supported by the Army Research Office under DAAD19-02-1-039.

‡Institut für Mathematik und Wissenschaftliches Rechnen, Universität Graz, Graz, Austria, and RICAM, Austrian Academy of Sciences, Linz, Austria (karl.kunisch@uni-graz.at).

While (1.2) is derived from (1.1) with A self-adjoint, this assumption is not essential in the remainder of this paper and we therefore drop it unless it is explicitly specified. If $\phi = -\infty$ or $\psi = \infty$, then (1.2) reduces to the unilaterally affine constraint case.

The primal-dual active set method uses the complementarity condition

$$\mu = \max(0, \mu + c\,(x - \psi)) + \min(0, \mu + c\,(x - \phi))$$

as a prediction strategy. Based on the current primal-dual pair (x, μ), the updates for the active and inactive sets are determined by

$$\mathcal{I} = \{\mu + c(x - \psi) \le 0\} \quad \text{and} \quad \mathcal{A} = \{\mu + c(x - \psi) > 0\},$$

where $\{\mu + c(x - \psi) \le 0\}$ is the abbreviation for $\{t : \mu(t) + c(x(t) - \psi(t)) \le 0\}$. This leads to the following Newton-like method.

Primal-dual active set method.

(1) Initialize x^0, μ^0. Set $k = 0$.

(2) Set $\mathcal{I}_k = \{\mu^k + c(x^k - \psi) \le 0 \le \mu^k + c(x^k - \phi)\}$, $\mathcal{A}_k^+ = \{\mu^k + c(x^k - \psi) > 0\}$, $\mathcal{A}_k^- = \{\mu^k + c(x^k - \phi) < 0\}$.

(3) Solve for (x^{k+1}, μ^{k+1})

$$Ax^{k+1} + \mu^{k+1} = a,$$

$$x^{k+1} = \psi \text{ in } \mathcal{A}_k^+, \quad x^{k+1} = \phi \text{ in } \mathcal{A}_k^-, \quad \text{and} \quad \mu^{k+1} = 0 \text{ in } \mathcal{I}_k.$$

(4) Stop, or set $k = k + 1$ and return to (4).

It was shown in [HIK] that the above algorithm can be interpreted as a semi-smooth Newton method for solving (1.2), and sufficient conditions were given for its local superlinear convergence. For related results in finite dimensions we refer to [FK], for example. The emphasis in this paper lies on providing sufficient conditions for global convergence without a globalization strategy such as a line search or a trust region method. This is motivated by the fact that global convergence was observed in many applications; see, e.g., [HIK, KR]. It appears to be difficult to find conditions which precisely describe this phenomenon. However, diagonal dominance and a structure which is close to the M-property appear to enhance this kind of unconditional convergence with respect to the initial condition.

We now describe the contributions of this paper. In section 2 we present motivating examples for the function space formulation of (1.1). We also consider the case where, in addition to the simple inequality constraints, equality constraints and a more general inequality constraint are present. A sufficient condition for the reduction of such problems to (1.1) is presented. Sufficient conditions for global convergence without globalization strategies of unilaterally constrained problems with arbitrary initialization are presented in section 3. In section 4 we analyze bilaterally constrained problems. Nonlinear problems are considered in section 5. As mentioned above, the primal-dual active set method converges for important practical problems without the necessity of introducing a globalization scheme. Of course, we cannot expect that this is universally true. Therefore, in section 6 we also consider a globalization strategy which can be utilized if the primal-dual active set strategy with full steps does not provide sufficient decrease. In particular we provide a sufficient condition, which guarantees that the direction supplied by the primal-dual active set strategy serves as a descent direction, and we describe alternative choices for obtaining descent directions.

2. Applications. Here we provide two motivating examples for studying the primal-dual active set strategy for (1.1) in function space. In Example 2.3, moreover, we consider a more general class of problems and their reduction to the form given by (1.1).

Example 2.1. Let us consider an optimal control problem with $\hat{\Omega} \subset \Omega$ as the control domain, Ω a bounded domain in a finite dimensional space, and $Z = L^2(\hat{\Omega})$:

$$\begin{cases} \min_{u\in X} \ \frac{1}{2}\int_\Omega |y-\bar{y}|^2\,dx + \frac{\alpha}{2}|u|_X^2 \\ \\ \text{subject to} \quad -\Delta y = Bu, \quad y = 0 \text{ on } \partial\Omega, \quad \text{and} \quad \phi \le u \le \psi, \end{cases}$$

where $\alpha > 0$, $\bar{y} \in L^2(\Omega), \phi \text{and} \psi \in Z$, and $B \in \mathcal{L}(Z, L^2(\Omega))$ is the extension-by-zero operator of the identity from $\hat{\Omega}$ to Ω. This problem can be formulated as (1.1), without equality constraint ($E = 0$), by setting

$$A = \alpha I + B^*(-\Delta)^{-2}B \quad \text{and} \quad a = B^*(-\Delta)^{-1}\bar{y},$$

where Δ denotes the Laplace operator with homogeneous Dirichlet boundary conditions. In this example A is an additive perturbation of a multiple of the identity operator, a situation which we shall return to in section 3. Note that A is also well defined from $L^p(\hat{\Omega})$ to $L^p(\hat{\Omega})$ for any $p \ge 1$. Moreover if B and B^* are positivity preserving, then A is positivity preserving, and discretizations of A have the property that off-diagonal elements are decaying at an α-dependent rate.

Example 2.2. Similarly we can consider the time-dependent problem

$$\begin{cases} \min_{u\in X} \ \frac{1}{2}\int_0^T \int_\Omega |y-\bar{y}|^2\,dx\,dt + \frac{\alpha}{2}|u|_X^2 \\ \\ \text{subject to} \quad \frac{d}{dt}y = \Delta y + Bu, \\ \\ y(0,\cdot) = y_0, \quad y = 0 \text{ on } (0,T)\times\partial\Omega, \quad \text{and} \quad \phi \le u \le \psi, \end{cases}$$

with $Z = L^2(0,T;L^2(\hat{\Omega})) = L^2((0,T)\times(\hat{\Omega})), \bar{y}$ and $y_0 \in L^2(\Omega)$, ϕ and $\psi \in Z$, $\alpha > 0$, and $B \in \mathcal{L}(Z, L^2(0,T;L^2(\Omega)))$ is the extension-by-zero operator of the identity from $(0,T)\times\hat{\Omega}$ to $(0,T)\times\Omega$. Again $A = \alpha I + ((\frac{d}{dt} - \Delta)^{-1}B)^*(\frac{d}{dt} - \Delta)^{-1}B$ is positivity preserving if B and B^* are positivity preserving, and off-diagonal elements of canonical discretizations of A are decaying at an α-dependent rate.

Example 2.3. Here we consider the quadratic programming problem

$$\begin{cases} \min_{x\in X} \ \frac{1}{2}\langle Ax, x\rangle_X - \langle a, x\rangle_X \\ \\ \text{subject to } Ex = b, \quad \phi \le Gx \le \psi, \end{cases} \tag{2.1}$$

where $a \in X, A \in \mathcal{L}(X)$ is a self-adjoint operator in the real Hilbert space X, $E \in \mathcal{L}(X,W)$, $G \in \mathcal{L}(X,Z)$, with W a real Hilbert space, and Z is as in (1.1). We assume that (2.1) admits a unique solution, denoted again by x. If x is a regular point [MZ] with respect to the constraints in (2.1), then there exists a Lagrange multiplier

$(\lambda, \mu) \in W^* \times Z$ such that

$$\begin{cases} Ax + E^*\lambda + G^*\mu = a, \\ Ex = b, \\ \mu = \max(0, \mu + c\,(Gx - \psi)) + \min(0, \mu + c\,(Gx - \phi)), \end{cases} \tag{2.2}$$

where $c > 0$ is a fixed constant as above.

We now derive sufficient conditions which allow us to transform (2.2) into (1.2). In a first step we assume that

$$G \text{ is surjective, } \operatorname{range}(G^*) \subset \ker E,\ E\bar{x} = b, \text{ where } \bar{x} \in (\ker E)^{\perp}. \tag{2.3}$$

Note that (2.3) implies that $G : N(E) \to Z$ is surjective. If not, then there exists a nonzero $z \in Z$ such that $(z, Gx)_Z = (G^*z, x)_X = 0$ for all $x \in \ker E$. If we let $x = G^*z$, then $|x|^2 = 0$ and $z = 0$, since G^* is injective. Let P_E denote the orthogonal projection in X onto $\ker E$. Then (2.2) is equivalent to

$$\mathcal{A}\hat{x} + G^*\mu = P_E(a - A\bar{x}), \quad \mu = \max(0, \mu + c\,(G\hat{x} - (\psi - G\bar{x}),$$

$$E^*\lambda = (I - P_E)(a - A(P_E\hat{x} + \bar{x})),$$

with $\mathcal{A} = P_E A P_E^*$ and $x = \hat{x} + \bar{x} \in \ker E + (\ker E)^{\perp}$. The first of the above equations is equivalent to the system

$$\begin{aligned} &(I - P_G)\mathcal{A}((I - P_G)\,\hat{x} + P_G\,\hat{x}) + G^*\mu = (I - P_G)P_E(a - A\,\bar{x}), \\ &P_G\,\mathcal{A}((I - P_G)\,\hat{x} + P_G\,\hat{x}) = P_G\,P_E\,(a - A\,\bar{x}), \end{aligned} \tag{2.4}$$

where $P_G = I - G^*(G\,G^*)^{-1}G$ is the orthogonal projection in $\ker E \subset X$ onto $\ker G$. Since G^* is injective, the first equation in (2.4) is equivalent to

$$(G\,G^*)^{-1}G\,\mathcal{A}(G^*(G\,G^*)^{-1}y + x_2) + \mu = (G\,G^*)^{-1}G\,P_E(a - A\,\bar{x}),$$

where $y = G\,x_1$ for $x_1 \in \ker E \cap (\ker G)^{\perp}$, $x_2 \in \ker G \cap \ker E$, and $\hat{x} = x_1 + x_2$. Let

$$A_{11} = (GG^*)^{-1}G\mathcal{A}G^*(GG^*)^{-1}, \ \ A_{12} = (GG^*)^{-1}G\mathcal{A}P_G, A_{22} = P_G\mathcal{A}P_G$$

and

$$a_1 = (GG^*)^{-1}GP_E(a - A\bar{x}),\ a_2 = P_GP_E(a - A\bar{x}).$$

Then (2.4) is equivalent to the following equation in $Z \times (\ker E \cap \ker G)$:

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{12}^* & A_{22} \end{pmatrix} \begin{pmatrix} y \\ x_2 \end{pmatrix} + \begin{pmatrix} \mu \\ 0 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}. \tag{2.5}$$

Let us summarize the discussion so far. If $A \in \mathcal{L}(Z)$ and (2.3) holds, then (2.5), together with

$$\mu = \max(0, \mu + c(y - (\psi - G\bar{x}))), \tag{2.6}$$

and $E^*\lambda = (I - P_E)(a - A(\hat{x} + \bar{x}))$ are equivalent to (2.2), where $x = \hat{x} + \bar{x}$, with $\hat{x} = x_1 + x_2 \in \ker E$, $x_2 \in \ker E \cap \ker G$, $x_1 \in \ker E \cap (\ker G)^\perp$, $y = Gx_1$.

Note that the system matrix in (2.5) is positive definite if A restricted to $\ker E$ is positive definite.

Let us now further assume that

$$A_{22} \text{ is nonsingular.} \tag{2.7}$$

Then (2.5), (2.6) are equivalent to

$$(A_{11} - A_{12}A_{22}^{-1}A_{12}^*)y + \mu = a_1 - A_{12}A_{22}^{-1}a_2 \tag{2.8}$$

and (2.6), which is the desired form (1.2). In the finite dimensional case, (1.2) admits a unique solution for every $a \in \mathbb{R}^n$, if and only if A is a P-matrix (see [BP, Theorem 10.2.15]). Recall that A is called a P-matrix, if all its principal minors are positive. In view of the fact that the reduction of (2.5) to (2.8) was achieved by taking the Schur complement with respect to A_{22}, it is also worthwhile to recall that the Schur complement of a P-matrix (resp., M-matrix) is again a P-matrix (resp., M-matrix); see [BP, page 292]. If one does not carry out the reduction step from (2.5) to (2.8), then the coordinates corresponding to x_2 can be treated as inactive ones in the algorithm and the convergence analysis that we carry out for (1.2) remains valid for (2.5)–(2.6).

Let us specify the primal-dual active set method for the extended problem (2.2).

Primal-dual active set method (extended problem).

(1) Initialize x^0, μ^0. Set $k = 0$.

(2) Set $\mathcal{I}_k = \{\mu^k + c(Gx^k - \psi) \le 0 \le \mu^k + c(Gx^k - \phi)\}$, $\mathcal{A}_k^+ = \{\mu^k + c(Gx^k - \psi) > 0\}$, $\mathcal{A}_k^- = \{\mu^k + c(Gx^k - \phi) < 0\}$.

(3) Solve for $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$

$$Ax^{k+1} + E^*\lambda^{k^+1} + G^*\mu^{k+1} = a,$$

$$Ex^{k+1} = b,$$

$$Gx^{k+1} = \psi \text{ in } \mathcal{A}_k^+, \quad Gx^{k+1} = \phi \text{ in } \mathcal{A}_k^-, \quad \text{and} \quad \mu^{k+1} = 0 \text{ in } \mathcal{I}_k.$$

(4) Stop, or set $k = k + 1$ and return to (2).

Applying the algorithm for the extended system to (2.2), or utilizing the primal-dual algorithm for the reduced system (1.2), results in algebraically equivalent systems if (2.3) holds. The relationship between the two approaches is given by

$$x^{k+1} = x_1^{k+1} + x_2^{k+1} + \bar{x}, \text{ where } x_1^{k+1} \in \ker E \cap (\ker G)^\perp, \quad x_2^{k+1} \in \ker E \cap \ker G,$$
$$Gx_1^{k+1} = y, \quad \mu = 0 \text{ in } \mathcal{I}_k, \quad y^{k+1} = \psi - G\bar{x} \text{ in } \mathcal{A}_k^+, \quad y^{k+1} = \phi - G\bar{x} \text{ in } \mathcal{A}_k^-,$$
and (2.4) holds with $(y, x_2) = (y^{k+1}, x_2^{k+1})$.

3. Diagonally dominated class: Unilateral case. In this section we discuss the global convergence of the primal-dual active set method in the unilateral case, i.e.,

$$Ax + \mu = a, \quad \mu = \max(0, \mu + c(x - \psi)). \tag{3.1}$$

For the convenience of the reader we recall the algorithm for this case.

PRIMAL-DUAL ACTIVE SET METHOD (unilateral problem).
(1) Initialize x^0, μ^0. Set $k = 0$.
(2) Set $\mathcal{I}_k = \{\mu^k + c(x^k - \psi) \le 0\}, \quad \mathcal{A}_k = \{\mu^k + c(x^k - \psi) > 0\}$.
(3) Solve for (x^{k+1}, μ^{k+1})

$$Ax^{k+1} + \mu^{k+1} = a,$$

$$x^{k+1} = \psi \text{ in } \mathcal{A}_k, \quad \text{and} \quad \mu^{k+1} = 0 \text{ in } \mathcal{I}_k.$$

(4) Stop, or set $k = k + 1$ and return to (4).

Here we abbreviate $\mathcal{A}_k^+$ by $\mathcal{A}_k$. Sufficient conditions for global convergence were established in [HIK] for the finite dimensional case $Z = \mathbb{R}^n$. The sufficient condition we discuss here is more general. It is related to diagonal dominance of A and will imply that

$$\mathcal{M}(x^{k+1}, \mu^{k+1}) = \max\left(\beta \int_\Omega (x^{k+1} - \psi)^+ \, dx, \int_\Omega (\mu^{k+1})^- \, dx\right)$$

with $\beta > 0$ acting as a merit functional for the primal-dual algorithm, i.e., $\mathcal{M}(x^{k+1}, \mu^{k+1}) \le \rho \mathcal{M}(x^k, \mu^k)$, for some $\rho < 1$. Here we set $\phi^+ = \max(\phi, 0)$ and $\phi^- = -\min(\phi, 0)$. Note that by step (3) of the algorithm we have

$$\mathcal{M}(x^{k+1}, \mu^{k+1}) = \max\left(\beta \int_{\mathcal{I}_k} (x^{k+1} - \psi)^+ \, dx, \int_{\mathcal{A}_k} (\mu^{k+1})^- \, dx\right). \tag{3.2}$$

The natural norm associated with this merit functional is the $L^1(\Omega)$-norm, and consequently we assume that

$$A \in \mathcal{L}(L^1(\Omega)), \; a \in L^1(\Omega), \text{ and } \psi \in L^1(\Omega). \tag{3.3}$$

The analysis of this section can also be used to obtain convergence in the $L^p(\Omega)$-norm for any $p \in (1, \infty)$, if the norms in the integrands of $\mathcal{M}$ are replaced with $|\cdot|^p$-norms and the $L^1(\Omega)$-norms below are replaced with $L^p(\Omega)$-norms as well.

The results also apply for $Z = \mathbb{R}^n$. In this case the integrals in (3.2) must be replaced with sums over the respective index sets.

We assume that there exist constants $\rho_i, i = 1, \ldots, 5$, such that for all partitions $\mathcal{A}$ and $\mathcal{I}$ of Ω and for all $\phi_\mathcal{A} \ge 0$ in $L^2(\mathcal{A})$ and $\phi_\mathcal{I} \ge 0$ in $L^2(\mathcal{I})$,

$$\begin{aligned} |[A_\mathcal{I}^{-1} \phi_\mathcal{I}]^-| &\le \rho_1 \, |\phi_\mathcal{I}|, \\ |[A_\mathcal{I}^{-1} A_{\mathcal{I}\mathcal{A}} \phi_\mathcal{A}]^+| &\le \rho_2 \, |\phi_\mathcal{A}| \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} |[A_\mathcal{A} \phi_\mathcal{A}]^-| &\le \rho_3 \, |\phi_\mathcal{A}|, \\ |[A_{\mathcal{A}\mathcal{I}} A_\mathcal{I}^{-1} \phi_\mathcal{I}]^-| &\le \rho_4 \, |\phi_\mathcal{I}|, \\ |[A_{\mathcal{A}\mathcal{I}} A_\mathcal{I}^{-1} A_{\mathcal{I}\mathcal{A}} \phi_\mathcal{A}]^+| &\le \rho_5 \, |\phi_\mathcal{A}|. \end{aligned} \tag{3.5}$$

Here $|\cdot|$ denotes the $L^1(\Omega)$-norm and $A_{\mathcal{I}} = R_{\mathcal{I}} A E_{\mathcal{I}}$, $A_{\mathcal{I}\mathcal{A}} = R_{\mathcal{A}} A E_{\mathcal{I}}$, where $E_{\mathcal{I}} : L^1(\mathcal{I}) \to L^1(\Omega)$ is the extension-by-zero operator and $R_{\mathcal{I}} : L^1(\Omega) \to L^1(\mathcal{I})$ the restriction operator from Ω to $\mathcal{I}$. The remaining symbols are defined by analogy. Assumption (3.4) requires in particular the existence of $A_{\mathcal{I}}^{-1}$. By a Schur-complement argument with respect to the sets $\mathcal{I}_k$ and $\mathcal{A}_k$ this implies existence of a solution to the linear systems in step (3) of the algorithm for every k.

THEOREM 3.1. *If* (3.3), (3.4), (3.5) *hold and* $\rho = \max(\beta\,\rho_1 + \rho_2, \frac{\rho_3}{\beta} + \rho_4 + \frac{\rho_5}{\beta}) < 1$, *then* $\mathcal{M}$ *is a merit function for the primal-dual algorithm of the reduced system, and* $\lim_{k\to\infty}(x^k, \mu^k) = (x^*, \mu^*)$ *in* $L^1(\Omega) \times L^1(\Omega)$, *with* (x^*, μ^*) *a solution to* (3.1).

Proof. Let $\delta x = x^{k+1} - x^k$ and $\delta\mu = \mu^{k+1} - \mu^k$. Then,

$$\begin{aligned} &A_{\mathcal{A}_k}\delta x_{\mathcal{A}_k} + A_{\mathcal{A}_k,\mathcal{I}_k}\delta x_{\mathcal{I}_k} + \delta\mu_{\mathcal{A}_k} = 0, \\ &A_{\mathcal{I}_k}\delta x_{\mathcal{I}_k} + A_{\mathcal{I}_k,\mathcal{A}_k}\delta x_{\mathcal{A}_k} - \mu^k_{\mathcal{I}_k} = 0. \end{aligned} \tag{3.6}$$

For every $k \geq 1$ we have $(x^{k+1} - \psi)^+ \leq (x^{k+1} - x^k)^+$ on $\mathcal{I}_k$ and $(\mu^{k+1})^- = (\delta\mu)^-$ on $\mathcal{A}_k$. Therefore

$$\mathcal{M}(x^{k+1}, \mu^{k+1}) \leq \max\left(\beta \int_{\mathcal{I}_k} (\delta x_{\mathcal{I}_k})^+, \int_{\mathcal{A}_k} (\delta\mu_{\mathcal{A}_k})^-\right). \tag{3.7}$$

From (3.6) we deduce that

$$\delta x_{\mathcal{I}_k} = -A^{-1}_{\mathcal{I}_k}(-\mu^k_{\mathcal{I}_k}) \,+\, A^{-1}_{\mathcal{I}_k} A_{\mathcal{I}_k\mathcal{A}_k}(-\delta x_{\mathcal{A}_k}),$$

with $\mu^k_{\mathcal{I}_k} \leq 0$ and $\delta x_{\mathcal{A}_k} \leq 0$. By (3.4) therefore

$$\begin{aligned} |(\delta x_{\mathcal{I}_k})^+| &\leq \rho_1 |\mu^k_{\mathcal{I}_k}| + \rho_2 |\delta x_{\mathcal{A}_k}| \\ &= \rho_1 \int_{\mathcal{I}_k \cap \mathcal{A}_{k-1}} |(\mu^k_{\mathcal{I}_k})^-| + \rho_2 \int_{\mathcal{A}_k \cap \mathcal{I}_{k-1}} (x^k - \psi)^+ \\ &\leq \left(\rho_1 + \frac{\rho_2}{\beta}\right) \mathcal{M}(x^k, \mu^k). \end{aligned} \tag{3.8}$$

Similarly by (3.6),

$$\delta\mu_{\mathcal{A}_k} = A_{\mathcal{A}_k}(-\delta x_{\mathcal{A}_k}) + A_{\mathcal{A}_k\mathcal{I}_k} A^{-1}_{\mathcal{I}_k}(-\mu^k_{\mathcal{I}_k}) - A_{\mathcal{A}_k\mathcal{I}_k} A^{-1}_{\mathcal{I}_k} A_{\mathcal{I}_k\mathcal{A}_k}(-\delta x_{\mathcal{A}_k}).$$

Since $\delta x_{\mathcal{A}_k} \leq 0$ and $\mu^k_{\mathcal{I}_k} \leq 0$, we find by (3.5)

$$|(\delta\mu_{\mathcal{A}_k})^-| \leq \rho_3 |\delta x_{\mathcal{A}_k}| + \rho_4 |\mu^k_{\mathcal{I}_k}| + \rho_5 |\delta x_{\mathcal{A}_k}| \leq \left(\frac{\rho_3 + \rho_5}{\beta} + \rho_4\right) \mathcal{M}(x^k, \mu^k), \tag{3.9}$$

and therefore

$$\mathcal{M}(x^{k+1}, \mu^{k+1}) \leq \max\left(\beta\rho_1 + \rho_2, \frac{\rho_3 + \rho_5}{\beta} + \rho_4\right) \mathcal{M}(x^k, \mu^k) = \rho\, \mathcal{M}(x^k, \mu^k).$$

Thus, if $\rho < 1$, then $\mathcal{M}$ is a merit functional. Furthermore $\mathcal{M}(x^{k+1}, \mu^{k+1}) \leq \rho^k \mathcal{M}(x^1, \mu^1)$. Together with (3.8), (3.9), and (3.6) it follows that (x^k, μ^k) is a Cauchy sequence. Hence there exists (x^*, μ^*) such that $\lim_{k\to\infty}(x^k, \mu^k) = (x^*, \mu^*)$ and

$Ax^* + \mu^* = a, \mu^*(x^* - \psi) = 0$ a.e. in Ω. Since $(x^k - \psi)^+ \to (x^* - \psi)^+$ as $k \to \infty$ and $\lim_{k\to\infty} \int_\Omega (x^{k+1} - \psi)^+ = 0$ it follows that $x^* \le \psi$. Similarly one argues that $\mu^* \ge 0$. Thus, (x^*, μ^*) is a solution to (3.1). ∎

Concerning the uniqueness of solutions to (3.1), assume that $A \in \mathcal{L}(L^2(\Omega))$ and that $(Ay, y)_{L^2(\Omega)} > 0$ for all $y \in L^2(\Omega)$ with $y \neq 0$. Assume further that (x^*, μ^*) and $(\hat{x}, \hat{\mu})$ are solutions to (3.1) with $\hat{x} - x^* \in L^2(\Omega)$. Then $(\hat{x} - x^*, A(\hat{x} - x^*))_{L^2(\Omega)} \le 0$ and therefore $\hat{x} - x^* = 0$.

Remark 3.1. In the finite dimensional case the integrals in the definition of $\mathcal{M}$ must be replaced with sums over the active/inactive index sets. If A is an M-matrix, then $\rho_1 = \rho_2 = 0$ and $\rho < 1$ if $\frac{\rho_3}{\beta} + \rho_4 + \frac{\rho_5}{\beta} < 1$. This is the case if A is diagonally dominant in the sense that $\rho_4 < 1$ and β is chosen sufficiently large. For such a matrix A the property $\rho < 1$ is stable under additive perturbations, which are not necessarily M-matrices.

Remark 3.2. Consider the infinite dimensional case with $A = \alpha I + K$, where $\alpha > 0$, $K \in \mathcal{L}(L^1(\Omega))$, and $K\phi \ge 0$ for all $\phi \ge 0$. This is the case for the operators in Example 2.1, as can be argued by using the maximum principle. Let $\|K\|$ denote the norm of $K \in \mathcal{L}(L^1(\Omega))$. For $\|K\| < \alpha$ and any $\mathcal{I} \subset \Omega$ we have $A_{\mathcal{I}}^{-1} = \frac{1}{\alpha} I_{\mathcal{I}} - \frac{1}{\alpha} K_{\mathcal{I}} A_{\mathcal{I}}^{-1}$, and hence $\rho_1 \le \frac{\|K\|}{\alpha(\alpha - \|K\|)}$. Moreover $\rho_3 = 0$. The conditions involving ρ_2, ρ_4, and ρ_5 are satisfied with $\rho_2 = \frac{\|K\|}{\alpha - \|K\|}$, $\rho_4 = \frac{\|K\|^2}{\alpha(\alpha - \|K\|)}$, and $\rho_5 = \frac{\|K\|^2}{\alpha - \|K\|}$, and $\rho < 1$ if α is sufficiently large.

4. Diagonally dominated class: Bilateral constraints. The primal-dual active set method for the bilateral constraint case was given in section 1. We now provide sufficient conditions for its global convergence. Analogously to section 3 we select the merit function $\mathcal{M}$ as

$$\mathcal{M}(x^{k+1}, \mu^{k+1}) = \max\Big(\int_{\mathcal{I}_k} ((x^{k+1} - \psi)^+ + (x^{k+1} - \varphi)^-)\, dx, \int_{\mathcal{A}_k^+} (\mu^{k+1})^-\, dx + \int_{\mathcal{A}_k^-} (\mu^{k+1})^+\, dx \Big). \tag{4.1}$$

In the finite dimensional case the integrals must be replaced with sums over the respective index sets. We note that step (3) of the algorithm implies the complementarity property

$$(x^k - \psi)(x^k - \varphi)\mu^k = 0 \text{ a.e. in } \Omega. \tag{4.2}$$

As in the previous section the merit function involves L^1-norms and accordingly we aim for convergence in $L^1(\Omega)$. We henceforth assume that

$$A \in \mathcal{L}(L^1(\Omega)),\ a \in L^1(\Omega),\ \psi \text{ and } \varphi \in L^1(\Omega). \tag{4.3}$$

Below, $\|\cdot\|$ denotes the norm of operators in $\mathcal{L}(L^1(\Omega))$. The following conditions will be used: There exist constants $\rho_i, i = 1, \dots, 5$, such that for arbitrary partitions $\mathcal{A} \cup \mathcal{I} = \Omega$ we have

$$\|A_{\mathcal{I}}^{-1}\| \le \rho_1, \qquad \|A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}}\| \le \rho_2 \tag{4.4}$$

and

$$\|A_{\mathcal{A}} - c\,I\| \le \rho_3,$$

$$\|A_{\mathcal{A}\mathcal{I}} A_{\mathcal{I}}^{-1}\| \le \rho_4, \tag{4.5}$$

$$\|A_{\mathcal{A}\mathcal{I}} A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}}\| \le \rho_5.$$

We further set $\rho = 2\max(\max(\rho_1, \rho_2, \frac{\rho_2}{c}), \max(\rho_3 + \rho_5, \rho_4), \frac{\rho_3+\rho_5}{c})$.

THEOREM 4.1. *If* (4.3), (4.4), (4.5) *hold and* $\rho < 1$, *then* $\mathcal{M}$ *is a merit function for the primal-dual algorithm applied to* (1.2) *and* $\lim_{k\to\infty}(x^k, \mu^k) = (x^*, \mu^*)$ *in* $L^1(\Omega) \times L^1(\Omega)$, *with* (x^*, μ^*) *a solution to* (1.2).

Proof. For $\delta x = x^{k+1} - x^k$ and $\delta\mu = \mu^{k+1} - \mu^k$ we have

$$\begin{aligned}
&A_{\mathcal{A}_k^+}\delta x_{\mathcal{A}_k^+} + A_{\mathcal{A}_k^+\mathcal{I}_k}\delta x_{\mathcal{I}_k} + A_{\mathcal{A}_k^+\mathcal{A}_k^-}\delta x_{\mathcal{A}_k^-} + \delta\mu_{\mathcal{A}_k^+} = 0,\\
&A_{\mathcal{I}_k}\delta x_{\mathcal{I}_k} + A_{\mathcal{I}_k\mathcal{A}_k^+}\delta x_{\mathcal{A}_k^+} + A_{\mathcal{I}_k\mathcal{A}_k^-}\delta x_{\mathcal{A}_k^-} - \mu^k_{\mathcal{I}_k} = 0, \qquad (4.6)\\
&A_{\mathcal{A}_k^-}\delta x_{\mathcal{A}_k^-} + A_{\mathcal{A}_k^-\mathcal{I}_k}\delta x_{\mathcal{I}_k} + A_{\mathcal{A}_k^-\mathcal{A}_k^+}\delta x_{\mathcal{A}_k^+} + \delta\mu_{\mathcal{A}_k^-} = 0
\end{aligned}$$

with

$$\mu^k_{\mathcal{A}_k^+} \begin{cases} > 0 & \text{on} \quad \mathcal{A}^+_{k-1} \cap \mathcal{A}^+_k, \\ = 0 & \text{on} \quad \mathcal{I}_{k-1} \cap A^+_k, \\ > c\,(\psi - \varphi) & \text{on} \quad \mathcal{A}^-_{k-1} \cap \mathcal{A}^+_k, \end{cases} \tag{4.7}$$

$$\mu^k_{\mathcal{I}_k} \in \begin{cases} [c(\varphi - \psi), 0) & \text{on} \quad \mathcal{A}^+_{k-1} \cap \mathcal{I}_k, \\ 0 & \text{on} \quad \mathcal{I}_{k-1} \cap \mathcal{I}_k, \\ (0, c(\psi - \varphi)] & \text{on} \quad \mathcal{A}^-_{k-1} \cap \mathcal{I}_k, \end{cases} \tag{4.8}$$

$$\mu^k_{\mathcal{A}_k^-} \begin{cases} < c\,(\varphi - \psi) & \text{on} \quad \mathcal{A}^+_{k-1} \cap \mathcal{A}^-_k, \\ = 0 & \text{on} \quad \mathcal{I}_{k-1} \cap A^-_k, \\ < 0 & \text{on} \quad \mathcal{A}^-_{k-1} \cap \mathcal{A}^-_k, \end{cases} \tag{4.9}$$

$$\delta x_{\mathcal{A}_k^+} \begin{cases} = 0 & \text{on} \quad \mathcal{A}^+_{k-1} \cap \mathcal{A}^+_k, \\ < 0 & \text{on} \quad \mathcal{I}_{k-1} \cap A^+_k, \\ = \psi - \varphi < \frac{\mu^k}{c} & \text{on} \quad \mathcal{A}^-_{k-1} \cap \mathcal{A}^+_k, \end{cases} \tag{4.10}$$

$$\delta x_{\mathcal{A}_k^-} \begin{cases} = \varphi - \psi > \frac{\mu^k}{c} & \text{on} \quad \mathcal{A}^+_{k-1} \cap \mathcal{A}^-_k, \\ > 0 & \text{on} \quad \mathcal{I}_{k-1} \cap A^-_k, \\ = 0 & \text{on} \quad \mathcal{A}^-_{k-1} \cap \mathcal{A}^-_k. \end{cases} \tag{4.11}$$

From (4.2)

$$(x^{k+1}_{\mathcal{I}_k} - \psi_{\mathcal{I}_k})^+ \le (\delta x_{\mathcal{I}_k})^+ \text{ and } (x^{k+1}_{\mathcal{I}_k} - \varphi_{\mathcal{I}_k})^- \le (\delta x_{\mathcal{I}_k})^-. \tag{4.12}$$

This implies that

$$\mathcal{M}(x^{k+1},\mu^{k+1}) \le \max\left(\int_{\mathcal{I}_k}|\delta x_{\mathcal{I}_k}|\,,\ \int_{\mathcal{A}_k^+}(\mu^{k+1})^- + \int_{\mathcal{A}_k^-}(\mu^{k+1})^+\right). \tag{4.13}$$

From (4.6), (4.4), (4.8), (4.10), and (4.11) we have

$$\begin{aligned}
|\delta x_{\mathcal{I}_k}| &\le \rho_1|\mu^k_{\mathcal{I}_k}| + \rho_2|\delta x_{\mathcal{A}_k}| \\
&\le \rho_1(|(\mu^k_{\mathcal{I}_k\cap\mathcal{A}^+_{k-1}})^-| + |(\mu^k_{\mathcal{I}_k\cap\mathcal{A}^-_{k-1}})^+|) + \rho_2(|\delta x_{\mathcal{A}^+_k}| + |\delta x_{\mathcal{A}^-_k}|) \\
&\le \rho_1(|(\mu^k_{\mathcal{I}_k\cap\mathcal{A}^+_{k-1}})^-| + |(\mu^k_{\mathcal{I}_k\cap\mathcal{A}^-_{k-1}})^+|) \\
&\quad + \rho_2\bigg(|(x^k-\psi)^+_{\mathcal{A}^+_k\cap\mathcal{I}_{k-1}}| + \frac{1}{c}|(\mu^k_{\mathcal{A}^+_k\cap\mathcal{A}^-_{k-1}})^+| + |(x^k-\varphi)^-_{\mathcal{A}^-_k\cap\mathcal{I}_{k-1}}| \\
&\qquad + \frac{1}{c}|(\mu^k_{\mathcal{A}^-_k\cap\mathcal{A}^+_{k-1}})^-|\bigg).
\end{aligned}$$

This implies

$$|\delta x_{\mathcal{I}_k}| \le 2\,\max\left(\rho_1,\rho_2,\frac{\rho_2}{c}\right)\,\mathcal{M}(x^k,\mu^k). \tag{4.14}$$

From (4.6) furthermore

$$\mu^{k+1}_{\mathcal{A}_k} - (\mu^k_{\mathcal{A}_k} - c\delta x_{\mathcal{A}_k}) = g, \tag{4.15}$$

where $g = (cI - A_{\mathcal{A}_k})\delta x_{\mathcal{A}_k} - A_{\mathcal{A}_k\mathcal{I}_k}A^{-1}_{\mathcal{A}_k}A_{\mathcal{I}_k\mathcal{A}_k}\delta x_{\mathcal{A}_k}$. By (4.7) and (4.10), we have

$$\mu^k_{\mathcal{A}^+_k} - c\,\delta x_{\mathcal{A}^+_k} \ge 0.$$

Similarly by (4.9) and (4.11), we have

$$\mu^k_{\mathcal{A}^-_k} - c\,\delta x_{\mathcal{A}^-_k} \le 0.$$

Consequently

$$\begin{aligned}
&|(\mu^{k+1}_{\mathcal{A}^+_k})^-| + |(\mu^{k+1}_{\mathcal{A}^-_k})^+| \le |g_{\mathcal{A}_k}| \le (\rho_3+\rho_5)|\delta x_{\mathcal{A}_k}| + \rho_4|\mu_{I_k}| \\
&\le 2\max\left(\rho_4,\rho_3+\rho_5,\frac{\rho_3+\rho_5}{c}\right)\mathcal{M}(x^k,\mu^k).
\end{aligned} \tag{4.16}$$

By (4.13), (4.14), and (4.16)

$$\mathcal{M}(x^{k+1},\mu^{k+1}) \le 2\max\left(\max\left(\rho_1,\rho_2,\frac{\rho_2}{c}\right),\max\left(\rho_4,\rho_3+\rho_5,\frac{\rho_3+\rho_5}{c}\right)\right)\mathcal{M}(x^k,\mu^k).$$

It follows that $\mathcal{M}(x^{k+1},\mu^{k+1}) \le \rho^k\mathcal{M}(x^1,\mu^1)$, and if $\rho < 1$, then $M(x^k,\mu^k) \to 0$ as $k \to \infty$. From the estimates leading to (4.14) it follows that x^k is a Cauchy sequence. Moreover μ^k is a Cauchy sequence by (4.6). Hence there exist (x^*,μ^*) such that $\lim_{k\to\infty}(x^k,\mu^k) = (x^*,\mu^*)$. By Lebesgue's bounded convergence theorem, and since $M(x^k,\mu^k) \to 0$, it follows that $\varphi \le x^* \le \psi$. Clearly $Ax^* + \mu^* = a$

and $(x^* - \psi)(x^* - \varphi)\mu^* = 0$ by (4.2). This last equation implies that $\mu^* = 0$ on $\mathcal{I}^* = \{\varphi < x^* < \psi\}$. It remains to show that $\mu^* \geq 0$ on $\mathcal{A}^{*,+} = \{x^* = \psi\}$ and $\mu^* \leq 0$ on $\mathcal{A}^{*,-} = \{x^* = \varphi\}$. Let $s \in \mathcal{A}^{*,+}$ be such that $x^k(s)$ and $\mu^k(s)$ converge. Then $\mu^*(s) \geq 0$. If not, then $\mu^*(s) < 0$ and there exists $\bar{k}$ such that $\mu^k(s) + c(x^k(s) - \psi(s)) \leq \frac{\mu^*(s)}{2} < 0$ for all $k \geq \bar{k}$. Then $s \in \mathcal{I}^k$ and $\mu^{k+1} = 0$ for $k \geq \bar{k}$, contradicting $\mu^*(s) < 0$. Analogously one shows that $\mu^* \leq 0$ on $\mathcal{A}^{*,-}$. ∎

We now specialize to perturbations A which are additive perturbations of a multiple of the identity operator.

THEOREM 4.2. *Assume that $A = cI + K$ with $K \in \mathcal{L}(L^1(\Omega))$ and $\|K\| < c$, and that (4.3), (4.4), (4.5) are satisfied. If $\bar{\rho} = 2\max(\max(\frac{\|K\|}{c}\rho_1, \rho_2), \max(\rho_3 + \rho_5, \rho_4), \frac{\rho_3+\rho_5}{c}) < 1$, then the conclusions of the previous theorem are valid.*

Proof. We follow the proof of Theorem 4.1 and eliminate the overestimate (4.12). Let $P = \{x^{k+1}_{\mathcal{I}_k} - \psi > 0\} \cap \mathcal{I}_k$. We find

$$x^{k+1} - \psi \begin{cases} \leq \delta x_{P \cap \mathcal{I}_{k-1}} & \text{on} \quad P \cap \mathcal{I}_{k-1}, \\ = \delta x_{P \cap \mathcal{A}^+_{k-1}} & \text{on} \quad P \cap \mathcal{A}^+_{k-1}, \\ \delta x_{P \cap \mathcal{A}^-_{k-1}} + (\varphi - \psi)_{P \cap \mathcal{A}^-_{k-1}} & \text{on} \quad P \cap \mathcal{A}^-_{k-1}. \end{cases}$$

This estimate, together with $A^{-1} = \frac{1}{c}I - \frac{1}{c}KA^{-1}$, implies that

$$\begin{aligned} \int_{\mathcal{I}_k} (x^{k+1} - \psi)^+ &\leq \int_P \delta x + \int_{P \cap A^-_{k-1}} \varphi - \psi \\ &\leq \int_P A^{-1}_{\mathcal{I}_k} \mu^k_{\mathcal{I}_k} - \int_P A^{-1}_{\mathcal{I}_k} A_{\mathcal{I}_k \mathcal{A}_k} \delta x_{\mathcal{A}_k} + \int_{P \cap A^-_{k-1}} \varphi - \psi \\ &= \frac{1}{c} \int_P \mu^k_{\mathcal{I}_k} + \int_{P \cap A^-_{k-1}} \varphi - \psi - \frac{1}{c} \int_P K_{\mathcal{I}_k} A^{-1}_{\mathcal{I}_k} \mu^k_{\mathcal{I}_k} - \int_P A^{-1}_{\mathcal{I}_k} A_{\mathcal{I}_k \mathcal{A}_k} \delta x_{\mathcal{A}_k} \\ &\leq -\frac{1}{c} \int_P K_{\mathcal{I}_k} A^{-1}_{\mathcal{I}_k} \mu^k_{\mathcal{I}_k} - \int_P A^{-1}_{\mathcal{I}_k} A_{\mathcal{I}_k \mathcal{A}_k} \delta x_{\mathcal{A}_k}, \end{aligned}$$

and hence

$$\int_{\mathcal{I}_k} (x^{k+1} - \psi)^+ \leq \frac{\|K\|}{c} \rho_1 |\mu^k_{\mathcal{I}_k}| + \rho_2 |\delta x_{\mathcal{A}_k}|.$$

An analogous estimate can be obtained for $\int_{\mathcal{I}_k} (x^{k+1} - \varphi)^-$, and we find

$$\int_{\mathcal{I}_k} (x^{k+1} - \psi)^+ \int_{\mathcal{I}_k} (x^{k+1} - \varphi)^- \leq \frac{\|K\|}{c} \rho_1 |\mu^k_{\mathcal{I}_k}| + \rho_2 |\delta x_{\mathcal{A}_k}|.$$

We can now proceed as in the proof of Theorem 4.1. ∎

Example 4.1. We apply Theorem 4.2 with $A = I + K \in \mathcal{L}(L^1(\Omega))$. By Neumann series arguments we find $\bar{\rho} = 2\max(\frac{\gamma}{1-\gamma}, \max(\gamma + \frac{\gamma^2}{1-\gamma}, \frac{\gamma}{1-\gamma})) = \frac{\gamma}{1-\gamma}$, where $\gamma = \|K\|$, and $\bar{\rho} < 1$ if $\|K\| < \frac{1}{3}$. If $A = I + K$ is replaced with $A = cI + K$, then $\bar{\rho} < 1$ if $\gamma < \frac{c}{2c+1}$, in case $c \geq 1$, and $\bar{\rho} < 1$ if $\frac{c^2}{c+2}$, in case $c \leq 1$.

Example 4.2. Consider the finite dimensional case $A = I + K \in \mathbb{R}^{n \times n}$, where $\mathbb{R}^n$ is endowed with the ℓ^1-norm. Again Theorem 4.2 is applicable and $\bar{\rho} < 1$, if $\|K\| < \frac{1}{3}$, where $\|\cdot\|$ denotes the matrix-norm subordinate to the ℓ^1-norm of $\mathbb{R}^n$. Recall that this norm is given by the maximum over the column sums of the absolute values of the matrix.

5. Nonlinear case. In this section we consider the nonlinear complementarity problem

$$x \in \mathcal{C}, \quad (f(x), y - x)_Z \geq 0 \text{ for } y \in \mathcal{C} = \{y \in Z : y \leq \psi\} \tag{5.1}$$

or equivalently,

$$f(x) + \mu = 0, \quad \mu = \max(0, \mu + x - \psi), \tag{5.2}$$

where f maps $L^2(\Omega)$ into itself and is C^1 from $L^1(\Omega)$ to itself. In this case step 3 of the primal-dual active set method is given by the following:

(3) Solve for (x^{k+1}, μ^{k+1})

$$\begin{aligned} &f'(x^k)(x^{k+1} - x^k) + f(x^k) + \mu^{k+1} = 0, \\ &x^{k+1} = \psi \text{ on } \mathcal{A}_k, \quad \mu^{k+1} = 0 \text{ on } \mathcal{I}_k. \end{aligned} \tag{5.3}$$

Let the pair (x^0, μ^0) satisfy $(x^0 - \psi)\mu^0 = 0$ (pointwise).

Throughout this section we assume that (5.3) admits a solution $(x^{k+1}, \mu^{k+1}) \in L^1(\Omega) \times L^1(\Omega)$, for every $k \geq 0$. We further assume that (3.4)–(3.5) hold and that

$$\|A_{\mathcal{I}}^{-1}\| \leq \hat{\rho}_1, \ \ \|A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}}\| \leq \hat{\rho}_2, \quad \text{and} \quad \|A_{\mathcal{A}\mathcal{I}} A_{\mathcal{I}}^{-1}\| \leq \hat{\rho}_4, \tag{5.4}$$

with $A = f'(x)$ for all $x \in B(x^0, r)$ and all partitions $\mathcal{A} \cup \mathcal{I} = \Omega$. Here $B(x^0, r)$ denotes the open ball with center x^0 and radius r. Let $A^0 = f'(x^0)$. Then, $\delta x = x^1 - x^0$ and $\delta\mu = \mu^1 - \mu^0$ satisfy

$$A^0_{\mathcal{A}_0} \delta x_{\mathcal{A}_0} + A^0_{\mathcal{A}_0 \mathcal{I}_0} \delta x_{\mathcal{I}_0} + \delta\mu_{\mathcal{A}_0} + (f(x^0) + \mu^0)_{\mathcal{A}_0} = 0,$$

$$A^0_{\mathcal{I}_0} \delta x_{\mathcal{I}_0} + A^0_{\mathcal{I}_0 \mathcal{A}_0} \delta x_{\mathcal{A}_0} - \mu^0_{\mathcal{I}_0} + (f(x^0) + \mu^0)_{\mathcal{I}_0} = 0.$$

Since $x^0 - \psi > 0$ on $\mathcal{A}_0$ and $\mu^0 \leq 0$ on $\mathcal{I}_0$, $\delta x = \psi - x^0 < 0$ on $\mathcal{A}_0$ and $\delta\mu = 0 - \mu^0 \geq 0$ on $\mathcal{I}_0$. Referring to the arguments in the proof of Theorem 3.1, it follows from (3.4)–(3.5) and (5.4) that

$$\begin{aligned} &|\delta x_{\mathcal{I}_0}| \leq \hat{\rho}_1 \int_{\mathcal{I}_0} (|\mu^0| + |(f(x^0) + \mu^0)|)\, dx + \hat{\rho}_2\, |(x^0 - \psi)_{\mathcal{A}_0}|, \\ &|(\delta x_{\mathcal{I}_0})^+| \leq C_1 = \rho_1 \int_{\mathcal{I}_0} |\mu^0|\, dx + \rho_2 \int_{\mathcal{A}_0} |x^0 - \psi|\, dx + \hat{\rho}_1\, |(f(x^0) + \mu^0)_{\mathcal{I}_0}|, \\ &|(\delta\mu_{\mathcal{A}_0})^-| \leq C_2 = (\rho_3 + \rho_5)|(x^0 - \psi)_{\mathcal{A}_0}| + \rho_4 |\mu^0_{\mathcal{I}_0}| + |(f(x^0) + \mu^0)_{\mathcal{A}_0}| \\ &\qquad\qquad + \hat{\rho}_4\, |(f(x^0) + \mu^0)_{\mathcal{I}_0}|. \end{aligned} \tag{5.5}$$

Thus,

$$\begin{aligned} &|x^1 - x^0| \leq C_3 = \hat{\rho}_1 |\mu^0_{\mathcal{I}_0}| + (\hat{\rho}_2 + 1)\, |(x^0 - \psi)_{\mathcal{A}_0}| + \hat{\rho}_1\, |(f(x^0) + \mu^0)_{\mathcal{I}_0}|, \\ &\mathcal{M}(x^1, \mu^1) \leq \max(\beta\, C_1, C_2). \end{aligned} \tag{5.6}$$

Next, for $k = 1, 2, \ldots$, let $A^k = f'(x^k)$ and $\delta x = x^{k+1} - x^k$, $\mu = \mu^{k+1} - \mu^k$, and note that

$$A^k_{\mathcal{A}_k} \delta x_{\mathcal{A}_k} + A^k_{\mathcal{A}_k \mathcal{I}_k} \delta x_{\mathcal{I}_1} + \delta\mu_{\mathcal{A}_k} + \Delta_{\mathcal{A}_k} = 0,$$

$$A^k_{\mathcal{I}_k} \delta x_{\mathcal{I}_k} + A^k_{\mathcal{I}_k \mathcal{A}_k} \delta x_{\mathcal{A}_k} - \mu^k_{\mathcal{I}_k} + \Delta_{\mathcal{I}_k} = 0,$$

where $\Delta = f(x^k) - f(x^{k-1}) - f'(x^{k-1})(x^k - x^{k-1})$. We now assume that there exists γ such that for $x^k,\ x^{k-1} \in B(x^0, r)$

$$|\Delta| = |f(x^k) - f(x^{k-1}) - f'(x^{k-1})(x^k - x^{k-1})| \le \gamma\, |x^k - x^{k-1}|.$$

Again, from (3.4)–(3.5) and (5.4)

$$|\delta x_{\mathcal{I}_k}| \le \hat\rho_1 \left(\int_{\mathcal{I}_k} |\mu^k|\, dx + |\Delta_{\mathcal{I}_k}| \right) + \hat\rho_2 \int_{\mathcal{A}_k} |x^k - \psi|\, dx,$$

$$|(\delta x_{\mathcal{I}_k})^+| \le \rho_1 \int_{\mathcal{I}_k} |\mu^k|\, dx + \rho_2 \int_{\mathcal{A}_k} |\delta x|\, dx + \hat\rho_1\, |\Delta_{\mathcal{I}_k}|,$$

$$|(\delta \mu_{\mathcal{A}_k})^-| \le (\rho_3 + \rho_5) \int_{\mathcal{A}_k} |x^k - \psi|\, dx + \rho_4 \int_{\mathcal{I}_k} |\mu^k|\, dx + |\Delta_{\mathcal{A}_k}| + \hat\rho_4\, |\Delta_{\mathcal{I}_k}|.$$

Thus,

$$|x^{k+1} - x^k| \le \left(\frac{1}{\beta} + \hat\rho_1 + \frac{\hat\rho_2}{\beta} \right) \mathcal{M}(x^k, \mu^k) + \gamma \hat\rho_1\, |x^k - x^{k-1}|,$$

(5.7)

$$\mathcal{M}(x^{k+1}, \mu^{k+1}) \le \rho\, \mathcal{M}(x^k, \mu^k) + 2\hat\rho_1\, \gamma\, \max(\beta\, \hat\rho_1, 1, \hat\rho_4)\, \frac{|x^k - x^{k-1}|}{2\hat\rho_1},$$

where we used

$$\int_{\mathcal{A}_k} |x^k - \psi|\, dx \le \int_{\mathcal{I}_{k-1}} |x^k - \psi|\, dx \ \text{ and } \ \int_{\mathcal{I}_k} |\mu^k|\, dx \le \int_{\mathcal{A}_{k-1}} |\mu^k|\, dx,$$

and ρ is defined as in Theorem 3.1.

If we set

$$\omega = \max \left(\rho + 2\hat\rho_1 \gamma\, \max(\beta\, \hat\rho_1, 1, \hat\rho_4), \frac{1}{2} + \frac{1 + \hat\rho_2}{2\beta\hat\rho_1} + \gamma\hat\rho_1 \right), \tag{5.8}$$

then

$$\max \left(\frac{|x^{k+1} - x^k|}{2\hat\rho_1}, \mathcal{M}(x^{k+1},\, \mu^{k+1}) \right) \le \omega\, \max \left(\frac{|x^k - x^{k-1}|}{2\hat\rho_1},\, \mathcal{M}(x^k, \mu^k) \right). \tag{5.9}$$

Hence if $\omega < 1$ and $x^j \in B(x^0, r)$, for $j \le k$, then

$$\max \left(\frac{|x^{k+1} - x^k|}{2\hat\rho_1}, \mathcal{M}(x^{k+1}, \mu^{k+1}) \right) \le \omega^k\, \max \left(\frac{|x^1 - x^0|}{2\hat\rho_1},\, \mathcal{M}(x^1, \mu^1) \right)$$

and thus,

$$|x^{k+1} - x^0| \le \frac{1}{1 - \omega}\, \max(\, |x^1 - x^0|, 2\hat\rho_1\, \mathcal{M}(x^1, \mu^1)\,). \tag{5.10}$$

Let C_1, C_2, C_3 be as defined in (5.5)–(5.6), and assume that

$$\frac{2\hat\rho_1}{1 - \omega}\, \max \left(\beta\, C_1, C_2, \frac{C_3}{2\hat\rho_1} \right) \le r. \tag{5.11}$$

Then it follows from (5.6) and (5.10) that $|x^{k+1} - x^0| \le r$, and thus by induction in k we have $x^k \in B(x^0, r)$ for all $k \ge 1$ and (5.9) holds for all k. Hence x^k is Cauchy, and by (5.3) μ^k is a Cauchy sequence as well. It follows that $\lim(x^k, \mu^k)$ converges to a limit (x^*, μ^*) in $L^1(\Omega) \times L^1(\Omega)$. As in the proof of Theorem 3.1 one argues that (x^*, μ^*) satisfies (5.2). Moreover from (5.9)

$$|x^m - x^k| \le \sum_{j=k}^{m-1} |x^{j+1} - x^j| \le \sum_{j=k+1}^{m} \omega^{j-k} \max(|x^k - x^{k-1}|, 2\hat{\rho}_1 \mathcal{M}(x^k, \mu^k))$$

$$\le \frac{\omega}{1-\omega} \max(|x^k - x^{k-1}|, 2\hat{\rho}_1 \mathcal{M}(x^k, \mu^k)),$$

and therefore

$$|x^k - x^*| \le \frac{\omega}{1-\omega} \max(|x^k - x^{k-1}|, 2\hat{\rho}_1 \mathcal{M}(x^k, \mu^k)). \tag{5.12}$$

In summary we obtain the following.

THEOREM 5.1. *Given (x^0, μ^0) and $r > 0$, we assume that* (3.4)–(3.5) *and* (5.4) *hold with $A = f'(x)$ for all $x \in B(x^0, r)$ and all partitions $\mathcal{A} \cup \mathcal{I} = \Omega$, and that for $x, y \in B(x^0, r)$*

$$|f(x) - f(y) - f'(x)(x - y)| \le \gamma |x - y|.$$

Suppose further that $\omega < 1$, with ω as defined in (5.8)*, that (x^0, μ^0) satisfies $(x^0 - \psi)\mu_0 = 0$, and that* (5.11) *holds. Then (x^k, μ^k) converges in $L^1(\Omega) \times L^1(\Omega)$ to a solution (x^*, μ^*) of* (5.2)*, and* (5.12) *holds.*

Remark 5.1. If f is C^2, then γ can be chosen proportionally to r. Thus, referring to the choice of A discussed in Remark 3.1, we can choose $r > 0$ sufficiently small and β sufficiently large so that $\omega < 1$ provided that $\rho < 1$, where ρ is as defined in Theorem 3.1. Alternatively, referring to the situation in Remark 3.2 we have $\omega < 1$ provided that $\rho < 1$, and r and $\hat{\rho}_2$ are sufficiently small.

6. Globalization. In this section we consider the globalization of the primal-dual algorithm for the unilateral constraint case in the finite dimensional case, i.e., $X = \mathbb{R}^d$. The extension to the bilateral case is straightforward. The variational inequality (5.1) is equivalent to (5.2), i.e., $F(x, \mu) = 0$, where

$$F(x, \mu) = \begin{cases} f(x) + \mu, \\ \\ \max(0, \mu + (x - \psi)) - \mu = \max(-\mu, x - \psi), \end{cases}$$

where we set $c = 1$. We introduce

$$\theta(x, \mu) = |F(x, \mu)|^2$$

as a merit function. If we let

$$\mathcal{A} = \{\mu + (x - \psi) > 0\}, \ \mathcal{I}_1 = \{\mu + (x - \psi) = 0\}, \ \text{and} \ \mathcal{I}_2 = \{\mu + (x - \psi) < 0\},$$

then the directional derivative of F is given by

$$F'((x,\mu);d) = \begin{cases} f'(x)\delta x + \delta\mu, \\ \begin{pmatrix} \delta x & \text{on } \mathcal{A} \\ -\delta\mu & \text{on } \mathcal{I}_2 \\ \max(-\delta\mu, \delta x) & \text{on } \mathcal{I}_1 \end{pmatrix} \end{cases}$$

with $d = (\delta x, \delta\mu)$. Here and throughout this section we assume that $f \in C^1(\mathbb{R}^d, \mathbb{R}^d)$. We find that

$$\frac{1}{2}\,\theta'(x,\mu;d) = -|f(x)+\mu|^2 + ((x-\psi),\delta x)_{\mathcal{A}} + (\mu,\delta\mu)_{\mathcal{I}_2} - (\mu, \max(-\delta\mu,\delta x))_{\mathcal{I}_1}$$

if the direction $d = (\delta x, \delta\mu)$ satisfies

$$f'(x)\delta x + \delta\mu + f(x) + \mu = 0.$$

This implies that

$$\frac{1}{2}(|f(x)+\mu|^2)'(d) = (f(x)+\mu, f'(x;\delta x) + \delta\mu) = -|f(x)+\mu|^2.$$

We assume that the following assumptions (6.1)–(6.4) are satisfied:

(6.1) $\qquad S = \{(x,\mu) \in \mathbb{R}^d \times \mathbb{R}^d : \theta(x,\mu) \le |F(x^0,\mu^0)|\}$ is bounded.

There exists $\bar\sigma,\ b > 0$ such that for all $(x,\mu) \in S$ there is a direction $d = (\delta x, \delta\mu)$ depending on (x,μ) such that

(6.2) $$\begin{aligned} &f'(x)\delta x + \delta\mu + f(x) + \mu = 0, \\ &\theta'((x,\mu);d) \le -\bar\sigma\theta(x,\mu), \quad \text{and} \quad |d| \le b\,|F(x,\mu)|. \end{aligned}$$

(6.3) $$\begin{aligned} &\theta \text{ is subdifferentiably regular for all } x \in S, \text{ i.e.,} \\ &\theta^o(x;d) = \theta'(x;d) \text{ for all } d \in \mathbb{R}^d. \end{aligned}$$

Moreover we assume the following closure property:

(6.4) $\qquad$ If $(x,\mu) \to (\bar x, \bar\mu)$ and $d = d(x,\mu) \to \bar d$, then $\theta'((\bar x,\bar\mu);\bar d) \le -\bar\sigma\theta(\bar x,\bar\mu)$.

In (6.3) above the Clarke generalized directional derivative $\theta^o(x;d)$ of θ at x in the direction d is defined by

$$\theta^o(x;d) = \limsup_{y\to x,\ t\to 0^+} \frac{\theta(y+t\,d) - \theta(y)}{t}.$$

In (6.4) the direction $d = d(x,\mu)$ is chosen according to (6.2), but $\bar d$ need not be related to $(\bar x,\bar\mu)$.

ALGORITHM.

(i) Let β, $\gamma \in (0,1)$, and $\sigma \in (0, \bar{\sigma})$.

(ii) Apply the primal-dual active set method to determine (x^{k+1}, μ^{k+1}):

$$f'(x^k)(x^{k+1} - x^k) + f(x^k) + \mu^{k+1} = 0,$$

$$x^{k+1} = \psi \text{ on } \mathcal{A}_k, \quad \mu^{k+1} = 0 \text{ on } \mathcal{I}_k.$$

If $|F(x^{k+1}, \mu^{k+1})| < \gamma\, |F(x^k, \mu^k)|$ and $|(x^{k+1}, \mu^{k+1}) - (x^k, \mu^k)| \le b|F(x^k, \mu^k)|$, set $k = k+1$ and skip (iii).

(iii) Let d^k be a descent direction for θ at (x^k, μ^k) according to (6.2) and set $\alpha_k = \beta^{m_k}$, where m_k is the first positive integer m for which

$$\theta((x^k, \mu^k) + \beta^m\, d^k) - \theta(x^k, \mu^k) \le -\sigma\beta^m\, \theta(x^k, \mu^k).$$

Set $(x^{k+1}, \mu^{k+1}) = (x^k, \mu^k) + \alpha^k\, d^k$, $k = k+1$, and go to (ii).

THEOREM 6.1. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}^d$ is C^1.*

(a) *Assume* (6.1)–(6.4) *hold. Then, the sequence $\{x^k\}$ generated by the algorithm is bounded, $|F(x^{k+1}, \mu^{k+1})| < |F(x^k, \mu^k)|$, for all $k \ge 0$, and each accumulation point (x^*, μ^*) of $\{(x^k, \mu^k)\}$ satisfies $F(x^*, \mu^*) = 0$.*

(b) *Moreover if for one such accumulation point*

$$|h| \le c\, |F'(x^*, \mu^*; h)| \quad \text{for all } h \in \mathbb{R}^d, \tag{6.5}$$

then the sequence (x^k, μ^k) converges superlinearly to (x^*, μ^*).

Proof. The proof follows from Theorem 2.1 in [IK2]. □

In [IK2] the reduced form of $F(x, \mu) = 0$, which is given by

$$-f(x) = \max(0, -f(x) + x - \psi),$$

is analyzed. In this case the new active set is determined on the basis of $\tilde{\mathcal{A}} = \{-f(x^k) + x^k - \psi > 0\}$. If the full step is taken, then this differs from the new active set $\mathcal{A}$ used in this paper in that $\mu^k = -f'(x^{k-1})(x^k - x^{k-1}) - f(x^{k-1})$ is replaced with $-f(x^k)$.

For related results on globalization of semismooth Newton methods applied to linear and nonlinear complementarity problems, we refer to [HP] and [DFK].

6.1. Descent directions. We turn to a discussion of directions which are descent directions satisfying (6.2)–(6.4). As for (6.3) we recall from [C] that convex locally Lipschitz continuous functions are subdifferentiably regular, and thus we concentrate on (6.2) and (6.4).

The direction $d = (\delta x, \delta\mu)$ defined by the primal-dual active set method satisfies

$$f'(x)\delta x + \delta\mu + f(x) + \mu = 0, \quad \delta x + x - \psi = 0 \text{ on } \mathcal{A}, \quad \delta\mu + \mu = 0 \text{ on } \mathcal{I}, \tag{6.6}$$

where we suppress the iteration index k. To estimate θ' observe that

$$((x - \psi), \delta x)_{\mathcal{A}} + (\mu, \delta\mu)_{\mathcal{I}_2} - (\mu, \max(-\delta\mu, \delta x))_{\mathcal{I}_1}$$

$$= -|x - \psi|^2_{\mathcal{A}} - |\mu|^2_{\mathcal{I}_2} - (\mu, \max(-\delta\mu, \delta x))_{\mathcal{I}_1}.$$

If $\delta\mu + \delta x \le 0$ on $\mathcal{I}_1$, then

$$-(\mu, \max(-\delta\mu, \delta x))_{\mathcal{I}_1} = (\mu, \delta\mu)_{\mathcal{I}_1} = -|\mu|^2_{\mathcal{I}_1}.$$

If $\delta\mu + \delta x > 0$ and $\delta\mu \le 0$ on $\mathcal{I}_1$, then

$$-(\mu, \max(-\delta\mu, \delta x))_{\mathcal{I}_1} = -(\mu, \delta x)_{\mathcal{I}_1} \le -|\mu|^2_{\mathcal{I}_1}.$$

Hence if

$$\{\delta\mu + \delta x > 0\} \cap \{\delta\mu > 0\} \cap \mathcal{I}_1 = \emptyset, \tag{6.7}$$

then

$$\theta'((x,\mu),(\delta x,\delta\mu)) \le -2\,|F(x,\mu)|^2,$$

and thus $d = (\delta x, \delta\mu)$ defined by (6.6) satisfies the first condition in (6.2). Note that if $(x,\mu) = (x^k,\mu^k)$ is chosen according to the primal-dual active set strategy without a subsequent line search, then $\mu^k = x^k - \psi = 0$ and $(\delta\mu)^{k+1} = 0$ on $(\mathcal{I}_k)_1$. In this case, (6.7) reduces to the assumption that $\{x^{k+1} > \psi\}$ on $(\mathcal{I}_k)_1$. As for the second condition in (6.2), let $R_{\mathcal{I}} : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{I}|}$ denote the restriction matrix from $\mathbb{R}^d$ to the coordinates of active indices, and set $A_{\mathcal{I}}(x) = R_{\mathcal{I}} f'(x) R_{\mathcal{I}}^T$. If $\|(A_{\mathcal{I}}(x))^{-1}\| \le K$ for a constant K independent of $x \in S$ and the combination of inactive indices $\mathcal{I}$, then a Schur complement argument shows that the second condition in (6.2) holds.

Turning to the closure property (6.4), let $(x,\mu) \to (\bar{x},\bar{\mu})$ and $d(x,\mu) \to \bar{d} = (\overline{\delta x}, \overline{\delta\mu})$. As a consequence of the assignment on the active and inactive sets as expressed in (ii) of the algorithm, we have

$$(\bar{x} + \overline{\delta x} - \psi)(\bar{\mu} + \overline{\delta\mu}) = 0. \tag{6.8}$$

Let $\bar{\mathcal{A}} = \{\bar{x} - \psi + \bar{\mu} > 0\}$, $\bar{\mathcal{I}}_1 = \{\bar{x} - \psi + \bar{\mu} = 0\}$, and $\bar{\mathcal{I}}_2 = \{\bar{x} - \psi + \bar{\mu} < 0\}$. We shall give a sufficient condition which guarantees that $\theta'((\bar{x},\bar{\mu}),\bar{d}) \le -2\,|F(\bar{x},\bar{\mu})|^2$. The estimates on $\bar{\mathcal{A}}$ and $\bar{\mathcal{I}}_2$ are simple, and we turn to $\bar{\mathcal{I}}_1$. We assume that

$$\begin{aligned} &\{\overline{\delta\mu} + \overline{\delta x} < 0\} \cap \{\overline{\delta x} < 0\} \cap \mathcal{I}_1 = \emptyset \quad \text{and} \\ &\{\overline{\delta\mu} + \overline{\delta x} > 0\} \cap \{\overline{\delta\mu} > 0\} \cap \mathcal{I}_1 = \emptyset. \end{aligned} \tag{6.9}$$

If $\overline{\delta\mu} + \overline{\delta x} \le 0$ on $\mathcal{I}_1$, we have

$$-(\bar{\mu}, \max(-\overline{\delta\mu}, \overline{\delta x}))_{\mathcal{I}_1} = (\bar{\mu}, \overline{\delta\mu})_{\mathcal{I}_1}.$$

If moreover the second factor in (6.8) is 0, then

$$-(\bar{\mu}, \max(-\overline{\delta\mu}, \overline{\delta x}))_{\mathcal{I}_1} = -|\bar{\mu}|^2_{\mathcal{I}_1},$$

as desired. Otherwise $(\bar{x} + \overline{\delta x} - \psi) = -\bar{\mu} + \overline{\delta x} = 0$, and hence

$$-(\bar{\mu}, \max(-\overline{\delta\mu}, \overline{\delta x}))_{\mathcal{I}_1} = (\bar{\mu}, \overline{\delta\mu})_{\mathcal{I}_1} = (\overline{\delta x}, \overline{\delta\mu}),$$

and by (6.9)

$$-(\bar{\mu}, \max(-\overline{\delta\mu}, \overline{\delta x}))_{\mathcal{I}_1} \le -|\overline{\delta x}|^2_{\mathcal{I}_1} = -|\bar{\mu}|^2_{\mathcal{I}_1}.$$

The case $\overline{\delta\mu} + \overline{\delta x} > 0$ can be treated analogously and we arrive at $\theta'((\bar{x},\bar{\mu}),\bar{d}) \le -2\,|F(\bar{x},\bar{\mu})|^2$, as announced.

So far we provided sufficient conditions which guarantee that the direction given by the primal-dual active set strategy can serve as a descent direction. We proceed by describing alternative methods for selecting descent directions d. This will allow us to circumvent an assumption of types (6.7) and (6.9) at the expense of solving a nonlinear equation.

We commence by defining the function $G((x,\mu);d)$, with $d=(\delta x,\delta\mu)$ as follows:

$$G((x,\mu);d)=\begin{cases} f'(x)\delta x+\delta\mu, \\ \begin{pmatrix} \delta x & \text{on } S_1=\mathcal{A}\cap\{\mu\geq 0\} \\ -\delta\mu & \text{on } S_2=\mathcal{I}_2\cap\{x\leq\psi\} \\ \max(-\delta\mu,\delta x) & \text{on } S_3=(\mathcal{A}\cap\{\mu<0\})\cup(\mathcal{I}_2\cap\{x>\psi\})\cup\mathcal{I}_1. \end{pmatrix}. \end{cases}$$

Note that for a full step based on the primal-dual active set strategy without a subsequent line search we have $S_1=\mathcal{A}$ and $S_2=\mathcal{I}_2$ since $(x-\psi)\mu=0$. This structure is not maintained if the line search of step (iii) is used.

The mapping $G((x,\mu);d)$ is a quasi-directional derivative of $F(x,\mu)$ in the sense that

(1) $(F(x,\mu),F'((x,\mu);d))\leq(F(x,\mu),G((x,\mu);d))$ for all $(x,\mu)\in S$ and $d=(\delta x,\delta\mu)$;
(2) $G((x,\mu),td)=t\,G((x,\mu);d)$ for $t\geq 0$ and all $(x,\mu)\in S$ and d;
(3) $\lim_{((x,\mu),d)\to((\bar x,\bar\mu),\bar d)}(F(x,\mu),G((x,\mu);d))\geq(F(\bar x,\bar\mu),G((\bar x,\bar\mu);\bar d))$.

In fact, to verify (1) note that on $(\mathcal{A}\cap\{\mu<0\})\cup(\mathcal{I}_2\cap\{x>\psi\})$ we have $\max(-\mu,x-\psi)>0$, and thus

$$\begin{aligned}(\max(\max(-\mu,x-\psi),-\delta\mu),(\max(-\mu,x-\psi),\delta x))\\ \leq(\max(-\mu,x-\psi),\max(-\delta\mu,\delta x)).\end{aligned}$$

Hence we obtain

$$(F(x,\mu),F'((x,\mu);d))\leq(F(x,\mu),G((x,\mu);d)).$$

Property (2) is obvious. Next suppose that $(x,\mu)\to(\bar x,\bar\mu)$, $d\to\bar d$. For indices i such that $(\bar\mu+\bar x-\psi)_i\neq 0$, we have

$$G((x,\mu);d)_i\to G((\bar x,\bar\mu);\bar d)_i.$$

For indices i such that $(\bar\mu+\bar x-\psi)_i=0$ the case $x-\psi>0,\mu>0$ (and analogously $x-\psi<0,\mu<0$) for infinitely many (x,μ) in a neighborhood of $(\bar x,\bar\mu)$ is trivial since in this case $(F^{(2)}(\bar x,\bar\mu))_i=0$. We need to examine two more cases. First, for i such that $(\mu+x-\psi)_i>0$ and $\mu_i\geq 0$, we have $\max(-\mu,x-\psi)_i=(x-\psi)_i$, and thus $\max(-\bar\mu,\bar x-\psi)_i=(\bar x-\psi)_i=-(\bar\mu)_i\leq 0$ and

$$F^{(2)}(x,\mu)_iG^{(2)}((x,\mu);d)_i\to F^{(2)}(\bar x,\bar\mu)_i\overline{\delta x}_i\geq F^{(2)}(\bar x,\bar\mu)_i\max(-\overline{\delta\mu},\overline{\delta x})_i.$$

Second, for indices i such that $(\mu+x-\psi)_i<0$ and $x_i\leq\psi_i$, $\max(-\mu,x-\psi)_i=-\mu_i\leq 0$, and thus $\max(-\bar\mu,\bar x-\psi)_i=-\bar\mu_i=(\bar x-\psi)_i\leq 0$ and

$$F^{(2)}(x,\mu)_iG^{(2)}((x,\mu);d)_i\to F^{(2)}(\bar x,\bar\mu)_i(-\overline{\delta\mu})_i\geq F^{(2)}(\bar x,\bar\mu)_i\max(-\overline{\delta\mu},\overline{\delta x})_i.$$

Hence we obtain

$$\lim_{(x,\mu)\to(\bar{x},\bar{\mu}),\ d\to\bar{d}} (F(x,\mu), G((x,\mu);d)) \geq (F(\bar{x},\bar{\mu}), G((\bar{x},\bar{\mu});\bar{d})),$$

(3) is satisfied, and G is a quasi-directional derivative.

We turn to the description of two methods for selecting a descent direction using the quasi-directional derivative G.

Method 1 (Bouligand direction). Find $d = (\delta x, \delta\mu)$ such that $G((x,\mu);d) + F(x,\mu) = 0$, i.e.,

$$\left\{ \begin{array}{l} f'(x)\delta x + \delta\mu + f(x) + \mu = 0, \\ \left(\begin{array}{ll} \delta x + x - \psi = 0 & \text{on } S_1 \\ \delta\mu + \mu = 0 & \text{on } S_2 \\ \max(-\mu, x-\psi) + \max(-\delta\mu, \delta x) = 0 & \text{on } S_3. \end{array} \right) \end{array} \right. \tag{6.10}$$

Assume that for each $(x,\mu) \in S$

$$G((x,\mu);d) + F(x,\mu) = 0 \text{ has a solution } d = d(x,\mu) \tag{6.11}$$

satisfying

$$|d| \leq b\,|G((x,\mu);d)|. \tag{6.12}$$

For such d we have using (1) that

$$\theta'(x;d) = 2\,(F(x,\mu), F'((x,\mu),d)) \leq 2\,(F(x,\mu), G((x,\mu);d)) = -2\,\theta(x).$$

Moreover the closure property holds, i.e.,

$$\theta'(\bar{x},\bar{d}) \leq 2\,(F(\bar{x},\bar{\mu}), G((\bar{x},\bar{\mu}),\bar{d})) \leq 2 \lim_{(x,\mu)\to(\bar{x},\bar{\mu}),d\to\bar{d}} (F(x), G((x,\mu);d))$$

$$= -2 \lim_{(x,\mu)\to(\bar{x},\bar{\mu})} |F(x,\mu)|^2 = -2\theta(\bar{x},\bar{\mu}),$$

where we used (3). Together with (6.12) this implies that (6.2) and (6.4) hold.

Method 2 (gradient direction). For $(x,\mu) \in S$, $d = (\delta x, \delta\mu)$ is chosen as the solution to

$$\min_d \quad (F(x,\mu), G((x,\mu);d)) + \frac{\hat{\sigma}}{2}\,|d|^2. \tag{6.13}$$

If d is an optimal solution, then $\alpha = 1$ is optimal for

$$\min_{\alpha>0} \quad \alpha\,(F(x), G((x,\mu);d)) + \frac{\alpha^2\hat{\sigma}}{2}\,|d|^2,$$

and thus differentiating this with respect α we have

$$(F(x), G(x,\mu);d) + \hat{\sigma}\,|d|^2 = 0.$$

Hence

$$\begin{aligned}\theta'((x,\mu);d) &= 2(F(x,\mu), F'((x,\mu);d)) \\ &\le 2(F(x), G((x,\mu);d)) \le -2\hat\sigma\,|d|^2\end{aligned} \tag{6.14}$$

and

$$\hat\sigma\,|d|^2 \le -(F(x,\mu), F'((x,\mu);d)) \le |F'((x,\mu);d)||F(x,\mu)| \le M\,|d||F(x,\mu)|, \tag{6.15}$$

where M denotes the bound of $F'(x,\mu)$ on S. This gives the second condition in (6.2) with $b = \frac{M}{\hat\sigma}$. Moreover, if (6.11)–(6.12) hold, then

$$\begin{aligned}\frac{1}{2}\theta'((x,\mu);d) &= (F(x,\mu), F'((x,\mu);d)) \le (F(x,\mu), G((x,\mu);d)) \\ &= (F(x,\mu), G((x,\mu);d)) + \frac{\hat\sigma}{2}\,|d|^2 \le (F(x), G((x,\mu);\hat d)) + \frac{\hat\sigma}{2}\,|\hat d|^2 \\ &\le |F(x)|^2 + \frac{\hat\sigma}{2}b^2\,|F(x)|^2 = -\frac{1}{2}(2-\hat\sigma b^2)\,\theta(x),\end{aligned} \tag{6.16}$$

where $G((x,\mu);\hat d) + F(x,\mu) = 0$ and we assume that $2-\hat\sigma b^2 > 0$. Thus, the direction d defined by (6.13) satisfies (6.2) with $\bar\sigma = (2-\bar\sigma b^2))$. To verify the closure property note that

$$\begin{aligned}\frac{1}{2}\theta'(\bar x,\bar d) &= (F(\bar x,\bar\mu), F'((\bar x,\bar\mu);\bar d)) \le (F(\bar x,\bar\mu), G((\bar x,\bar\mu);\bar d)) \\ &= (F(\bar x,\bar\mu), G((\bar x,\bar\mu);\bar d)) + \frac{\hat\sigma}{2}|\bar d|^2 \le \lim_{(x,d)\to(\bar x,\bar\mu),d\to\bar d}(F(x,\mu), G((x,\mu),d)) + \frac{\hat\sigma}{2}|d|^2.\end{aligned}$$

On the other hand, we have

$$2(F(x,\mu), G((x,\mu);d)) + \hat\sigma|d|^2 \le 2(F(x,\mu), G((x,\mu);\hat d)) + \hat\sigma|\hat d|^2 \le -(2-\hat\sigma b^2)\,\theta(x).$$

It thus follows that

$$\theta'((\bar x,\bar\mu);\bar d) \le -(2-\hat\sigma b^2)\,\theta(\bar x).$$

Let $A = f'(x)$. Then, condition (6.10) reduces to a variational inequality for δx on S_3:

$$\max(B\delta x + b, \delta x) + \max(-\mu, x-\psi) = 0, \tag{6.17}$$

where

$$B = A_{S_3S_3} - A_{S_3S_2}A^{-1}_{S_2S_2}A_{S_2S_3}$$

and

$$b = -A_{S_3S_1}(x-\psi)_{S_1} - A_{S_3S_2}A^{-1}_{S_2S_2}f(x)_{S_2} + (f(x)+\mu)_{S_3}.$$

Hence (6.10) has a unique solution if A is symmetric and positive definite, for example. In fact (6.17) is equivalent to

$$\max(B\xi + z, \xi) = 0, \tag{6.18}$$

where $\xi = \delta x + c$, $z = b + c - Bc$, and $c = \max(-\mu, x - \psi)$. The variational inequality (6.18) in turn is equivalent to the equation for the Lagrange multiplier for the constrained problem

$$\begin{cases} \min \frac{1}{2}(B^{-1}y, y) + (B^{-1}z, y), \\ y \geq 0, \end{cases}$$

which clearly has a unique solution.

REFERENCES

[BP] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Computer Science and Applied Mathematics Series, Academic Press, New York, 1979.

[C] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.

[DFK] T. De Luca, F. Facchinei, and C. Kanzow, *A theoretical and numerical comparison of some semismooth algorithms for complementarity problems*, Comput. Optim. Appl., 16 (2000), pp. 173–205.

[FK] A. Fischer and C. Kanzow, *On finite termination of an iterative method for linear complementarity problems*, Math. Programming, 74 (1996), pp. 279–292.

[HIK] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.

[HP] P. T. Harker and J.-S. Pang, *A damped-Newton method for the linear complementarity problem*, in Computational Solution of Nonlinear Systems of Equations, E. L. Allgower and K. Georg, eds., Lectures in Appl. Math. 26, American Math. Soc., Providence, RI, 1990, pp. 265–284.

[IK1] K. Ito and K. Kunisch, *The primal-dual active set method for nonlinear optimal control problems with bilateral constraints*, SIAM J. Control Optim., 43 (2004), pp. 357–376.

[IK2] K. Ito and K. Kunisch, *On the semi-smooth Newton method and its globalization*, submitted.

[KR] K. Kunisch and F. Rendl, *An infeasible active set method for quadratic problems with simple bounds*, SIAM J. Optim., 14 (2003), pp. 35–52.

[MZ] H. Maurer and J. Zowe, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 35–60

MORSE DECOMPOSITIONS FOR GENERAL DYNAMICAL SYSTEMS AND DIFFERENTIAL INCLUSIONS WITH APPLICATIONS TO CONTROL SYSTEMS*

DESHENG LI†

Abstract. The Morse decomposition theory for single-valued dynamical systems is extended to general dynamical systems and differential inclusions. Relations between chain recurrent sets, chain recurrent components (chain control sets), and Morse decompositions for open-loop systems are discussed. The robustness of feedback laws of closed-loop systems with respect to small time delays and sample-hold controls is also addressed.

Key words. Morse decomposition, general dynamical system, differential inclusion, control system

AMS subject classifications. 37B25, 34E10, 93C10, 93C15

DOI. 10.1137/060662101

1. Introduction.

General dynamical systems (GDSs), also referred to as general control systems or set-valued dynamical systems, are basically used to describe the behavior of differential equations without uniqueness, differential inclusions, and control systems [6, 9, 12, 18, 28, 24, 31, 34, 36, 41] as well as dynamical economic phenomena [10]. They have been widely studied since the pioneering work in the 1960s [41, 42, 45].

In this article we want to take a step towards the Morse decomposition theory for GDSs as well as differential inclusions. Morse decomposition theory is a powerful tool for analyzing dynamical behavior of nonlinear systems inside invariant sets and attractors, and it has aroused an increasing interest in recent years. For single-valued dynamical systems, the Morse decomposition theory can be found in the original work of Conley [15] (see also [2] and [43]). Extensions to random dynamical systems can be found in the recent work of Crauel, Duc, and Siegmund [16] and Ochs [37], and extensions to nonautonomous dynamical systems can be found in Rasmussen [40].

Our purposes here are as follows. First, we extend the Morse decomposition theory for single-valued dynamical systems to GDSs. A key point in this part is to introduce a suitable notion of an attractor-repeller pair and overcome technical difficulties due to the lack of invariance properties of repellers, Morse sets, and limit sets (fortunately attractors of GDSs are invariant). One will see that the definition of an attractor-repeller pair given here for GDSs is somewhat different from that for single-valued systems. However, the two definitions coincide when we come back to the latter case.

Second, we are interested in the stability of Morse decompositions of attractors. Specifically, we prove that Morse decompositions of attractors for GDSs are stable under parameter perturbations (upper semicontinuity of Morse sets).

*Received by the editors June 5, 2006; accepted for publication (in revised form) September 26, 2006; published electronically February 23, 2007. This work was supported by NNSF of China (grant 10251002) and NSF of Gansu (grant 325041-A25-006).

http://www.siam.org/journals/sicon/46-1/66210.html

†Department of Mathematics, Tianjin University, Tianjin 300072, People's Republic of China (lidsmath@hotmail.com).

Third, we discuss Morse decompositions and their stability with respect to perturbations for differential inclusions by applying the abstract results on GDSs. Differential inclusions have a very rich background and successful applications in many different areas; see, for instance, [3, 4, 5, 11, 12, 17, 29], etc. In contrast to differential equations, the understanding of dynamical behavior for differential inclusions seems to be more difficult, and it has aroused an increasing interest in recent years.

Finally, we give some applications to control systems. First, we investigate relationships between chain recurrent sets, chain recurrent components (chain control sets), and Morse decompositions for open-loop systems under weaker assumptions. Then we discuss the robustness of feedback laws of closed-loop systems with respect to small time delays and sample-hold controls from the point of view of the stability of Morse decompositions.

This paper is organized as follows. Section 2 is concerned with preliminary work on GDSs. Section 3 is devoted to the Morse decompositions for GDSs. In section 4 we consider upper semicontinuity of Morse decompositions of attractors with respect to parameter perturbations for GDSs, and in section 5 we establish a Morse decomposition theory for differential inclusions. Section 6 consists of some applications to open-loop control systems. Robustness results on feedback laws of closed-loop systems are contained in section 7.

2. Preliminary work on GDSs. Let X be a complete locally compact metric space with metric $d(\cdot,\cdot)$. For any nonempty subsets A and B of X, define the Hausdorff semidistance and distance, respectively, as

$$d_{\mathrm{H}}(A,B)=\sup_{x\in A} d(x,B),\quad \delta_{\mathrm{H}}(A,B)=\max\left\{d_{\mathrm{H}}(A,B),d_{\mathrm{H}}(B,A)\right\},$$

where $d(x,B)=\inf_{b\in B} d(x,b)$. For convenience, we also assign $d_{\mathrm{H}}(\emptyset,B)=0$.

Let $A\subset X$. The closure of A is denoted by $\overline{A}$ or $\mathrm{cl}A$, and the interior of A is denoted by $\operatorname{int}A$. We denote by $\mathrm{B}(A,r)$ the neighborhood $\{y\in X:\ d(y,A)<r\}$ of radius $r>0$ of A. We say that a subset V of X is a neighborhood of A if $\overline{A}\subset \operatorname{int}V$.

Definition 2.1 (see [27]). *A set-valued mapping $G:\mathbb{R}^+\times X\to X$ with nonempty closed images is said to be a GDS if the following axioms hold:*

(1) *The semigroup property is*

$$G(0,x)=x,\quad G(t,G(s,x))=G(t+s,x)\qquad \forall x\in X,\ s,t\in\mathbb{R}^+;$$

(2) *$G(t,x)$ is continuous in t for each fixed x in the sense of Hausdorff distance;*

(3) *$G(t,x)$ is upper semicontinuous in x uniformly in t on any compact interval J.*

For a GDS G, we will also write $G(t,x)$ as $G(t)x$.

From now on we always assume that there has been given a GDS G on X.

For convenience, we denote by $G(I)V$ the set $\bigcup_{(t,x)\in I\times V} G(t)x$ for $V\subset X$ and $I\subset\mathbb{R}^+$.

Let A and V be two subsets of X. We say that A *attracts* V if

$$d_{\mathrm{H}}\left(G(t)V,\,A\right)\to 0\quad \text{as } t\to\infty.$$

The *attraction region* $\Omega(A)$ and *uniform attraction region* $\Omega_u(A)$ of A are defined, respectively, as

$$\Omega(A):=\{x\in X|\ \ A \text{ attracts } x\},$$

$$\Omega_u(A):=\{x\in X|\ \ A \text{ attracts a neighborhood } U \text{ of } x\}.$$

A is said to be *Lyapunov stable* if for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$G(\mathbb{R}^+)\mathrm{B}(A,\delta) \subset \mathrm{B}(A,\varepsilon).$$

If A is Lyapunov stable with $\Omega(A)$ (resp., $\Omega_u(A)$) being a neighborhood of A, then we say that A is *asymptotically stable* (resp., *uniformly asymptotically stable*).

Remark 2.2 (see [30]). If A is compact and asymptotically stable, then $\Omega(A) = \Omega_u(A)$. Therefore asymptotic stability and uniform asymptotic stability for compact sets are actually equivalent. We also know that A attracts each compact subset K of $\Omega_u(A)$.

Let $I \subset \mathbb{R}$ be an interval. A *trajectory* γ of G on I is a mapping $\gamma : I \to X$ satisfying $\gamma(t_2) \in G(t_2 - t_1)\gamma(t_1)$ for any $t_1, t_2 \in I$ with $t_1 \le t_2$.

In case $I = \mathbb{R}$, we will simply say that γ is a *complete trajectory*. A complete trajectory γ through $x \in X$ means a complete trajectory with $\gamma(0) = x$.

THEOREM 2.3 (see [41]). *Every trajectory is continuous. Further, let $y \in G(t_1 - t_0)x$, where $t_0 \le t_1$. Then there is a trajectory γ of G on $[t_0, t_1]$ such that $\gamma(t_0) = x$ and $\gamma(t_1) = y$.*

THEOREM 2.4 (see[41], Barbashin's theorem). *Let $[t_0, t_1]$ be a compact interval and γ_n be a sequence of trajectories of G on $[t_0, t_1]$ with $\gamma_n(t_0) \to x_0$. Then there is a subsequence γ_{n_i} and a trajectory γ_0 such that γ_{n_i} converges uniformly on $[t_0, t_1]$ to γ_0.*

A set $A \subset X$ is said to be *positively invariant*, *negatively invariant*, or *invariant*, if $G(t)A \subset A$, $G(t)A \supset A$, or $G(t)A = A$), respectively, for all $t \ge 0$.

A set A is said to be *weakly invariant* if, for any $x \in A$, there passes through x a complete trajectory γ with $\gamma(\mathbb{R}) \subset A$. In case A is compact, we infer from [30] that this amounts to saying that A is negatively invariant with $G(t)A \cap A \neq \emptyset$ for all $t \ge 0$.

The following basic fact is well known. See also [30].

PROPOSITION 2.5. *Let A be a nonempty compact subset of X. If A is negatively (resp., positively) invariant under G, then for any $x \in A$, there is a trajectory γ of G on $(-\infty, 0]$ (resp. $[0, \infty)$) which lies in A such that $\gamma(0) = x$.*

Let $A \subset X$, and let γ be a trajectory of G on $[a, +\infty)$. The *ω-limit sets* $\omega(A)$ and $\omega(\gamma)$ are defined, respectively, as

$$\omega(A) := \{y \in X : \ \exists t_n \to \infty \text{ and } y_n \in G(t_n)A \text{ such that } y_n \to y\},$$

$$\omega(\gamma) := \{x \in X| \ \ \exists t_n \to +\infty \text{ such that } \gamma(t_n) \to x\}.$$

Similarly one can define the *α-limit set* $\alpha(\gamma)$ of a trajectory γ on $(-\infty, a]$.

Remark 2.6. In case $G(\mathbb{R}^+)A$, $\gamma([a, \infty))$, and $\gamma((-\infty, a])$ are precompact, it can be easily shown that the limit sets defined above are nonempty compact weakly invariant sets [30].

DEFINITION 2.7. *Let $\mathcal{A}$ be a compact subset of X. If there is a neighborhood U of $\mathcal{A}$ such that $\mathcal{A} = \omega(U)$, then we say that $\mathcal{A}$ is an attractor of G, and U is called an attractor neighborhood of $\mathcal{A}$. A global attractor is an attractor $\mathcal{A}$ with $\Omega_u(\mathcal{A}) = X$.*

We allow the empty set $\emptyset$ to be an attractor of G with $\Omega_u(\emptyset) = \emptyset$.

Remark 2.8. In general limit sets of GDSs may fail to be invariant. However, if $\mathcal{A} = \omega(U)$ is an attractor of G, then it is invariant (see [30]).

An attractor $\mathcal{A}$ is necessarily uniformly asymptotically stable [30]. Thus by Remark 2.2 we have $\Omega(\mathcal{A}) = \Omega_u(\mathcal{A})$. So from now on we will not distinguish $\Omega(\mathcal{A})$ and $\Omega_u(\mathcal{A})$.

Let G_λ ($\lambda \in \Lambda$) be a family of GDSs, where Λ is a metric space with metric $\rho(\cdot, \cdot)$.

THEOREM 2.9 (see [31]). *Let $\lambda_0 \in \Lambda$. Assume the following continuity assumption holds:*

(C0) *For any $\varepsilon, T > 0$ and compact set A, there exists $\delta > 0$ such that, when $\rho(\lambda, \lambda_0) < \delta$,*

$$d_{\mathrm{H}}\left(G_\lambda(t)x,\, G_{\lambda_0}(t)B(x,\varepsilon)\right) < \varepsilon \qquad \forall x \in A,\ t \in [0,T].$$

Suppose G_{λ_0} has an attractor $\mathcal{A}$. Then the following hold:

(1) *there exists $\mu > 0$ such that, when $\rho(\lambda, \lambda_0) < \mu$, G_λ has an attractor $\mathcal{A}_\lambda$;*

(2) *$d_{\mathrm{H}}(\mathcal{A}_\lambda, \mathcal{A}) \to 0$ as $\lambda \to \lambda_0$;*

(3) *if $K \subset \Omega(\mathcal{A})$ is compact, then $K \subset \Omega(A_\lambda)$ provided $\rho(\lambda, \lambda_0)$ sufficiently small.*

3. Morse decompositions of invariant sets for GDSs.

3.1. Attractor-repeller pair. As usual, let X be a complete locally compact metric space, and let G be a GDS on X. We also assume that there has been given a compact invariant set S of G.

Denote by $G|_S$ the restriction of G on S. Since S is invariant, $G|_S$ is also a GDS on S.

We say that a compact set $\mathcal{A}$ is an *attractor of G in S*; by this we mean that it is an attractor of $G|_S$ in S. We denote by $\Omega^S(\mathcal{A})$ the attraction region of $\mathcal{A}$ in S (under the GDS $G|_S$). Then for any compact subset $K \subset \Omega^S(\mathcal{A})$, as $\mathcal{A}$ attracts K, we necessarily have $\omega(K) \subset \mathcal{A}$.

Let $\mathcal{A}$ be an attractor of G in S. Define

$$\mathcal{A}^* = \{x \in S |\ \omega(x) \setminus \mathcal{A} \neq \emptyset\}. \tag{3.1}$$

$\mathcal{A}^*$ is said to be the *repeller* of G in S dual to $\mathcal{A}$, and $(\mathcal{A}, \mathcal{A}^*)$ is said to be an *attractor-repeller pair* in S.

Remark 3.1. For single-valued dynamical systems, the repeller $\mathcal{A}^*$ of an attractor $\mathcal{A}$ in an invariant set S is defined by

$$\mathcal{A}^* = \{x \in S |\ \omega(x) \cap \mathcal{A} = \emptyset\}; \tag{3.2}$$

see [15, 43]. Here we have used a slightly relaxed condition "$\omega(x) \setminus \mathcal{A} \neq \emptyset$" to define a repeller for GDSs. One easily checks that the two definitions coincide when we come back to the situation of the single-valued case. Our definition seems to be more suitable for set-valued systems. This can be seen from the following easy example.

Example 3.1. Consider the scalar differential inclusion,

$$x'(t) \in f(x(t)), \qquad \text{where } f(x) = \begin{cases} [-8x(x+1)^2,\, 2], & -1 \le x \le 0; \\ -8x(x+1)^2, & \text{otherwise.} \end{cases}$$

This system generates a GDS G on $\mathbb{R}$ which has two equilibria -1 and 0, with 0 being asymptotically stable and -1 being unstable; see Figure 3.1.

The global attractor of G is the interval $[-1, 0]$. Define γ on $[0, \infty)$ as

$$\gamma(t) = -1 + 2t \quad (0 \le t \le 1/2) \quad \text{and} \quad \gamma(t) = 0 \quad (t > 1/2).$$

Then γ is a trajectory of G, which implies $0 \in \omega(-1)$.

Let $S = [-1, 0]$. It is easy to see that $\mathcal{A} = \{0\}$ is an attractor of G with $\Omega^S(\mathcal{A}) = (-1, 0]$. If we define the repeller $\mathcal{A}^*$ as in (3.1), then we have $\mathcal{A}^* = \{-1\}$. However, if we define $\mathcal{A}^*$ as in (3.2), then since $0 \in \omega(-1) \cap \mathcal{A} \neq \emptyset$, we have $\mathcal{A}^* = \emptyset$.

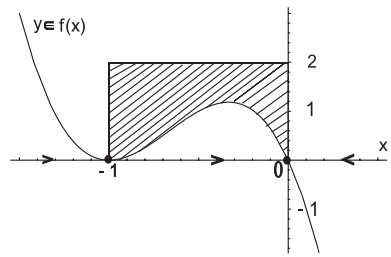


FIG. 3.1.

PROPOSITION 3.2. *Let $(\mathcal{A}, \mathcal{A}^*)$ be an attractor-repeller pair in S, and let U be a neighborhood of $\mathcal{A}^*$. Then for any $\varepsilon > 0$, there is a $T > 0$ such that*

$$G(t)x \subset B(\mathcal{A}, \varepsilon) \qquad \forall t > T,\ x \in S \setminus U.$$

Proof. Let $K = S \setminus \text{int}U$. Then $K \subset \Omega^S(\mathcal{A})$ and is compact. Hence $\mathcal{A}$ attracts K, and the conclusion follows. □

As a direct consequence of the above result, we have the following.

PROPOSITION 3.3. *Let $(\mathcal{A}, \mathcal{A}^*)$ be an attractor-repeller pair in S, and let B be a closed set distinct from $\mathcal{A}$. Then for every $\varepsilon > 0$, there is a $T > 0$ such that, whenever $x \in S$ and $t \geq T$ is such that $G(t)x \cap B \neq \emptyset$, we have $d(x, \mathcal{A}^*) < \varepsilon$.*

PROPOSITION 3.4. *Let $\mathcal{A}$ be an attractor of G in S. Then the following hold:*

(1) $\mathcal{A}^* = S \setminus \Omega^S(\mathcal{A}) = S \setminus \Omega(\mathcal{A})$;

(2) $\mathcal{A}^*$ *is compact and weakly invariant.*

Proof. (1) Assume $x \in S$. Then $\omega(x)$ is a nonempty compact set, and $\omega(x) \subset \mathcal{A}$ if and only if $x \in \Omega^S(\mathcal{A})$. This implies $\mathcal{A}^* = S \setminus \Omega^S(\mathcal{A})$.

Since S is invariant, we have $\Omega^S(\mathcal{A}) = S \cap \Omega(\mathcal{A})$. Hence the second equality follows.

(2) The compactness of $\mathcal{A}^*$ follows from (1).

To check the weak invariance of $\mathcal{A}^*$, we first show that $\mathcal{A}^*$ is negatively invariant. Indeed, we observe that

$$S = G(t)S = G(t)\left(\Omega^S(\mathcal{A}) \cup \mathcal{A}^*\right) = \left(G(t)\Omega^S(\mathcal{A})\right) \cup G(t)\mathcal{A}^* \subset \Omega^S(\mathcal{A}) \cup G(t)\mathcal{A}^*.$$

Since $\mathcal{A}^* \subset S$ and $\mathcal{A}^* \cap \Omega^S(\mathcal{A}) = \emptyset$, we necessarily have $\mathcal{A}^* \subset G(t)\mathcal{A}^*$.

Now assume that $x \in \mathcal{A}^*$. Then by Proposition 2.5 and the negative invariance of $\mathcal{A}^*$, there is a trajectory γ^- on $(-\infty, 0]$ such that $\gamma^-((-\infty, 0]) \subset \mathcal{A}^*$ and $\gamma^-(0) = x$.

To complete the proof, it suffices to show that there is also a trajectory γ^+ on $[0, \infty)$ such that $\gamma([0, \infty)) \subset \mathcal{A}^*$ with $\gamma(0) = x$.

Take an open neighborhood V of $\mathcal{A}^*$ and $\varepsilon > 0$ sufficiently small so that $\overline{V} \cap \mathrm{B}(\mathcal{A}, \varepsilon) = \emptyset$. By virtue of Proposition 3.2 there is a $T > 0$ such that

$$G(t)y \subset \mathrm{B}(\mathcal{A}, \varepsilon) \qquad \forall t > T,\ y \in S \setminus V. \tag{3.3}$$

Since $\omega(x)$ is compact, if $\omega(x) \subset \Omega^S(\mathcal{A})$, then $\mathcal{A}$ attracts $\omega(x)$ and consequently $\mathcal{A}$ attracts x, which implies $x \in \Omega^S(\mathcal{A})$ and leads to a contradiction. Therefore we deduce that $\omega(x) \cap \mathcal{A}^* \neq \emptyset$. It follows that there exist $t_n \to +\infty$ and $y_n \in G(t_n)x$ such that $y_n \in V$ for all n. Let γ_n be a trajectory of G on $[0, t_n]$ satisfying $\gamma_n(0) = x$ and $\gamma_n(t_n) = y_n$. Define

$$\tau_n = \sup\{\tau > 0 |\ \gamma_n([0, \tau]) \subset V\}.$$

Then $\gamma_n(\tau_n) \in \partial V \subset S \setminus V$. We claim that there is a subsequence of τ_n, still denoted by τ_n, such that $\tau_n \to +\infty$. Indeed, if τ_n is bounded, say, $\tau_n \le s$ for all n, then by (3.3) we have

$$y_n = \gamma_n(t_n) \in G(t_n - \tau_n)\gamma_n(\tau_n) \in \mathrm{B}(\mathcal{A}, \varepsilon)$$

for large n with $t_n > T + s$, which leads to a contradiction and thus proves our claim.

Now thanks to Barbashin's theorem (Theorem 2.4), one concludes that there is a trajectory γ^+ on $[0,\infty)$ such that $\gamma^+(0) = x$ and $\gamma^+([0,\infty)) \subset \overline{V}$. Note that we actually have $\gamma^+([0,\infty)) \subset \mathcal{A}^*$; otherwise, one will find $d(\gamma^+(t), \mathcal{A}) \to 0$ as $t \to \infty$, a contradiction!

The proof is complete. □

PROPOSITION 3.5. *Let $\mathcal{A}$ be an attractor of G in S, and let $\gamma : \mathbb{R} \to S$ be a complete trajectory through $x \in S$. Then the following properties hold:*

(1) *If $\omega(\gamma) \cap \mathcal{A}^* \neq \emptyset$, then $\gamma(\mathbb{R}) \subset \mathcal{A}^*$, and if $\alpha(\gamma) \cap \mathcal{A} \neq \emptyset$, then $\gamma(\mathbb{R}) \subset \mathcal{A}$;*

(2) *If $x \notin \mathcal{A}$, then $\alpha(\gamma) \subset \mathcal{A}^*$, and if $x \notin \mathcal{A}^*$, then $\omega(\gamma) \subset \mathcal{A}$.*

Proof. The proof is a slight modification of the corresponding one for single-valued systems (see, e.g., [43]). It is thus omitted. □

3.2. Morse decompositions of invariant sets.

DEFINITION 3.6. *Let S be a compact invariant set. An ordered collection $\mathcal{M} = \{M_1, \dots, M_n\}$ of subsets of S is called a Morse decomposition of S if there exists an increasing sequence $\emptyset = \mathcal{A}_0 \subset \mathcal{A}_1 \subset \cdots \subset \mathcal{A}_n = S$ of attractors of G in S such that*

$$M_k = \mathcal{A}_k \cap \mathcal{A}^*_{k-1}, \qquad 1 \le k \le n.$$

The sets M_k in Definition 3.6 are called *Morse sets*. For convenience in statement, we allow Morse sets to be empty. This occurs, say, for instance, in case $\mathcal{A}_j = \mathcal{A}_{j-1}$ for j (in which case $M_j = \mathcal{A}_j \cap \mathcal{A}^*_{j-1} = \emptyset$). However, if two Morse decompositions $\mathcal{M}$ and $\mathcal{M}'$ have the same nonempty Morse sets, they will be regarded as the same.

THEOREM 3.7. *Let $\mathcal{M} = \{M_1, \dots, M_n\}$ be a Morse decomposition of S with the corresponding attractor sequence $\emptyset = \mathcal{A}_0 \subset \mathcal{A}_1 \subset \cdots \subset \mathcal{A}_n = S$. Then the following hold:*

(1) *For each k, $(\mathcal{A}_{k-1}, M_k)$ is an attractor-repeller pair in $\mathcal{A}_k$;*

(2) *M_k are pairwise disjoint weakly invariant compact sets;*

(3) *If γ is a complete trajectory, then either $\gamma(\mathbb{R}) \subset M_k$ for some Morse set M_k or else there are indices $i < j$ such that $\alpha(\gamma) \subset M_j$ and $\omega(\gamma) \subset M_i$;*

(4) *The attractors $\mathcal{A}_k$ are uniquely determined by the Morse sets, i.e.,*

$$\mathcal{A}_k = \bigcup_{1 \le i \le k} W^u(M_i), \qquad 1 \le k \le n,$$

where $W^u(M_i) = \{x |$ there is a trajectory $\gamma : \mathbb{R} \to S$ through x with $\alpha(\gamma) \subset M_i\}$;

(5) *If S is isolated, then so is each $\mathcal{A}_k$.*

Proof. The proof of (1)–(5) except for the weak invariance of Morse sets M_k can be given in a quite similar manner as in the situation of single-valued dynamical systems (see [43, Chapter 3, Theorem 1.7]). We thus omit the details of the argument. □

The weak invariance of M_k is a consequence of (1) and the weak invariance of repellers.

The following result seems to be interesting, which demonstrates that (2) and (3) in Theorem 3.7 uniquely characterize a Morse decomposition.

THEOREM 3.8. *Let S be a compact invariant set of G, and let $\mathcal{M} = \{M_1, \dots, M_n\}$ be an ordered collection of pairwise disjoint compact and weakly invariant subsets of S. Suppose that for every $x \in S$ and every complete trajectory γ through x, we have either $\gamma(\mathbb{R}) \subset M_i$ for some i or else there are indices $i < j$ such that $\alpha(\gamma) \subset M_j$ and $\omega(\gamma) \subset M_i$. Then $\mathcal{M}$ is a Morse decomposition of S.*

Proof. Set $\mathcal{A}_0 = \emptyset$, and for $1 \le k \le n$ define $\mathcal{A}_k$ as

$$\mathcal{A}_k = \{x \in S| \quad \text{there is a complete trajectory } \gamma : \mathbb{R} \to S \text{ through } x \text{ satisfying } \alpha(\gamma) \subset (M_1 \cup \cdots \cup M_k)\}.$$

We will show that $\mathcal{A}_0 \subset \mathcal{A}_1 \subset \cdots \subset \mathcal{A}_n = S$ is an attractor sequence in S such that $\mathcal{A}_k \cap \mathcal{A}^*_{k-1} = M_k$, thus proving the result.

Step 1. We show that the sets $\mathcal{A}_k$ are closed.

It is clear that $\mathcal{A}_n = S$ is closed. We now proceed inductively and assume that $\mathcal{A}_{k+1}$ is closed. We prove that $\mathcal{A}_k$ is closed. Let $x_m \in \mathcal{A}_k$ with $x_m \to x \in S$. Then $x \in \mathcal{A}_{k+1}$, since $\mathcal{A}_k \subset \mathcal{A}_{k+1}$ and $\mathcal{A}_{k+1}$ is closed. For each x_m there is a complete trajectory $\gamma_m : \mathbb{R} \to S$ with $\gamma_m(0) = x_m$ and $\alpha(\gamma_m) \subset (M_1 \cup \cdots \cup M_k)$. Using Barbashin's compactness theorem one finds a subsequence of γ_m, still denoted by γ_m, such that γ_m converges uniformly on any compact interval of $\mathbb{R}$ to a complete trajectory $\gamma : \mathbb{R} \to S$ through x. We claim that $\alpha(\gamma) \subset (M_1 \cup \cdots \cup M_k)$, and hence $x \in \mathcal{A}_k$.

Indeed, since $\gamma_m(\mathbb{R}) \subset \mathcal{A}_{k+1}$ and $\mathcal{A}_{k+1}$ is closed, it follows that $\gamma(\mathbb{R}) \subset \mathcal{A}_{k+1}$ and so $\alpha(\gamma) \subset \mathcal{A}_{k+1}$. We first show that

$$M_j \cap \mathcal{A}_{k+1} = \emptyset \qquad \text{for } j > k+1. \tag{3.4}$$

Suppose the contrary, then there is a $y \in M_j \cap \mathcal{A}_{k+1}$ for some $j > k+1$. Since $y \in \mathcal{A}_{k+1}$, there is a complete trajectory $\sigma_1 : \mathbb{R} \to S$ such that $\alpha(\sigma_1) \subset (M_1 \cup \cdots \cup M_{k+1})$ and $\sigma_1(0) = y$. By $y \in M_j$ and the weak invariance of M_j, there is also a complete trajectory $\sigma_2 : \mathbb{R} \to M_j$ such that $\sigma_2(0) = y$. Define $\sigma : \mathbb{R} \to S$ as

$$\sigma(t) = \sigma_1(t) \text{ (for } t \le 0) \quad \text{and} \qquad \sigma(t) = \sigma_2(t) \text{ (for } t > 0).$$

Then σ is a complete trajectory of G. Note that $\alpha(\sigma) \subset (M_1 \cup \cdots \cup M_{k+1})$ and $\omega(\sigma) \subset M_j$. On the other hand by assumption in the theorem and $\alpha(\sigma) \subset (M_1 \cup \cdots \cup M_{k+1})$, we deduce that $\omega(\sigma) \subset M_i$ for some $i \le k+1$. This leads to a contradiction, as $M_i \cap M_j = \emptyset$.

Now because $\alpha(\gamma) \subset M_j$ for some j (by assumption in the theorem), we necessarily have by $\alpha(\gamma) \subset \mathcal{A}_{k+1}$ and (3.4) that $\alpha(\gamma) \subset M_1 \cup \cdots \cup M_k \cup M_{k+1}$. Consequently, either $\alpha(\gamma) \subset M_1 \cup \cdots \cup M_k$, in which case we are done, or else $\alpha(\gamma) \subset M_{k+1}$.

We prove that the latter case will not occur. Suppose the contrary. We take two open neighborhoods U, V of M_{k+1} with $\overline{U}$ and $\overline{V}$ compact such that $\overline{U} \subset V$ and $\overline{V} \cap M_j = \emptyset$ for $j \ne k+1$. Since $\alpha(\gamma) \subset M_{k+1}$, there is a $\tau > 0$ such that $\gamma((-\infty, -\tau]) \subset U$. Since $\gamma_m(-\tau) \to \gamma(-\tau)$, we may assume that $\gamma_m(-\tau) \in V$ for all m. Let

$$s_m = \sup\{s > \tau| \ \gamma_m([-s, -\tau]) \subset V\}.$$

Then $\gamma_m(-s_m) \in \partial V$. We claim that $s_m \to \infty$. Otherwise, say $s_m \le s$ for some $s > \tau$, due to the uniform convergence of γ_m to γ on $[-s, -\tau]$, one will find that $\gamma_m(-s_m) \in V$ for m sufficiently large, a contradiction! Now set

$$\sigma_m(t) = \gamma(-s_m + t), \qquad t \in [0, s_m - \tau_m].$$

Then by Barbashin's theorem there is a subsequence of σ_m, still denoted by σ_m, such that σ_m converges uniformly on any compact interval of $\mathbb{R}^+$ to a trajectory $\sigma : \mathbb{R}^+ \to \overline{V}$. Note that $\sigma(0) \in \partial V$ as $\sigma_m(0) = \gamma_m(-s_m) \in \partial V$. Since $\mathcal{A}_{k+1}$ is closed and $\gamma_m(\mathbb{R}) \subset \mathcal{A}_k \subset \mathcal{A}_{k+1}$, we conclude that $\sigma(t) \in \mathcal{A}_{k+1}$ for $t \geq 0$. In particular, $\sigma(0) \in \mathcal{A}_{k+1}$. Therefore by the definition of $\mathcal{A}_{k+1}$, there is also a complete trajectory $\widetilde{\sigma} : \mathbb{R} \to S$ through $\sigma(0)$ such that $\alpha(\widetilde{\sigma}) \subset M_1 \cup \cdots \cup M_k \cup M_{k+1}$. Define a complete trajectory $\sigma' : \mathbb{R} \to S$ as

$$\sigma'(t) = \widetilde{\sigma}(t) \text{ (for } t \leq 0) \quad \text{and} \qquad \sigma'(t) = \sigma(t) \text{ (for } t > 0).$$

Then $\alpha(\sigma') \subset M_1 \cup \cdots \cup M_k \cup M_{k+1}$. By $\sigma'(\mathbb{R}^+) \subset \overline{V}$ we also have $\omega(\sigma') \subset M_{k+1}$. Therefore by the assumption in the theorem we necessarily have $\sigma'(\mathbb{R}) \subset M_{k+1}$, which contradicts the fact $\sigma'(0) = \sigma(0) \in \partial V$.

Step 2. $\mathcal{A}_k$ is invariant. Indeed, if $y \in \mathcal{A}_k$, then there is a complete trajectory $\gamma : \mathbb{R} \to S$ through y such that $\alpha(\gamma) \subset M_1 \cup \cdots \cup M_k \subset \mathcal{A}_k$. It then follows that $\gamma(\mathbb{R}) \subset \mathcal{A}_k$. Consequently $y \in G(t)\gamma(\mathbb{R}) \subset G(t)\mathcal{A}_k$ for any $t \geq 0$. Hence $\mathcal{A}_k \subset G(t)\mathcal{A}_k$.

We check that the converse inclusion $G(t)\mathcal{A}_k \subset \mathcal{A}_k$ also holds true. Let $y \in G(t)\mathcal{A}_k$. Then $y \in G(t)x$ for some $x \in \mathcal{A}_k$. Since S is invariant, there is a trajectory $\gamma_3 : [t, \infty) \to S$ such that $\gamma_3(t) = y$. Thanks to Theorem 2.3, one can find a trajectory γ_2 on $[0, t]$ such that $\gamma_2(0) = x$ and $\gamma_2(t) = y$. By invariance of S, we necessarily have $\gamma_2([0, t]) \subset S$. As $x \in \mathcal{A}_k$, there is also a complete trajectory $\gamma_1 : \mathbb{R} \to S$ such that $\gamma_1(0) = x$ and $\alpha(\gamma) \subset M_1 \cup \cdots \cup M_k$. Now we define a complete trajectory $\gamma : \mathbb{R} \to S$ such that

$$\gamma|_{(-\infty,0]} = \gamma_1, \quad \gamma|_{[0,t]} = \gamma_2, \quad \gamma|_{[t,\infty)} = \gamma_3.$$

Then $\gamma(t) = y$ and $\alpha(\gamma) \subset M_1 \cup \cdots \cup M_k$, which implies $y \in \mathcal{A}_k$ and proves the conclusion.

Step 3. $\mathcal{A}_k$ is an attractor of G in S.

This is clearly true for $k = n$. We proceed by induction and assume $\mathcal{A}_{k+1}$ to be an attractor in S. Choose a neighborhood U_{k+1} of $\mathcal{A}_{k+1}$ such that $\overline{U}_{k+1}$ is compact and $\omega(U_{k+1} \cap S) = \mathcal{A}_{k+1}$. Since $\mathcal{A}_k$ is closed, $M_{k+1} \cup \mathcal{A}_k \subset \mathcal{A}_{k+1}$, and $M_{k+1} \cap \mathcal{A}_k = \emptyset$, we can choose a neighborhood U of $\mathcal{A}_k$ and neighborhoods V_0, V_1 of M_{k+1} contained in U_{k+1} such that $\overline{V}_0 \subset V_1$ and $\overline{U} \cap \overline{V}_1 = \emptyset$. We show that $\omega(U \cap S) = \mathcal{A}_k$ when U is chosen sufficiently small. Indeed, since $\mathcal{A}_k$ is invariant, we have $\mathcal{A}_k \subset \omega(U \cap S)$. There remains to check

$$\mathcal{A}_k \supset \omega(U \cap S). \tag{3.5}$$

Suppose $\omega(U \cap S) \setminus \mathcal{A}_k \neq \emptyset$, and choose a $y \in \omega(U \cap S) \setminus \mathcal{A}_k \neq \emptyset$. Then there are sequences $x_m \in U \cap S$, $t_m \to \infty$, and $y_m \in G(t_m)x_m$ such that $y_m \to y$. Let γ_m be a trajectory on $[0, t_m]$ with $\gamma_m(0) = x_m$ and $\gamma_m(t_m) = y_m$. We can extract a subsequence of γ_m, still denoted by γ_m, such that $\gamma_m(t_m + t)$ converges uniformly on any compact interval of $\mathbb{R}$ to a complete trajectory $\gamma(t)$ with $\gamma(\mathbb{R}) \subset S$ and $\gamma(0) = y$. Now $\omega(U \cap S) \subset \omega(U_{k+1} \cap S) = \mathcal{A}_{k+1}$ implies $\gamma(\mathbb{R}) \subset \mathcal{A}_{k+1}$, and hence $\alpha(\gamma) \subset \mathcal{A}_{k+1}$ (by Step 1). Therefore $\alpha(\gamma) \subset M_1 \cup \cdots \cup M_k \cup M_{k+1}$. As $y \notin \mathcal{A}_k$, we necessarily have $\alpha(\gamma) \subset M_{k+1}$ and so there is a $T > 0$ such that, when $t > T$, we have $\gamma(-t) \in V_0$. Since $\gamma_m(t_m + t)$ converges uniformly on any compact interval of $\mathbb{R}$ to $\gamma(t)$ for any $\tau > 0$, we can find a m_τ sufficiently large such that $\gamma_{m_\tau}([t_{m_\tau} - T - \tau,\, t_{m_\tau} - T]) \subset V_1$. Rewriting $t_{m_\tau} - T - \tau$ as s_{m_τ}, we obtain

$$\gamma_{m_\tau}([s_{m_\tau}\, s_{m_\tau} + \tau]) \subset V_1.$$

Now suppose (3.5) fails to be true for any U. Pick a sequence $\varepsilon_m \downarrow 0$ with

$$\overline{\mathrm{B}}(\mathcal{A}_k, \varepsilon_m) \cap \overline{V}_1 = \emptyset, \quad \mathrm{B}(\mathcal{A}_k, \varepsilon_m) \subset U_{k+1}.$$

Applying what we have just proved above for each $U = \mathrm{B}(\mathcal{A}_k, \varepsilon_m)$ and $\tau = 1/\varepsilon_m$, one obtains a sequence of trajectories denoted by γ_m such that, for each m,

$$\gamma_m(0) \in \mathrm{B}(\mathcal{A}_k, \varepsilon_m) \cap S, \qquad \gamma_m([s_m,\, s_m + 1/\varepsilon_m]) \subset V_1$$

for some $s_m \geq 0$. Let $\tau_m = \inf\{0 \leq \tau \leq s_m : \ \gamma_m([\tau,\, s_m + 1/\varepsilon_m]) \subset V_1\}$. Then $\gamma_m(\tau_m) \in \partial V_1$. We claim that $\tau_m \to \infty$. Otherwise, say $\tau_m \leq T < \infty$, γ_m will converge uniformly on $[0, T]$ to a trajectory γ. Since $d(\gamma_m(0), \mathcal{A}_k) \to 0$ and $\mathcal{A}_k$ is invariant, one necessarily has $\gamma([0, T]) \subset \mathcal{A}_k$. It then follows that $d(\gamma_m(\tau_m), \mathcal{A}_k) \to 0$, which contradicts $\gamma_m(\tau_m) \in \partial V_1$ and proves the claim.

We may assume that $\gamma_m(\tau_m) \to z$. Then since $\gamma_m(0) \in \mathrm{B}(\mathcal{A}_k, \varepsilon_m) \cap S \subset U_{k+1} \cap S$ and $\tau_m \to \infty$, we see that $z \in \omega(U_{k+1} \cap S) = \mathcal{A}_{k+1}$. Of course we also have $z \in \partial V_1$ (by $\gamma_m(\tau_m) \in \partial V_1$). Set $\widetilde{\gamma}_m(t) = \gamma_m(\tau_m + t)$. Noting that $s_m + 1/\varepsilon_m - \tau_m \geq 1/\varepsilon_m$, we see that $\widetilde{\gamma}_m$ is well defined at least on $[-\tau_m,\, 1/\varepsilon_m]$ with $\widetilde{\gamma}_m([0,\, 1/\varepsilon_m]) \subset V_1$. By Barbashin's theorem one can easily find a complete trajectory $\widetilde{\gamma} : \mathbb{R} \to S$ such that $\widetilde{\gamma}(0) = z \in \partial V_1$ and $\widetilde{\gamma}(\mathbb{R}^+) \subset \overline{V}_1$. Due to the assumption of the theorem we conclude

$$\omega(\widetilde{\gamma}) \subset M_{k+1}. \tag{3.6}$$

It is clear that $\widetilde{\gamma}(\mathbb{R}) \subset \mathcal{A}_{k+1}$ and so $\alpha(\widetilde{\gamma}) \subset M_1 \cup \cdots \cup M_k \cup M_{k+1}$. This and (3.6) as well as the assumption in the theorem imply that $\widetilde{\gamma}(\mathbb{R}) \subset M_{k+1}$, which contradicts $\widetilde{\gamma}(0) \in \partial V_1$.

Step 4. $M_k = \mathcal{A}_k \cap \mathcal{A}^*_{k-1}$ for all $1 \leq k \leq n$.

Indeed, if $x \in M_k$, then there is a complete trajectory $\gamma : \mathbb{R} \to M_k$ through x, and thus by definition of $\mathcal{A}_k$ we have $x \in \mathcal{A}_k$. If $x \notin \mathcal{A}^*_{k-1}$, then $\omega(x) \cap \mathcal{A}^*_{k-1} = \emptyset$, and hence $\omega(x) \subset \mathcal{A}_{k-1}$. It follows that $\omega(\gamma) \subset \mathcal{A}_{k-1}$; therefore, $\omega(\gamma) \subset M_i$ for some $1 \leq i \leq k-1$. This contradicts $\omega(\gamma) \subset M_k$. Thus $x \in \mathcal{A}^*_{k-1}$, i.e., $M_k \subset \mathcal{A}_k \cap \mathcal{A}^*_{k-1}$.

Conversely, let $x \in \mathcal{A}_k \cap \mathcal{A}^*_{k-1}$. Then by Proposition 3.4 there is a complete trajectory $\gamma : \mathbb{R} \to \mathcal{A}^*_{k-1}$. We have $\omega(\gamma) \cap (M_1 \cup \cdots \cup M_{k-1}) = \emptyset$ and so $\omega(\gamma) \subset M_j$ for some $j \geq k$. On the other hand $x \in \mathcal{A}_k$ implies $\alpha(\gamma) \subset M_1 \cup \cdots \cup M_k$. Consequently by the assumption of the theorem, we have $\gamma(\mathbb{R}) \subset M_k$, which implies $x \in M_k$. The proof is complete. □

3.3. The finest Morse decomposition. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two Morse decompositions of a compact invariant $\mathcal{S}$. We say that $\mathcal{M}_2$ is *finer* than $\mathcal{M}_1$; this means that each Morse set $M \in \mathcal{M}_2$ is contained in a Morse set $M' \in \mathcal{M}_1$. A Morse decomposition $\mathcal{M}$ is said to be the *finest Morse decomposition* of $\mathcal{S}$ if $\widetilde{\mathcal{M}}$ is a finer Morse decomposition implies $\widetilde{\mathcal{M}} = \mathcal{M}$.

In general the finest Morse decomposition may fail to exist (see [13, Appendix B]).

LEMMA 3.9. *Let $\mathcal{S}$ be a compact invariant set, and let $\mathcal{M} = \{M_1, \ldots, M_m\}$ and $\mathcal{M}' = \{M'_1, \ldots, M'_n\}$ be two Morse decompositions of $\mathcal{S}$. Let M_{ij} be the maximal weakly invariant set contained in $M_i \cap M'_j$. Then the ordered collection*

$$\{M_{ij} |\ 1 \leq i \leq m,\ 1 \leq j \leq n\}$$

is a Morse decomposition of $\mathcal{S}$, where the order "$\preceq$" is given by

$$M_{ij} \preceq M_{kl} \iff \text{either } i < k \text{ or } i = k,\ j \leq l.$$

Proof. Let $A_{ij} = M_i \cap M_j$. Since $\mathcal{M}$ and $\mathcal{M}'$ are both Morse decompositions, it is trivial to check that, for any complete trajectory γ in $\mathcal{S}$, either $\gamma(\mathbb{R})$ is contained in some A_{ij} or there are indices $1 \le i, j \le m$ and $1 \le k, l \le n$ with $i \le k$ and $j \le l$ (in case $i = k$) such that $\omega(\gamma) \subset A_{ij},\ \alpha(\gamma) \subset A_{kl}$. Let $M_{ij} = \operatorname{cl} \widetilde{M}_{ij}$, where $\widetilde{M}_{ij}$ is the set of the union of all ω-limit sets and α-limit sets of complete trajectories in $\mathcal{S}$ that are contained in A_{ij} and the orbit of complete trajectories in A_{ij}. It is clear that $\widetilde{M}_{ij}$ is weakly invariant (see Remark 2.6). Therefore by Barbashin's theorem we easily know that M_{ij} is weakly invariant. By definition of M_{ij}, it is necessarily the maximal weakly invariant set in A_{ij}. It is also clear that the ordered collection $\{M_{ij} |\ 1 \le i \le m,\ 1 \le j \le n\}$ satisfies all the assumptions in Theorem 3.8; hence, it is a Morse decomposition of $\mathcal{S}$. □

Remark 3.10. We infer from Lemma 3.9 that the finest Morse decomposition, if it exists, is finer than any other Morse decomposition of $\mathcal{S}$.

PROPOSITION 3.11. *Morse sets in the finest Morse decomposition are connected.*

Proof. Let $\mathcal{M} = \{M_1, \dots, M_n\}$ be the finest Morse decomposition of $\mathcal{S}$. Suppose that there is a Morse set M_k which is not connected. Then there exist disjoint open sets U, V such that $M_k \subset U \cup V$ and $M_k \cap U \ne \emptyset \ne M_k \cap V$. Let $M_k^+ = M_k \cap U$, and let $M_k^- = M_k \cap V$. We claim that, for any complete trajectory γ in $\mathcal{S}$, none of the following cases will occur:

(1) $\omega(\gamma) \cap M_k^+ \ne \emptyset \ne \omega(\gamma) \cap M_k^-$, or $\alpha(\gamma) \cap M_k^+ \ne \emptyset \ne \alpha(\gamma) \cap M_k^-$;

(2) $\omega(\gamma) \subset M_k^+$ and $\alpha(\gamma) \subset M_k^-$, or $\omega(\gamma) \subset M_k^-$ and $\alpha(\gamma) \subset M_k^+$.

Indeed, the first case does not occur because the limit sets are connected. Suppose the latter case. Then we must have $\gamma(\mathbb{R}) \cap U \ne \emptyset \ne \gamma(\mathbb{R}) \cap V$ and $\gamma(\mathbb{R}) \subset M_k \subset U \cup V$. (The first inclusion is due to the fact that both ω- and α-limit sets of γ are contained in M_k.) This contradicts the connectedness of $\gamma(\mathbb{R})$. Hence the claim holds true.

Now let $\mathcal{M}' = \{M1, \dots, M_{k-1}, M_k^-, M_k^+, M_{k+1}, \dots, M_n\}$. One easily checks that $\mathcal{M}'$ is a Morse decomposition which is finer than $\mathcal{M}$, a contradiction! □

4. Stability of Morse decompositions of attractors for GDSs. We first state and prove the following basic results.

THEOREM 4.1. *Let $\mathcal{A}$ be an attractor of G, and let $\mathcal{A}_0 \subset \mathcal{A}$ be an attractor of G in $\mathcal{A}$. Then $\mathcal{A}_0$ is also an attractor of G (in X).*

Proof. We first show that $\mathcal{A}_0$ is Lyapunov stable in X. Suppose the contrary. Then one would find a $\delta > 0$ and an $\varepsilon_0 > 0$ such that, for any $0 < \varepsilon < \varepsilon_0$, there is an $x_\varepsilon \in \mathrm{B}(\mathcal{A}_0, \varepsilon)$ and a $t_\varepsilon > 0$ such that $G(t_\varepsilon)x_\varepsilon \setminus \mathrm{B}(\mathcal{A}_0, \delta) \ne \emptyset$. We may assume δ sufficiently small so that $\overline{\mathrm{B}}(\mathcal{A}_0, \delta) \subset \Omega(\mathcal{A})$, and $\overline{\mathrm{B}}(\mathcal{A}_0, \delta) \cap \mathcal{A}$ is a compact subset of $\Omega^{\mathcal{A}}(\mathcal{A}_0)$.

By Theorem 2.3 there is a trajectory γ_ε on $[0, t_\varepsilon]$ with $\gamma_\varepsilon(0) = x_\varepsilon \in \mathrm{B}(\mathcal{A}_0, \varepsilon)$ and $d\,(\gamma_\varepsilon(t_\varepsilon), \mathcal{A}_0) \ge \delta$. We may assume that

$$\gamma_\varepsilon([0, t_\varepsilon]) \subset \overline{\mathrm{B}}(\mathcal{A}_0, \delta);$$

otherwise, we can choose t_ε as $t_\varepsilon = \inf\{t > 0 |\ d\,(\gamma_\varepsilon(t), \mathcal{A}_0) \ge \delta\}$. We claim that $t_\varepsilon \to +\infty$ as $\varepsilon \to 0$. Indeed, if this is not the case, then there would exist a sequence $x_n := x_{\varepsilon_n} \to x_0 \in \mathcal{A}_0$ such that the sequence $t_n := t_{\varepsilon_n}$ is bounded. By axiom (3) in Definition 2.1,

$$d\,(\gamma_{\varepsilon_n}(t_n), \mathcal{A}_0) = d\,(\gamma_{\varepsilon_n}(t_n), G(t_n)\mathcal{A}_0) \le d_{\mathrm{H}}\,(G(t_n)x_n, G(t_n)x_0) \to 0$$

as $n \to \infty$, which leads to a contradiction and proves our claim. Now let

$$\sigma_\varepsilon(t) = \gamma_\varepsilon(t_\varepsilon + t) \qquad \text{for } t \in [-t_\varepsilon, 0].$$

Then σ_ε is a trajectory of G on $[-t_\varepsilon, 0]$. Invoking Barbashin's theorem, we can extract a subsequence $\sigma_n := \sigma_{\varepsilon_n}$ with $\varepsilon_n \to 0$ such that σ_n converges to some trajectory $\sigma : (-\infty, 0] \to \overline{\mathrm{B}}(\mathcal{A}_0, \delta)$ uniformly on any compact interval $[t, 0]$. Since $d(\sigma_n(0), \mathcal{A}_0) = \delta$, we necessarily have $d(\sigma(0), \mathcal{A}_0) = \delta$. Now we extend σ to a complete trajectory, still denoted by σ. Then since $\sigma(0) \in \Omega(\mathcal{A})$, we see that $\sigma(\mathbb{R}) \subset \Omega(\mathcal{A})$. We infer from

$$\sigma((-\infty, 0]) \subset \overline{\mathrm{B}}(\mathcal{A}_0, \delta) \subset \Omega(\mathcal{A})$$

that $\sigma(\mathbb{R}) \subset \mathcal{A}$. It follows that $\alpha(\sigma) \subset \overline{\mathrm{B}}(\mathcal{A}_0, \delta) \cap \mathcal{A}$. Using the same argument as in the proof of Proposition 3.5(1), one can easily check that $\sigma(\mathbb{R}) \subset \mathcal{A}_0$, which contradicts to $d(\sigma(0), \mathcal{A}_0) = \delta > 0$. Hence $\mathcal{A}_0$ is Lyapunov stable.

Now we proceed to prove that $\mathcal{A}_0$ is an attractor of G in X. Take $0 < \delta_0 < \delta_1 < \delta_2$ such that $\overline{\mathrm{B}}(\mathcal{A}_0, \delta_2) \subset \Omega(\mathcal{A})$ and $\overline{\mathrm{B}}(\mathcal{A}_0, \delta_2) \cap \mathcal{A}$ is a compact subset of $\Omega^{\mathcal{A}}(\mathcal{A}_0)$; moreover,

$$G(\mathbb{R}^+)\overline{\mathrm{B}}(\mathcal{A}_0, \delta_0) \subset \overline{\mathrm{B}}(\mathcal{A}_0, \delta_1), \quad G(\mathbb{R}^+)\overline{\mathrm{B}}(\mathcal{A}_0, \delta_1) \subset \overline{\mathrm{B}}(\mathcal{A}_0, \delta_2).$$

Note that $\mathcal{A}_0$ is an isolated invariant, as $\mathcal{A}$ is an attractor of G. We show that $\mathrm{B}(\mathcal{A}_0, \delta_0) \subset \Omega(\mathcal{A}_0)$, and thus $\mathcal{A}_0$ is an attractor of G in X.

For this purpose, we first check that

$$\omega(x) \subset \Omega(\mathcal{A}_0), \qquad x \in \mathrm{B}(\mathcal{A}_0, \delta_0). \tag{4.1}$$

Let $x \in \mathrm{B}(\mathcal{A}_0, \delta_0)$. Then $\omega(x)$ is a nonempty compact set, and $\omega(x) \subset \overline{\mathrm{B}}(\mathcal{A}_0, \delta_1)$. Let $y \in \omega(x)$. Since $\omega(x)$ is negatively invariant, there is a trajectory $\gamma : (-\infty, 0] \to \omega(x)$ with $\gamma(0) = y$. We extend γ to a complete trajectory, still denoted by γ. Then $\gamma(\mathbb{R}) \subset \overline{\mathrm{B}}(\mathcal{A}_0, \delta_2)$. As above, we can again show that γ is in fact contained in $\mathcal{A}_0$. Thus $y \in \mathcal{A}_0$.

Now $x \in \mathrm{B}(\mathcal{A}_0, \delta_0) \subset \Omega(\mathcal{A})$ implies $\omega(x) \subset \mathcal{A}$. This and (4.1) imply that $\omega(x) \subset \mathcal{A}_0$. □

THEOREM 4.2. *Let $\mathcal{A}$ be an attractor, and let $\mathcal{M} = \{M_1, \dots, M_n\}$ be a Morse decomposition of $\mathcal{A}$. Then for each trajectory $\gamma : \mathbb{R}^+ \to X$ with $\gamma(0) \in \Omega(\mathcal{A})$, there is a k such that*

$$\lim_{t \to \infty} d(\gamma(t), M_k) = 0.$$

Proof. Let $\emptyset = \mathcal{A}_0 \subset \mathcal{A}_1 \subset \cdots \subset \mathcal{A}_n = \mathcal{A}$ be the attractor sequence that corresponds to $\mathcal{M}$. Then there is a smallest k such that $\omega(\gamma) \subset \mathcal{A}_k$. A direct inspection shows that $\omega(\gamma) \subset M_k$, and the conclusion follows. □

The main result in this section is the following theorem.

THEOREM 4.3. *Let G_λ ($\lambda \in \Lambda$) be a family of GDS on X, where Λ is a metric space with metric $\rho(\cdot, \cdot)$. Assume G_λ satisfies the continuity assumption* (C0) *in Theorem* 2.9 *at $\lambda_0 \in \Lambda$.*

Let $\mathcal{A}$ be an attractor of $G = G_{\lambda_0}$ with Morse decomposition $\mathcal{M} = \{M_1, \dots, M_n\}$. Then when $\rho(\lambda, \lambda_0)$ is sufficiently small, G_λ has an attractor $\mathcal{A}(\lambda)$ with Morse decomposition $\mathcal{M}(\lambda) = \{M_1(\lambda), \dots, M_n(\lambda)\}$; moreover, for each $1 \le k \le n$, we have

$$\lim_{\lambda \to \lambda_0} d_{\mathrm{H}}(M_k(\lambda), M_k) = 0.$$

Proof. Let $\emptyset = \mathcal{A}_0 \subset \mathcal{A}_1 \subset \cdots \subset \mathcal{A}_n = \mathcal{A}$ be the increasing sequence of attractors associated with the Morse decomposition $\mathcal{M}$. Then $\mathcal{A}_k$ are also attractors of G in X.

Let $\varepsilon > 0$ be given arbitrarily. We may assume that ε is small so that $\overline{\mathrm{B}}(\mathcal{A}_k, \varepsilon) \subset \Omega(\mathcal{A}_k)$ for all $1 \le k \le n$ and are compact. Let $\widetilde{\mathcal{A}}^*_k = X \setminus \Omega(\mathcal{A}_k)$. Then it is trivial to verify that $M_k = \mathcal{A}_k \cap \widetilde{\mathcal{A}}^*_{k-1}$. We observe that

$$\mathrm{B}(\mathcal{A}_k, \varepsilon) \cap \mathrm{B}\left(\widetilde{\mathcal{A}}^*_{k-1}, \varepsilon\right) \subset \mathrm{B}(M_k, \varepsilon). \tag{4.2}$$

Let $V_k = \overline{\mathrm{B}}(\mathcal{A}_k, \varepsilon) \setminus \mathrm{B}(\widetilde{\mathcal{A}}^*_{k-1}, \varepsilon)$. Note that V_k is a compact subset of $\Omega(\mathcal{A}_{k-1})$. Thanks to Theorem 2.9, there is a $\delta > 0$ such that, when $\rho(\lambda, \lambda_0) < \delta$, G_λ has for each k an attractor $\mathcal{A}_k(\lambda)$ with

$$\mathcal{A}_k(\lambda) \subset \mathrm{B}(\mathcal{A}_k, \varepsilon), \quad V_k \subset \Omega\left(\mathcal{A}_{k-1}(\lambda)\right). \tag{4.3}$$

Since $\overline{\mathrm{B}}(\mathcal{A}_k, \varepsilon) \subset \Omega(\mathcal{A}_k)$ and is compact, we can also assume $\delta > 0$ is sufficiently small so that by Theorem 2.9 we also have $\overline{\mathrm{B}}(\mathcal{A}_k, \varepsilon) \subset \Omega(\mathcal{A}_k(\lambda))$ for $0 \le k \le n$, and consequently for each k fixed,

$$\mathcal{A}_i(\lambda) \subset \mathrm{B}(\mathcal{A}_i, \varepsilon) \subset \mathrm{B}(\mathcal{A}_k, \varepsilon) \subset \Omega(\mathcal{A}_k(\lambda)) \qquad \text{for } i \le k,$$

which implies $\mathcal{A}_i(\lambda) \subset \mathcal{A}_k(\lambda)$ for all $i \le k$. Therefore

$$\emptyset = \mathcal{A}_0(\lambda) \subset \mathcal{A}_1(\lambda) \subset \cdots \subset \mathcal{A}_n(\lambda) = \mathcal{A}(\lambda)$$

is an increasing sequence of attractors of G_λ.

Let $M_k(\lambda) = \mathcal{A}_k(\lambda) \cap \mathcal{A}^*_{k-1}(\lambda)$. Then $\mathcal{M}(\lambda) = \{M_1(\lambda), \ldots, M_n(\lambda)\}$ is a Morse decomposition of $\mathcal{A}(\lambda)$. There remains to check that $M_k(\lambda) \subset \mathrm{B}(M_k, \varepsilon)$. Indeed, by Theorem 3.7 we know that $M_k(\lambda)$ is the repeller of G_λ in $\mathcal{A}_k(\lambda)$ dual to $\mathcal{A}_{k-1}(\lambda)$. Therefore by Proposition 3.4(1),

$$M_k(\lambda) = \mathcal{A}_k(\lambda) \setminus \Omega^{\mathcal{A}_k(\lambda)}(\mathcal{A}_{k-1}(\lambda)) = \mathcal{A}_k(\lambda) \setminus \Omega(\mathcal{A}_{k-1}(\lambda)).$$

Further by (4.3) we find that

$$\begin{aligned} M_k(\lambda) &\subset \mathrm{B}(\mathcal{A}_k, \varepsilon) \setminus V_k = \mathrm{B}(\mathcal{A}_k, \varepsilon) \setminus \left(\overline{\mathrm{B}}(\mathcal{A}_k, \varepsilon) \setminus \mathrm{B}\left(\widetilde{\mathcal{A}}^*_{k-1}, \varepsilon\right)\right) \\ &\subset \mathrm{B}(\mathcal{A}_k, \varepsilon) \setminus \left(\mathrm{B}(\mathcal{A}_k, \varepsilon) \setminus \mathrm{B}\left(\widetilde{\mathcal{A}}^*_{k-1}, \varepsilon\right)\right) \\ &= \mathrm{B}(\mathcal{A}_k, \varepsilon) \cap \mathrm{B}\left(\widetilde{\mathcal{A}}^*_{k-1}, \varepsilon\right) \subset (\text{by (4.2)}) \subset \mathrm{B}(M_k, \varepsilon). \end{aligned}$$

The proof of the theorem is complete. □

Some results concerning the robustness of Morse decompositions under discretization for specific functional differential equations can be found in Gedeon and Hines [20, 21], etc.

Remark 4.4. We remark that some of the perturbed Morse sets $M_k(\lambda)$ in Theorem 4.3 may be $\emptyset$, even if each M_k is nonvoid. This can be seen by considering the scalar equation:

$$x'(t) = -(x+2)(x^2-1)x^2,$$

which has an equilibrium $E = 0$ that disappears when we add an arbitrary small positive number $\lambda > 0$ in the right-hand side.

THEOREM 4.5. *Assume the hypothesis in Theorem* 4.3. *Let $\mathcal{A}$ be an attractor of G_{λ_0} with Morse decomposition $\mathcal{M} = \{M_1, \ldots, M_n\}$. Then for any compact subset K*

of $\Omega(\mathcal{A})$ *and* $\varepsilon > 0$*, there exists* $\delta > 0$ *such that when* $\rho(\lambda, \lambda_0) < \delta$*, for any trajectory* γ *of* G_λ *with* $\gamma(0) \in K$*,*

$$\limsup_{t\to\infty} d\left(\gamma(t),\, M_k\right) \le \varepsilon$$

for some M_k*.*

Proof. It is a consequence of Theorems 2.9, 4.2, and 4.3. □

5. Morse decompositions for differential inclusions. Consider the differential inclusion on $\mathbb{R}^m$

$$x'(t) \in f(x(t)), \qquad t \ge 0, \tag{5.1}$$

where f is always assumed to satisfy the following assumptions:

(H1) $f(x)$ is a nonempty convex compact subset of $\mathbb{R}^m$ for every $x \in \mathbb{R}^m$;

(H2) f is upper semicontinuous.

Let I be an interval. A mapping $x(\cdot) : I \to \mathbb{R}^m$ is said to be a *solution* of (5.1) on I if it is absolutely continuous on any compact interval $J \subset I$ and solves (5.1) at a.e. $t \in I$.

A solution on $\mathbb{R}$ will be simply called a *complete solution*.

The *reachable mapping* $\mathcal{F}$ of (5.1) is defined as, for all $(t, x) \in \mathbb{R}^+ \times \mathbb{R}^m$,

$$\mathcal{F}(t)x = \{x(t)|\ x(\cdot) \text{ is a solution of (5.1) with } x(0) = x\}.$$

Although $\mathcal{F}$ is a GDS on $\mathbb{R}^m$ provided no solutions of (5.1) blow up in finite time [31], we prefer to introduce the dynamical concepts for (5.1) while disregarding the blowup of solutions which may be outside our interest. For instance, for two subsets A and V of $\mathbb{R}^m$ we say that A attracts V; this means no solutions $x(\cdot)$ with $x(0) \in V$ blow up in finite time, and $d_{\mathrm{H}}(\mathcal{F}(t)V,\, A) \to 0$ as $t \to \infty$. The other concepts such as attraction regions, asymptotic stability, limit sets, invariant sets, attractors, and Morse decompositions can also be defined in the same manner as we do in the situation of GDSs. We omit the details.

PROPOSITION 5.1. *Let* Ω *be a locally compact subset of* $\mathbb{R}^m$*. If* Ω *is positively invariant, then* $\mathcal{F}$ *is a GDS on* Ω*.*

Proof. The proof is a slight modification of the one for Theorem 5.4 in Li and Zhang [31]. □

Remark 5.2. Thanks to Proposition 5.1, one can immediately apply the abstract results in sections 3 and 4 (except Theorems 4.3 and 4.5) to system (5.1) restricted on any locally compact positively invariant sets to establish a Morse decomposition theory for differential inclusions (without any additional assumptions on $f(x)$). Since the results are just copies of those for GDSs, their statements are thus omitted. When referring to these results, we will directly consult the corresponding ones in previous sections.

Now let us give two simple examples on Morse decompositions.

Example 5.1 (gradient system). Let Ω be an open subset of $\mathbb{R}^m$. We say that (5.1) is a *gradient system* on Ω; there exists a function $V(x)$ on Ω such that, for any solution $x(\cdot)$ of (5.1) contained in Ω, $V(x(t))$ strictly decreases whenever $x(\cdot)$ is not an equilibrium. $V(x)$ is called a *Lyapunov function* of (5.1) on Ω.

An example for gradient systems is the one in which $f(x) = -\partial V(x)$, where $V : \mathbb{R}^m \to \mathbb{R}$ is a locally Lipschitz sleek function or a lower semicontinuous convex function and $\partial V(x)$ is the generalized gradient of $V(x)$; see [3, p. 341] and [4, p. 158] for details.

Suppose that (5.1) is a gradient system on Ω with a Lyapunov function $V(x)$ satisfying $V(x) \to +\infty$ as $x \to \partial\Omega$; i.e., for any $a > 0$ there is a compact subset K of Ω such that

$$V(x) > a \qquad \forall\, x \in \Omega \setminus K.$$

Then the set $\mathcal{E}$ of equilibria of the system in Ω is nonvoid, as the minimum point x_* of V in Ω exists and is necessarily an equilibrium. Furthermore, it can be easily shown that the system has an attractor $\mathcal{A}$ in Ω. If we further assume $\mathcal{E}$ to be finite,

$$\mathcal{E} = \{E_k|\ k = 1, 2, \ldots, n\},$$

where E_k $(k = 1, 2, \ldots, n)$ are ordered so that $V(E_i) \le V(E_j)$ when $i < j$, then by Theorem 3.8 (in fact, a copy of this theorem to differential inclusions) we easily deduce that $\{E_1, \ldots, E_n\}$ forms a natural Morse decomposition of $\mathcal{A}$.

Example 5.2. Consider the following differential inclusion which relates to the generalized equations governing Chua's circuit [8]:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} \in A \begin{pmatrix} x_1 - k\,\mathrm{Sgn}(x_1) \\ x_2 \\ x_3 + k\,\mathrm{Sgn}\ (x_1) \end{pmatrix}, \quad A = \begin{pmatrix} -\alpha(b+1) & \alpha & 0 \\ 1 & -1 & 1 \\ 0 & -\beta & 0 \end{pmatrix}, \tag{5.2}$$

where $\dot{x}_i = x_i'(t)$ and $\mathrm{Sgn}(x)$ corresponds to the signal function,

$$\mathrm{Sgn}(x) = 1\ (x > 0), \quad \mathrm{Sgn}(x) = -1\ (x < 0), \quad \mathrm{Sgn}(0) = [-1, 1].$$

Taking $\alpha = -1$, $\beta = 288$, $b = -36$, and $k = 1$, the system reads as follows:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} \in A_0 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 35\mathrm{Sgn}(x_1) \\ 0 \\ 0 \end{pmatrix}, \quad A_0 = \begin{pmatrix} -35 & -1 & 0 \\ 1 & -1 & 1 \\ 0 & -288 & 0 \end{pmatrix}. \tag{5.3}$$

Simple computations show that all the eigenvalues of A_0 are negative, so the system (5.3) is dissipative and has a global attractor $\mathcal{A}$. (5.3) has three equilibria:

$$E_1 = (-1, 0, 1), \quad E_2 = (1, 0, -1), \quad E_3 = (0, 0, 0),$$

where E_1 and E_2 are asymptotically stable. Let $\mathcal{A}_0 = \emptyset$, $\mathcal{A}_1 = \{E_1\}$, $\mathcal{A}_2 = \{E_1, E_2\}$, and $\mathcal{A}_3 = \mathcal{A}$. Then $\{\mathcal{A}_k\}$ is an increasing attractor sequence which yields a Morse decomposition $\mathcal{M} = \{M_1, M_2, M_3\}$ of $\mathcal{A}$ with

$$M_1 = \{E_1\}, \quad M_2 = \{E_2\}, \quad E_3 \in M_3.$$

Concerning the stability of Morse decompositions, we have the following.

THEOREM 5.3. *Assume* (5.1) *has an attractor* $\mathcal{A}$ *with Morse decomposition* $\{M_1, \ldots, M_n\}$. *Then for any compact subset* K *of* $\Omega(\mathcal{A})$, *there exists a* $\delta > 0$ *such that, when* $0 \le \lambda < \delta$, *the inflated system*

$$x'(t) \in f_\lambda(x(t)), \qquad f_\lambda(x) = \overline{\mathrm{conv}}\ f\left(x + \lambda\overline{B}_1\right) + \lambda\overline{B}_1, \tag{5.4}$$

where B_r *denotes the ball* $B(0, r)$ *in* $\mathbb{R}^m$, *has an attractor* $\mathcal{A}(\lambda)$ *with* $K \subset \Omega(\mathcal{A}(\lambda))$ *and a Morse decomposition* $\{M_1(\lambda), \ldots, M_n(\lambda)\}$; *moreover,*

$$\lim_{\lambda \to 0} \delta_H\left(M_k(\lambda),\, M_k\right) = 0 \qquad \forall\ 1 \le k \le n. \tag{5.5}$$

Proof. The proof of the theorem can be obtained by slightly modifying the proof of Theorem 2.10 in [32] and applying Theorem 4.3. We omit the details. □

As a consequence of the above theorem and Theorem 4.2, we have the following.

THEOREM 5.4. *Assume* (5.1) *has an attractor* $\mathcal{A}$ *with Morse decomposition* $\{M_1, \dots, M_n\}$. *Let* $K \subset \Omega(\mathcal{A})$ *be compact. Then for any compact subset* K *of* $\Omega(\mathcal{A})$ *and* $\varepsilon > 0$, *there is a* $\delta > 0$ *such that when* $\lambda < \delta$, *for any solution* $x(\cdot)$ *of the perturbed system* (5.4) *with* $x(0) \in K$, *we have*

$$\limsup_{t\to\infty} d\,(x(t),\, M_k) \le \varepsilon$$

for some M_k.

6. Chain recurrent sets and Morse decompositions. This section is concerned with the open-loop system on $\mathbb{R}^m$:

$$x'(t) = f(x(t), u(t)), \qquad u \in \mathscr{U}, \tag{6.1}$$

where $\mathscr{U} = \{u \in L^\infty(\mathbb{R}; \mathbb{R}^d)|\ u(t) \in U \text{ for a.e. } t \in \mathbb{R}\}$ is the set of admissible controls (inputs), and $U \subset \mathbb{R}^d$ is the control range.

It is well known that chain recurrent sets, chain recurrent components, and chain control sets play a crucial role in the dynamical theory of control. As one of our main purposes here, we give a detailed discussion on the relations between these sets and Morse decompositions. In case one could associate the control system with a single-valued control flow defined on the lifted phase space $\mathscr{X} = \mathscr{U} \times \mathbb{R}^m$, where $\mathscr{U}$ is equipped with the weak*-topology [13, 14], all the results presented below can be obtained by applying some known ones in the theory of single-valued dynamical systems. However, to do so we have to impose on the system some strong assumptions to guarantee the continuity of the control flow. For instance, a typical assumption in this line is to assume that the system is affine in control [13], i.e.,

$$f(x, u) = h(x) + \sum_{1\le i\le d} u_i g_i(x).$$

Our approach here is to connect control systems directly with differential inclusions, which enables us to work under weaker conditions.

We first recall that a mapping $x(\cdot) : I \to \mathbb{R}^m$ is said to be a *solution* of (6.1) on I with input $u \in \mathscr{U}$, if it is absolutely continuous on any compact interval $J \subset I$ and the pair $(x(\cdot), u)$ solves (6.1) at a.e. $t \in I$.

We will denote by $\phi(t, x, u)$ a solution of (6.1) with input u and $\phi(0, x, u) = x$.

THEOREM 6.1 (see [36]). *Assume that the following assumptions hold:*

(C1) *The control* U *is compact;*

(C2) $f(x, u)$ *is continuous with* $F(x) := \{f(x, u)|\ u \in U\}$ *being convex for each* $x \in \mathbb{R}^m$.

Then the open-loop system (6.1) *is equivalent to the following differential inclusion:*

$$x'(t) \in F(x(t)). \tag{6.2}$$

Remark 6.2. Note that under the assumptions (C1) and (C2), $F(x)$ is a continuous set valued mapping with compact convex images.

There is an additional assumption "$f(x, u) \le c(1 + |x|)$ for all $x \in \mathbb{R}^m$" in [36]. However, we point out that this can be removed by using appropriate cutoff functions. In fact, if $x(\cdot)$ is a solution of (6.2) on $[0, T)$, then for any $\tau < T$, $x(\cdot)$ is bounded on

$[0,\tau]$. Assume that $|x(t)| \le R$ on $[0,\tau]$. Let $a(x)$ be a continuous function on $\mathbb{R}^m$ with $a(x) \equiv 1$ for $|x| \le R$ and $a(x) \equiv 0$ for $|x| > 2R$. Then $x(\cdot)$ solves $x'(t) \in a(x(t))F(x(t))$ on $[0,\tau]$. Noting that

$$a(x)F(x) = \{a(x)f(x,u)|\ u \in U\}$$

and that $a(x)f(x,u)$ is bounded, by the original result in [36] one immediately concludes that $x(\cdot)$ is a solution (6.1) on $[0,\tau]$. Since τ is arbitrary, we see that $x(\cdot)$ is a solution of (6.1) on $[0,T)$. Conversely, it is clear that any solution of (6.1) is a solution of (6.2).

Due to Theorem 6.1, we will not distinguish the control system (6.1) with differential inclusion (6.2) in the following argument, and all the dynamical concepts without definitions below are understood with respect to (6.2).

Throughout this section we will always assume and only assume that (C1), (C2), and a local Lipschitz continuity condition (C3) are satisfied, where (C3) is given as follows:

(C3) $f(x,u)$ is locally Lipschitz in x in a uniform manner with respect to $u \in U$; i.e., for any bounded set $B \subset \mathbb{R}^m$, there exists $L_B > 0$ such that

$$|f(x,u) - f(y,u)| \le L_B|x-y| \qquad \forall\, x,y \in B,\ u \in U. \tag{6.3}$$

We emphasize that in our case it is difficult to define a continuous control flow on the lifted space mentioned above. The main difficulty lies in verifying the continuity of the flow due to the nonlinearity of $f(x,u)$ in u (recall that $\mathscr{U}$ is equipped with the weak*-topology).

We denote by $\mathcal{F}$ the reachable mapping of (6.2).

DEFINITION 6.3. *Let K be a closed subset of $\mathbb{R}^m$. For $x,y \in \mathbb{R}^m$ and $\varepsilon, T > 0$, an (ε,T)-chain ζ in K from x to y is given by a positive integer $n \in \mathbb{N}$, $x_0,\dots,x_n \in \mathbb{R}^m$, $u_0,\dots,u_{n-1} \in \mathscr{U}$, and $T_0,\dots,T_{n-1} \ge T$ with $x_0 = x$, $x_n = y$ and*

$$\phi([0,T_i],\, x_i, u_i) \subset K, \quad |\phi(T_i,x_i,u_i) - x_{i+1}| < \varepsilon, \qquad i = 0,1,\dots,n-1.$$

If for all $\varepsilon, T > 0$ there is an (ε,T)-chain from x to y in K, then we say that x is chain controllable to y in K.

Remark 6.4. The set K can be viewed as a constraint. In case $K = \mathbb{R}^m$, we will simply drop the words "in K" in the above definition.

Let $K \subset \mathbb{R}^m$ be closed, $A \subset K$. Define the *K-chain limit set* $\mathscr{C}_K(A)$ as

$$\mathscr{C}_K(A) = \{y \in \mathbb{R}^m|\ \exists\, x \in A \text{ such that } x \text{ is chain controllable to } y \text{ in } K\}.$$

LEMMA 6.5. *Let K be a compact positively invariant set, $A \subset K$. Then $\mathscr{C}_K(A)$ is a compact invariant set with $\omega(A) \subset \mathscr{C}_K(A)$.*

Proof. We prove only the negative invariance of $\mathscr{C}_K(A)$. The closedness and positive invariance as well as inclusion "$\omega(A) \subset \mathscr{C}_K(A)$" can be easily examined.

Let $z \in \mathscr{C}_K(A)$ and $s > 0$. We need to check that there exits $w \in \mathscr{C}_K(A)$ such that $z \in F(s)w$.

Let $\varepsilon_k = 1/k$. Take a sequence of positive numbers $s \le \tau_k \to +\infty$. Then for each k there exists $0 < \delta_k < \varepsilon_k$ such that

$$|\phi(t,x,u) - \phi(t,y,u)| < \varepsilon_k \qquad \forall\, t \in [0,4\tau_k]\ \forall\, u \in \mathscr{U} \tag{6.4}$$

for all $x,y \in K$ with $|x-y| < \delta_k$ (this follows from the Lipschitz continuity property of f). Since $z \in \mathscr{C}_K(A)$, there is a $\widehat{x} \in A$ such that for each k one can find a

$(\delta_k, 2\tau_k)$-chain from $\widehat{x}$ to z in K given by

$$\widehat{x} = x_0, x_1, \ldots, x_{n_k-1}, x_{n_k} = z, \quad u_0, u_1, \ldots, u_{n_k-1} \in \mathscr{U}, \quad T_0, T_1, \ldots, T_{n_k-1} \geq 2\tau_k.$$

We may assume that $T_i \leq 4\tau_k$. Otherwise, since there is no limitation on the number n_k of "jumps," one could modify the original chain to such a one by just dividing the time intervals with length larger than $4\tau_k$.

Set $w_k = \phi(T_{n_k-1} - s, x_{n_k-1}, u_{n_k-1})$. Then $d\,(z, \mathcal{F}(s)w_k) \leq \delta_k$. Since K is compact, we may assume that $w_k \to w$. By upper semicontinuity of $\mathcal{F}$ we have $z \in \mathcal{F}(s)w$. There remains to verify that $w \in \mathscr{C}_K(A)$. For this purpose we first formulate for each w_k an (ε_k, τ_k)-chain from $\widehat{x}$ to w_k in K.

If $n_k = 1$, then clearly

$$\widetilde{x}_0 = \widehat{x}, \quad \widetilde{x}_1 = w_k \in \mathbb{R}^m, \quad \widetilde{T}_0 = T_0 - s \geq \tau_k, \quad \text{and} \quad \widetilde{u}_0 = u_0 \in \mathscr{U}$$

give an (ε_k, τ_k)-chain from $\widehat{x}$ to w_k. So we assume that $n_k \geq 2$. Define $\widetilde{u}_{n_k-2} \in \mathscr{U}$ as

$$\widetilde{u}_{n_k-2}(t) = u_{n_k-2}(t) \quad (t < T_{n_k-2}), \quad \widetilde{u}_{n_k-2}(t) = u_{n_k-1}(t - T_{n_k-2}) \quad (t \geq T_{n_k-2}).$$

Let $\widetilde{T}_{n_k-2} = T_{n_k-2} + (T_{n_k-1} - s)$. Then

$$\begin{aligned} & |\phi(\widetilde{T}_{n_k-2},\, x_{n_k-2}, \widetilde{u}_{n_k-2}) - w_k| \\ = \;& |\phi\left(T_{n_k-1} - s,\, \phi\left(T_{n_k-2},\, x_{n_k-2}, u_{n_k-2}\right),\, u_{n_k-1}\right) - \phi(T_{n_k} - s, x_{n_k-1}, u_{n_k-1})| \end{aligned}$$

Because

$$|\phi\left(T_{n_k-2},\, x_{n_k-2}, u_{n_k-2}\right) - x_{n_k-1}| < \delta_k,$$

by (6.4) we deduce that $|\phi(\widetilde{T}_{n_k-2},\, x_{n_k-2}, \widetilde{u}_{n_k-2}) - w_k| < \varepsilon_k$. It follows that

$$\widehat{x} = x_0, \ldots, x_{n_k-2}, w_k \in \mathbb{R}^m, \quad u_0, \ldots, u_{n_k-3}, \widetilde{u}_{n_k-2} \in \mathscr{U}, \quad T_0, \ldots, T_{n_k-3}, \widetilde{T}_{n_k-2} \geq \tau_k$$

is an (ε_k, τ_k)-chain from $\widehat{x}$ to w_k in K.

Now since $\omega_k \to w$, one easily sees that, for any $\varepsilon, T > 0$, there is an (ε, T)-chain from $\widehat{x}$ to w in K. Hence $w \in \mathscr{C}_K(A)$. □

THEOREM 6.6. *Let K be a compact positively invariant set. Let A be a closed subset of K. Then $\mathscr{C}_K(A)$ is the intersection of all attractors of* (6.2) *in K containing $\omega(A)$.*

Proof. For $\varepsilon, T > 0$, define

$$\mathscr{C}_K(A, \varepsilon, T) = \{y \in K|\ \ \exists\, x \in A \text{ and an } (\varepsilon, T)\text{-chain from } x \text{ to } y \text{ in } K\}.$$

Then $\mathscr{C}_K(A, \varepsilon, T)$ is open in K. Note that $\mathscr{C}_K(A) = \bigcap_{\varepsilon, T>0} \mathscr{C}_K(A, \varepsilon, T)$. Indeed, it is clear that $\mathscr{C}_K(A) \subset \bigcap_{\varepsilon, T>0} \mathscr{C}_K(A, \varepsilon, T)$. Suppose that $y \in \bigcap_{\varepsilon, T>0} \mathscr{C}_K(A, \varepsilon, T)$. Then for each $k \in \mathbb{N}$, there is a $(1/k, k)$-chain from some x_k to y. As K is compact, we can assume $x_k \to x \in A$. Now using some techniques used above, for any $\varepsilon, T > 0$ one can easily formulate an (ε, T)-chain in K from x to y. Hence $y \in \mathscr{C}_K(A)$.

For $\varepsilon, T > 0$, let $V := \overline{\mathscr{C}}_K(A, \varepsilon, T)$. We claim that

$$\omega(V) \subset \mathscr{C}_K(A, \varepsilon, T) \subset \operatorname{int}_K V, \tag{6.5}$$

where $\text{int}_K V$ is the interior of V with respect to K. The second inclusion in (6.5) is in fact obvious. Now let $z \in \omega(V)$. Then there are $x_n \in V$, $t_n \to +\infty$, and $u_n \in \mathscr{U}$ such that $\phi(t_n, x_n, u_n) \to z$. Choose $n_0 \in \mathbb{N}$, $\delta > 0$, and $p \in \mathscr{C}_K(A, \varepsilon, T)$ such that

$$|p - x_{n_0}| < \delta, \quad t_{n_0} > T, \quad \text{and} \quad |\phi(t_{n_0}, x_{n_0}, u_{n_0}) - z| < \varepsilon/2,$$

$$|\phi(t_{n_0}, x, u_{n_0}) - \phi(t_{n_0}, x_{n_0}, u_{n_0})| < \varepsilon/2 \qquad \forall\, x \in K \text{ with } |x - x_{n_0}| < \delta.$$

By definition of p there is an (ε, T)-chain from some $y \in Y$ to p in K. We observe that

$$\begin{aligned} |\phi(t_{n_0}, p, u_{n_0}) - z| &\le |\phi(t_{n_0}, p, u_{n_0}) - \phi(t_{n_0}, x_{n_0}, u_{n_0})| + |\phi(t_{n_0}, x_{n_0}, u_{n_0}) - z| \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

Thus concatenation yields an (ε, T)-chain from y to z in K. This proves (6.5).

(6.5) implies that $\mathcal{A} = \omega(V)$ is an attractor in K. By invariance of $\mathscr{C}_K(A)$, we find that

$$\mathcal{A} = \omega(\overline{\mathscr{C}}_K(A, \varepsilon, T)) \supset \omega(\mathscr{C}_K(A)) = \mathscr{C}_K(A) \supset \omega(A).$$

Therefore we have shown that, for any $\varepsilon, T > 0$, there is an attractor $\mathcal{A}$ of (6.2) in K containing $\omega(A)$ such that $\mathcal{A} \subset \mathscr{C}_K(A, \varepsilon, T)$. Consequently the intersection of all attractors of (6.2) in K containing $\omega(A)$ is necessarily contained in $\mathscr{C}_K(A)$.

Now suppose that $\mathcal{A}$ is an attractor in K containing $\omega(A)$. Let $\delta > 0$ be given arbitrarily but sufficiently small so that

$$V_k := \overline{\mathrm{B}}(\mathcal{A}, k\delta) \cap K \subset \Omega^K(\mathcal{A}), \qquad k = 1, 2.$$

Then V_k are neighborhoods of $\mathcal{A}$ in K and $\mathcal{A}$ attracts V_2. Therefore there is a $t_0 > 0$ such that $\mathcal{F}(t)V_2 \subset V_1$ for all $t > t_0$. Choose a $T > t_0$ such that $\mathcal{F}(T)A \subset V_1$. Then it is trivial to check that, when $\varepsilon < \delta$, any (ε, T)-chain in K from A must end in V_1. It follows that $\mathscr{C}_K(A) \subset V_1$. Since δ is arbitrary, we find $\mathscr{C}_K(A) \subset \mathcal{A}$. The proof is complete. □

Definition 6.7. *Let K be a closed subset of $\mathbb{R}^m$. A K-chain control set $\mathcal{D}$ is a maximal set in K with the property that, for any $x, y \in \mathcal{D}$, x is controllable to y in K.*

The K-chain recurrent set $\mathcal{R}_K$ is defined as

$$\mathcal{R}_K = \{x \in K |\ x \in \mathscr{C}_K(x)\}.$$

The connected components of $\mathcal{R}_K$ is said to be K-chain recurrent components.

Proposition 6.8. *Let K be a compact positively invariant set. Then the following hold:*

(1) *$\mathcal{D}$ is a K-chain control set if and only if it is a K-chain recurrent component;*

(2) *any K-chain recurrent component $\mathcal{D}$ is compact and weakly invariant.*

Proof. (1) The proof of this conclusion is just a copy of the one for [13, Proposition B.2.21] and is thus omitted.

(2) We check only the weak invariance of $\mathcal{D}$. The verification of compactness is trivial.

Let $x \in \mathcal{D}$. We need to show that there is a complete solution through x which lies in $\mathcal{D}$. For this purpose, it suffices to prove that there is a solution $x(\cdot) : [0, 1] \to \mathcal{D}$ as well as a solution $\widetilde{x}(\cdot) : [-1, 0] \to \mathcal{D}$ of (6.2) with $x(0) = x = \widetilde{x}(0)$.

Since $x \in \mathcal{D}$, for each $k \in \mathbb{N}$ there is a $(1/k, k)$-chain from x to x in K given by

$$x_0^k, \dots, x_{n_k}^k \in K, \quad u_0^k, \dots, u_{n_k-1}^k \in \mathscr{U}, \quad T_0^k, \dots, T_{n_k-1}^k \geq k,$$

with $x_0^k = x = x_{n_k}^k$. We can assume that $x_k(\cdot) = \phi(\cdot, x, u_0^k)$ converges uniformly on $[0,1]$ to a solution $x(\cdot)$. We show that $x(s) \in \mathcal{D}$ for all $s \in [0,1]$.

Let $\varepsilon, T > 0$ be given arbitrarily. We formulate an (ε, T)-chain ζ in K from $x(s)$ to $x(s)$ as follows. First, since each solution $x_k(\cdot)$ is defined on $[0, T+1]$ for k sufficiently large, by Barbashin's theorem one can also assume $x(\cdot)$ is defined on $[0, T+1]$ and that $x_k(\cdot)$ converges uniformly on $[0, T+1]$ to $x(\cdot)$. Let $x(t) = \phi(t, x, \widetilde{u})$ for some $\widetilde{u} \in \mathscr{U}$. We choose a $k > 2T + 1$ with $1/k < \varepsilon$ such that

$$|\phi(t, y, u) - \phi(t, z, u)| < \varepsilon \qquad \forall t \in [0, T+1],\ u \in \mathscr{U} \tag{6.6}$$

for all $y, z \in K$ with $|y - z| < 1/k$ and

$$|x_k(t) - x(t)| < \varepsilon \qquad \forall t \in [0, T+1]. \tag{6.7}$$

Define ζ as

$$x_0 = x(s), \quad x_1 = \phi(T + s, x_0^k, u_0^k), \quad x_i = x_{i-1}^k \text{ (for } 2 \leq i \leq n_k), \quad x_{n_k+1} = x(s);$$

$$u_0 = \widetilde{u}, \quad u_1(t) = u_0^k(T + s + t), \quad u_i = u_{i-1}^k \text{ (for } 2 \leq i \leq n_k - 1),$$

$$u_{n_k}(t) = u_{n_k-1}^k(t) \text{ (for } t \leq T_{n_k-1}^k), \quad u_{n_k}(t) = \widetilde{u}(t - T_{n_k-1}^k) \text{ (for } t > T_{n_k-1}^k),$$

$$T_0 = T, \quad T_1 = T_0^k - T - s, \quad T_i = T_{i-1}^k \text{ (for } 2 \leq i \leq n_k - 1), \quad T_{n_k} = T_{n_k-1}^k + s.$$

We claim that ζ is an (ε, T)-chain from $x(s)$ to $x(s)$.

Indeed, simple computations show that

$$|\phi(T_0, x_0, u_0) - x_1| = |x(T + s) - x_k(T + s)| < (\text{by (6.7)}) < \varepsilon,$$

$$|\phi(T_1, x_1, u_1) - x_2| = |\phi(T_0^k, x_0^k, u_0^k) - x_1^k| < 1/k < \varepsilon.$$

We also have

$$\begin{aligned} |\phi(T_{n_k}, x_{n_k}, u_{n_k}) - x_{n_k+1}| &= |\phi\left(s, \phi(T_{n_k-1}^k, x_{n_k-1}^k, u_{n_k-1}^k), \widetilde{u}\right) - x(s)| \\ &= |\phi\left(s, \phi(T_{n_k-1}^k, x_{n_k-1}^k, u_{n_k-1}^k), \widetilde{u}\right) - \phi(s, x, \widetilde{u})|. \end{aligned}$$

Because $|\phi(T_{n_k-1}^k, x_{n_k-1}^k, u_{n_k-1}^k) - x| < 1/k$, by (6.6) we find

$$|\phi(T_{n_k}, x_{n_k}, u_{n_k}) - x_{n_k+1}| < \varepsilon.$$

This finishes the proof of our claim. □

The proof for the existence of a solution $\widetilde{x}(\cdot) : [-1, 0] \to \mathcal{D}$ with $\widetilde{x}(0) = x$ is parallel.

Remark 6.9. Since K-chain recurrent components are compact, they are pairwise disjoint.

PROPOSITION 6.10. *Let K be a compact positively invariant set, and let $x(\cdot)$ be a complete solution in K. Then there are K-chain recurrent components $\mathcal{D}_1$ and $\mathcal{D}_2$ of $\mathcal{R}_K$ such that $\omega(x(\cdot)) \subset \mathcal{D}_1$ and $\alpha(x(\cdot)) \subset \mathcal{D}_2$. Further if $\mathcal{D}_1 = \mathcal{D}_2 = \mathcal{D}$, then $x(\mathbb{R}) \subset \mathcal{D}$.*

Proof. One easily checks that $\omega(x(\cdot)), \alpha(x(\cdot)) \subset \mathcal{R}_K$. Since they are connected, each of them is naturally contained in a K-chain recurrent component of $\mathcal{R}_K$.

In case $\mathcal{D}_1 = \mathcal{D}_2 = \mathcal{D}$, it can be shown that $x(\cdot)$ is contained in $\mathcal{R}_K$ with $\mathcal{D} \cup x(\mathbb{R})$ being a K-chain recurrent component (using the fact that $\mathcal{D}$ is a K-chain control set). Hence by maximality of chain recurrent components we must have $x(\mathbb{R}) \subset \mathcal{D}$.

For K-chain recurrent components $\mathcal{D}_1, \mathcal{D}_2 \subset \mathcal{R}_K$, we denote by $[\mathcal{D}_1, \mathcal{D}_2]$ the set

$$\{x \in K| \quad \text{there is a complete solution } x(\cdot) \text{ in } K \text{ with } x(0) = x \text{ and } \alpha(x(\cdot)) \subset \mathcal{D}_1,\ \omega(x(\cdot)) \subset \mathcal{D}_2\ \}.$$

Then by Proposition 6.10 we have $[\mathcal{D}, \mathcal{D}] = \mathcal{D}$ if K is a compact positively invariant set. □

THEOREM 6.11. *Assume $\mathcal{S}$ is a compact invariant set of* (6.2). *Let $\mathscr{A}$ be the family of attractors of* (6.2) *in $\mathcal{S}$. Then $\mathcal{R}_{\mathcal{S}} = \bigcap_{\mathcal{A} \in \mathbb{A}} (\mathcal{A} \cup \mathcal{A}^*)$.*

Proof. The proof is a slight modification of that of Theorem B.2.26 in [13] (Theorem 6.6 here plays a key role). We omit the details. □

Now we state and prove the following theorem.

THEOREM 6.12. *Assume $\mathcal{S}$ is a compact invariant set of* (6.2). *Let $\mathscr{M}$ be the family of Morse decompositions of $\mathcal{S}$. Then the following assertions hold.*

(1) $\mathcal{R}_{\mathcal{S}} = \bigcap_{\mathcal{M} \in \mathbb{M}} \left(\bigcup_{M \in \mathcal{M}} M\right)$.

(2) *Let $\mathcal{M}$ be a Morse decomposition of $\mathcal{S}$. Then each $\mathcal{S}$-chain recurrent component is contained in some Morse set $M \in \mathcal{M}$. Furthermore, for each $M \in \mathcal{M}$, there is a family $\mathscr{D}$ of $\mathcal{S}$-chain recurrent components such that $M = \bigcup_{\mathcal{D}_1, \mathcal{D}_2 \in \mathbb{D}} [\mathcal{D}_1, \mathcal{D}_2]$.*

(3) *If $\mathcal{R}_{\mathcal{S}}$ has only a finite number of $\mathcal{S}$-chain recurrent components, then $\mathcal{S}$ has the finest Morse decomposition $\mathcal{M}$ with each Morse set being precisely the $\mathcal{S}$-chain recurrent components of $\mathcal{R}_{\mathcal{S}}$.*
Conversely if $\mathcal{S}$ has the finest Morse decomposition $\mathcal{M}$, then $\mathcal{R}_{\mathcal{S}}$ has only a finite number of $\mathcal{S}$-chain recurrent components.

Proof. (1) Since, for any attractor $\mathcal{A}$ in $\mathcal{S}$, $\mathcal{M} = \{\mathcal{A}, \mathcal{A}^*\}$ is a Morse decomposition of $\mathcal{S}$, we see that the intersection is contained in $\mathcal{R}_{\mathcal{S}}$.

Conversely, let $\mathcal{M} = \{M_1, \ldots, M_n\} \in \mathscr{M}$. Then a direct inspection shows that

$$\bigcup_{1 \le k \le n} M_k = \bigcap_{1 \le k \le n} (A_k \cup \mathcal{A}_k^*),$$

where $\emptyset = \mathcal{A}_0 \subset \mathcal{A}_1 \subset \cdots \subset \mathcal{A}_n = \mathcal{A}$ is the increasing attractor sequence which corresponds to $\mathcal{M}$. Due to Theorem 6.11 we see that $\mathcal{R}_{\mathcal{S}} \subset \bigcup_{1 \le k \le n} M_k$. Hence $\mathcal{R}_{\mathcal{S}}$ is also contained in the intersection.

(2) The first conclusion follows from (1) and the fact that Morse sets in the same Morse decomposition are disjoint. Now assume $M \in \mathcal{M}$, and let

$$\mathscr{D} = \{\mathcal{D} \subset M| \ \mathcal{D} \text{ is a } \mathcal{S}\text{-chain recurrent component}\}.$$

Then by Proposition 6.10 we easily see that $M = \bigcup_{\mathcal{D}_1, \mathcal{D}_2 \in \mathbb{D}} [\mathcal{D}_1, \mathcal{D}_2]$.

(3) Suppose that $\mathcal{R}_{\mathcal{S}}$ has n $\mathcal{S}$-chain recurrent components. Let $\mathcal{M}$ be the set of these chain recurrent components.

We first prove that there is a $M \in \mathcal{M}$ which will be marked as M_n such that, for any complete solution $x(\cdot)$ of (6.2), if $\omega(x(\cdot)) \subset M$, then we necessarily have $\alpha(x(\cdot)) \subset M$. Indeed, if this fails to be true, then by Proposition 6.10 one can easily

find $\mathcal{D}_1, \dots, \mathcal{D}_k \in \mathcal{M}$ for some $2 \le k \le n$ and complete solutions $x_i(\cdot)$ $(1 \le i \le k)$ in $\mathcal{S}$ such that

$$\omega(x_i(\cdot)) \subset \mathcal{D}_i \ (1 \le i \le k), \quad \alpha(x_i(\cdot)) \subset \mathcal{D}_{i+1} \ (1 \le i \le k-1), \quad \text{and} \quad \alpha(x_k(\cdot)) \subset \mathcal{D}_1.$$

However this implies that $\bigcup_{1 \le i \le k} \mathcal{D}_i$ is a $\mathcal{S}$-chain recurrent component, a contradiction!

The same argument applies to show that there is a $M_{n-1} \in \mathcal{M}$ such that if a complete solution $x(\cdot)$ satisfies $\omega(x(\cdot)) \subset M_{n-1}$, then $\alpha(x(\cdot)) \subset M_{n-1} \cup M_n$. Continuing the procedure, we finally find that $\mathcal{M}$ can be ordered as $\mathcal{M} = \{M_1, \dots, M_n\}$ with the following property holding: for any complete solution $x(\cdot)$ in $\mathcal{S}$, if $\omega(x(\cdot)) \subset M_k$, then $\alpha(x(\cdot)) \subset M_k \cup \cdots \cup M_n$. Now Theorem 3.8 and Proposition 6.10 imply that $\mathcal{M}$ is a Morse decomposition of $\mathcal{S}$.

We proceed to show that $\mathcal{M}$ is the finest Morse decomposition. We argue by contradiction and suppose that there is a finer Morse decomposition $\widetilde{\mathcal{M}}$. Then using the increasing attractor sequence corresponding to $\widetilde{\mathcal{M}}$, one would find an attractor A in $\mathcal{S}$ such that $M_k \cap A$ is a proper subset of M_k. Take a $y \in M_k \setminus A$. Then $\delta = d(y, A) > 0$. Choose a positive number $\varepsilon_0 < \delta/3$ such that $V := \mathcal{S} \cap \overline{\mathrm{B}}(A, 2\varepsilon_0)$ is a compact subset of $\Omega^{\mathcal{S}}(A)$. Then A attracts V. Thus there exists a $T > 0$ such that $\mathcal{F}(t)V \subset V_0 := \mathcal{S} \cap \mathrm{B}(A, \varepsilon_0)$ for all $t \ge T$. This implies that, when $\varepsilon < \varepsilon_0$, any (ε, T)-chain in $\mathcal{S}$ from V_0 will end in V_0. Consequently there is no (ε, T)-chain in $\mathcal{S}$ from $x \in M_k \cap A$ to y. This contradicts the chain controllability of M_k (see Proposition 6.8).

Now we turn to the proof of the converse conclusion. Assume $\mathcal{S}$ has the finest Morse decomposition $\mathcal{M} = \{M_1, \dots, M_n\}$. Then each M_k is connected. By Remark 3.10 we deduce that, for any Morse decomposition $\mathcal{M}'$, $\bigcup_{1 \le k \le n} M_k \subset \bigcup_{M \in \mathcal{M}'} M$. It then follows from (1) that $\mathcal{R}_{\mathcal{S}} = \bigcup_{1 \le k \le n} M_k$. Therefore each Morse set M_k is a $\mathcal{S}$-chain recurrent component.

The proof of the theorem is complete. □

The following result seems to be interesting, too.

THEOREM 6.13. *Assume $\mathcal{S}$ is a compact invariant set of* (6.2). *Let $\mathscr{M}$ be the family of Morse decompositions of $\mathcal{S}$. Then for all $\varepsilon > 0$, there exists $\mathcal{M} \in \mathscr{M}$ such that $\bigcup_{M \in \mathcal{M}} M \subset B(\mathcal{R}_{\mathcal{S}}, \varepsilon)$.*

Proof. As in [2, Proposition 4.8], it is not difficult to show that there are at most countably many attractors in $\mathcal{S}$. Consequently $\mathscr{M}$ is at most countable. If $\mathscr{M}$ is finite, then the finest Morse decomposition exists, and the conclusion clearly holds true. So we assume $\mathscr{M}$ is infinite. Let $\mathscr{M} = \{\mathcal{M}_1, \mathcal{M}_2, \dots\}$. We choose a sequence $\{\mathcal{M}_{n_k}\}$ as follows. First, we let $\mathcal{M}_{n_1} = \mathcal{M}_1$. Then by using Lemma 3.9 we pick a $\mathcal{M}_{n_2}$ such that $\mathcal{M}_{n_2}$ is finer than both $\mathcal{M}_{n_1}$ and $\mathcal{M}_2$. We proceed by induction and assume that $\mathcal{M}_{n_k}$ is chosen. Then we next pick a $\mathcal{M}_{n_{k+1}}$ from $\mathscr{M}$ such that it is finer than both $\mathcal{M}_{n_k}$ and $\mathcal{M}_{k+1}$ (by Lemma 3.9, this is always available).

Set $A_k = \bigcup_{M \in \mathcal{M}_{n_k}} M$. Then $\{A_k\}$ is a decreasing sequence of compact sets; moreover, $A_k \subset \bigcup_{M \in \mathcal{M}_k} M$ for each k. Thus we conclude that $\mathcal{R}_{\mathcal{S}} = \bigcap_{k \ge 1} A_k = \lim_{k \to \infty} A_k$. Now for any $\varepsilon > 0$, one can easily verify that $A_k \subset \mathrm{B}(\mathcal{R}_{\mathcal{S}}, \varepsilon)$ for k sufficiently large, which proves the validity of the conclusion. □

7. Robustness of feedback laws to small time delays and sample-hold controls. Consider the control system on $\mathbb{R}^m$:

$$x'(t) = f(x(t), u(t)), \quad u(t) \in U, \tag{7.1}$$

where the control range U is a metric space. Given a target set $\mathcal{A} \subset \mathbb{R}^m$, a basic task in feedback control is to design a feedback law $\kappa : \mathbb{R}^m \to U$ so that $\mathcal{A}$ is asymptotically stable (asymptotic controllability) under the closed-loop system

$$x'(t) \in f(x(t),\, \kappa(x(t))). \tag{7.2}$$

(Here we have written the closed-loop system in the form of a differential inclusion instead of a differential equation; this is mainly due to the fact that in many situations the feedback laws may be set-valued [2] or discontinuous [7, 25].) Such a feedback law can be designed via control-Lyapunov functions; see [1, 25]. One can also consult [11] and [29], etc., in the case where $\mathcal{A}$ is an equilibrium.

A very natural and interesting problem is to ask whether the feedback κ (or equivalently the asymptotic behavior of the closed-loop system) possesses some nice robustness properties so that it still works when small time delays are involved in the feedback loop, when one uses a sample-hold control, or when there are measurement errors as well as external disturbances. This problem has important practical sense and has been discussed in the literature by many authors (especially in the case where $\mathcal{A}$ is an equilibrium); see, e.g., [1, 11, 22, 23, 25, 29, 32] and references therein. Here we will try to readdress the problem and establish some new results from the point of view of stability of Morse decompositions. We hope that these results provide more detailed information and thus help us to have a better understanding of the robustness of feedback laws.

7.1. Robustness with respect to small time delays. Consider the closed-loop system with small-time-delay r:

$$x'(t) \in f\left(x(t), \kappa(x_t)\right), \qquad t \geq 0, \tag{7.3}$$

where $x_t = x(t - r(t))$ and $r \in C(\mathbb{R}^+;\, [0, \tau])$ for some $\tau > 0$.

Let $|\cdot|$ be the usual Euclidean norm on $\mathbb{R}^m$. We write $\mathcal{C}_K = C([-\tau, 0]; K)$ for any $K \subset \mathbb{R}^m$. In particular, $\mathcal{C} = \mathcal{C}_{\mathbb{R}^m}$, which is equipped with the norm $\|\cdot\|$ defined by $\|\xi\| = \max_{[-\tau,0]} |\xi(t)|$ for any $\xi \in \mathcal{C}$.

A *solution* of (7.3) with initial value $\xi \in \mathcal{C}$ is a function $x(\cdot) : [-\tau, T) \to \mathbb{R}^m$ which is absolutely continuous on any compact interval $J \subset [0, T)$ and solves (7.3) at a.e. $t \in (0, T)$ with $x(t) = \xi(t)$ for $t \in [-\tau, 0]$. The reader is referred to [4] for existence results on delay differential inclusions.

We denote by $\psi(t, \xi)$ any solution $x(\cdot)$ of (7.3) with initial value ξ.

THEOREM 7.1. *Assume that $f(x) := f(x, \kappa(x))$ satisfies* (H1) *and* (H2). *Suppose that $\mathcal{A}$ is an attractor of the nondelayed system* (7.2) *with Morse decomposition $\{M_1, \ldots, M_n\}$.*

Then for any compact subset K of $\Omega(\mathcal{A})$ and $\varepsilon > 0$, there exist $\tau > 0$ such that, for any delay $r \in C(\mathbb{R}^+;\, [0, \tau])$ and any solution $\psi(t, \xi)$ of (7.3) *with $\xi \in \mathcal{C}_K$, we have*

$$\limsup_{t \to \infty} d\left(\psi(t, \xi),\, M_k\right) \leq \varepsilon \tag{7.4}$$

for some $1 \leq k \leq n$.

Proof. Take an $\eta > 0$ sufficiently small so that $V := \overline{\mathrm{B}}(K, \eta) \subset \Omega(\mathcal{A})$ and is compact. Then $\mathcal{A}$ attracts V under (7.2). Invoking [31, Lemma 5.2], there exists a $R > 0$ such that

$$|x(t)| \leq R \qquad \forall t \geq 0 \tag{7.5}$$

for any solution $x(\cdot)$ of (7.2) with $x(0) \in V$. Hence $K \subset V \subset \overline{\mathrm{B}}_R$.

Let $a(x)$ be a continuous cutoff function on $\mathbb{R}^m$ satisfying

$$a(x) = 1 \text{ when } |x| \le 3R, \qquad a(x) = 0 \text{ when } |x| \ge 4R.$$

Consider the modified systems

$$x'(t) \in h\,(x(t), x_t)\,, \qquad h(x,y) = a(x)a(y)f(x,\kappa(y)), \tag{7.6}$$

$$x'(t) \in H(x(t)), \qquad H(x) = h(x,x). \tag{7.7}$$

Clearly $\mathcal{A}$ is an attractor of (7.7) with $V \subset \Omega_H(\mathcal{A})$. Here $\Omega_H(\mathcal{A})$ is the attraction region of $\mathcal{A}$ under (7.7). For $\lambda > 0$, consider the inflation of (7.7):

$$x'(t) \in H_\lambda(x(t)), \qquad H_\lambda(x) = \overline{\text{conv}}\ H\left(x + \lambda\overline{\text{B}}_1\right) + \lambda\overline{\text{B}}_1. \tag{7.8}$$

Let $\varepsilon > 0$ be given arbitrarily. We may assume $\varepsilon < R/2$. Thanks to Theorem 5.4, there is a $\delta > 0$ such that when $\lambda < \delta$, for any solution $x(\cdot)$ of (7.8) with $x(0) \in V$, we have

$$\limsup_{t\to\infty} d\,(x(t),\, M_k) \le \varepsilon \tag{7.9}$$

for some k. Moreover, by (7.5) we can restrict δ sufficiently small (as in [32, equation (3.6)]) so that

$$|x(t)| \le 2R \qquad \forall t \ge 0. \tag{7.10}$$

We fix a $\lambda = \lambda_1 < \delta$ such that both (7.9) and (7.10) hold.

Note that h is bounded on $\mathbb{R}^{2m}$, so there is a $c_0 > 0$ such that $|h(x,y)| \le c_0$ for all $x, y \in \mathbb{R}^m$. Let $x(t) = \widetilde{\psi}(t,\xi)$ be any solution of (7.6) with initial value $\xi \in \mathcal{C}_K$. Then

$$|x_t - x(t)| = |x'(\theta)|r(t) \le c_0\tau \qquad \text{for } t \ge \tau,$$
$$|x(t) - x(0)| = |x'(\theta)|t \le c_0\tau \qquad \text{for } t \in [0,\tau].$$

Therefore if $\tau > 0$ is taken sufficiently small so that $c_0\tau < \min\{\eta, \lambda_1\}$, then

$$x(t) \in V \text{ (for } 0 \le t \le \tau), \qquad x_t \in \overline{\text{B}}(x(t), \lambda_1) \text{ (for } t \ge \tau). \tag{7.11}$$

Setting $z(t) = x(t+\tau)$ for $t \ge 0$, by (7.11) and the definition of h and H, one sees that $z(t)$ is a solution of the system $(7.8)_{\lambda=\lambda_1}$ with initial value $z(0) \in V$ on any interval $[0,T)$ where $|z(t)| = |x(t+\tau)| \le 2R$. This, together with (7.10), in turn implies that $|z(t)| \le 2R$ for all $t \ge 0$, i.e.,

$$|x(t+\tau)| \le 2R \qquad \forall\ t \ge 0. \tag{7.12}$$

Thus $z(t)$ is a solution of $(7.8)_{\lambda=\lambda_1}$ on $\mathbb{R}^+$. By (7.9) we conclude that, for some $1 \le k \le n$,

$$\limsup_{t\to\infty} d\,(x(t),\, M_k) = \limsup_{t\to\infty} d\,(z(t-\tau),\, M_k) \le \varepsilon. \tag{7.13}$$

We infer from (7.11) and (7.12) that, for any solution $\widetilde{\psi}(t,\xi)$ of (7.6) with $\xi \in \mathcal{C}_K$,

$$|\widetilde{\psi}(t,\xi)| \le 2R \qquad \forall t \ge 0,$$

and hence it is a solution of (7.3). Conversely, using this basic fact one can also easily examine that any solution $\psi(t,\xi)$ of the original system (7.3) with initial value $\xi \in \mathcal{C}_K$ exists on $\mathbb{R}^+$ and is a solution of (7.6). This and (7.13) complete the proof of what we desired. □

Remark 7.2. The results can be extended without any difficulty to differential inclusion

$$x'(t) \in f(x(t), x(t-r_1(t)), \dots, x(t-r_k(t))), \qquad t \geq 0$$

with multiple small time delays. Here we omit the details.

Remark 7.3. A particular but interesting case is the one where each Morse set M_k consists of an equilibrium E_k, as in the situation of a gradient system. In such a case (7.4) reads

$$\limsup_{t\to\infty} d\left(\psi(t,\xi),\, E_k\right) \leq \varepsilon. \tag{7.14}$$

The robustness of asymptotic stability with respect to small time delays for scalar differential equations with multiple equilibria can be found in [38, 39, 44], etc., where the authors used some monotonicity method to show that each bounded solution of the small-time-delayed system approaches one of the equilibria. Similar results were also established in Friesecke [19] for a scalar parabolic equation with small time delays by using the special Lyapunov function of the system.

Remark 7.4. We refer the reader to [23, 26, 32, 33, 35] and the large number of references cited therein for the works and related discussions on robustness of asymptotic stability of a single equilibrium or a compact set with respect to small time delays.

7.2. Robustness with respect to sample-hold controls. We first recall the concept of π-solutions of the closed-loop system. Let

$$\pi:\ 0 = t_0 < t_1 < \cdots < t_i < t_{i+1} < \cdots, \quad \text{where } t_i \to \infty,$$

be a partition of $\mathbb{R}^+$, $||\pi|| = \sup_{i\geq 0} |t_{i+1} - t_i|$. Given an initial value $x_0 \in \mathbb{R}^m$, a *π-solution* of the closed-loop system (7.2) on $[0,T)$ is a function $x(\cdot) : [0,T) \to \mathbb{R}^m$ which is absolutely continuous on any compact interval $J \subset [0,T)$ and satisfies

$$x'(t) = f\left(x(t),\, \kappa\left(x(t_i)\right)\right), \quad x(0) = x_0, \qquad \text{a.e. } t \in [t_i, t_{i+1}] \cap [0,T)$$

for all $i \geq 0$ such that $t_{i+1} \leq T$.

THEOREM 7.5. *Assume that $f(x) := f(x,\kappa(x))$ satisfies* (H1) *and* (H2). *Suppose that $\mathcal{A}$ is an attractor of the closed-loop system* (7.2) *with a Morse decomposition $\{M_1, \dots, M_k\}$.*

Then for any compact subset $K \subset \Omega(\mathcal{A})$ and $\varepsilon > 0$, there is a $\delta > 0$ such that, when $||\pi|| < \delta$, any π-solution $x(\cdot)$ of (7.2) *with $x(0) \in K$ exists on $\mathbb{R}^+$ and satisfies*

$$\limsup_{t\to\infty} d(x(t), M_k) < \varepsilon$$

for some M_k.

Proof. We may assume that f is bounded on $\mathbb{R}^m \times U$; otherwise, we can employ the cutoff techniques used in the proof of Theorem 7.1.

Let $|f(x,u)| \leq c_0$ for all $(x,u) \in \mathbb{R}^m \times U$. Then for any $\lambda > 0$, when $||\pi|| < \lambda/c_0$ we find that any π-solution $x(\cdot)$ of (7.2) satisfies

$$|x(t_i) - x(t)| \leq c_0||\pi|| < \lambda \qquad \forall i \geq 0\ \forall t \in [t_i, t_{i+1}),$$

and hence $x(\cdot)$ solves

$$x'(t) \in f\left(x(t),\, \kappa(x(t) + \lambda \overline{\mathrm{B}}_1)\right), \qquad t \geq 0.$$

Now the conclusion follows immediately from Theorem 5.4. □

Remark 7.6. Similarly we could consider robustness of feedback laws with respect to measurement errors and external disturbances. In fact, results along this line are readily contained in Theorems 5.3 and 5.4.

Acknowledgments. First, I highly appreciate the work of the anonymous referees whose comments and suggestions have helped me greatly improve the paper in many ways. Second, I would like to express my gratitude to Professor P. E. Kloeden and Dr. Y. J. Wang for many helpful discussions. Finally, special thanks go to my wife, Q. P. Ren, Dr. A. Y. Zhang, and Z. Li for their constant support and encouragement in the past years. This work was carried out while I was visiting Chern Institute of Mathematics. I would like to express my sincere thanks to the support and hospitality of the Institute.

REFERENCES

[1] F. Albertini and E. D. Sontag, *Continuous control-Lyapunov functions for asymtotically controllable time-varying systems*, Internat. J. Control, 72 (1999), pp. 1630–1641.
[2] E. Akin, *The General Topology of Dynamical Systems*, Grad. Stud. Math. 1, Amer. Math. Soc., Providence, RI, 1993.
[3] J. P. Aubin, *Viability Theory*, Birkhäuser Boston, Cambridge, MA, 1991.
[4] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.
[5] J. P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Cambridge, MA, 1990.
[6] A. Bacciotti and N. Kalouptsidis, *Topological dynamics of control system: Stability and attraction*, Nonlinear Anal., 10 (1986), pp. 547–565.
[7] R. W. Brochett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brochett, R. S. Millman, and H. J. Sussmann, eds., Birkhäuser Boston, Cambridge, MA, 1983, pp. 181–191.
[8] R. Brown, *Generalizations of the Chua equations*, IEEE Trans. Circuits Syst. Fund. Theory Appl., 40 (1993), pp. 878–883.
[9] T. Caraballo and J. A. Langa, *Global attractors for multivalued random dynamical systems generated by random differential inclusions with multiplicative noise*, J. Math. Anal. Appl., 260 (2001), pp. 602–625.
[10] L. J. Cherene, Jr., *Set Valued Dynamical System and Economic Flow*, Lecture Notes in Econom. and Math. Systems 158, Springer-Verlag, Berlin, 1978.
[11] F. H. Clarke, Yu. S. Ledyaev, L. Rifford, and R. J. Stern, *Feedback stabilization and Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 25–48.
[12] F. H. Clarke, Yu. S. Ledyaev, and R. J. Stern, *Asymptotic stability and smooth Lyapunov functions*, J. Differential Equations, 149 (1998), pp. 69–114.
[13] F. Colonius and W. Kliemann, *The Dynamics of Control*, Birkhäuser Boston, Cambridge, MA, 2000.
[14] F. Colonius and W. Kliemann, *Limits of input-to-state stability*, Systems Control Lett., 49 (2003), pp. 111–120.
[15] C. Conley, *Isolated Invariant Sets and the Morse Index*, Regional Conf. Ser. Math. 38, Amer. Math. Soc., Providence, RI, 1978.
[16] H. Crauel, L. H. Duc, and S. Siegmund, *Towards a Morse theory for random dynamical systems*, Stoch. Dyn., 4 (2004), pp. 277–296.
[17] K. Deimling, *Multi-valued Differential Equations*, De Gruyter, Berlin, 1998.
[18] A. F. Filippov, *Differential Equations with Discontinuous Righthand Side*, Kluwer Academic Publishers, Dordrecht, 1998.
[19] C. R. Friesecke, *Convergence to equilibrium for delay-diffusion equations with small delay*, J. Dynam. Differential Equations, 5 (1993), pp. 89–103.
[20] G. T. Gedeon and G. Hines, *Upper semicontinuity of Morse sets of a discretization of a delay-differential equation*, J. Differential Equations, 151 (1999), pp. 36–78.

[21] G. T. Gedeon and G. Hines, *Upper semicontinuity of Morse sets of a discretization of a delayed-differential equation: An improvement*, J. Differential Equations, 179 (2002), pp. 369–383.
[22] L. Grüne, *Asymptotic Behavior of Dynamical and Control Systems under Perturbation and Discretization*, Springer-Verlag, Berlin, 2002.
[23] J. K. Hale and S. M. Verduyn Lunel, *Effects of Small Delays on Stability and Control*, Operator Theory and Analysis 122 (The M.A. Kaashoek Anniversary Volume), H. Bart, I. Gohberg, and A. C. M. Ran, eds., Birkhäuser Boston, Cambridge, MA, 2001, pp. 275–301.
[24] A. V. Kapustyan and J. Valero, *Attractors of multivalued semiflows generated by differential inclusions and their approximations*, Abstr. Appl. Anal., 5 (2000), pp. 33–46.
[25] C. M. Kellett and A. R. Tell, *Weak convergence Lyapunov theorems and control-Lyapunov functions*, SIAM J. Control Optim., 42 (2004), pp. 1934–1959.
[26] V. L. Kharitonov, *Robust stability analysis of time delay systems: A Survey*, Ann. Rev. Control, 23 (1999), pp. 185–196.
[27] P. E. Kloeden, *Asymptotic invariance and limit sets of general control systems*, J. Differential Equations, 19 (1975), pp. 91–105.
[28] P. E. Kloeden, *Eventual stability in general control systems*, J. Differential Equations, 19 (1975), pp. 106–124.
[29] Yu. S. Ledyaev and E. D. Sontag, *A Lyapunov charaterization of robust stability*, Nonlinear Anal., 37 (1999), pp. 813–840.
[30] D. S. Li, *On dynamical stability in general dynamical systems*, J. Math. Anal. Appl., 263 (2001), pp. 455–478.
[31] D. S. Li and X. X. Zhang, *On the stability in general dynamical systems and differential inclusions*, J. Math. Anal. Appl., 274 (2002), pp. 705–724.
[32] D. S. Li and P. E. Kloeden, *Robustness of asymptotic stability to small time delays*, Discrete Contin. Dyn. Syst., 13 (2005), pp. 1007–1034.
[33] H. Logemann, R. Rebarber, and G. Weiss, *Conditions for robustness and nonrobustness of the stability of feedback systems with respect to small delays in the feedback loop*, SIAM J. Control Optim., 37 (1996), pp. 572–600.
[34] V. S. Melnik and J. Valero, *On attractors of multivalued semi-flows and differential inclusions*, Set-Valued Anal., 6 (1998), pp. 83–111.
[35] X. R. Mao, *Exponential stability of nonlinear differential delay equations*, Systems Control Lett., 28 (1996), pp. 159–165.
[36] J. W. Nieuwenhuis, *Some remarks on set-valued dynamical systems*, J. Aust. Math. Soc., 22 (1981), pp. 308–313.
[37] G. Ochs, *Weak Random Attractors*, Technical report 449, Institut für Dynamische Systeme, Universität Bremen, Bremen, Germany, 1999.
[38] M. Pituk, *Convergence to equilibria in a differential equation with small delay*, Math. Bohem., 127 (2002), pp. 293–299.
[39] M. Pituk, *Convergence to equilibria in scalar nonquasimonotone functional differential equations*, J. Differential Equations, 193 (2003), pp. 95–130.
[40] M. Rasmussen, *Morse decompositions of nonautonomous dynamical systems*, Trans. Amer. Math. Soc., to appear.
[41] E. Roxin, *Stability in general control systems*, J. Differential Equations, 1 (1965), pp. 115–150.
[42] E. Roxin, *On generalized dynamical systems defined by contingent equations*, J. Differential Equations, 1 (1965), pp. 188–205.
[43] K. P. Rybakowski, *The Homotopy Index and Partial Differential Equations*, Springer-Verlag, Berlin, 1987.
[44] H. L. Smith and H. R. Thieme, *Monotone semiflows in scalar non-quasi-monotone functional differential equations*, Amer. Math. Anal. Appl., 150 (1990), pp. 289–306.
[45] G. P. Szegö and G. Treccani, *Semigruppi di Trasformazioni Multivoche*, Lecture Notes in Math. 101, Springer-Verlag, Berlin, 1969.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 61–83

ON CONVERGENCE IN ELLIPTIC SHAPE OPTIMIZATION*

KARSTEN EPPLER†, HELMUT HARBRECHT‡, AND REINHOLD SCHNEIDER‡

Abstract. The present paper aims at analyzing the existence and convergence of approximate solutions in shape optimization. Motivated by illustrative examples, an abstract setting of the underlying shape optimization problem is suggested, taking into account the so-called two norm discrepancy. A Ritz–Galerkin-type method is applied to solve the associated necessary condition. Existence and convergence of approximate solutions are proved, provided that the infinite dimensional shape problem admits a stable second order optimizer. The rate of convergence is confirmed by numerical results.

Key words. shape optimization, shape calculus, existence and convergence of approximate solutions, optimality conditions

AMS subject classifications. 49Q10, 49K20, 49M15, 65K10

DOI. 10.1137/05062679X

1. Introduction.

Shape optimization is quite important for aircraft design, bridge construction, electromagnetic shaping, etc. Many problems that arise in applications, particularly in structural mechanics, can be formulated as the minimization of functionals defined over a class of admissible domains. Such problems have been intensively studied in the literature in the past 25–30 years (see [14, 33, 35, 44, 47] and the references therein). In the majority of papers, the undiscretized problem has been studied. Only a few papers deal with the convergence of approximate solutions to the solution of the original shape optimization problem. For example, in [6, 7, 8] the question of convergence is considered on the fully discretized level. Therein, a grid is fixed in advance on the hold all and the admissible shapes are allowed to vary only *on* this predefined grid. Consequently, a *discrete* optimization problem has to be solved next. Further investigations on convergence of approximate solutions have been reported in [33, 44].

In [18, 19, 20, 22, 23, 24, 25], we considered the numerical solution of several elliptic shape optimization problems. Boundary variations were used to derive boundary integral representations of the shape gradient and the shape Hessian. This approach allows the embedding of a shape problem into a Banach space by identifying the domain with the parametrization of its boundary, i.e., with a function. Solving the shape optimization problem becomes equivalent to finding the parametrization of the minimizing domain. We applied a Ritz–Galerkin-type method to approximate this parametrization. All ingredients of the shape gradient and Hessian that arise from the state equation were computed with sufficiently high accuracy by a fast wavelet boundary element method. In this way, the discretization of the shape is decoupled

*Received by the editors March 14, 2005; accepted for publication (in revised form) September 1, 2006; published electronically March 2, 2007. This research was carried out during the second author's visit to the Department of Mathematics, Utrecht University, The Netherlands, and was supported by the EU-IHP project *Nonlinear Approximation and Adaptivity: Breaking Complexity in Numerical Modelling and Data Representation.*

http://www.siam.org/journals/sicon/46-1/62679.html

†Institute of Numerical Mathematics, TU Dresden, Zellescher Weg 12–14, 01069 Dresden, Germany (karsten.eppler@tu-dresden.de).

‡Institut für Informatik und Praktische Mathematik, Christian–Albrechts–Universität zu Kiel, Olshausenstr. 40, 24098 Kiel, Germany (hh@numerik.uni-kiel.de, rs@numerik.uni-kiel.de).

from the discretization of the state equation. Consequently, we may distinguish two types of errors.

First, the discretization error of the shape refers to the approximation error and determines the best possible rate of convergence. The present paper mainly tackles this issue by proving *existence* and *convergence* of approximate solutions. To this end, it is assumed that the objective, the constraints, and the state are given exactly.

Second, solving the state equation numerically induces a consistency error. Consistency errors are also caused by the approximate computation of the objective and constraints by, e.g., numerical quadrature. We present a Strang-type lemma to incorporate the error arising from numerical approximation. It gives a sufficient condition for realizing the best order of convergence.

When identifying the boundary of the regular $C^{k,\alpha}$-domain with its parametrization with respect to a fixed reference manifold $\widehat{\Gamma}$, a shape calculus based on boundary variational fields of prescribed smoothness leads to a second order Frechét calculus in a Banach space. For applications of interest, the space $C^{2,\alpha}(\widehat{\Gamma})$ for a certain $\alpha \in (0,1]$ is appropriate; cf. [15, 16, 17]. Since shape optimization problems are highly nonlinear, we are looking for domains that satisfy the first order necessary condition. These solutions are called stationary domains. To verify their local optimality, the second order Frechét derivative has to be coercive. However, for *integral objectives* in elliptic shape optimization it turns out that coercivity cannot be expected in the norm of the space of differentiation $C^{2,\alpha}(\widehat{\Gamma})$. Instead, coercivity of the shape Hessian at $\Omega^\star$ can be usually shown only in a *weaker* Sobolev space $H^s(\widehat{\Gamma})$. This *lack of coercivity* is known from other PDE-constrained optimal control problems as the so-called *two norm discrepancy*; cf. [4, 5, 28, 29]. The two norm discrepancy in shape optimization was first observed in [10, 11, 12, 15, 17]. It will play a key role in our convergence analysis.

Our investigations concentrate on the optimization of shapes and are not applicable to dealing with topological changes. Certainly, dealing with variable topologies is of enormous practical interest, and much important work has been done for the theoretical foundation and development of algorithms; see the monograph [3] for the state of the art. The so-called topological derivative has been addressed in [27, 31, 37, 39, 48, 49, 39] (we mention only some of the related papers). Related necessary optimality conditions for simultaneous shape and topology optimization have been investigated in [50], but the study of sufficient optimality conditions seems to be a challenging problem.

Concerning the present paper, section 2 is dedicated to a summary of second order shape calculus. Additionally, some examples are presented to illustrate the two norm discrepancy. First, we consider shape functionals based on a simple domain or boundary integral. Then, we treat PDE-constrained shape optimization problems by means of elliptic free boundary problems. In addition to the problem with simple constraints on the domain, we also discuss shape optimization problems subject to further functional constraints.

Motivated by these examples, we present in section 3 an abstract setting for the investigation of the second order sufficient optimality condition to verify stable minimizers. Then we introduce suitable trial spaces to discretize the shape optimization problem by means of a Ritz–Galerkin method for solving the necessary condition. The Ritz–Galerkin method solves a finite dimensional optimization problem that arises from restricting the class of admissible domains to domains given by the trial space. We show that there exist approximate solutions, provided that the level of discretization is sufficiently large, and prove convergence of the approximate solutions $\Omega_N^\star$ to

$\Omega^\star$, the optimal solution of the original infinite dimensional shape problem. The approximate solution behaves like the best approximation in the trial space to $\Omega^\star$, with respect to the natural space of coercivity of the shape Hessian. Therefore, the computation of the rate of convergence is along the lines of conventional approximation theory.

In section 4, we present two numerical examples that confirm our analysis. The first one is a simple shape problem based on a domain integral minimization, which is mainly incorporated for illustration. The second is a more advanced PDE-constrained shape optimization problem, with several additional functional constraints. Both examples are chosen such that the optimal domain is known a priori. We observe rates of convergence which verify the present theory.

2. Motivation and background.

2.1. Shape calculus. Shape optimization is concerned with the minimization of the shape functional

$$J(\Omega) = \int_\Omega j(u, \nabla u, \mathbf{x}) d\mathbf{x} \to \min, \qquad \Omega \in \Upsilon, \tag{2.1}$$

where Υ is a suitable class of admissible domains $\Omega \in \mathbb{R}^n$. The so-called *state* u satisfies an abstract boundary value problem

$$\mathcal{A}u = f \text{ in } \Omega, \qquad \mathcal{B}u = g \text{ on } \Gamma, \tag{2.2}$$

where $\mathcal{A}$ corresponds to a well-posed elliptic partial differential operator in the domain Ω, and $\mathcal{B}$ operates on the functions supported at the free boundary $\Gamma \subset \partial\Omega$. For the sake of simplicity, we restrict ourselves to finding solutions with known topology and assume that all involved functions and data are sufficiently smooth.

Generally, problem (2.1) is highly implicit, with respect to the shape of the domain, and has to be solved iteratively. The canonical way to solve the minimization problem is to determine its stationary points. Then, via the second order optimality condition, regular minimizers of second order are verified. To this end, we will briefly survey shape calculus. In particular, we refer the reader to Murat and Simon [38], Simon [46], Pironneau [44], Sokołowski and Zolésio [47], Delfour and Zolésio [14], and the references therein. Herein, two basic concepts are considered, namely, the perturbation of identity (Murat and Simon) and the speed method (Sokołowski and Zolésio).

For example, the perturbation of identity exploits a smooth perturbation field $\mathbf{U} : \Omega \to \mathbb{R}^n$ and defines the standard domain perturbation as

$$\Omega_\varepsilon[\mathbf{U}] := \{(\mathbf{I} + \varepsilon\mathbf{U})(\mathbf{x}) : \mathbf{x} \in \Omega\}.$$

Then the directional derivative of $J(\Omega)$ is computed as

$$\nabla J(\Omega)[\mathbf{U}] := \lim_{\varepsilon \to 0} \frac{J(\Omega_\varepsilon[\mathbf{U}]) - J(\Omega)}{\varepsilon}.$$

Ever since Hadamard [32] it has been known that $\nabla J(\Omega)[\mathbf{U}]$ is a distribution living only on the free boundary of the domain Ω, provided that $J(\Omega)$ is shape differentiable; see also [13].

The latter observation leads to the idea of considering only boundary variations for the update in the optimization algorithm. Therefore, we shall directly apply

boundary variations for the computation of the boundary integral representations of the shape gradient and Hessian. To this end, we introduce a reference manifold $\widehat{\Gamma} \subset \mathbb{R}^n$ and consider a fixed boundary perturbation field, for example, in the direction of the outer normal $\widehat{\mathbf{n}}$. We suppose that the free boundary of each domain $\Omega \in \Upsilon$ can be parametrized via a sufficiently smooth function r in terms of

$$\boldsymbol{\gamma} : \widehat{\Gamma} \to \Gamma, \quad \boldsymbol{\gamma}(\mathbf{x}) = \mathbf{x} + r(\mathbf{x})\widehat{\mathbf{n}}(\mathbf{x}).$$

That is, we can identify a domain with the scalar function r. Defining the standard variation

$$\boldsymbol{\gamma}_\varepsilon : \widehat{\Gamma} \to \Gamma_\varepsilon, \quad \boldsymbol{\gamma}_\varepsilon(\mathbf{x}) := \boldsymbol{\gamma}(\mathbf{x}) + \varepsilon dr(\mathbf{x})\widehat{\mathbf{n}}(\mathbf{x}),$$

where dr is again a sufficiently smooth scalar function, we obtain the perturbed domain Ω_ε. Consequently both the shape and its increment can be seen as elements of a Banach space X. We will specify the notion of "sufficiently smooth" in the next subsections.

2.2. Optimization of domain or boundary integrals. First, we introduce some notation. For a given domain $D \in \mathbb{R}^n$, the space $C^2(\overline{D})$ consists of all two times continuously differentiable functions $f : \overline{D} \to \mathbb{R}^m$. A function $f \in C^2(\overline{D})$ belongs to $C^{2,\alpha}(\overline{D})$ if the (spatial) Hessian $\nabla^2 f$ is Hölder continuous with coefficient $0 < \alpha \leq 1$. A domain $D \in \mathbb{R}^n$ is of class $C^{2,\alpha}$ if for each $\mathbf{x} \in \partial D$ a neighborhood $U(\mathbf{x}) \subseteq \partial\Omega$ and a diffeomorphism $\boldsymbol{\gamma} : [0,1]^{n-1} \to \overline{U(\mathbf{x})}$ exist such that $\boldsymbol{\gamma} \in C^{2,\alpha}([0,1]^{n-1})$; see [52], for example.

For the sake of clearness, we present here two elementary shape problems, since both the shape calculus and the analysis become much more evident in comparison with the more advanced shape optimization problems presented in the subsequent subsections. To this end, let $n = 2$, $\Omega \in C^1$, and consider the following shape optimization problem of domain integral type:

$$J(\Omega) = \int_\Omega h(\mathbf{x}) d\mathbf{x} \to \min, \tag{2.3}$$

where $h \in C^1(\mathbb{R}^2)$ are given data. We choose the class of admissible domains as the set of all domains that are star-shaped with respect to the origin. Then we can choose $\widehat{\Gamma}$ as the unit circle. Equivalently, we can parametrize $\Gamma = \partial\Omega$ via polar coordinates

$$\Gamma := \Big\{ \boldsymbol{\gamma}(\phi) = r(\phi) \left[\begin{smallmatrix} \cos\phi \\ \sin\phi \end{smallmatrix} \right] : \phi \in [0, 2\pi] \Big\},$$

where $r \in C^1_{\mathrm{per}}([0,2\pi])$ is a positive function. Here and in what follows, the space $C^{k,\alpha}_{\mathrm{per}}$ is defined as

$$C^{k,\alpha}_{\mathrm{per}}([0,2\pi]) = \{ f \in C^{k,\alpha}([0,2\pi]) : f^{(i)}(0) = f^{(i)}(2\pi) \text{ for all } i = 0, \ldots, k \},$$

and likewise $C^k_{\mathrm{per}}([0,2\pi])$. Let us further remark that the tangent and the outward normal at Γ are computed by

$$\mathbf{t} = \frac{r' \left[\begin{smallmatrix} \cos\phi \\ \sin\phi \end{smallmatrix} \right] + r \left[\begin{smallmatrix} -\sin\phi \\ \cos\phi \end{smallmatrix} \right]}{\sqrt{r^2 + r'^2}}, \qquad \mathbf{n} = \frac{r' \left[\begin{smallmatrix} \sin\phi \\ -\cos\phi \end{smallmatrix} \right] + r \left[\begin{smallmatrix} \cos\phi \\ \sin\phi \end{smallmatrix} \right]}{\sqrt{r^2 + r'^2}}. \tag{2.4}$$

We consider $dr \in C^1_{\mathrm{per}}([0,2\pi])$ as a standard variation for perturbed domains Ω_ε, respectively, boundaries Γ_ε, defined by $r_\varepsilon(\phi) = r(\phi) + \varepsilon dr(\phi)$, where $\boldsymbol{\gamma}_\varepsilon(\phi) = r_\varepsilon(\phi)\widehat{\mathbf{n}}(\phi)$

is always a Jordan curve. Herein, $\widehat{\mathbf{n}}(\phi) = [\cos\phi, \sin\phi]^T$ denotes the outward normal vector to the reference manifold $\widehat{\Gamma}$.

LEMMA 2.1 (see [16]). *The shape functional from* (2.3) *is twice Frechét differentiable with respect to* $C^1_{\mathrm{per}}([0,2\pi])$*, where the shape gradient and Hessian read as*

$$\nabla J(\Omega)[dr] = \int_0^{2\pi} r(\phi) dr(\phi) h\big(r(\phi),\phi\big) d\phi,$$
$$\nabla^2 J(\Omega)[dr_1, dr_2] = \int_0^{2\pi} dr_1(\phi) dr_2(\phi) \left\{ h\big(r(\phi),\phi\big) + r(\phi) \frac{\partial h}{\partial \widehat{\mathbf{n}}}\big(r(\phi),\phi\big) \right\} d\phi.$$

Consider now a stationary domain $\Omega^\star$, which means $\nabla J(\Omega^\star)[dr] = 0$ for all $dr \in C^1([0,2\pi])$. Of course, the latter equation implies that $h|_{\Gamma^\star} \equiv 0$. Hence, as one readily verifies, it holds that

$$\nabla^2 J(\Omega^\star)[dr_1, dr_2] = \int_0^{2\pi} dr_1(\phi) dr_2(\phi) \left\{ \frac{r^{\star 2}(\phi)}{\sqrt{r^{\star 2}(\phi) + r^{\star\prime 2}(\phi)}} \frac{\partial h}{\partial \mathbf{n}}\big(r^\star(\phi),\phi\big) \right\} d\phi.$$

Optimality usually can be guaranteed by coercivity of the second order Frechét derivative. However, it is impossible to realize coercivity with respect to $C^1_{\mathrm{per}}([0,2\pi])$; only an estimate

$$\nabla^2 J(\Omega^\star)[dr, dr] \geq c_E \|dr\|^2_{L^2([0,2\pi])}$$

for some $c_E > 0$ can be expected. Note that we have such an estimate if $(\partial h/\partial \mathbf{n})|_{\Gamma^\star} \geq c_E > 0$. This lack of regularity is known from other control problems as the so-called *two norm discrepancy*. Nevertheless, the bilinear form imposed by the shape Hessian $\nabla^2 J(\Omega)$ is obviously also continuous on $L^2([0,2\pi]) \times L^2([0,2\pi])$, that is,

$$|\nabla^2 J(\Omega)[dr_1, dr_2]| \leq c_S(\Omega) \|dr_1\|_{L^2([0,2\pi])} \|dr_2\|_{L^2([0,2\pi])}$$

for all $dr_1, dr_2 \in L^2([0,2\pi])$. Notice that it is generally impossible to extend the domain of definition $C^1([0,2\pi])$ to $L^2([0,2\pi])$. In other words, J is only densely defined with respect to $L^2([0,2\pi])$.

Also, in the case of a shape optimization problem of boundary integral type

$$J(\Omega) = \int_\Gamma g(\mathbf{x}) d\sigma \to \min, \tag{2.5}$$

where $g \in C^2(\mathbb{R}^2)$ are given data, one makes the above observations concerning the coercivity. Similarly to above, coercivity cannot be realized in $C^1_{\mathrm{per}}([0,2\pi])$. The energy space of the bilinear form imposed by the shape Hessian $\nabla^2 J(\Omega)$ is the Sobolev space $H^1_{\mathrm{per}}([0,2\pi])$; see [16] for details.

2.3. PDE-constrained shape optimization problems. We shall consider free elliptic boundary problems as the most illustrative model problem for PDE-constrained shape optimization problems. Let $T \subset \mathbb{R}^n$ denote a bounded domain with boundary $\partial T = \Gamma$. Inside the domain T we assume the existence of a simply connected subdomain $S \subset T$ with fixed boundary $\partial S = \Sigma$. We denote the annular domain $T \setminus \overline{S}$ by Ω; see also Figure 2.1.

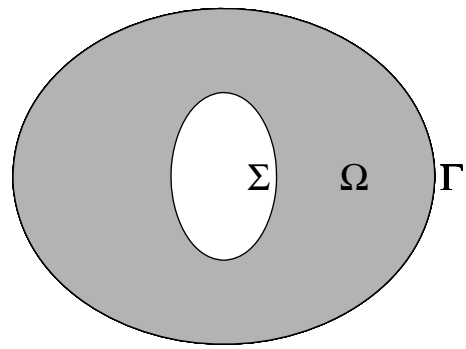


FIG. 2.1. *The domain* Ω *and its boundaries* Γ *and* Σ.

We consider the following overdetermined boundary value problem in the annular domain Ω:

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ \|\nabla u\| &= g && \text{on } \Gamma, \\ u &= 0 && \text{on } \Gamma, \\ u &= h && \text{on } \Sigma, \end{aligned} \tag{2.6}$$

where $f \geq 0$ and $g, h > 0$ are sufficiently smooth functions such that the shape differentiability of the objective (2.7) is provided up to second order. We like to stress that the positivity of the data implies that u is positive in Ω. Hence, there holds the identity

$$\|\nabla u\| \equiv -\frac{\partial u}{\partial \mathbf{n}} \qquad \text{on } \Gamma$$

since u admits homogeneous Dirichlet data on Γ.

We arrive at a free boundary problem if the boundary Γ is the unknown. In other words, we seek a domain Ω with fixed boundary Σ and unknown boundary Γ such that the overdetermined boundary value problem (2.6) is solvable. For the existence of solutions we refer the reader to, e.g., [1, 26].

Shape optimization provides an efficient tool for solving such free boundary value problems; cf. [14, 34, 47, 51]. Considering the cost functional

$$J(\Omega) = \int_\Omega \|\nabla u\|^2 - 2fu + g^2 d\mathbf{x} \tag{2.7}$$

with underlying *state equation*

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma, \\ u &= h && \text{on } \Sigma, \end{aligned} \tag{2.8}$$

the solution of the free boundary problem is equivalent to the shape optimization problem

$$J(\Omega) \to \min. \tag{2.9}$$

This issues from the necessary condition of a minimizer to the cost functional (2.7); that is,

$$\nabla J(\Omega)[\mathbf{U}] = \int_\Gamma \langle \mathbf{U}, \mathbf{n}\rangle \left\{ g^2 - \left[\frac{\partial u}{\partial \mathbf{n}}\right]^2 \right\} d\sigma = 0 \tag{2.10}$$

has to be valid for all sufficiently smooth perturbation fields $\mathbf{U}$. Hence, shape optimization induces a variational formulation of the condition

$$\frac{\partial u}{\partial \mathbf{n}} = -g \qquad \text{on } \Gamma. \tag{2.11}$$

However, a stationary domain $\Omega^\star$ of the minimization problem (2.7), (2.8) will be a stable minimum if and only if the shape Hessian is strictly $H^{1/2}([0,2\pi])$-coercive at this domain (see below).

It suffices to consider $S \in C^{0,1}$, but due to a second order boundary perturbation calculus, we have to assume $T \in C^{2,\alpha}$ for some fixed $\alpha \in (0,1)$. We assume, similarly to the previous subsection, that the domain T is star-shaped with respect to $\mathbf{0}$, and we apply the same shape calculus. The shape gradient of the cost functional in (2.7) becomes, in polar coordinates,

$$\langle \nabla J(\Omega), dr\rangle = \int_0^{2\pi} dr\, r \left\{ g^2 - \left[\frac{\partial u}{\partial \mathbf{n}}\right]^2 \right\} d\phi. \tag{2.12}$$

According to [15, 16] the shape Hessian reads as

$$\begin{aligned} \langle \nabla^2 J(\Omega)\cdot dr_1, dr_2\rangle = \int_0^{2\pi} dr_1 dr_2 \Bigg\{ & g^2 - \left[\frac{\partial u}{\partial \mathbf{n}}\right]^2 + 2rg\langle \nabla g, \widehat{\mathbf{n}}\rangle \\ & - \frac{2r}{\sqrt{r^2+r'^2}} \frac{\partial u}{\partial \mathbf{n}} \left[r\frac{\partial^2 u}{\partial \mathbf{n}^2} + r' \frac{\partial^2 u}{\partial \mathbf{n}\partial \mathbf{t}} \right] \Bigg\} - 2r\, dr_1 \frac{\partial u}{\partial \mathbf{n}} \cdot \frac{\partial du[dr_2]}{\partial \mathbf{n}} d\phi. \end{aligned} \tag{2.13}$$

Herein, the *local shape derivative* $du = du[dr_2]$ of the state function satisfies

$$\begin{aligned} \Delta du &= 0 && \text{in } \Omega, \\ du &= 0 && \text{on } \Sigma, \\ du &= -dr_2\, \langle \widehat{\mathbf{n}}, \mathbf{n}\rangle \tfrac{\partial u}{\partial \mathbf{n}} && \text{on } \Gamma. \end{aligned} \tag{2.14}$$

Notice that $\partial^2 u/\partial \mathbf{n}^2 := \langle \nabla^2 u \cdot \mathbf{n}, \mathbf{n}\rangle$ and $\partial^2 u/(\partial \mathbf{n}\partial \mathbf{t}) := \langle \nabla^2 u \cdot \mathbf{n}, \mathbf{t}\rangle$.

LEMMA 2.2 (see [15, 25]). *The shape Hessian $\nabla^2 J(\Omega)$ defines a continuous bilinear form on $H^{1/2}([0,2\pi]) \times H^{1/2}([0,2\pi])$; that is, there exists a constant $c_S(\Omega)$ depending only on the actual domain Ω such that*

$$|\nabla^2 J(\Omega)[dr_1, dr_2]| \le c_S(\Omega) \|dr_1\|_{H^{1/2}([0,2\pi])} \|dr_2\|_{H^{1/2}([0,2\pi])}.$$

In accordance with this lemma, we observe that the shape Hessian is a pseudo-differential operator of order one, i.e., $\nabla^2 J(\Omega) : H^{1/2}([0,2\pi]) \to H^{-1/2}([0,2\pi])$. In particular the last term in (2.13) implies that the shape Hessian is a nonlocal operator.

According to [25] the following sufficient criterion concerning the $H^{1/2}([0,2\pi])$-coercivity holds.

LEMMA 2.3. *The shape Hessian $\nabla^2 J(\Omega^\star)$ is $H^{1/2}([0,2\pi])$-coercive; that is, there exists a constant $c_E > 0$ such that*

$$\nabla^2 J(\Omega^\star)[dr, dr] \ge c_E \|dr\|^2_{H^{1/2}([0,2\pi])}$$

if

$$\kappa + \left[\frac{\partial g}{\partial \mathbf{n}} - f\right] \Big/ g \geq 0 \qquad \text{on } \Gamma^\star.$$

In particular, in the case when $f \equiv 0$ and $g \equiv$ const., the shape Hessian is $H^{1/2}([0, 2\pi])$-coercive if the boundary $\Gamma^\star$ is convex (seen from inside).

The problem under consideration can be viewed as the prototype of a free boundary problem arising in many applications. For example, the growth of anodes in electrochemical processing might be modeled as above with $f \equiv 0$ and $g, h \equiv 1$.

In the two-dimensional exterior magnetic shaping of liquid metals, the state equation is an exterior Poisson equation and the uniqueness is ensured by a volume constraint of the domain Ω [9, 20, 41, 43]; see also the following subsection. However, since the shape functional involves the perimeter, which corresponds to the surface tension of the liquid, the energy space of the shape Hessian will be $H^1([0, 2\pi])$.

The detection of voids or inclusions in two- or three-dimensional electrical impedance tomography is slightly different since the roles of Σ and Γ are interchanged [23, 24, 45]. Particularly, this inverse problem is severely ill-posed, in contrast to the present class of problems. It has been proven in [23] that the shape Hessian is *not* strictly coercive in any $H^s([0, 2\pi])$ for all $s \in \mathbb{R}$.

2.4. Shape problems with additional functional constraints. We consider the following shape optimization problem:

$$J(\Omega) = \int_\Omega j(u, \nabla u, \mathbf{x}) d\mathbf{x} \to \min,$$

subject to L domain or boundary integral equality constraints

$$J_i(\Omega) = \int_\Omega h_i(\mathbf{x}) d\mathbf{x} = c_i, \quad 1 \leq i \leq K,$$
$$J_i(\Omega) = \int_\Gamma g_i(\mathbf{x}) d\sigma = c_i, \quad K < i \leq L.$$

We suppose that all functionals J and J_i, $1 \leq i \leq L$, are twice Frechét differentiable in a certain Banach space X. Moreover, let the Sobolev space H^s denote the strongest energy space of the bilinear forms imposed by the shape Hessians of *all* the above shape functionals.

Along the lines of standard optimization theory, one considers the free minimization of the Lagrangian

$$L(\Omega, \lambda_1, \ldots, \lambda_L) := J(\Omega) + \sum_{i=1}^{L} \lambda_i \big(J_i(\Omega) - c_i\big)$$

if Kuhn–Tucker regularity is provided. Hence, it is well known that the necessary and sufficient optimality condition for a regular local optimal shape $\Omega^\star$ reads as

LEMMA 2.4. *Let $\Omega^\star \in X$ satisfy*

$$\nabla L(\Omega^\star, \lambda_1^\star, \ldots, \lambda_L^\star)[dr] = 0 \quad \text{for all } dr \in X$$

for certain $\lambda_i^\star \in \mathbb{R}$. Moreover, define the linearizing cone

$$Y := \{dr \in X : \nabla J_i(\Omega^\star)[dr] = 0 \text{ for all } 1 \leq i \leq L\} \subset X,$$

and assume the linear independence of all gradients $\nabla J_i(\Omega^\star)$ at $\Omega^\star$.

Then $\Omega^\star$ is a regular local minimizer of second order if and only if the following coercivity condition is satisfied:

$$\nabla^2 L(\Omega^\star, \lambda_1^\star, \dots, \lambda_L^\star)[dr, dr] \geq c_E \|dr\|_{H^s}^2 \quad \textit{for all } dr \in Y.$$

Here, the techniques of the proof from [17, subsection 4.3] remain directly applicable, including the case of integral constraints that depend again on a PDE solution.

Remark 2.5. The linear independent constraint qualification (LICQ) implies in particular that the (vector valued) gradient of the constraints is a mapping onto $\mathbb{R}^L$. Hence, Y is a closed subspace of X of finite codimension L.

Consequently, the general concept developed in section 3 keeps applicable with respect to the Banach space Y. We mention that the treatment of inequality constraints is obvious and related modifications are well established in theory.

3. Approximation theory in shape optimization.

3.1. Assumptions on the optimization problem. Let us first introduce the abstract setting needed for our theory. To this end, let X denote a Banach space, where we shall denote the ball $\{h \in X : \|r - h\|_X < \delta\}$ by $B_\delta^X(r)$.

We consider the following optimization problem in the Banach space X:

$$(P) \qquad J(r) \to \min, \qquad r \in X.$$

Herein, $J : X \mapsto \mathbb{R}$ defines a two times continuously differentiable functional; i.e., the gradient $\nabla J(r) \in X^\star$ as well as the Hessian $\nabla^2 J(r) \in L(X, X^\star)$ exist for all $r \in X$, and the mappings $\nabla J(\cdot) : X \to X^\star$, $\nabla J^2(\cdot) : X \to L(X, X^\star)$ are continuous.

We assume that the necessary first order optimality condition holds in $r^\star$:

$$(\text{A1}) \qquad \nabla J(r^\star)[dr] = 0 \qquad \text{for all } dr \in X.$$

As illustrated in the previous section, we have to take the two norm discrepancy into account; i.e., the coercivity estimate holds only in a weaker Sobolev space $H^s \supset X$, $s \geq 0$. Therefore, we shall assume that there is a constant $c_S > 0$, depending *continuously* on the actual variable r, such that the continuous bilinear form imposed by the shape Hessian on $X \times X$ extends continuously to a bilinear form on $H^s \times H^s$, i.e.,

$$(\text{A2}) \qquad |\nabla^2 J(r)[h_1, h_2]| \leq c_S(r) \|h_1\|_{H^s} \|h_2\|_{H^s} \quad \text{for all } h_1, h_2 \in H^s,$$

if $r \in \overline{B_\delta^X(r^\star)}$. Of course, there exists an absolute constant C_S, defined by

$$(3.1) \qquad C_S := \max \left\{ c_S(r) : r \in \overline{B_\delta^X(r^\star)} \right\},$$

such that $c_S(r) \leq C_S$ for all $r \in \overline{B_\delta^X(r^\star)}$. Moreover, we assume that $\nabla^2 J$ is strongly coercive at $r^\star$, that is,

$$(\text{A3}) \qquad \nabla^2 J(r^\star)[h, h] \geq c_E \|h\|_{H^s}^2 \quad \text{for all } h \in H^s$$

for some $c_E > 0$.

Remark 3.1. The existence of a continuous extension for the objective J from X to H^s is not assumed throughout this paper since this is, in general, not realistic for shape problems; cf. subsection 2.2. That is, J remains only "densely defined" with respect

to H^s; this holds similarly for ∇J, $\nabla^2 J$. As it turns out, by our investigations a complete convergence analysis is possible without assuming a continuation property.

As a first consequence of our assumptions we have the following lemma concerning Lipschitz continuity of the shape gradient with respect to the topology that is induced by the coercivity space of the shape Hessian.

LEMMA 3.2. *The gradient is locally Lipschitz as a mapping in the* (H^{-s}, H^s)*-duality* ($H^{-s} := (H^s)'$)*, that is,*

$$\|\nabla J(r+h) - \nabla J(r)\|_{H^{-s}} \leq C_S \|h\|_{H^s} \tag{3.2}$$

for all $r, r+h \in \overline{B_\delta^X(r^\star)}$*. Herein, the constant* C_S *is given by* (3.1).

Proof. The assertion follows immediately from the following estimate:

$$|\nabla J(r+h)[dr] - \nabla J(r)[dr]| = \left| \int_0^1 \langle \nabla^2 J(r+th)\cdot h, dr\rangle dt \right| \leq C_S \|h\|_{H^s} \|dr\|_{H^s}$$

for all $r, r+h \in \overline{B_\delta^X(r^\star)}$, and $dr \in H^s$. ∎

Notice that the twice differentiability of J provides only the Lipschitz continuity of the shape gradient in the $(X^\star, X)$-duality, i.e.,

$$\|\nabla J(r+h) - \nabla J(r)\|_{X^\star} \leq C_S \|h\|_X$$

for all $r, r+h \in \overline{B_\delta^X(r^\star)}$.

3.2. Sufficient conditions. The above assumptions allow the following statement on the regular local optimality of second order of $r^\star$. Although this is rather standard, we recall it for convenience.

THEOREM 3.3 (sufficient second order optimality condition). *Let the necessary condition* (A1) *hold for a certain* $r^\star \in X$*. For all* $r \in \overline{B_\delta^X(r^\star)}$ *suppose that the bilinear form imposed by the shape Hessian satisfies* (A2) *and the following remainder estimate:*

$$\begin{aligned} &|\nabla^2 J(r)[h_1,h_2] - \nabla^2 J(r^\star)[h_1,h_2]| \\ &\qquad \leq \eta(\|r-r^\star\|_X)\|h_1\|_{H^s}\|h_2\|_{H^s} \quad \text{for all } h_1, h_2 \in H^s, \end{aligned} \tag{A4}$$

where $\eta : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ *is a decreasing function that satisfies* $\eta(t) \to 0$ *as* $t \to 0$*. Then, the domain* $r^\star$ *is a strong regular local optimum of second order with respect to certain constants* $\widehat{c}_E > 0$*,*

$$J(r) - J(r^\star) \geq \widehat{c}_E \|r - r^\star\|_{H^s}^2 \quad \text{for all } r \in \overline{B_{\widehat{\delta}}^X(r^\star)}, \tag{3.3}$$

if and only if the shape Hessian satisfies the strong coercivity estimate (A3).

Proof. For all $r = r^\star + h \in \overline{B_\delta^X(r^\star)}$ the following Taylor expansion holds:

$$J(r) - J(r^\star) = 0 + \frac{1}{2}\nabla^2 J(r^\star + \xi h)[h,h], \quad \xi \in (0,1).$$

According to (A3) and (A4), one infers on the one hand,

$$\begin{aligned} J(r) - J(r^\star) &\geq \frac{1}{2}\nabla^2 J(r^\star)[h,h] - |\nabla^2 J(r^\star+\xi h)[h,h] - \nabla^2 J(r^\star)[h,h]| \\ &\geq \frac{1}{2}\nabla^2 J(r^\star)[h,h] - \eta(\|h\|_X)\|h\|_{H^s}^2 \\ &\geq \frac{1}{2}\big(c_E - \eta(\|h\|_X)\big)\|h\|_{H^s}^2. \end{aligned}$$

Supposing $0 < \widehat{\delta} \leq \delta$ to be chosen such that $\eta(\|r - r^\star\|_X) \leq c_E/2$ for all $r \in \overline{B^X_{\widehat{\delta}}(r^\star)}$, we arrive at

$$J(r) - J(r^\star) \geq \frac{c_E}{4}\|r - r^\star\|^2_{H^s} \quad \text{for all } r \in \overline{B^X_{\widehat{\delta}}(r^\star)}.$$

On the other hand, we choose $r = r^\star + h \in \overline{B_{\widehat{\delta}}(r^\star)}$ arbitrarily but fixed. Combining the Taylor expansion

$$J(r) - J(r^\star) = \frac{1}{2}\nabla^2 J(r^\star + \xi h)[h,h] \geq \widehat{c}_E\|h\|^2_{H^s}, \quad \xi \in (0,1),$$

with (A4) yields

$$\begin{aligned}\nabla^2 J(r^\star)[h,h] &= \nabla^2 J(r^\star + \xi h)[h,h] + \nabla^2 J(r^\star)[h,h] - \nabla^2 J(r^\star + \xi h)[h,h] \\ &\geq \big(2\widehat{c}_E - \eta(\|h\|_X)\big)\|h\|^2_{H^s}.\end{aligned}$$

Fixing, similarly to the above, $0 < \delta \leq \widehat{\delta}$ such that $\eta(\|h\|_X) \leq \widehat{c}_E/2$ yields the coercivity estimate (A3) with $c_E := 3\widehat{c}_E/2$ for all $h \in \overline{B^X_\delta(0)}$. This finishes the proof since X is dense in H^s and $\nabla^2 J(r^\star) : H^s \times H^s \to \mathbb{R}$ is bilinear. ∎

Let us remark that the verification of (A4) turns out to be rather technical in the case of PDE-constrained shape optimization problems. For the presented model problems, (A4) has been proven in [10, 11, 12], whereas the verification of (A2) is much simpler (see, e.g., [15]) but already an indicator of the two norm discrepancy.

Combining the assumptions (A2) (together with (3.1)), (A3), and (A4) leads to the following corollary by repeating a portion of the preceding proof.

COROLLARY 3.4. *For $\widehat{\delta} > 0$ sufficiently small, the shape Hessian is strongly coercive in the whole ball $\overline{B^X_{\widehat{\delta}}(r^\star)}$, that is,*

$$\nabla^2 J(r)[h,h] \geq \frac{c_E}{2}\|h\|^2_{H^s} \quad \textit{for all } h \in H^s,\ r \in \overline{B^X_{\widehat{\delta}}(r^\star)}. \tag{3.4}$$

Moreover, with respect to the objective, the following upper and lower quadratic bound

$$\frac{c_E}{4}\|r - r^\star\|^2_{H^s} \leq J(r) - J(r^\star) \leq \frac{C_S}{2}\|r - r^\star\|^2_{H^s} \tag{3.5}$$

holds for all $r \in \overline{B^X_{\widehat{\delta}}(r^\star)}$.

3.3. Ritz–Galerkin discretization. We shall consider a Ritz–Galerkin scheme to solve the necessary condition (A1); i.e., we replace the given infinite dimensional optimization problem with a finite dimensional problem. The trial space should provide sufficient regularity in order to approximate functions in X. To this end, we introduce an appropriate Hilbert space $H^k \subset X$, continuously embedded in X, i.e.,

$$\|r\|_X \leq c_{H^k \to X}\|r\|_{H^k} \qquad \text{for all } r \in H^k. \tag{V1}$$

Then we shall consider a sequence of nested finite dimensional trial spaces,

$$V_0 \subset V_1 \subset \cdots \subset V_N \subset \cdots \subset H^k \subset X, \quad \bigcap_{N\geq 0} V_N = V_0, \quad \overline{\bigcup_{N\geq 0} V_N}^{H^k} = H^k, \tag{V2}$$

providing the following inverse estimate:

$$\|r_N\|_{H^k} \leq E(N)\|r_N\|_{H^s} \quad \text{for all } r_N \in V_N. \tag{V3}$$

Moreover, we assume that there exists an $L > k$ such that the following approximation property holds:

$$\text{(V4)} \qquad \inf_{r_N \in V_N} \|r - r_N\|_{H^s} = o\left(\frac{1}{E(N)}\right) \|r\|_{H^\ell} \quad \text{if } r \in H^\ell \ (k < \ell \leq L).$$

Herein, the Landau symbol $g(x) = o(f(x))$ means that $\lim_{x\to\infty} g(x)/f(x) = 0$.

Remark 3.5. Suppose $X = C^{2,\alpha}([0,1])$ for some $\alpha \in (0,1)$. Then the Sobolev space $H^k([0,1])$ with $3 \geq k > 2 + \alpha$ provides a continuous embedding in accordance with (V1). Choosing $V_N \subset C^{2,1}([0,1])$ as the space of smoothest cubic splines on the uniform subdivision with step width $h_N := 2^{-N}/4$, we have the approximation property

$$\inf_{r_N \in V_N} \|r - r_N\|_{H^s} \lesssim h_N^{\ell-s} \|r\|_{H^\ell} \quad \text{if } r \in H^\ell \ (k < \ell \leq 4)$$

uniformly in N, provided that $s < k$. The inverse estimate reads as

$$\|r_N\|_{H^k} \lesssim h_N^{s-k} \|r_N\|_{H^s} \quad \text{for all } r_N \in V_N$$

uniformly in N, provided that $s \leq k$. Hence, we conclude that the trial spaces $(V_N)_{N\geq 0}$ satisfy (V2)–(V4).

The Ritz–Galerkin scheme reads as follows. In order to solve

$$(P_N) \qquad J(r_N) \to \min, \qquad r_N \in V_N,$$

one seeks an approximate solution $r_N^\star \in V_N$ such that the discretized necessary condition

$$\nabla J(r_N^\star)[q_N] = 0 \tag{3.6}$$

holds for all $q_N \in V_N$.

There exist different strategies for finding $r_N \in V_N$ such that (3.6) holds. In general, suppose that r_N has N degrees of freedom; i.e., there exist $\varphi_1, \varphi_2, \ldots, \varphi_N$ such that

$$V_N = \operatorname{span}\{\varphi_1, \varphi_2, \ldots, \varphi_N\}.$$

One makes the ansatz $r_N = \sum_{i=1}^N r_i \varphi_i$ and considers an iterative scheme

$$\mathbf{r}^{(n+1)} = \mathbf{r}^{(n)} - h^{(n)} \mathbf{M}^{(n)} \mathbf{G}^{(n)}, \tag{3.7}$$

where $h^{(n)}$ is a suitable step width and

$$\mathbf{r}^{(n)} = \big(r_i^{(n)}\big)_{i=1,\ldots,N}, \quad \mathbf{G}^{(n)} := \big(\nabla J(r_N^{(n)})[\varphi_i]\big)_{i=1,\ldots,N}.$$

First order methods are the gradient method ($\mathbf{M}^{(n)} := \mathbf{I}$) and the quasi-Newton method, where $\mathbf{M}^{(n)}$ denotes a suitable approximation to the inverse shape Hessian. Choosing

$$\mathbf{M}^{(n)} := \big(\nabla^2 J(r_N^{(n)})[\varphi_i, \varphi_j]\big)_{i,j=1,\ldots,N}^{-1}$$

yields the Newton method, which converges much faster compared to first order methods; see [19] for an example.

3.4. Existence of approximate solutions. We will consider the existence of solutions of (3.6) and the question of the accuracy of approximate solutions $r_N^\star$. Since the solutions of (3.6) are only stationary points, it is reasonable to consider only local optimization problems. Therefore, we replace the global problems (P) and (P_N) with the local optimization problem

$$(P^\delta) \qquad J(r) \to \min, \quad r \in \overline{B_\delta^X(r^\star)},$$

and its discrete variant

$$(P_N^\delta) \qquad J(r_N) \to \min, \qquad r_N \in V_N \cap \overline{B_\delta^X(r^\star)},$$

where $\delta = \widehat{\delta}$ is chosen in accordance with the estimates (3.4), (3.5) and is *independent* of N. Obviously, the solution of (P^δ) is $r^\star$, since J is strictly coercive on the convex set $\overline{B_\delta^X(r^\star)}$. Moreover, we have as a first consequence the following lemma.

LEMMA 3.6. *Problem (P_N^δ) always admits a solution $r_N^\star \in V_N \cap \overline{B_\delta^X(r^\star)}$. Any point $r_N^\star \in V_N \cap \overline{B_\delta^X(r^\star)}$ satisfying* (3.6) *is a local regular optimizer of second order. Moreover, the coercivity implies the uniqueness of $r_N^\star$.*

Proof. The existence of $r_N^\star$ is obvious since the admissible set $V_N \cap \overline{B_\delta^X(r^\star)}$ is compact. It follows for all $r_N := r_N^\star + h_N \in V_N \cap \overline{B_\delta^X(r^\star)}$ that $r_N^\star + \xi h_N \in V_N \cap \overline{B_\delta^X(r^\star)}$ is always satisfied for all $\xi \in (0,1)$ by convexity of the admissible set. Consequently, if $r_N^\star$ also satisfies (3.6), we deduce from (3.4) that

$$J(r_N) - J(r_N^\star) = \frac{1}{2}\nabla^2 J(r_N^\star + \xi h_N)[h_N, h_N] \geq \frac{c_E}{4}\|r_N - r_N^\star\|_{H^s}^2, \quad \xi \in (0,1),$$

for all $r_N = r_N^\star + h_N \in V_N \cap \overline{B_\delta^X(r^\star)}$. Uniqueness of $r_N^\star$ is an immediate consequence of the strict convexity of J (ensured again by (3.4)) on the convex set $V_N \cap \overline{B_\delta^X(r^\star)}$. ∎

Nevertheless, if $r_N^\star$ remains at the "artificial" boundary $\partial\{V_N \cap \overline{B_\delta^X(r^\star)}\} = V_N \cap \partial B_\delta^X(r^\star)$, only a related variational inequality holds instead of (3.6). Furthermore, $\|r_N^\star - r^\star\|_X = \delta$ for $N \to \infty$ contradicts convergence on its own. Consequently, we have to ensure that $r_N^\star$ is an *interior* point of the set $V_N \cap \overline{B_\delta^X(r^\star)}$, i.e.,

$$\|r_N^\star - r^\star\|_X < \delta,$$

at least for all sufficiently large $N \geq N_0$. This result is provided by the next theorem.

THEOREM 3.7. *Let* (A1)–(A4) *and* (V1)–(V4) *hold. Then, if $r^\star \in H^\ell$ for some $\ell > k$, there exists an N_0 such that*

$$r_N^\star \in V_N \cap B_\delta^X(r^\star) \quad \textit{for all } N \geq N_0.$$

Proof. We split the proof into four parts.

(i) We define $P_N : L^2 \to V_N$ as the L^2-orthogonal projection onto V_N. Then, by our assumptions (V1), (V2) we have

$$\|P_N(r^\star) - r^\star\|_X \leq c_{H^k \to X}\|P_N(r^\star) - r^\star\|_{H^k} \lesssim \inf_{r_N \in V_N} \|r_N - r^\star\|_{H^k} \overset{N\to\infty}{\longrightarrow} 0$$

and likewise by (V4),

$$\|P_N(r^\star) - r^\star\|_{H^s} \lesssim \inf_{r_N \in V_N} \|r_N - r^\star\|_{H^s} \overset{N\to\infty}{\longrightarrow} 0.$$

Hence, we deduce that there exists an N_0 such that $V_N \cap \overline{B_\delta^X(r^\star)} \neq \emptyset$ for all $N \geq N_0$. Without loss of generality we assume that $N_0 = 0$.

(ii) Recall that

$$J(r^\star) = \inf\big\{J(r) : r \in \overline{B_\delta^X(r^\star)}\big\},$$
$$J(r_N^\star) = \inf\big\{J(r_N) : r_N \in V_N \cap \overline{B_\delta^X(r^\star)}\big\},$$

and define $J_\delta(N) \geq J(r_N^\star) \geq J(r^\star)$ via

$$J_\delta(N) := \inf\big\{J(r_N) : r_N \in V_N \cap \partial B_\delta^X(r^\star)\big\}.$$

Since $J\big(P_N(r^\star)\big) \geq J(r_N^\star)$, we conclude the assertion $\|r_N^\star - r^\star\|_X < \delta$ if we can prove

$$J_\delta(N) > J\big(P_N(r^\star)\big) \quad \text{for all } N \geq N_0. \tag{3.8}$$

On the one hand, (3.5) implies

$$J\big(P_N(r^\star)\big) - J(r^\star) \leq \frac{C_S}{2}\|P_N(r^\star) - r^\star\|_{H^s}^2. \tag{3.9}$$

On the other hand, by introducing the quantity

$$\begin{aligned} F_\delta^X(N) &:= \inf\big\{\|r_N - r^\star\|_{H^s} : r_N \in V_N \cap \partial B_\delta^X(r^\star)\big\} \\ &= \inf\big\{\|r_N - r^\star\|_{H^s} : r_N \in V_N \setminus B_\delta^X(r^\star)\big\}, \end{aligned}$$

we derive from (3.5)

$$J_\delta(N) - J(r^\star) \geq \frac{c_E}{4} F_\delta^X(N)^2. \tag{3.10}$$

Combining (3.9) and (3.10), we see that the inequality

$$\|P_N(r^\star) - r^\star\|_{H^s} < C^\star \cdot F_\delta^X(N), \quad C^\star := \sqrt{\frac{c_E}{2C_S}}, \tag{3.11}$$

will imply (3.8) and, thus, $\|r_N^\star - r^\star\|_X < \delta$.

(iii) We shall establish a relation between $F_\delta^X(N)$, $\|r^\star - P_N(r^\star)\|_{H^s}$, and $E(N)$ from the inverse estimate (V3). For the sake of simplicity, we assume without loss of generality that the constant $c_{H^k \to X}$ from (V1) is less than one such that

$$B_\delta^{H^k}(r^\star) \subseteq B_\delta^X(r^\star). \tag{3.12}$$

Introducing

$$\begin{aligned} F_\delta^{H^k}(N) &:= \inf\big\{\|r_N - r^\star\|_{H^s} : r_N \in V_N \cap \partial B_\delta^{H^k}(r^\star)\big\} \\ &= \inf\big\{\|r_N - r^\star\|_{H^s} : r_N \in V_N \setminus B_\delta^{H^k}(r^\star)\big\}, \end{aligned}$$

there follows from (3.12) the relation

$$F_\delta^{H^k}(N) \leq F_\delta^X(N).$$

We shall now compute a lower bound for $F_\delta^{H^k}(N)$. From

$$\|r_N - P_N(r^\star)\|_{H^s} - \|P_N(r^\star) - r^\star\|_{H^s} \leq \|r_N - r^\star\|_{H^s}$$

one infers the inequality

$$F_\delta^{H^k}(N) \geq \inf\big\{\|r_N - P_N(r^\star)\|_{H^s} : r_N \in V_N \setminus B_\delta^{H^k}(r^\star)\big\} - \|P_N(r^\star) - r^\star\|_{H^s}. \tag{3.13}$$

We choose N_0 sufficiently large to ensure

$$\|P_N(r^\star) - r^\star\|_{H^k} \leq \delta/2 \quad \text{for all } N \geq N_0.$$

Then it holds that $B_{\delta/2}^{H^k}\big(P_N(r^\star)\big) \subset B_\delta^{H^k}(r^\star)$, and we arrive at

$$\begin{aligned}
&\inf\big\{\|r_N - P_N(r^\star)\|_{H^s} : r_N \in V_N \setminus B_\delta^{H^k}(r^\star)\big\} \\
&\quad \geq \inf\big\{\|r_N - P_N(r^\star)\|_{H^s} : r_N \in V_N \setminus B_{\delta/2}^{H^k}\big(P_N(r^\star)\big)\big\} \\
&\quad \geq \inf_{r_N \in V_N} \{\|r_N\|_{H^s} : \|r_N\|_{H_k} = \delta/2\} \\
&\quad \geq \frac{\delta}{2E(N)}.
\end{aligned}$$

Inserting this estimate into (3.13), we deduce

$$F_\delta^X(N) \geq F_\delta^{H^k}(N) \geq \frac{\delta}{2E(N)} - \|P_N(r^\star) - r^\star\|_{H^s} \quad \text{for all } N \geq N_0. \tag{3.14}$$

(iv) Observing

$$\|P_N(r^\star) - r^\star\|_{H^s} \lesssim \inf_{r_N \in V_N} \|r_N - r^\star\|_{H^s},$$

we infer from (V4) that we can increase N_0 such that

$$\|P_N(r^\star) - r^\star\|_{H^s} < \frac{\delta}{2E(N)} \cdot \frac{C^\star}{C^\star + 1} \quad \text{for all } N \geq N_0.$$

Thus, in view of (3.14), we arrive at

$$\|P_N(r^\star) - r^\star\|_{H^s} < C^\star \left(\frac{\delta}{2E(N)} - \|P_N(r^\star) - r^\star\|_{H^s} \right) < C^\star F_\delta^X(N),$$

that is, (3.11), for all $N \geq N_0$, which finishes the proof according to part (ii). ∎

Remark 3.8. Obviously, by means of standard optimization theory, (3.3) and (3.6) imply well-posedness of the finite dimensional optimization problem; that is, existence and (local) uniqueness of minimizers are ensured. In particular, the strict coercivity of (P_N^δ), induced by the coercivity of (P^δ), provides the convergence

$$r_N^{(n)} \to r_N^\star \quad \text{as } n \to \infty$$

of the iterative scheme (3.7); see, e.g., [30, 40].

3.5. Convergence. The above theorem ensures the *existence* of an approximate solution $r_N^\star$ to the finite dimensional problems (P_N^δ) that satisfies the necessary condition (3.6), provided that N is sufficiently large. The next theorem estimates the distance $\|r_N^\star - r^\star\|_{H^s}$.

THEOREM 3.9. *The approximate solution $r_N^\star$ of the finite dimensional problem (P_N^δ) satisfies the error estimate*

$$\|r_N^\star - r^\star\|_{H^s} \leq \frac{2C_S}{c_E} \inf_{r_N \in V_N} \|r_N - r^\star\|_{H^s}$$

uniformly with the number of unknowns N.

Proof. For the sake of clearness in the representation, let $\langle\cdot,\cdot\rangle$ denote the duality pairing between H^s and its dual space H^{-s}.

On the one hand, observing (3.2), Galerkin orthogonality implies

$$\begin{aligned}\langle\nabla J(r_N^\star)-\nabla J(r^\star),r_N^\star-r^\star\rangle &= \langle\nabla J(r_N^\star)-\nabla J(r^\star),r_N-r^\star\rangle\\ &\le C_S\|r_N^\star-r^\star\|_{H^s}\|r_N-r^\star\|_{H^s}\end{aligned}$$

for all $r_N\in V_N$. On the other hand, by introducing

$$j(t):=\langle\nabla J\big(tr_N^\star+(1-t)r^\star\big),r_N^\star-r^\star\rangle,$$

we derive the estimate

$$\begin{aligned}\langle\nabla J(r_N^\star)-\nabla J(r^\star),r_N^\star-r^\star\rangle &= j(1)-j(0)=\int_0^1 j'(t)\,dt\\ &=\int_0^1\langle\nabla^2 J(tr_N^\star+(1-t)r^\star\big)\cdot(r_N^\star-r^\star),r_N^\star-r^\star\rangle\,dt\ge\frac{c_E}{2}\|r_N^\star-r^\star\|_{H^s}^2.\end{aligned}$$

Combining both estimates yields

$$\|r_N^\star-r^\star\|_{H^s}^2\le\frac{2C_S}{c_E}\|r_N^\star-r^\star\|_{H^s}\|r_N-r^\star\|_{H^s}$$

for all $r_N\in V_N$, which is equivalent to the assertion. ∎

Of course, from this theorem one can determine the *rate of convergence* if one estimates $\inf_{r_N\in V_N}\|r_N-r^\star\|_{H^s}$.

3.6. The fully discretized problem. Up to now, we investigated only the discretization with respect to the shape. Hence, we neglected consistency errors arising from the approximate solution of the state equation or from computing the objective and constraints by, e.g., quadrature. Consequently, we shall focus on the following further modification of problem (P_N^δ):

$(P_{N\epsilon}^\delta)$ seek $r_{N\epsilon}^\star\in V_N\cap\overline{B_\delta^X(r_N^\star)}$ such that $\langle\nabla J_\epsilon(r_N^\star),q_N\rangle=0$ for all $q_N\in V_N$,

where ϵ is an approximation parameter referring to the inexact computation of the gradient. We prove the following Strang-type lemma which estimates the consistency error induced by solving $(P_{N\epsilon}^\delta)$.

LEMMA 3.10. *Assume that the estimate*

$$|\langle[\nabla J_\epsilon(r_N)-\nabla J(r_N)]-[\nabla J_\epsilon(q_N)-\nabla J(q_N)],s_N\rangle|\le\epsilon\|r_N-q_N\|_{H^s}\|s_N\|_{H^s}\tag{3.15}$$

holds for all $r_N,q_N\in V_N\cap\overline{B_\delta^X(r_N^\star)}$ and $s_N\in V_N$. Then, provided that ϵ is sufficiently small, $(P_{N\epsilon}^\delta)$ admits a unique solution $r_{N\epsilon}^\star\in V_N\cap\overline{B_\delta^X(r_N^\star)}$ which satisfies the a priori estimate

$$\|r^\star-r_{N\epsilon}^\star\|_{H^s}\le\left(1+\frac{2\max\{1,C_S\}}{c_E-2\epsilon}\right)\left\{\|r^\star-r_N\|_{H^s}+\sup_{q_N\in V_N}\frac{\langle\nabla J(r_N)-\nabla J_\epsilon(r_N),q_N\rangle}{\|q_N\|_{H^s}}\right\}.$$

Proof. Due to our assumptions from the previous subsections, the unperturbed Richardson iteration

$$r_N^{(n+1)}=r_N^{(n)}-\alpha\sum_{i=1}^N\nabla J(r_N^{(n)})[\varphi_i]\varphi_i,\quad n=0,1,\dots,$$

defines a contraction of $V_N \cap \overline{B_\delta^X(r_N^\star)}$ onto itself for a whole range of $\alpha \in [\underline{\alpha}, \overline{\alpha}]$. Estimate (3.15) ensures that the perturbed Richardson iteration

$$r_{N\epsilon}^{(n+1)} = r_{N\epsilon}^{(n)} - \alpha \sum_{i=1}^{N} \nabla J_\epsilon(r_{N\epsilon}^{(n)})[\varphi_i]\varphi_i, \quad n = 0, 1, \ldots,$$

is still a contraction of $V_N \cap \overline{B_\delta^X(r_N^\star)}$ onto itself for $\alpha := (\underline{\alpha} + \overline{\alpha})/2$, provided that ϵ is sufficiently small. This proves existence and uniqueness of the perturbed solution $r_{N\epsilon}^\star$.

Next, using again (3.15), we find

$$\begin{aligned}
&\langle \nabla J_\epsilon(r_N) - \nabla J_\epsilon(q_N), r_N - q_N \rangle \\
&\quad \geq \langle \nabla J(r_N) - \nabla J(q_N), r_N - q_N \rangle - \epsilon \|r_N - q_N\|_{H^s}^2 \\
&\quad \geq \Big(\frac{c_E}{2} - \epsilon \Big) \|r_N - q_N\|_{H^s}^2,
\end{aligned}$$

where $\widetilde{c}_E := c_E/2 - \epsilon > 0$ holds if ϵ is sufficiently small.

Due to Galerkin orthogonality, the Ritz–Galerkin solution $r_{N\epsilon}^\star$ of $(P_{N\epsilon}^\delta)$ satisfies

$$\begin{aligned}
\widetilde{c}_E \|r_{N\epsilon}^\star - r_N\|_{H^s}^2 &\leq \langle \nabla J_\epsilon(r_{N\epsilon}^\star) - \nabla J_\epsilon(r_N), r_{N\epsilon}^\star - r_N \rangle \\
&\leq \langle \nabla J(r^\star) - \nabla J(r_N), r_{N\epsilon}^\star - r_N \rangle + \langle \nabla J(r_N) - \nabla J_\epsilon(r_N), r_{N\epsilon}^\star - r_N \rangle \\
&\leq C_S \|r^\star - r_N\|_{H^s} \|r_{N\epsilon}^\star - r_N\|_{H^s} + \langle \nabla J(r_N) - \nabla J_\epsilon(r_N), r_{N\epsilon}^\star - r_N \rangle,
\end{aligned}$$

that is,

$$\|r_{N\epsilon}^\star - r_N\|_{H^s} \leq \frac{\max\{1, C_S\}}{\widetilde{c}_E} \left\{ \|r^\star - r_N\|_{H^s} + \sup_{q_N \in V_N} \frac{\langle \nabla J(r_N) - \nabla J_\epsilon(r_N), q_N \rangle}{\|q_N\|_{H^s}} \right\}.$$

Since $r_N \in V_N \cap \overline{B_\delta^X(r_N^\star)}$ is arbitrary, we arrive at the assertion using the triangle inequality

$$\|r^\star - r_{N\epsilon}^\star\|_{H^s} \leq \|r^\star - r_N\|_{H^s} + \|r_N - r_{N\epsilon}^\star\|_{H^s}. \qquad \square$$

4. Numerical results.

4.1. An unconstrained shape optimization problem. For comparison we shall employ model problems, where the solution is known analytically. To this end, we choose the shape optimization problem (2.3) based on the domain integral

$$J(\Omega) = \int_\Omega \Big(\frac{x^2}{8} + \frac{y^2}{4} - 2 \Big) d\mathbf{x}$$

as our first numerical example. In accordance with subsection 2.2, the ellipse centered in $\mathbf{0}$ with semiaxes $2\sqrt{2}$ and 2 is a strict minimizer of second order.

The numerical setting is as follows. We subdivide the parameter interval $[0, 2\pi]$ equidistantly into N intervals. With respect to this subdivision, the radial function $r \in X := C_{\mathrm{per}}^1([0, 2\pi])$ is then approximated periodically by N cubic B-splines B_i^3, $i = 1, \ldots, N$, that is,

$$r_N = \sum_{i=1}^{N} a_i B_i^3 \in C_{\mathrm{per}}^{2,1}([0, 2\pi]).$$

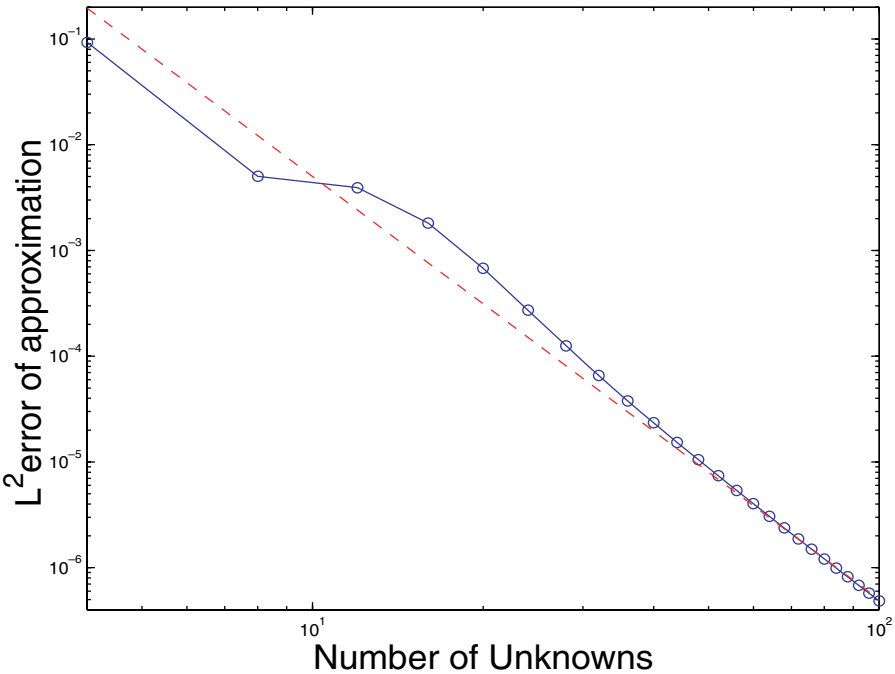


FIG. 4.1. *L^2-error of the approximate solution.*

We employ a Newton method to iteratively solve the necessary condition $\nabla J(\Omega) \equiv 0$, using the circle with radius 2 as an initial guess.

Since the energy space for the shape Hessian is $L^2([0, 2\pi])$, we measure the L^2-norm of the approximation error given by

$$\|r - r_N\|^2_{L^2([0,2\pi])} = \int_0^{2\pi} |r - r_N|^2 d\phi.$$

The measurements are shown in Figure 4.1. We observe, as predicted, the rate of convergence N^{-4}, indicated by the dashed line.

4.2. A constrained shape optimization problem. We consider next a cylindric circular bar which is homogeneous and isotropic with a planar, simply connected cross section $\Omega \in \mathbb{R}^2$. We follow Banichuk and Karihaloo [2], but normalize the shear modulus $G = 1$ and the elastic modulus $E = 1$. We want to solve the problem of maximizing the torsional rigidity of the bar subject to given equality constraints on the bending stiffness and the volume.

First, we briefly recall the mathematical formulation of the quantities. The torsional rigidity is calculated by

$$T(\Omega) = 2 \int_\Omega u(\mathbf{x}) d\mathbf{x},$$

where the stress function $u = u(\Omega)$ satisfies

$$-\Delta u = 2 \text{ in } \Omega, \qquad u = 0 \text{ on } \Gamma.$$

The bending rigidity with respect to a fixed barycenter in the origin is given by

$$B(\Omega) = \int_\Omega y^2 d\mathbf{x}.$$

The volume of the domain and its (simplified) barycenter coordinates read as

$$V(\Omega) = \int_\Omega d\mathbf{x}, \qquad S_x(\Omega) = \int_\Omega x d\mathbf{x}, \qquad S_y(\Omega) = \int_\Omega y d\mathbf{x}.$$

Consequently, we arrive at the following constraint shape optimization problem:

$$J(\Omega) := -T(\Omega) \to \min$$

subject to

$$B(\Omega) = B_0, \qquad V(\Omega) = V_0, \qquad S_x(\Omega) = 0, \qquad S_y(\Omega) = 0.$$

Choosing $B_0 = \sqrt{2}\pi/4$, $V_0 = \pi$, we see that the necessary condition is fulfilled by the ellipse with semiaxes $h_x = 2^{-1/4}$ and $h_y = 2^{1/4}$. The associated Lagrange multipliers are $\lambda_B = -4/9$, $\lambda_V = 8\sqrt{2}/9$, and $\lambda_{S_x} = \lambda_{S_y} = 0$; cf. [2]. From the identity

$$T(\Omega) = \int_\Omega \|\nabla u(\mathbf{x})\|^2 d\mathbf{x},$$

we deduce that $\nabla T(\Omega)[dr]$ and $\nabla^2 T(\Omega)[dr_1, dr_2]$ are given as in (2.12) and (2.13) with $g \equiv 0$ and

$$\Delta du = 0 \text{ in } \Omega, \qquad du = -dr_2 \, \langle \widehat{\mathbf{n}}, \mathbf{n} \rangle \frac{\partial u}{\partial \mathbf{n}} \text{ on } \Gamma.$$

Recall that twice differentiability needs $r \in X := C^{2,\alpha}_{\mathrm{per}}([0, 2\pi])$; cf. subsection 2.3. The computation of the other gradients and Hessians is straightforward; see [18, 19] for the details.

We approximate the radial function r similarly to our first example by periodic cubic splines on the interval $[0, 2\pi]$. Even though the sufficient optimality condition has not yet been proven, our experience indicates coercivity in the energy space $H^{1/2}([0, 2\pi])$; cf. [18, 19, 21]. More precisely, coercivity of the Lagrangian at $(\Omega^\star, \lambda^\star)$ has to hold on the closed subspace $Y \subseteq C^{2,\alpha}_{\mathrm{per}}([0, 2\pi])$, where

$$\begin{aligned} Y := \{dr \in C^{2,\alpha}_{\mathrm{per}}([0, 2\pi]) : \nabla B(\Omega^\star)[dr] = 0 \wedge \nabla V(\Omega^\star)[dr] = 0 \\ \wedge \nabla S_x(\Omega^\star)[dr] = 0 \wedge \nabla S_y(\Omega^\star)[dr] = 0\}. \end{aligned}$$

However, the pure Lagrangian is introduced only for investigating the sufficient optimality condition. In order to numerically solve the discretized constrained shape optimization problem, we need to find the stationary points of the following augmented Lagrange functional:

$$L_c(\Omega, \boldsymbol{\lambda}) := -T(\Omega) + \boldsymbol{\lambda}^T \begin{bmatrix} B(\Omega) - B_0 \\ V(\Omega) - V_0 \\ S_x(\Omega) \\ S_y(\Omega) \end{bmatrix} + \frac{c}{2} \left\| \begin{bmatrix} B(\Omega) - B_0 \\ V(\Omega) - V_0 \\ S_x(\Omega) \\ S_y(\Omega) \end{bmatrix} \right\|^2,$$

where $\boldsymbol{\lambda} := (\lambda_B, \lambda_V, \lambda_{S_x}, \lambda_{S_y})$ and $c > 0$ is an appropriate chosen penalty parameter. The optimization algorithm then reads as follows:

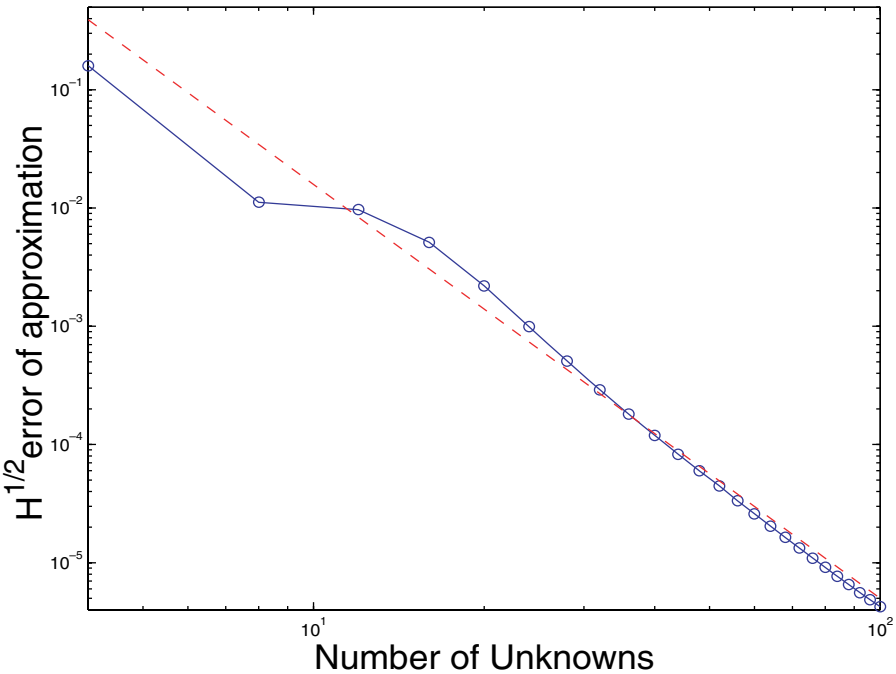


FIG. 4.2. *$H^{1/2}$-error of the approximate solution.*

- initialization: choose initial guess $(\Omega^{(0)}, \boldsymbol{\lambda}^{(0)})$ for $(\Omega^\star, \boldsymbol{\lambda}^\star)$.
- inner iteration: solve $\Omega^{(n+1)} := \operatorname{argmin} L_c(\Omega, \boldsymbol{\lambda}^{(n)})$ with initial guess $\Omega^{(n)}$.
- outer iteration: update

$$\lambda^{(n+1)} := \lambda^{(n)} - c \begin{bmatrix} B(\Omega^{(n+1)}) - B_0 \\ V(\Omega^{(n+1)}) - V_0 \\ S_x(\Omega^{(n+1)}) \\ S_y(\Omega^{(n+1)}) \end{bmatrix}.$$

In the inner iteration, we employ a Newton scheme combined with a quadratic line-search. Instead of the first order update rule described above, we use a second order Lagrange multiplier method introduced in [36] (see also [21]), which provides faster convergence of the dual parameters. The state equation is solved by using a boundary element method; cf. [18, 19] for the details. Notice that about 2000 boundary elements are required to solve the state equation sufficiently accurately if we discretize the free boundary by $N = 100$ B-splines.

According to our convergence result we shall observe the rate of convergence

$$\|r - r_N\|_{H^{1/2}([0,2\pi])} \lesssim N^{-3.5} \|r\|_{H^4([0,2\pi])}.$$

We measure this norm via the approximation

$$\|r - r_N\|^2_{H^{1/2}([0,2\pi])} \sim \|r - r_N\|^2_{L^2([0,2\pi])} + \int_0^{2\pi} |r - r_N||r' - r'_N| d\phi.$$

The results are presented in Figure 4.2. As predicted, the error decreases like $N^{-3.5}$, which is indicated by the dashed line.

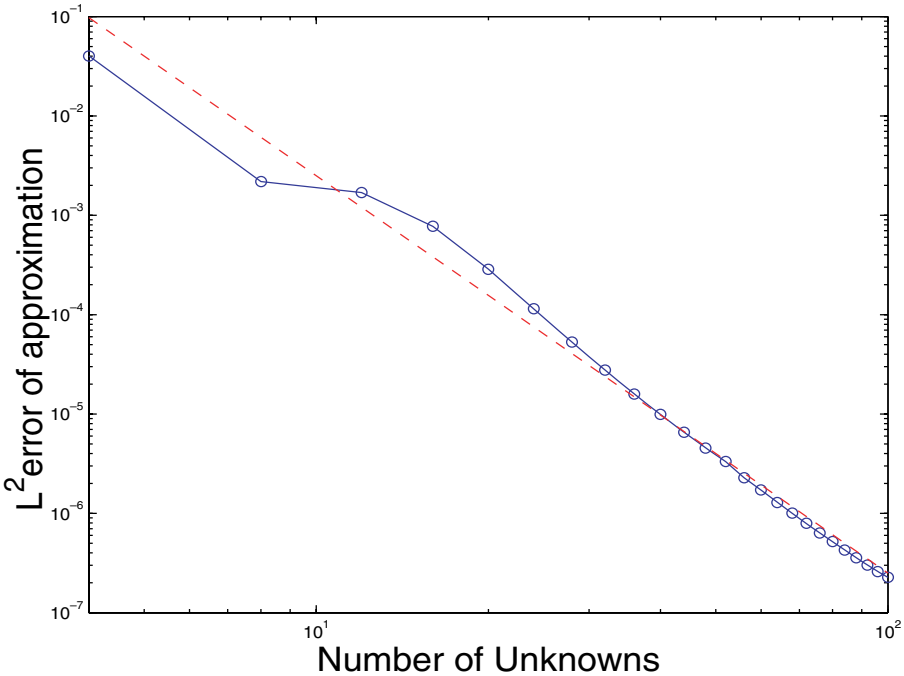


FIG. 4.3. *L^2-error of the approximate solution.*

In addition we also measured the L^2-norm of the approximation error, visualized in Figure 4.3. In fact, even though we have not proven the Aubin–Nitsche trick, we observe the higher rate of convergence N^{-4}, indicated by the dashed line.

5. Concluding remarks. In the present paper we established a complete convergence analysis for approximate solutions of shape optimization problems. In particular, we incorporated the two norm discrepancy. We presented numerical results which verify the predicted rates of convergence. We would like to point out that our analysis applies also to p-discretizations of the domain's parametrization, for example, finite dimensional Fourier sequences for the discretization of the radial function. For several applications we refer to [9, 18, 19, 20]; see also [22, 24, 42] for related problems in three dimensions.

REFERENCES

[1] H. W. ALT AND L. A. CAFFARELLI, *Existence and regularity for a minimum problem with free boundary*, J. Reine Angew. Math., 325 (1981), pp. 105–144.

[2] N. V. BANICHUK AND B. L. KARIHALOO, *Minimum-weight design of multi-purpose cylindrical bars*, Internat. J. Solids Structures, 12 (1976), pp. 267–273.

[3] M. P. BENDSOE AND O. SIGMUND, *Topology Optimization. Theory, Methods and Applications*, Springer, New York, 2003.

[4] E. CASAS, F. TRÖLTZSCH, AND A. UNGER, *Second order sufficient optimality conditions for a nonlinear elliptic boundary control problem*, Z. Anal. Anwend., 15 (1996), pp. 687–707.

[5] E. CASAS, F. TRÖLTZSCH, AND A. UNGER, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim., 38 (2000), pp. 1369–1391.

[6] D. CHENAIS AND E. ZUAZUA, *Controllability of an elliptic equation and its finite difference apporoximation by the shape of the domain*, Numer. Math., 95 (2003), pp. 63–99.

[7] D. Chenais and E. Zuazua, *Finite Element Approximation on Elliptic Optimal Design*, C. R. Acad. Sci. Paris Ser. I, 338 (2004), pp. 729–734.
[8] D. Chenais and E. Zuazua, *Finite element approximation of 2D elliptic optimal design*, J. Math. Pures Appl. (9), 85 (2006), pp. 225–249.
[9] O. Colaud and A. Henrot, *Numerical approximation of a free boundary problem arising in electromagnetic shaping*, SIAM J. Numer. Anal., 31 (1994), pp. 1109–1127.
[10] M. Dambrine and M. Pierre, *About stability of equilibrium shapes*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 811-834.
[11] M. Dambrine, *Hessiennes de forme et stabilité des formes critiques*, Ph.D. thesis, ENS Cachan Bretagne, Rennes, 2000 (in French).
[12] M. Dambrine, *On variations of the shape Hessian and sufficient conditions for the stability of critical shapes*, RACSAM Rev. R. Acad. Cienc. Exactas Fis. Nat. Ser. A. Mat., 96 (2002), pp. 95–121.
[13] M. C. Delfour and J.-P. Zolésio, *Velocity method and Lagrangian formulation for the computation of the shape Hessian*, SIAM J. Control Optim., 29 (1991), pp. 1414–1442.
[14] M. C. Delfour and J.-P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, SIAM, Philadelphia, 2001.
[15] K. Eppler, *Boundary integral representations of second derivatives in shape optimization*, Discuss. Math. Differ. Incl. Control Optim., 20 (2000), pp. 63–78.
[16] K. Eppler, *Optimal shape design for elliptic equations via BIE-methods*, J. Appl. Math. Comput. Sci., 10 (2000), pp. 487–516.
[17] K. Eppler, *Second derivatives and sufficient optimality conditions for shape functionals*, Control Cybernet., 29 (2000), pp. 485–512.
[18] K. Eppler and H. Harbrecht, *Numerical solution of elliptic shape optimization problems using wavelet-based BEM*, Optim. Methods Softw., 18 (2003), pp. 105–123.
[19] K. Eppler and H. Harbrecht, *2nd order shape optimization using wavelet BEM*, Optim. Methods Softw., 21 (2006), pp. 135–153.
[20] K. Eppler and H. Harbrecht, *Exterior electromagnetic shaping using wavelet BEM*, Math. Methods Appl. Sci., 28 (2005), pp. 387–405.
[21] K. Eppler and H. Harbrecht, *Second order Lagrange multiplier approximation for constrained shape optimization problems*, in Control and Boundary Analysis, Lect. Notes Pure Appl. Math., J. Cagnol and J.-P. Zolésio, eds., Chapman & Hall/CRC, Boca Raton, FL, 2005, pp. 107–118.
[22] K. Eppler and H. Harbrecht, *Fast wavelet BEM for 3d electromagnetic shaping*, Appl. Numer. Math., 54 (2005), pp. 537–554.
[23] K. Eppler and H. Harbrecht, *A regularized Newton method in electrical impedance tomography using shape Hessian information*, Control Cybernet., 34 (2005), pp. 203–225.
[24] K. Eppler and H. Harbrecht, *Shape optimization for 3D electrical impedance tomography*, in Free and Moving Boundaries: Analysis, Simulation and Control, Lecture Notes Pure Appl. Math. 252, R. Glowinski and J.-P. Zolésio, eds., Chapman & Hall/CRC, Boca Raton, FL, to appear.
[25] K. Eppler and H. Harbrecht, *Efficient treatment of stationary free boundary problems*, Appl. Numer. Math., 56 (2006), pp. 1326–1339.
[26] M. Flucher and M. Rumpf, *Bernoulli's free-boundary problem, qualitative theory and numerical approximation*, J. Reine Angew. Math., 486 (1997), pp. 165–204.
[27] S. Garreau, Ph. Guillaume, and M. Masmoudi, *The topological asymptotic for PDE systems: The elasticity case*, SIAM J. Control Optim., 39 (2001), pp. 1756–1778.
[28] H. Goldberg and F. Tröltzsch, *Second order optimality conditions for a class of control problems governed by nonlinear integral equations with application to parabolic boundary control*, Optimization, 20 (1989), pp. 687–698.
[29] H. Goldberg and F. Tröltzsch, *Second-order sufficient optimality conditions for a class of nonlinear parabolic boundary control problems*, SIAM J. Control Optim., 31 (1993), pp. 1007–1025.
[30] C. Grossmann and J. Terno, *Numerik der Optimierung*, Teubner, Stuttgart, 1993.
[31] Ph. Guillaume and K. Sid Idris, *The topological asymptotic expansion for the Dirichlet problem*, SIAM J. Control Optim., 41 (2002), pp. 1042–1072.
[32] J. Hadamard, *Lessons on Calculus of Variations*, Gauthier–Villiars, Paris, 1910 (in French).
[33] J. Haslinger and P. Neitaanmäki, *Finite Element Approximation for Optimal Shape, Material and Topology Design*, 2nd ed., Wiley, Chichester, 1996.
[34] J. Haslinger, T. Kozubek, K. Kunisch, and G. Peichl, *Shape optimization and fictitious domain approach for solving free boundary value problems of Bernoulli type*, Comput. Optim. Appl., 26 (2003), pp. 231–251.

[35] A. M. Khludnev and J. Sokołowski, *Modelling and Control in Solid Mechanics*, Birkhäuser, Basel, 1997.
[36] K. Mårtensson, *A new approach to constrained function optimization*, J. Optim. Theory Appl., 12 (1973), pp. 531–554.
[37] W. G. Mazja, S. A. Nazarov, and B. A. Plamenevsky, *Asymptotic Theory of Elliptic Boundary Value Problems in Singularly Perturbed Domains*, I, II, Birkhäuser, Basel, 2000.
[38] F. Murat and J. Simon, *Étude de problémes d'optimal design*, in Optimization Techniques, Modeling and Optimization in the Service of Man, J. Céa, ed., Lect. Notes Comput. Sci. 41, Springer-Verlag, Berlin, 1976, pp. 54–62.
[39] S. A. Nazarov and J. Sokołowski, *Asymptotic analysis of shape functionals*, J. Math. Pures Appl., 82 (2003), pp. 125–196.
[40] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.
[41] A. Novruzi and J. R. Roche, *Second Derivatives, Newton Method, Application to Shape Optimization*, Tech. report 2555, INRIA, Lecheshay, France, 1995.
[42] A. Novruzi and J.-R. Roche, *Newton's method in shape optimisation: A three-dimensional case*, BIT, 40 (2000), pp. 102–120.
[43] M. Pierre and J.-R. Roche, *Computation of free surfaces in the electromagnetic shaping of liquid metals by optimization algorithms*, Eur. J. Mech. B Fluids, 10 (1991), pp. 489–500.
[44] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer, New York, 1983.
[45] J.-R. Roche and J. Sokołowski, *Numerical methods for shape identification problems*, Control Cybernet., 25 (1996), pp. 867–894.
[46] J. Simon, *Differentiation with respect to the domain in boundary value problems*, Numer. Funct. Anal. Optim., 2 (1980), pp. 649–687.
[47] J. Sokołowski and J.-P. Zolésio, *Introduction to Shape Optimization*, Springer, Berlin, 1992.
[48] J. Sokołowski and A. Żochowski, *On the topological derivative in shape optimization*, SIAM J. Control Optim., 37 (1999), pp. 1251–1272.
[49] J. Sokołowski and A. Żochowski, *Topological derivatives for elliptic problems*, Inverse Problems, 15 (1999), pp. 123–134.
[50] J. Sokołowski and A. Żochowski, *Optimality conditions for simultaneous topology and shape optimization*, SIAM J. Control Optim., 42 (2003), pp. 1198–1221.
[51] T. Tiihonen, *Shape optimization and trial methods for free-boundary problems*, RAIRO Model. Math. Anal. Numér., 31 (1997), pp. 805–825.
[52] J. Wloka, *Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1987.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 84–115

LYAPUNOV-BASED DESIGN OF iISS FEEDFORWARD SYSTEMS WITH UNCERTAINTY AND NOISY MEASUREMENTS*

STEFANO BATTILOTTI†

Abstract. We study the problem of achieving integral input to state stability (iISS) with respect to noise for a class of upper triangular nonlinear systems with uncertainty and measurement noise. We propose a novel step-by-step Lyapunov-based design, consisting of (1) splitting an n-dimensional system into n one-dimensional systems, each with its own state, inputs, and measurement, (2) constructing a one-dimensional measurement feedback controller for each one-dimensional system, according to a certainty equivalence principle, and (3) selecting the parameters of these controllers so that their interconnection gives a measurement feedback controller for the n-dimensional system. The stability analysis is performed through *filtered* Lyapunov functions, which are Lyapunov functions with parameters being the output of suitable dynamical filters.

Key words. measurement feedback, bounded feedback, feedforward systems, uncertain systems

AMS subject classifications. 93D05, 93D15, 93D20

DOI. 10.1137/050631501

1. Introduction and main results. In practical applications the controlled output of a system is in general different from the measured output (*measurement*), and, moreover, the measured output is affected by noise and uncertainty. In this paper we study the problem of achieving integral input to state stability (iISS) in the sense of [1] with respect to noise by means of measurement feedback. To this aim, consider the class of systems

$$\dot{x}_j = x_{j+1} + \psi_{js}(x, x_{n+1}, w), \quad \mu_j = x_j + \psi_{jm}(x, x_{n+1}, w), \quad j = 1, \dots, n, \tag{1.1}$$

with states $x = (x_1 \ \cdots \ x_n)^T$, control $u := x_{n+1}$, noise $w \in \mathcal{L}_2^r(\mathbb{R}^{\geq})$ (space of functions $w : [0, \infty) \to \mathbb{R}^r$ such that $\int_0^\infty \|w(s)\|^2 ds < \infty$), uncertainties $\psi_{js}(x, x_{n+1}, w)$ and $\psi_{jm}(x, x_{n+1}, w)$, and measurements $\mu = (\mu_1 \ \cdots \ \mu_n)^T$, with continuous functions $\psi_{ji} : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^r \to \mathbb{R}$ such that for $j = 1, \dots, n$, $i = s, m$, and $h = j+1, \dots, n+1$,

$$\begin{aligned} &|\psi_{ji}|_{w=0} - \psi_{ji}|_{x_h=\overline{x}_h, w=0}|^2 \leq |x_h - \overline{x}_h|^2 a_{jih}(\overline{x}_h, x_{j+1}, \dots, x_{n+1}), \\ &|\psi_{ji}(x, x_{n+1}, w) - \psi_{ji}|_{w=0}|^2 \leq \|w\|^2 b_{ji}(x_{j+1}, \dots, x_{n+1}), \\ &a_{js,j+1}(0, 0, \dots, 0) = 0, \quad \psi_{ji}|_{x_h=0, h=j+1,\dots,n+1; w=0} = 0 \end{aligned} \tag{1.2}$$

for all $x \in \mathbb{R}^n, w \in \mathbb{R}^r$, and $x_{n+1}, \overline{x}_h \in \mathbb{R}$ and for some smooth functions $a_{jih} : \mathbb{R}^{n-j+2} \to \mathbb{R}^{\geq}$ and $b_{ji} : \mathbb{R}^{n-j+1} \to \mathbb{R}^{\geq}$, where $\psi_{ji}|_{x_h=\overline{x}_h}$ denotes the function $\psi_{ji}(x, x_{n+1}, w)$ evaluated for $x_h = \overline{x}_h$. In particular, if each function $\psi_{ji}(x, x_{n+1}, w)$, $i = s, m$, does not depend on $x_1, \dots, x_j$, then the first two inequalities of (1.2) are satisfied as long as $\psi_{ji}(x, x_{n+1}, w)$, $i = s, m$, is smooth and grows at most linearly with respect to w. The linear growth with respect to w is required only for simplicity and can be relaxed, by replacing $\|w\|^2$ with $\alpha(\|w(s)\|)$, $\alpha \in \mathcal{K}$, such that $\int_0^\infty \alpha(\|w(s)\|) ds < \infty$. On the other hand, $a_{js,j+1}(0, 0, \dots, 0) = 0$ requires that for

*Received by the editors May 13, 2005; accepted for publication (in revised form) September 25, 2006; published electronically March 22, 2007.
http://www.siam.org/journals/sicon/46-1/63150.html

†Dipartimento di Informatica e Sistemistica "Antonio Ruberti," Università di Roma La Sapienza, Via Eudossiana 18, 00184 Rome, Italy (battilotti@dis.uniroma1.it).

$x_l = 0$, $l = j+2, \dots, n+1$, the incremental ratio of $\psi_{ji}(x, x_{n+1}, 0)$ between x_{j+1} and $\overline{x}_{j+1}$ can be made in norm as small as desired by making small x_{j+1} and $\overline{x}_{j+1}$. For smooth ψ_{ji} this means that ψ_{ji} does not contain, for $x_l = 0$, $l = j+2, \dots, n+1$, linear terms in x_{j+1}. The example

$$\begin{aligned} \dot{x}_1 &= x_2 + x_2^2 \sin(x_1 x_2) + x_2 x_3 w, \quad \mu_1 = x_1, \\ \dot{x}_2 &= x_3, \quad \mu_2 = x_2 + x_3^4 \sin(x_2) + x_3^2 w, \\ \dot{x}_3 &= u + u^2 \cos(x_1 x_3), \quad \mu_3 = x_3 + w \end{aligned} \tag{1.3}$$

clearly satisfies (1.2) with $a_{1s2} = 2(x_2 + \overline{x}_2)^2 + 2\overline{x}_2^4 \max_{x_1, x_2, \overline{x}_2}[(\sin(x_1 x_2) - \sin(x_1 \overline{x}_2)) / (x_2 - \overline{x}_2)]^2$, $a_{2m3} = (x_3 + \overline{x}_3)^2 (x_3^2 + \overline{x}_3^2)^2$, $a_{3s4} = (u + \overline{u})^2$, $b_{1s} = x_2^2 x_3^2$, $b_{2m} = x_3^4$, and $b_{3m} = 1$. The main result of this paper for the class of systems (1.1) is the following theorem. Let $\widetilde{x}_1 := 0$, $\widetilde{x}_{n+1} := x_{n+1} = u$, and $G(s) = \frac{s}{\sqrt{1+s^2}}$.

THEOREM 1.1. *Under assumptions* (1.2) *there exist* $h_j \in (0,1)$ *and* $R_j \geq 1$, $j = 1, \dots, n$, *such that the feedback controller* $\mathcal{C}$,

$$\widetilde{x}_{j+1} = -\frac{G(\sigma_j)}{2R_j},\ \dot{\sigma}_j = \frac{1}{2R_j}\Big[(h_j - 1)G(\sigma_j) + \frac{1}{h_j} G(\mu_j - \widetilde{x}_j - \sigma_j)\Big],\ j = 1, \dots, n, \tag{1.4}$$

renders (1.1) *iISS with respect to* w.

We recall that a system $\Sigma : \dot{z} = f(z, w)$ is *iISS* if there exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ and $\beta \in \mathcal{KL}$ such that $\alpha_1(\|z(t)\|) \leq \beta(\|z(0)\|, t) + \int_0^t \alpha_2(\|w(s)\|) ds$ holds for all $t \geq 0$ along the trajectories of Σ. Note the "nested" structure of the controller (1.4), in the sense that the dynamics of $\dot{\sigma}_j$ depends on σ_{j-1} through the term $\widetilde{x}_j$.

With full state information (i.e., $\mu = x$) and without noise (i.e., $w = 0$) systems (1.1) have been studied in [7], [12] and extensively in the textbook [13]. On the other hand, with full state information iISS properties with respect to w have been obtained in [14]: we remark that Theorem 1.1 gives an improvement over [14] even with state feedback, since we do not require in (1.2) that $b_{ji}(0, \dots, 0) = 0$. However, these results cannot be applied to (1.1)–(1.2) since the measurements μ may differ significantly from the states x due to unbounded noise w and large uncertainty and, moreover, the coordinate change, as introduced in section 6.2 of [13], is a function of the noise and uncertainty and can hardly be implemented as a (smooth) function of the measurements. Moreover, even an observer-based control design is not feasible since the observer design for (1.1) is highly nontrivial. Note also that the systems (1.1)–(1.2) may be not in feedforward form, which is a substantial improvement over the above results.

Following [3] we implement a "dynamic" backstepping design in which $\widetilde{x}_{j+1}$ is used as the control for each dynamics $\dot{z}_j$, $j = 1, \dots, n$, where $z_j := x_j - \widetilde{x}_j$ is the backstepping coordinate, and for each state z_j we implement a robust observer to obtain an estimate of z_j to be used in $\widetilde{x}_{j+1}$ according to a certainty equivalence strategy (see (1.4)). In accordance with previous contributions on the subject, the boundedness of $\widetilde{x}_{j+1}$ in (1.4) is crucial due to the presence of terms $\mathcal{O}(|x_{j+1}|)$ in ψ_{js} (assumption $a_{js,j+1}(0, 0, \dots, 0) = 0$ in (1.2)), such as x_2^2 and u^2 in $\dot{x}_1$ and, respectively, $\dot{x}_3$ of (1.3) ($\mathcal{O}(|x_{j+1}|)$ means infinitesimals with order greater than $|x_{j+1}|$). Bounded backstepping with full state information has been also studied in [6] and [11].

The novelty with respect to previous contributions is the boundedness of the function $G(\mu_j - \widetilde{x}_j - \sigma_j)$ in $\dot{\sigma}_j$ of (1.4). This boundedness is crucial due to the nonlinear terms containing x_j fed back in $\dot{z}_j$ through $\dot{\widetilde{x}}_j$ (and thus, according to (1.4), through μ_{j-1}) and which could not be otherwise counteracted by the bounded control

$\widetilde{x}_{j+1}$ of $\dot{z}_j$. In example (1.3) with a controller (1.4) the term $x_3^4 \sin x_2$ in $\dot{\sigma}_2$ cannot be counteracted by the bounded control term $u + u^2 \cos(x_1 x_3)$ in $\dot{z}_3$ unless $G(s)$ is bounded for all s.

The proof of the main Theorem 1.1 is organized along the following lines.

• Define a system as characterized by states z, inputs ι, measurements ς, and uncertainty (section 3); both the inputs and measurements can be distinguished as endogenous (i.e., from inside the system) and exogenous (i.e., from outside the system) type. The endogenous inputs are the controls, and the endogenous measurements are the measured outputs. The rate of the uncertainty with respect to each state and input is quantified by the *incremental rates*.

• Split (1.1) into n one-dimensional systems Σ_j, $j = 1, \dots, n$ (section 4), each one with state $z_j := x_j - \widetilde{x}_j$, inputs ι_j (of which only one controls $\widetilde{x}_{j+1}$), and measurements ς_j (of which only one is an endogenous measurement $\widetilde{\mu}_j := x_j - \widetilde{x}_j + \psi_{jm}$).

• Find a one-dimensional measurement feedback controller $\mathcal{C}_j$ (each $\widetilde{x}_{j+1}$ and $\dot{\sigma}_j$ in (1.4)) and a Lyapunov function W_j for each one-dimensional system $\Sigma_j \circ \mathcal{C}_j$ (section 5) according to a certainty equivalence design (Theorem 3.1 of [4]).

• Take the interconnection $\mathcal{C}_j$, $j = 1, \dots, n$ (the controller (1.4)), as candidate controller $\mathcal{C}$ for (1.1) (section 8).

• Prove the iISS properties of the closed-loop system $\Sigma_j \circ \mathcal{C}_j$, $j = 1, \dots, n$, by suitably selecting the parameters of the controllers $\mathcal{C}_j$, $j = 1, \dots, n$, in such a way as to enlarge the stability margins of each $\Sigma_j \circ \mathcal{C}_j$ and compensate for the incremental rates of the exogenous inputs in the time derivatives $\dot{W}_i$, $i = 1, \dots, n$ and $i \neq j$ (section 8). A Lyapunov function W for the closed-loop system Σ is obtained from W_j, $j = 1, \dots, n$, according to Theorem 7.4 in section 7. Theorem 7.4 gives the procedure for obtaining a Lyapunov function under the quite general assumptions that the state trajectories are bounded and definitely enter a neighborhood of the origin, where a small gain condition is met. With respect to the small gain theorem of [8] which assesses only the existence of a Lyapunov function W for the interconnection, application of Theorem 7.4 to systems (1.1)–(1.4) gives a Lyapunov function W consisting of a "filtered" combination of $W_1, \dots, W_n$, i.e., $\theta[W_1 + \tau_2[W_2 + \tau_3[\cdots + \cdots]]]$, where $\tau_2, \dots, \tau_n > 0$ and θ is the output of a filter implemented by explicitly using the stability margins and incremental rates of each $\Sigma_j \circ \mathcal{C}_j$, $j = 1, \dots, n$, in the time derivatives $\dot{W}_j$, $j = 1, \dots, n$. This leads to the notion of *filtered Lyapunov functions* (section 6) which give a new tool for the stability analysis of interconnected systems like (1.1)–(1.4). These systems are not triangular and, thus, the results for the design of composite Lyapunov functions given in [7] cannot be applied (Example 6.1 in section 6).

2. Notation. • $\|v\| = \sqrt{v^T v}$ denotes the euclidean norm of any given vector v and $\|v\|_A := \sqrt{v^T A v}$ for any positive semidefinite matrix A. $\mathbb{R}^s$ is the vector space of s-dimensional real column vectors; $\mathbb{R}^+$ (resp., $\mathbb{R}^{\geq}$) denotes the set of positive (resp., nonnegative) real numbers; I_n is the $n \times n$ identity matrix; and $\mathbb{R}^{n \times n}$ denotes the set of $n \times n$ matrices.

• For any continuous function $f : \mathbb{R}^q \times \mathbb{R}^l \to \mathbb{R}^r$, $(s, r) \mapsto f(s, r)$, we denote by $f(z, r)$ or $f|_{s=z}$ the function $f(s, r)$ with $s = z$. A continuous function $\alpha : \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$ is said to be of class $\mathcal{K}$ (or $\alpha \in \mathcal{K}$) if $\alpha(0) = 0$ and it is increasing; a function $\alpha : \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$ is said to be of class $\mathcal{K}_\infty$ (or $\alpha \in \mathcal{K}_\infty$) if $\alpha \in \mathcal{K}$ and $\lim_{s \to +\infty} \alpha(s) = +\infty$; a function $\alpha : \mathbb{R}^{\geq} \times \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$ is said to be of class $\mathcal{KL}$ (or $\alpha \in \mathcal{KL}$) if for each r, $\alpha(s, r)$ is of class $\mathcal{K}$ and for each s it is decreasing and $\lim_{r \to +\infty} \alpha(s, r) = 0$; $\mathcal{K}_\infty^0$ is the class of continuous functions $f : \mathbb{R} \to \mathbb{R}^{\geq}$ such that $f \in \mathcal{K}_\infty$ when restricted to $[0, \infty)$.

• $\mathcal{L}_2^s(\mathbb{R}^{\geq})$ is the class of measurable and square integrable functions $\chi : \mathbb{R}^{\geq} \to \mathbb{R}^s$, and the norm of $\chi(t)$ in $\mathcal{L}_2^s(\mathbb{R}^{\geq})$ is $\|\chi\|_2 = \sqrt{\int_0^\infty \|\chi(\tau)\|^2 d\tau}$. We simply write $\mathcal{L}_2(\mathbb{R}^{\geq})$ when $s = 1$.

• For any given functions $h_j : \mathbb{R}^q \to \mathbb{R}$, $j = 1, 2$, we say that $h_1(s)$ *is of the same order as* $h_2(s)$ (in $\mathcal{S} \subset \mathbb{R}^q$) if there exists $\kappa_1, \kappa_2 > 0$ such that $\kappa_1 h_2(s) \leq h_1(s) \leq \kappa_2 h_2(s)$ for all s ($\in \mathcal{S}$) and we write $h_1 \sim h_2$. Moreover, we say that $h_1(s)$ *is less than or equal to* $h_2(s)$ (in $\mathcal{S} \subset \mathbb{R}^q$) if there exists $\kappa > 0$ such that $h_1(s) \leq \kappa h_2(s)$ for all s ($\in \mathcal{S}$) and we write $h_1 \preceq h_2$.

• For any smooth functions $V : \mathbb{R} \to \mathbb{R}^{\geq}$, $s \mapsto V(s)$, we denote by $\nabla_r V(r)$ the derivative of $V(r)$ with respect to r and by $\nabla_s V|_r$ the derivative of $V(s)$ with respect to s evaluated for $s = r$.

3. Complex dynamics as interconnection of simpler dynamics. In view of a general approach to the stabilization of an interconnected system such as (1.3) we consider it as the interconnection of three one-dimensional systems:

$$\Sigma_1 : \dot{x}_1 = x_2 + x_2^2 \sin(x_1 x_2) + x_2 x_3 w, \quad \mu_1 = x_1$$

with state x_1, control x_2, and measurement μ_1;

$$\Sigma_2 : \dot{x}_2 = x_3, \quad \mu_2 = x_2 + x_3^4 \sin(x_2) + x_3^2 w$$

with state x_2, control x_3, and measurement μ_2; and

$$\Sigma_3 : \dot{x}_3 = u + u^2 \cos(x_1 x_3), \quad \mu_3 = x_3 + w$$

with state x_3, control u, and measurement μ_3. Note that in each system Σ_j the following hold.

– We can distinguish the inputs into *control (or endogenous) inputs* v (such as x_2 for Σ_1) and *exogenous inputs* χ (such as x_3 and w for Σ_1). We denote all the inputs, endogenous and exogenous, by ι.

– We can distinguish the measurements into *endogenous measurements* μ (available from the system itself such as μ_3 for Σ_3) and *exogenous measurements* ν (available from other systems such as μ_3 for Σ_1). We denote all the measurements, endogenous and exogenous, by ς.

– State, inputs and measurements satisfy some constraints, given by the measurement equations and differential equations of the other systems such as $x_2 = \mu_2 - x_3^4 \sin(x_2) - x_3^2 w$ and $x_3 = \mu_3 - w$ for Σ_2 or $\dot{x}_2 = x_3$ for Σ_1. Note that on account of these constraints $|x_2| \leq |\mu_2| + 8\mu_3^4 + 4w^4 + w^2$. We will refer to this entire set of constraints by saying that *a system* Σ *with states* $x \in \mathbb{R}^n$*, inputs* $\iota = (v^T \ \chi^T)^T \in \mathcal{I} \subseteq \mathbb{R}^m \times \mathbb{R}^r$*, and measurements* $\varsigma = (\mu^T \ \nu^T)^T \in \mathcal{Z} \subseteq \mathbb{R}^p \times \mathbb{R}^s$ *satisfies a set of constraints* $\mathcal{M}$ *among the state, inputs, and measurements or* $(x, \iota, \varsigma) \in \mathcal{M}$.

– State and measurement uncertainty ψ_{js} and ψ_{jm} can be accommodated into a vector Ψ, which denotes the "uncertainty" of a system. The uncertainty can be seen as a locally Lipschitz continuous function $\Psi : \mathbb{R}^n \times \mathcal{I} \to \mathbb{R}^q$, $(x, \iota) \mapsto \Psi(x, \iota)$, such that $\Psi(0, 0) = 0$. We will denote the uncertainty by Ψ or, when needed, more explicitly with ψ_{js} and ψ_{jm}.

Thus, any interconnected n-dimensional system such as (1.3) can be viewed as the interconnection of n one-dimensional systems, each one with a state z, a control v, an endogenous measurement μ, uncertainty Ψ, some exogenous inputs χ, some exogenous measurements ν, and a set of constraints $\mathcal{M}$ among the state, inputs, and measurements. It is important for the controller design to evaluate the effect of the

uncertainty Ψ on the system under the constraints $\mathcal{M}$. Indeed, the constraints $\mathcal{M}$ allow us to either get suitable bounds depending only on the measurements, which can be used in the design of a controller for the system itself (internal stability properties), or bound some state-dependent terms by means of the exogenous inputs which can be related to the stability properties of other interconnected systems (external stability properties). In the case of Σ_1 with state x_1, measurements ς_1 (endogenous μ_1 and exogenous $\nu_1 = (\mu_2\ \mu_3)^T$), inputs ι_1 (control $v = x_2$ and exogenous $\chi_1 = (x_3\ w)^T$), uncertainty $\Psi_1 = (x_2^2 \sin(x_1 x_2) + x_2 x_3 w \quad 0)^T$, and constraint $\mathcal{M}_1 = \{\mu_3 = x_3 + w,\ \mu_2 = x_2 + x_3^4 \sin(x_2) + x_3^2 w,\ |w| \le 1\}$, we have $(\|\Psi_1\| - \|\Psi_1|_{w=0}\|)^2 \le a[\mu_3^2 + 1][\mu_2^2 + \mu_3^8 + 1]^2 w^2$ for some $a > 0$ and for all $(x_1, \iota_1, \varsigma_1) \in \mathcal{M}_1$. Note that the function $\gamma_1^w(\varsigma_1) = a[\mu_3^2+1][\mu_2^2+\mu_3^8+1]^2$ gives a "worst case" bound of the "incremental term" $(\|\Psi_1\| - \|\Psi_1|_{w=0}\|)^2/w^2$ under the constraints $\mathcal{M}_1$. Note also that the constraint $\mathcal{M}_1$ has been used in γ_1^w to bound the inputs x_2, x_3 in terms of the measurements μ_2, μ_3 so that γ_1^w can be used in the design of a controller for Σ_1. This motivates the following definition, which we recall here from [4] for extensive use.

DEFINITION 3.1 (incremental rate). *We will say that a system Σ with states $x \in \mathbb{R}^n$, inputs $\iota \in \mathcal{I} \subseteq \mathbb{R}^m \times \mathbb{R}^r$, measurements $\varsigma \in \mathcal{Z} \subseteq \mathbb{R}^p \times \mathbb{R}^s$, uncertainty $\Psi \in \mathbb{R}^q$, and constraints $\mathcal{M}$ has (smooth) incremental rate γ^z if there exist a nonempty subvector z of $(x^T\ \iota^T)^T$ and a (smooth) nonnegative function $\gamma^z : \mathbb{R}^n \times \mathcal{I} \times \mathcal{Z} \to \mathbb{R}^{\ge}$ such that $(\|\Psi\| - \|\Psi|_{z=0}\|)^2 \le \gamma^z(x, \iota, \varsigma)\|z\|^2$ for all $(x, \iota, \varsigma) \in \mathcal{M}$.*

Since $\Psi|_{x,v,\chi=0} = 0$ and $\|\Psi\| \le \|\Psi - \Psi|_{x=0}\| + \|\Psi|_{x=0} - \Psi|_{x,v=0}\| + \|\Psi|_{x,v=0} - \Psi|_{x,v,\chi=0}\|$, for a system Σ with states x, inputs ι, measurements ς, uncertainty Ψ, and constraints $\mathcal{M}$ we expect to have the following general relation among Ψ on one side, and x, ι, ς on the other, under the constraints $\mathcal{M}$:

$$\gamma^2(\varsigma)\|\Psi\|^2 \le \gamma^x(x,\varsigma)\|x\|^2 + \gamma^v(\varsigma)\|v\|^2 + \sum_{j\in J} \gamma^{\chi_j}(x,\iota,\varsigma)\chi_j^2 \quad \forall (x,\iota,\varsigma) \in \mathcal{M}, \tag{3.1}$$

where χ_j is the jth element of χ, $j \in J := \{1, \dots, r\}$, and $\gamma^x : \mathbb{R}^n \times \mathcal{Z} \to \mathbb{R}^{\ge}$, $\gamma^v : \mathcal{Z} \to \mathbb{R}^+$, and $\gamma^{\chi_j} : \mathbb{R}^n \times \mathcal{I} \times \mathcal{Z} \to \mathbb{R}^{\ge}$ are (smooth) incremental rates ("rescaled" by the square of a smooth function $\gamma : \mathcal{Z} \to \mathbb{R}^+$). The inequality (3.1) means that the uncertainty Ψ is known up to the square of the states and the inputs, weighted by the corresponding incremental rates evaluated under some constraints $\mathcal{M}$. This approach is inspired by an $\mathcal{H}_\infty$-control problem formulation, with Ψ having the role of a "disturbance," the right-hand side of (3.1) having the role of a "penalty index," and γ having the role of an "attenuation level" [2]. Note that the incremental rate γ^v is assumed (without loss of generality) to be a positive real function. Note also that, by use of the constraints $\mathcal{M}$ in (3.1), the incremental rate γ^v depends only on the measurements, so that it can be used directly in designing the controller gains.

DEFINITION 3.2 (incremental rates and scaling of Σ). *We will say that a system Σ with states $x \in \mathbb{R}^n$, inputs $\iota \in \mathcal{I} \subseteq \mathbb{R}^m \times \mathbb{R}^r$, measurements $\varsigma \in \mathcal{Z} \subseteq \mathbb{R}^p \times \mathbb{R}^s$, uncertainty $\Psi \in \mathbb{R}^q$, and constraints $\mathcal{M}$ has (smooth) incremental rates γ^x, γ^v, and γ^{χ_j}, $j \in J$, with scaling γ if* (3.1) *holds for some (smooth) nonnegative functions $\gamma^x : \mathbb{R}^n \times \mathcal{Z} \to \mathbb{R}^{\ge}$, $\gamma^{\chi_j} : \mathbb{R}^n \times \mathcal{I} \times \mathcal{Z} \to \mathbb{R}^{\ge}$ and positive $\gamma^v, \gamma : \mathcal{Z} \to \mathbb{R}^+$.*

Throughout the paper, *when an incremental rate is not explicitly cited in the context, we consider it equal to zero.*

4. Splitting systems into simpler dynamics. In this section, using the general framework introduced in section 3, we split the dynamics (1.1) into n one-dimensional systems having a "canonical" form. To do this we change coordinate x_j into *backstepping coordinate* $z_j := x_j - \widetilde{x}_j$, $j = 1, \dots, n$, $\widetilde{x}_1 := 0$, to use $\widetilde{x}_{j+1}$ as the

control for $\dot z_j$ instead of x_{j+1} (which is a state for (1.1)) as the control for $\dot x_j$. Also we change each measurement μ_j into $\widetilde\mu_j := \mu_j - \widetilde x_j$, $j = 1, \dots, n$, to have z_j as the "nominal part" of the measurement $\widetilde\mu_j$. Since by (1.2) each ψ_{ji} contains terms $\mathcal{O}(|\widetilde x_{j+1}|)$, to control $\dot z_j$ through $\widetilde x_{j+1}$ it is thus important to keep $\widetilde x_{j+1}$ as small as possible (see [14]). This accounts for introducing the *constraints* $\{|\widetilde x_{i+1}| \le \Delta_i \in (0,1], i = 1, \dots, n\}$. We see that in *backstepping coordinates* z_j, $j = 1, \dots, n$, and with *measurement change* $\widetilde\mu_j$, $j = 1, \dots, n$, (1.1) can be split into n one-dimensional systems of the form

$$\Sigma_j : \dot z_j = \widetilde x_{j+1} + \widetilde\psi_{js},\ \widetilde\mu_j = z_j + \psi_{jm} \tag{4.1}$$

with state z_j, control $\widetilde x_{j+1}$ with $\widetilde x_{n+1} := u$, (endogenous) measurement $\widetilde\mu_j$, and exogenous inputs $z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_n, \widetilde x_2, \dots, \widetilde x_j, \dot{\widetilde x}_j, \widetilde x_{j+2}, \dots, \widetilde x_{n+1}$ and w; uncertainties ψ_{jm} and $\widetilde\psi_{js} := \psi_{js} + z_{j+1} - \dot{\widetilde x}_j$, with $\dot{\widetilde x}_1 := 0$ and $z_{n+1} := 0$; and constraints $\widetilde M_j := \{|\widetilde x_{i+1}| \le \Delta_i \in (0,1], i = 1, \dots, n\}$. Moreover, by using (1.2) and Lemma A.1 we find $\gamma_{ji}^{\widetilde x_h}(\Delta_j, \dots, \Delta_{h-1}) \ge 0$, $h = j+1, \dots, n+1$, $i = s, m$, and smooth functions $\gamma_{ji}^{z_h} : \mathbb{R}^{n-h+1} \to \mathbb{R}^{\ge}$, $h = j+1, \dots, n$, and $\gamma_{ji}^{w} : \mathbb{R}^{n-j} \to \mathbb{R}^{\ge}$, $i = s, m$, such that (A.1)–(A.3) hold true for $j = 1, \dots, n$ and for all $w \in \mathbb{R}^r$, $x_h, \widetilde x_h \in \mathbb{R}$, $h = 1, \dots, n+1$, $\dot{\widetilde x}_h \in \mathbb{R}$, $h = 1, \dots, n$, such that $|\widetilde x_h| \le \Delta_{h-1}$. As a consequence of (A.2)–(A.3) $\gamma_{js}^{\dot{\widetilde x}_j} := 2$, $\gamma_{jm}^{\dot{\widetilde x}_j} := 0$, $\gamma_{ji}^{\widetilde x_h}$, $h = j+1, \dots, n+1$, $i = s, m$, $\gamma_{ji}^{z_h}$, $h = j+1, \dots, n$, and γ_{ji}^{w}, $i = s, m$, are the incremental rates of Σ_j. We remark that the incremental rates of $z_1, \dots, z_{j-1}$ and $\widetilde x_2, \dots, \widetilde x_j$ are set to zero on account of the same Lemma A.1. Note also that the input $\dot{\widetilde x}_j$ will be specified later. When in the proof of Theorem 1.1 we will define $\widetilde x_j$ as a function of σ_{j-1} (see (1.4)), we will consider z_{j-1} and $z_{j-1} - \sigma_{j-1}$ as exogenous inputs for Σ_j and express $\dot{\widetilde x}_j$ and $\widetilde x_j$ in terms of the incremental rates of z_{j-1} and $z_{j-1} - \sigma_{j-1}$.

Example 4.1. In the case of (1.3) we get the three one-dimensional systems Σ_1 : $\dot z_1 = \widetilde x_2 + \widetilde\psi_{1s}$, $\widetilde\mu_1 = z_1$, with $\widetilde\psi_{1s} = z_2 + (z_2 + \widetilde x_2)^2 \sin(z_1(z_2 + \widetilde x_2)) + (z_2 + \widetilde x_2)(z_3 + \widetilde x_3)w$ and incremetal rates $\gamma_{1s}^{z_2} \sim 1 + z_2^2$, $\gamma_{1s}^{\widetilde x_2} \sim \Delta_1^2$, and $\gamma_{1s}^{w} \sim (z_2^2 + 1)(z_3^2 + 1)$ under the constraints $\{|\widetilde x_{i+1}| \le \Delta_i \in (0,1], i = 1, 2, 3\}$; $\Sigma_2 : \dot z_2 = \widetilde x_3 + \widetilde\psi_{2s}$, $\widetilde\mu_2 = z_2 + \psi_{2m}$, with $\widetilde\psi_{2s} = z_3 - \dot{\widetilde x}_2$, $\psi_{2m} = (z_3 + \widetilde x_3)^4 \sin(z_2 + \widetilde x_2) + (z_3 + \widetilde x_3)^2 w$ and incremetal rates $\gamma_{2s}^{z_3}, \gamma_{2s}^{\dot{\widetilde x}_2} \sim 1$, $\gamma_{2m}^{w} \sim (z_3 + 1)^4$, $\gamma_{2m}^{z_2} \sim 1 + z_3^6$, and $\gamma_{2m}^{\widetilde x_3} \sim \Delta_2^6$ under the constraints $\{|\widetilde x_{i+1}| \le \Delta_i \in (0,1], i = 1, 2, 3\}$; and $\Sigma_3 : \dot z_3 = u + \widetilde\psi_{3s}$, $\widetilde\mu_3 = z_3 + \psi_{3m}$, with $\widetilde\psi_{3s} = u^2 \cos(z_1(z_3 + \widetilde x_3)) - \dot{\widetilde x}_3$, $\psi_{3m} = w$ and incremetal rates $\gamma_{3s}^{u} \sim \Delta_3^2$ and $\gamma_{3s}^{\dot{\widetilde x}_3}, \gamma_{3m}^{w} \sim 1$ under the constraints $\{|\widetilde x_{i+1}| \le \Delta_i \in (0,1], i = 1, 2, 3\}$. As already stipulated regarding the other incremental rates which are not explicitly cited in the context, we consider them equal to zero.

In the next section we address the problem of controlling simple dynamics like (4.1).

5. Controlling simple dynamics. In this section we apply Theorem 3.1 of [4] for obtaining a measurement feedback controller $\mathcal{C}_j$ and a Lyapunov function W_j for any one-dimensional system Σ_j like (4.1). This result, although applied in our case to one-dimensional systems, can be used to extend Theorem 1.1 to block-feedforward systems, for which each system Σ_j in (4.1) has dimension possibly greater than one. By smooth *measurement feedback controller* $\mathcal{C}_j$ for Σ_j we mean

$$\dot\sigma_j = H_j(\sigma_j) + G_j(\widetilde\mu_j - \sigma_j), \quad \widetilde x_{j+1} = F_j(\sigma_j) \tag{5.1}$$

with $\sigma_j \in \mathbb{R}$ and smooth $H_j : \mathbb{R} \to \mathbb{R}$ and $G_j : \mathbb{R} \to \mathbb{R}$, vanishing at the origin and satisfying the *constraints*

$$|F_j(\sigma_j)| \le \Delta_{jf} \in (0,\infty], \quad |G_j(\widetilde{\mu}_j - \sigma_j)| \le \Delta_{jm} \in (0,\infty] \quad \forall \sigma_j, \widetilde{\mu}_j, \tag{5.2}$$

for some given $\Delta_{jf}, \Delta_{jm} \in (0,\infty]$. This definition extends without further remarks to the case of n-dimensional systems Σ_j. Note that the structure of the controller (5.1) is based on a *certainty equivalence principle*, consisting of replacing in the state feedback controller $\widetilde{x}_{j+1} = F_j(z_j)$ the state z_j with an estimate σ_j. The numbers Δ_{jf} and Δ_{jm} characterize, respectively, the maximum level allowed for the control input $\widetilde{x}_{j+1}$ and for the innovations $\widetilde{\mu}_j - \sigma_j$ fed back into the control loop by (5.1).

Definition 5.1 (controller levels). *We say that a smooth measurement feedback controller* (5.1) *has control input level* Δ_{jf} *and innovations level* Δ_{jm} *(or simply levels* $(\Delta_{jf}, \Delta_{jm})$*) if there exist* $\Delta_{jf}, \Delta_{jm} \in (0,\infty]$ *such that* (5.2) *holds true.*

We recall that any system Σ_j like (4.1) has state z_j, control $\widetilde{x}_{j+1}$, (endogenous) measurement $\widetilde{\mu}_j$, and among its exogenous inputs $z_{j+1}, \dots, z_n, \widetilde{x}_j, \dot{\widetilde{x}}_j, \widetilde{x}_{j+2}, \dots, \widetilde{x}_{n+1}$ and w. Throughout, we will denote by χ_{ji}, $i \in J_j$, any one of these exogenous inputs. Also, set $\gamma_{js}^{\dot{\widetilde{x}}_j} = 2$ and $\gamma_{jm}^{\dot{\widetilde{x}}_j} = 0$, and let $G(s)$ be as in Theorem 1.1.

Theorem 5.2. *For any system* Σ_j *like* (4.1) *satisfying* (A.1) *and* (A.2)–(A.3) *for* $j = 1, \dots, n$ *and for all* $w \in \mathbb{R}^r$*;* $x_h, \widetilde{x}_h \in \mathbb{R}$*,* $h = 1, \dots, n+1$*; and* $\dot{\widetilde{x}}_h \in \mathbb{R}$*,* $h = 1, \dots, n$*, such that* $|\widetilde{x}_h| \le \Delta_{h-1}$*, there exist* $\Delta_j, h_{js}, h_{jm} \in (0,1)$ *and* $R'_{j1} \ge 1$ *such that for all* $R_{j1} \ge R'_{j1}$ *the smooth measurement feedback controller*

$$\mathcal{C}_j : \widetilde{x}_{j+1} = -\frac{1}{2R_{j1}} G(\sigma_j), \quad \dot{\sigma}_j = \frac{1}{2R_{j1}} \Big[(h_{js} - 1) G(\sigma_j) + \frac{1}{h_{js}} G(\widetilde{\mu}_j - \sigma_j) \Big] \tag{5.3}$$

has levels $(\Delta_j, \Delta_j/h_{js})$ *and* $W_j(z_j, \sigma_j) = \sqrt{1+z_j^2} + \sqrt{1+(z_j-\sigma_j)^2} - 2$ *is smooth, proper, and positive definite and satisfies along the trajectories of* $\Sigma_j \circ \mathcal{C}_j$

$$\dot{W}_j \le -\varphi_{js}(z_j) z_j^2 - \varphi_{jm}(z_j - \sigma_j)(z_j - \sigma_j)^2 + \sum_{i \in J_j} \gamma_j^{\chi_{ji}} \chi_{ji}^2 \tag{5.4}$$

with stability margins

$$\varphi_{js}(z_j) := \frac{1 - h_{js}}{8R_{j1}(1+z_j^2)}, \quad \varphi_{jm}(z_j - \sigma_j) := \frac{1 - 16h_{js}}{32R_{j1}h_{js}[1+(z_j-\sigma_j)^2]} \tag{5.5}$$

and incremental rates

$$\gamma_j^{\chi_{ji}} := \frac{R_{j1}}{h_{js}} [\gamma_{js}^{\chi_{ji}} + h_{jm} \gamma_{jm}^{\chi_{ji}}]. \tag{5.6}$$

Proof. First we rewrite Σ_j as a system Σ of the form (1) in [4]; then we check the assumptions of Theorem 3.1 of [4] on Σ. By (A.1) there exist $\Delta_j \in (0,1)$ and $h_{js}, h_{jm} > 0$ such that

$$\gamma_{js}^{\widetilde{x}_{j+1}}(\Delta_j) + h_{jm} \gamma_{jm}^{\widetilde{x}_{j+1}}(\Delta_j) \le h_{js} \le 1/80. \tag{5.7}$$

Let

$$R_{j1} > R'_{j1} := \max \left\{ \frac{1}{\Delta_j}, \frac{1}{\sqrt{h_{jm}}} \right\}, \quad \gamma_j^2 = \frac{R_{j1}}{h_{js}}, \tag{5.8}$$

$$C_2 = B_2 = 1, \quad B_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad C_1 = \begin{pmatrix} 0 & 1 \end{pmatrix} R_{j1}, \quad \Psi_j = \begin{pmatrix} \widetilde{\psi}_{js} & \dfrac{\psi_{jm}}{R_{j1}} \end{pmatrix}^T. \tag{5.9}$$

Along with positions (5.9), Σ_j reads out as a system Σ of the form (1) of [4] with $B_1C_1^T = 0$, $R_2 := C_1C_1^T = R_{j1}^2 > 0$, with state $x = z_j$, control $v = \widetilde{x}_{j+1}$, uncertainty $\Psi = \Psi_j$, and exogenous inputs χ_{ji}, $i \in J_j$. Moreover, from (A.2)–(A.3), the first inequality in (5.7), the fact that $h_{jm} > 1/R_{j1}^2$ and $h_{js}\gamma_j^2 = R_{j1}$ by (5.8), and having "rescaled" ψ_{jm} as in (5.9), we obtain under the constraints $\widetilde{M}_j := \{|\widetilde{x}_{i+1}| \le \Delta_i \in (0,1], i = 1,\dots,n\}$

$$\gamma_j^2\|\Psi_j\|^2 = \gamma_j^2\left[\widetilde{\psi}_{js}^2 + \frac{\psi_{jm}^2}{R_{j1}^2}\right] \le R_{j1}\widetilde{x}_{j+1}^2 + \frac{R_{j1}}{h_{js}}\Big[(\gamma_{js}^{\dot{\widetilde{x}}_j} + h_{jm}\gamma_{jm}^{\dot{\widetilde{x}}_j})|\dot{\widetilde{x}}_j|^2$$

$$(5.10)\quad + \sum_{h=j+1}^{n} z_h^2(\gamma_{js}^{z_h} + h_{jm}\gamma_{jm}^{z_h}) + \sum_{h=j+2}^{n+1} \widetilde{x}_h^2(\gamma_{js}^{\widetilde{x}_h} + h_{jm}\gamma_{jm}^{\widetilde{x}_h}) + (\gamma_{js}^{w} + h_{jm}\gamma_{jm}^{w})\|w\|^2\Big];$$

i.e., the incremental rates of Σ are $\gamma^v := R_{j1}$ and $\gamma^{\chi_{ji}} := \gamma_j^{\chi_{ji}}$ as in (5.6) with scaling $\gamma := \gamma_j$. We check the assumptions of Theorem 3.1 of [4] on Σ. Let $P_m = 1$ and $V_s(s) = V_m(s) = \widetilde{V}_m(s) = \frac{1}{2}[\sqrt{1+s^2} - 1]$. The functions $V_s(r)$ and $V_m(r)$ are proper and positive definite, and, moreover, $\nabla_s\widetilde{V}_m(s)/s := 1/[2\sqrt{1+s^2}] \in (0,1]$ for all s, and is even and nonincreasing for all $s \ge 0$. By direct calculations with $R_2 = R_{j1}^2$, we obtain that the inequalities (3) and (10) of [4] with $n = 1$ are satisfied, respectively, with $\varphi_s(s) = \varphi_{js}(s)$ and $\varphi_m(s) = \varphi_{jm}(s)$, where $\varphi_{js}(s)$ and $\varphi_{jm}(s)$ are as in (5.5) and, by the second inequality of (5.7), are positive for all z_j and σ_j. Also (12) and (13) of [4] follow from direct calculation of $\nabla_s\widetilde{V}_m(s)$, $\nabla_{ss}^2\widetilde{V}_m(s)$, and $\nabla_{sss}^3\widetilde{V}_m(s)$. Finally, since $R_{j1} > 1/\Delta_j > 1$ by (5.8) and since $\Delta_j \in (0,1)$, also the feedback constraints $|F(s,\varsigma)| \le \Delta_f$ and $|G(s,\varsigma)| \le \Delta_m$ in Theorem 3.1 of [4] are met for all s, with

$$(5.11)\qquad F(s,\varsigma) := -\frac{s}{2R_{j1}\sqrt{1+s^2}},$$
$$G(s,\varsigma) := \frac{s}{2R_{j1}h_{js}\sqrt{1+s^2}}, \quad \Delta_f = \Delta_j,\ \Delta_m = \frac{\Delta_j}{h_{js}}.$$

Finally, it is not difficult to see that

$$f_1(s_1,s_2) = \left[\frac{(s_1-s_2)}{\sqrt{1+(s_1-s_2)^2}} - \frac{s_1}{\sqrt{1+s_1^2}}\right]^2$$

has for each s_2 a global maximum for $s_1 = s_2/2$. Using this fact and the second inequality of (5.7), for all z_j, e_j

$$\frac{2}{\gamma_j^2}\nabla_{e_j}V_m(e_j)B_1B_1^T\Big[\nabla_{z_j}V_s(z_j) - \nabla_{z_j-e_j}V_s(z_j-e_j) - \nabla_{e_j}V_s(e_j)\Big]^T$$
$$+ \gamma^v(\|\ F(z_j-e_j) - F(z_j)\|^2 - \|F(e_j)\|^2)$$
$$\le \left[\frac{h_{js}}{R_{j1}} + \frac{1}{4R_{j1}}\right]\left[\frac{(z_j-e_j)}{\sqrt{1+(z_j-e_j)^2}} - \frac{z_j}{\sqrt{1+z_j^2}}\right]^2$$
$$\le \left[\frac{h_{js}}{R_{j1}} + \frac{1}{4R_{j1}}\right]\frac{e_j^2}{1+e_j^2/4} \le \frac{2}{R_{j1}}\frac{e_j^2}{1+e_j^2} \le \varphi_{jm}(e_j)e_j^2,$$

which implies (14) of [4] with $n = 1$. Application of Theorem 3.1 of [4] and (16) therein gives the controller $\mathcal{C}_j$ in (5.3) with levels $(\Delta_j, \Delta_j/h_{js})$ by (5.12) and that the smooth, proper, and positive definite $W_j(z_j,\sigma_j)$ satisfies (5.4) along the trajectories of $\Sigma_j \circ \mathcal{C}_j$. □

6. Filtered Lyapunov functions. The Lyapunov functions used in Theorem 5.2 cover with enough generality the stability analysis of systems of the form $\Sigma_j \circ \mathcal{C}_j$. However, when interconnecting more systems $\Sigma_i \circ \mathcal{C}_i$, $i = 1, \dots, n$, each one with Lyapunov function W_i, $i = 1, \dots, n$, a simple combination $\sum_{i=1}^{r} W_i \theta_i$, $\theta_i > 0$, may be not satisfactory for being a candidate Lyapunov function for the system interconnection Σ_i, $i = 1, \dots, n$ (see [7]). However, these interconnected systems are not necessarily triangular and, thus, even the results for the design of composite Lyapunov functions given in [7], [13] or [12] cannot be applied.

Example 6.1. Consider the system Σ:

$$\dot{z}_1 = -\frac{2z_1}{\sqrt{1+z_1^2}} + z_2^2 \sin(z_1), \quad \dot{z}_2 = -\frac{2z_2}{\sqrt{1+z_2^2}} + \varepsilon_2 \frac{z_1}{\sqrt{1+z_1^2}} \cos(z_1 z_2) \tag{6.1}$$

with $\varepsilon_2 \in (0, 1/2)$. If $W_j(z_j) = (1+z_j^2)^{1/2} - 1$, $j = 1, 2$, then after some computations

$$\begin{aligned} &\dot{W}_j \le -\varphi_j(z_j) z_j^2 + \gamma_j^{z_i}(z_i) z_i^2, \ j \neq i, \\ &\varphi_1(z_1) = \frac{1}{(1+z_1^2)}, \quad \gamma_1^{z_2}(z_2) = z_2^2, \quad \varphi_2(z_2) = \frac{1}{(1+z_2^2)}, \quad \gamma_2^{z_1}(z_1) = \frac{\varepsilon_2}{(1+z_1^2)}. \end{aligned} \tag{6.2}$$

Note that neither $\theta_1 W_1(z_1) + \theta_2 W_2(z_2)$, $\theta_1, \theta_2 > 0$, is a Lyapunov function for (6.1) nor the results of [7] can be applied to derive from $W_1(z_1)$ and $W_2(z_2)$ a composite Lyapunov function for (6.1).

Thus, we look for a "filtered" combination $\sum_{i=1}^{r} W_i \theta_i$, where θ_i are dynamical parameters which may depend on the system trajectories. The parameter $\theta_i(t)$, $i = 1, \dots, n$, should be positive along the system trajectories for $\sum_{i=1}^{r} W_i(t)\theta_i(t)$ being positive as well, and its time derivative should be nonpositive along the system trajectories for dominating the cross terms $\gamma_j^{z_i}(z_i) z_i^2$ in $\dot{W}_j$ and $\gamma_i^{z_j}(z_j) z_j^2$ in $\dot{W}_i$, $i, j = 1, \dots, n$, $i \neq j$.

Example 6.1 *(continued).* Let $c_2 := [1/(1-2\varepsilon_2)^{1/2} - 1]$ and assume that ε_2 is sufficiently small so that $(1+c_2)^2[(1+c_2)^2 - 1] < 1/2$. Moreover, let

$$\begin{aligned} &c_1 := 1/(1 - 2(1+c_2)^2[(1+c_2)^2 - 1])^{1/2} - 1, \\ &\kappa_1 := \tau_1(c_2), \quad \tau_1(s) := (1+s)^2[(1+s)^2 - 1], \quad \kappa_2 := \varepsilon_2. \end{aligned} \tag{6.3}$$

It can be easily seen that the trajectories of (6.1) are bounded, enter in finite time the set $\mathcal{R} = \{(z_1, z_2) : W_j(z_j) \le c_j, j = 1, 2\}$, and remain thereinafter. Boundedness of the trajectories of (6.1) follows from the fact that for each $i \neq j$ while $z_j(t)$ approaches the set $\{W_j(z_j) \le c_j\}$, $z_i(t)$ has infinite escape time. If, in addition, ε_2 is sufficiently small so that $\kappa_1 \kappa_2 < 1$, the "filtered" linear combination of W_1 and W_2,

$$\widetilde{W}(z_1, z_2, \theta) = \theta(W_1(z_1) + d_2 W_2(z_2)), \ d_2 \in (\kappa_1, 1/\kappa_2), \tag{6.4}$$

$$\dot{\theta} = -[\theta / \min\{c_1, d_2 c_2\}] \max\{\tau_1(W_2(z_2)) - \tau_1(c_2), 0\} \varphi_2(z_2) z_2^2, \ \theta(0) = 1, \tag{6.5}$$

satisfies

$$\dot{\widetilde{W}} \le -\theta[(1 - \kappa_2 d_2)\varphi_1 z_1^2 + (d_2 - \kappa_1)\varphi_2 z_2^2] \tag{6.6}$$

along the trajectories of (6.1) and (6.5). Also, since the trajectories of (6.1) are captured in finite time by $\mathcal{R}$ and $\tau_1(s)$ is nondecreasing for all $s \ge 0$, for each trajectory $(z_1(t), z_2(t))$ of (6.1)

$$\theta(t) = \exp\left\{ -[1/\min\{c_1, d_2 c_2\}] \int_0^t \max\{\tau_1(W_2(z_2(s))) - \tau_1(c_2), 0\} \varphi_2(z_2(s)) z_2^2(s) ds \right\}$$

for all $t \geq 0$, and, moreover, there exists $T > 0$ such that $\theta(t) = \theta(T)$ for all $t \geq T$. Thus, $\theta(t)$ is bounded and positive for all $t \geq 0$ and definitely approaches a constant positive value. For this reason, the filtered Lyapunov function $\widetilde{W}$ is a linear combination of W_1 and W_2 locally around the origin. Moreover, since $\theta(t)$ is positive for all $t \geq 0$ and $1 - \kappa_2 d_2 > 0$ and $d_2 - \kappa_1 > 0$ by (6.5), it follows from (6.6) that $\widetilde{W}(z_1(t), z_2(t), \theta(t))$ decreases along the trajectories of (6.1) and (6.5).

Although once the trajectories are trapped closely to the origin the stability analysis can be performed locally with a quadratic Lyapunov function, filtered Lyapunov functions unify the local and global dynamic behavior of interconnected systems and can be used in a Lyapunov based controller design to stabilize the interconnection itself (if not stable) or in small gain theorems for the stability analysis. The above discussion motivates the following definition. Let Σ be any given system with controller $\mathcal{C}$ and let z and σ be their corresponding (n-dimensional) states, with components z_j and, respectively, σ_j, $j = 1, \ldots, n$, and let χ be the exogenous inputs of Σ, with component χ_j, $j \in J$.

Definition 6.1 (filtered Lyapunov functions). *We say that* $\widetilde{W} : \mathbb{R}^n \times \mathbb{R}^n \times \Theta \to \mathbb{R}^{\geq}$, $(z, \sigma, \theta) \mapsto \widetilde{W}(z, \sigma, \theta) = \theta W(z, \sigma)$, $\Theta \subseteq \mathbb{R}$, *is a smooth filtered Lyapunov function for* $\Sigma \circ \mathcal{C}$ *with stability margins* $\theta\varphi_{sj}$, $\theta\varphi_{mj}$, $j = 1, \ldots, n$, *and incremental rates* $\theta\gamma^{\chi_j}$, $j \in J$, *if*

(i) $W(z, \sigma)$ *is smooth, proper and positive definite,*

(ii) $\dot{\theta}(t) \leq 0$ *and* $\theta(t) > 0$ *along the trajectories of* $\Sigma \circ \mathcal{C}$, *and*

(iii) *along the trajectories of* $\Sigma \circ \mathcal{C}$

$$\dot{\widetilde{W}} \leq -\theta \left\{ \sum_{j=1}^{n} [\varphi_{sj}(z) z_j^2 + \varphi_{mj}(z - \sigma)(z_j - \sigma_j)^2] + \sum_{j \in J} \gamma^{\chi_j}(z, \iota, \varsigma) \chi_j^2 \right\} \tag{6.7}$$

for some continuous positive (definite) functions $\varphi_{sj}, \varphi_{mj} : \mathbb{R}^n \to \mathbb{R}^{\geq}$ *and continuous functions* $\gamma^{\chi_j} : \mathbb{R}^n \times \mathcal{I} \times \mathcal{Z} \to \mathbb{R}^{\geq}$.

Clearly, by Theorem 5.2 W_j is a (filtered) Lyapunov function for $\Sigma_j \circ \mathcal{C}_j$ with stability margins (5.5) and incremental rates as in (5.6). If $\theta = 1$ and γ^{χ_j} are functions only of χ, then our filtered Lyapunov functions are *iISS Lyapunov functions* [1]. A smooth, proper, and positive definite function $\widetilde{W} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{\geq}$ for which there exist positive definite $\alpha : \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$ and $\beta \in \mathcal{K}_\infty$ such that $\dot{\widetilde{W}} \leq -\alpha(\widetilde{W}) + \beta(\|\chi\|)$ along the trajectories of $\Sigma \circ \mathcal{C}$ is said to be an *iISS Lyapunov function* for $\Sigma \circ \mathcal{C}$.

As already remarked, (ii) and (6.7) imply that $\widetilde{W}(z(t), \sigma(t), \theta(t))$ decreases along the trajectories of $\Sigma \circ \mathcal{C}$. However, this is not enough for inferring stability properties on the system trajectories of $\Sigma \circ \mathcal{C}$, since $\theta(t)$ varies itself along the trajectories of $\Sigma \circ \mathcal{C}$. The following results clarify the stability issues related to the existence of a filtered Lyapunov function and will be used to prove Theorem 1.1. Let χ be the exogenous inputs of Σ. We will say that $\Sigma \circ \mathcal{C}$ is 0-*GAS* if there exists $\beta \in \mathcal{KL}$ such that $\|z(t), \sigma(t)\| \leq \beta(\|z(0), \sigma(0)\|, t)$ for all $t \geq 0$ along the trajectories of $\Sigma \circ \mathcal{C}$ with $\chi = 0$, i.e., the origin of $\Sigma \circ \mathcal{C}$ is globally asymptotically stable. The following stability result can be proved as Theorem 4.1 of [9].

Lemma 6.2 (0-GAS). *Assume the existence of* $\alpha_1 \in \mathcal{K}$ *and* $\alpha_2, \alpha_3 \in \mathcal{K}_\infty$ *for which a smooth filtered Lyapunov function* $\widetilde{W} : \mathbb{R}^n \times \mathbb{R}^n \times \Theta \to \mathbb{R}^{\geq}$ *for* $\Sigma \circ \mathcal{C}$ *satisfies*

$$\dot{\widetilde{W}}(z(t), \sigma(t), \theta(t)) \leq -\alpha_1(\widetilde{W}(z(t), \sigma(t), \theta(t))), \tag{6.8}$$

$$\widetilde{W}(z(0), \sigma(0), \theta(0)) \leq \alpha_2(\|z(0), \sigma(0)\|), \tag{6.9}$$

$$\widetilde{W}(z(t),\sigma(t),\theta(t)) \geq \alpha_3(\|z(t),\sigma(t)\|) \tag{6.10}$$

for all $t \geq 0$ along the trajectories of $\Sigma \circ \mathcal{C}$. Then $\Sigma \circ \mathcal{C}$ is 0-GAS.

The iISS properties are then inferred through the following lemma. Let $\chi \in \mathcal{L}_2^r(\mathbb{R}^{\geq})$ be the exogenous inputs of Σ. We will say that $\Sigma \circ \mathcal{C}$ is *UBEBS* if there exist $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ and $\alpha_3 > 0$ such that $\|z(t),\sigma(t)\| \leq \alpha_1(\|z(0),\sigma(0)\|) + \alpha_2(\|\chi\|_2) + \alpha_3$ for all $\chi \in \mathcal{L}_2^r(\mathbb{R}^{\geq})$ and $t \geq 0$ along the trajectories of $\Sigma \circ \mathcal{C}$ [1].

LEMMA 6.3 (iISS [1]). *Assume that $\Sigma \circ \mathcal{C}$ is 0-GAS and UBEBS. Then $\Sigma \circ \mathcal{C}$ is iISS.*

In the next section we show how to construct filtered Lyapunov functions for complex systems, resulting from interconnecting dynamics like (4.1), using the filtered Lyapunov functions of the simpler dynamics in which these systems can be decomposed.

7. Filtered Lyapunov functions for interconnected systems. Let Σ_j, $j = 1, 2$, be given systems with controller $\mathcal{C}_j$ and smooth filtered Lyapunov function $\widetilde{W}_j(z_j,\sigma_j,\theta_j) = \theta_j W_j(z_j,\sigma_j)$ with stability margins $\theta_j\varphi_{jls}$, $\theta_j\varphi_{jlm}$, $l = 1,\dots,n_j$, and incremental rates $\theta_j\gamma_j^{\chi_{jl}}$, $l \in J_j := \{1,\dots,r_j\}$ (n_j is the dimension of the state vector z_j). In other words, along the trajectories of $\Sigma_j \circ \mathcal{C}_j$

$$\dot{\widetilde{W}}_j \leq -\theta_j \left\{ \sum_{l=1}^{n_j} [\varphi_{jls}(z_j) z_{jl}^2 + \varphi_{jlm}(e_j) e_{jl}^2] + \sum_{l \in J_j} \gamma_j^{\chi_{jl}}(z_j,\iota_j,\varsigma_j)\chi_{jl}^2 \right\}, \tag{7.1}$$

where $e_j := z_j - \sigma_j$ and e_{jl} and z_{jl} are the lth elements of e_j and z_j, respectively. We also assume that z_j, e_j are exogenous inputs of Σ_i and that μ_j, σ_j are exogenous measurements of Σ_i for $j \neq i$. Thus, z_j and e_j are elements of χ_i, and μ_j and σ_j are elements of ν_i for $j \neq i$. Moreover, wherever possible we will denote $\widetilde{W}_j(z_j(t),\sigma_j(t),\theta_j(t))$, $W_j(z_j(t),\sigma_j(t))$, φ_{jls} $(z_j(t))$, and $\varphi_{jlm}(e_j(t))$ simply by $\widetilde{W}_j(t)$, $W_j(t)$, $\varphi_{jls}(t)$, and $\varphi_{jlm}(t)$. We will also omit the arguments $z_j \in \mathbb{R}^{n_j}, \varsigma_j \in \mathcal{Z}_j, \iota_j \in \mathcal{I}_j$ of the functions whenever there is no ambiguity. In this section, we study the problem of finding a filtered Lyapunov function for the interconnection of $\Sigma_j \circ \mathcal{C}_j$, $j = 1, 2$. We start with the following definition.

DEFINITION 7.1 (local saturation). *Let $j, i = 1, 2$ with $i \neq j$. We say that φ_{jls} (resp., φ_{jlm}) locally saturates $\gamma_i^{z_{jl}}$ (resp., $\gamma_i^{e_{jl}}$) with levels (c_j, κ_i) if there exist $c_j > 0$ and a continuous nondecreasing function $\tau_i : \mathbb{R}^{\geq} \to \mathbb{R}^+$ such that $\tau_i(c_j) \leq \kappa_i$ and $\gamma_i^{z_{jl}} \leq \tau_i(W_j)\varphi_{jls}$ (resp., $\gamma_i^{e_{jl}} \leq \tau_i(W_j)\varphi_{jlm}$) for all ι_l, ς_l, $l = 1, 2$.*

Thus, "local saturation" in our context means that the ratio between $\gamma_i^{z_{jl}}$ and φ_{jls}, $j \neq i$, is, for $W_j \leq c_j$, locally bounded by κ_i and globally bounded by $\tau_i(W_j)$. If the function $\tau_i : \mathbb{R}^{\geq} \to \mathbb{R}^+$ can be taken constant, we have the following stronger property.

DEFINITION 7.2 (saturation). *Let $j, i = 1, 2$ with $i \neq j$. We say that φ_{jls} (resp., φ_{jlm}) saturates $\gamma_i^{z_{jl}}$ (resp., $\gamma_i^{e_{jl}}$) with level κ_i if there exist $\kappa_i > 0$ such that $\gamma_i^{z_{jl}} \leq \kappa_i\varphi_{jls}$ (resp., $\gamma_i^{e_{jl}} \leq \kappa_i\varphi_{jlm}$) for all ι_l, ς_l, $l = 1, 2$.*

In Example 6.1, $\gamma_2^{z_1}$ is saturated by φ_1 with levels (c_1, κ_2) and $\gamma_1^{z_2}$ is locally saturated by φ_2 with levels (c_2, κ_1), where $\tau_2(s) = \kappa_2$ and κ_1, κ_2, and $\tau_1(s)$ are as in (6.3).

The main result of this section points out the construction of a filtered Lyapunov function for the interconnection of $\Sigma_j \circ \mathcal{C}_j$, $j = 1, 2$, when each $\gamma_i^{z_{jl}}$ (resp., $\gamma_i^{e_{jl}}$) is locally saturated by φ_{jls} (resp., φ_{jlm}), $j \neq i$, with levels (c_j, κ_i) satisfying the small gain condition $\kappa_1\kappa_2 < 1$ and under the assumption that the trajectories of $\Sigma_j \circ \mathcal{C}_j$,

$j = 1, 2$, are bounded, enter in finite time a set in which $W_j \le c_j$, $j = 1, 2$, and remain thereinafter. This result will be applied in section 8 for constructing a filtered Lyapunov function for the interconnection of $\Sigma_j \circ \mathcal{C}_j$, $j = 1, 2$. It gives a filtered Lyapunov function W as a "filtered" linear combination of W_j, $j = 1, 2$. The design of composite Lyapunov functions has been widely studied in [8] for general systems and [12], [7], and [13] in the case of triangular systems $\dot z = f(z) + \psi(z, \xi)$, $\dot\xi = a(\xi)$. However, while the result of [8] does not lead to a constructive procedure, the constructive procedures of [12], [7], and [13] cannot be applied here since, as already remarked, the control design is not performed in our case by using feedforwarding [13] and rather relies on a backstepping-like strategy. As a consequence of this, in backstepping coordinates $z_j := z_j - \widetilde x_j$, $j = 1, \dots, n$, from (1.1)–(1.4) we get interconnected systems of the more general form $\dot z = f(z) + \psi(z, \xi)$, $\dot\xi = a(\xi) + \zeta(z, \xi)$. As a particular case, if $\zeta(z, \xi) \equiv 0$, we obtain the same class of interconnected systems considered in [7] and [13], and our filtered Lyapunov functions are alternative to the composite Lyapunov functions proposed in [7] and [13].

We have already seen in Lemma 6.2 that from the point of view of the asymptotic stability properties of the system trajectories it is important that $\widetilde W(t)$ be lower bounded uniformly with respect to $\theta(t)$ and that $\dot{\widetilde W}(t)$ be bounded by a definite negative function uniformly with respect to $\theta(t)$. In the case of the interconnection of two systems as in Example 6.1, $\widetilde W(t)$ is the filtered combination of two Lyapunov functions and has uniform (with respect to θ) upper and lower bounds since the system trajectories are bounded for all times and enter some "invariant" set, where a small gain condition is met. To formalize these "capture" and "invariance" properties we introduce the notions of traps and recurrence. A set $\mathcal{W}$ is a *trap* relative to a system Σ if the trajectories of Σ are captured by $\mathcal{W}$ for all times, while a system Σ is *recurrent* relative to a set $\mathcal{W}$ if each trajectory of Σ ensuing from outside $\mathcal{W}$ hits $\mathcal{W}$ at some time T.

DEFINITION 7.3 (recurrence and traps). *Let Σ be a given system with state $z \in \mathbb{R}^n$ and exogenous inputs $\chi \in \mathcal{X} \subseteq \mathbb{R}^r$. We say that $\mathcal{R} \subseteq \mathbb{R}^n$ is a trap for $\mathcal{W} \subseteq \mathcal{R}$ if for each $z_0 \in \mathcal{W}$, exogenous input $\chi(t)$, and $T \ge 0$ the trajectory $z(t)$ of Σ ensuing from $z(T) = z_0$ satisfies $z(t) \in \mathcal{R}$ for all $t \ge T$. We say that a system Σ is recurrent relative to a closed set $\mathcal{W} \subseteq \mathbb{R}^n$ if for each $z_0 \in \mathbb{R}^n \backslash \mathcal{W}$ and exogenous input $\chi(t)$ the trajectory $z(t)$ of Σ ensuing from $z(0) = z_0$ is defined for all $t \ge 0$ and there exists $T > 0$ (recurrence time) such that $z(T) \in \mathcal{W}$. If, in addition, $\mathcal{R} \subseteq \mathbb{R}^n$ is a trap for $\mathcal{W} = \mathcal{R}$, then we say that Σ is recurrent relative to the trap $\mathcal{R}$.*

We are ready to state and prove the main result of this section. For any $c_j, d_j > 0$ and continuous nondecreasing functions $\tau_j : \mathbb{R}^{\ge} \to \mathbb{R}^+$ and $I_j \subseteq \{1, \dots, n_j\}$, $j = 1, 2$, let

$$
\begin{aligned}
a(z_1, z_2, e_1, e_2) &:= d_1 \max\{\tau_1(W_2(z_2, \sigma_2)) - \tau_1(c_2), 0\} \sum_{l \in I_2} (\varphi_{2ls}(z_2) z_{2l}^2 + \varphi_{2lm}(e_2) e_{2l}^2) \\
&+ d_2 \max\{\tau_2(W_1(z_1, \sigma_1)) - \tau_2(c_2), 0\} \sum_{r \in I_1} (\varphi_{1rs}(z_1) z_{1r}^2 + \varphi_{1rm}(e_1) e_{1r}^2). \qquad (7.2)
\end{aligned}
$$

In what follows we will denote $a(z_1(t), z_2(t), e_1(t), e_2(t))$ simply by $a(t)$.

THEOREM 7.4. *Assume that $\Sigma_j \circ \mathcal{C}_j$, $j = 1, 2$, has smooth (filtered) Lyapunov function $\widetilde W_j$ with stability margins $\theta_j \varphi_{jls}$, $\theta_j \varphi_{jlm}$, $l = 1, \dots, n_j$, and incremental rates $\theta_j \gamma_j^{\chi_{jl}}$, $l \in J_j$, and let $I_j \subseteq \{1, \dots, n_j\}$, $j = 1, 2$. Assume also that*

(i) *φ_{1rs} and φ_{1rm}, $r \in I_1$, locally saturate $\gamma_2^{z_{1r}}$ and $\gamma_2^{e_{1r}}$, with levels (c_1, κ_2),*

(ii) φ_{2ls} *and* φ_{2lm}, $l \in I_2$, *locally saturate* $\gamma_1^{z_{2l}}$ *and* $\gamma_1^{e_{2l}}$, *with levels* (c_2, κ_1),
(iii) $\kappa_1\kappa_2 < 1$,
(iv) *for each* $j = 1, 2$, $\Sigma_j \circ \mathcal{C}_j$ *is recurrent relative to the trap* $\mathcal{R}_j = \{(z_j, \sigma_j) : W_j(z_j, \sigma_j) \le c_j\}$.
There exist $d_1 > 0$ *and* $d_2 \in (d_1\kappa_1, d_1/\kappa_2)$ *such that* $\widetilde{W} = \theta[d_1W_1 + d_2W_2]$, *with* $\theta = \theta_0\theta_1\theta_2$ *and*

$$\dot\theta_0 = -[a(z_1, z_2, e_1, e_2)/\min_j\{d_jc_j\}]\theta_0, \theta_0(0) > 0, \tag{7.3}$$

is a smooth filtered Lyapunov function for the interconnection Σ *of* $\Sigma_j \circ \mathcal{C}_j$, $j = 1, 2$, *with stability margins*

$$\begin{aligned} &\theta\delta_{jl}\varphi_{jls}, \quad \theta\delta_{jl}\varphi_{jlm}, \quad l = 1, \dots, n_j,\ j = 1, 2, \\ &\delta_{1r} = d_1 - \kappa_2 d_2,\ r \in I_1, \quad \delta_{2l} = d_2 - \kappa_1 d_1,\ l \in I_2, \quad \delta_{jl} = d_j \text{ otherwise} \end{aligned} \tag{7.4}$$

and incremental rates

$$\theta d_j\gamma_j^{\chi_{jl}},\ \chi_{jl} \notin H := \{z_{1r}, z_{2l}, e_{1r}, e_{2l} : r \in I_1, l \in I_2\};$$

i.e.,

$$\dot{\widetilde{W}} \le \theta \sum_{j=1,2} \left\{ -\sum_{j=1}^{n_j} \delta_{jl}[\varphi_{jls}z_{jl}^2 + \varphi_{jlm}e_{jl}^2] + \sum_{\chi_{jl}\notin H} d_j\gamma_j^{\chi_{jl}}\chi_{jl}^2 \right\} \tag{7.5}$$

along the trajectories of Σ *and* (7.3). *Moreover, the interconnection* Σ *is recurrent relative to the trap* $\mathcal{R} = \{(z_1, z_2, \sigma_1, \sigma_2) : d_1W_1 + d_2W_2 \le d_1c_1 + d_2c_2\}$, *and along each trajectory of* Σ *and* (7.3) *there exists* $T, \overline\theta > 0$ *such that* $\theta_0(t) = \overline\theta$ *for all* $t \ge T$. *In particular,*

$$\begin{aligned} \theta_0(0) = \exp\left(\int_0^\infty a(s)ds\right) \\ \Rightarrow a(t) \ge 0, \quad \theta_0(t) \ge 1\ \forall t \ge 0, \quad a(t) = 0, \quad \theta_0(t) = 1\ \forall t \ge T. \end{aligned} \tag{7.6}$$

Proof. We will prove the theorem with $\theta_0(0) = \exp(\int_0^\infty a(s)ds)$ in (7.3). By (iii) it is possible to select $d_1 > 0$ and $d_2 \in (d_1\kappa_1, d_1/\kappa_2)$ in such a way that

$$\delta_{1r}, \delta_{2l} > 0, \quad r \in I_1, \quad l \in I_2. \tag{7.7}$$

Since $W_j(z_j, \sigma_j)$, $j = 1, 2$, is smooth, proper, and positive definite, then also $W = \sum_{j=1}^2 d_jW_j$ is smooth, proper, and positive definite. Let $\tau_j : \mathbb{R}^{\ge} \to \mathbb{R}^+$, $j = 1, 2$, be continuous nondecreasing functions such that

$$\begin{aligned} \gamma_2^{z_{1r}} \le \tau_2(W_1)\varphi_{1rs}, \quad \gamma_2^{e_{1r}} \le \tau_2(W_1)\varphi_{1rm}, \\ \gamma_1^{z_{2r}} \le \tau_1(W_2)\varphi_{2rs}, \quad \gamma_1^{e_{2r}} \le \tau_1(W_2)\varphi_{2rm}, \end{aligned} \tag{7.8}$$

which indeed exist by (i) and (ii). By (iv) each trajectory $(z_j(t), \sigma_j(t))$ of $\Sigma_j \circ \mathcal{C}_j$, $j = 1, 2$, is defined for all $t \ge 0$, and for each such trajectory there exists $T_j > 0$ such that $W_j(z_j(t), \sigma_j(t)) \le c_j$ for all $t \ge T_j$. This, together with $\tau_j(s)$, $j = 1, 2$, being nondecreasing and positive for all $s \ge 0$, implies that

$$a(t) \ge 0\ \forall t \ge 0, \quad a(t) = 0\ \forall t \ge T := \max\{T_1, T_2\} \tag{7.9}$$

and, thus, $0 \le \int_0^\infty a(s)ds = \int_0^T a(s)ds < \infty$. This gives from (7.3), together with $\theta_0(0) = \exp(\int_0^\infty a(s)ds)$,

$$\theta_0(t) = \theta_0(0)\exp\left\{-\int_0^t a(s)ds\right\} = \exp\left\{-\int_{\min\{t,T\}}^T a(s)ds\right\}, \quad t \ge 0. \tag{7.10}$$

Thus, $\theta_0(t)$ is bounded and positive for all $t \ge 0$ along each trajectory of Σ and (7.7) holds true. Also, by (iv) the interconnection Σ is recurrent relative to the trap $\mathcal{R}$.

We are left with proving (7.5). Taking into account that for all $t \ge 0$ along the trajectories of $\Sigma_j \circ \mathcal{C}_j$, we have $\theta_j(t) > 0, \dot\theta_j(t) \le 0$, $j = 1, 2$, since W_j, $j = 1, 2$, is a filtered Lyapunov function for $\Sigma_j \circ \mathcal{C}_j$, and $\theta_0(t) > 0, \dot\theta_0(t) \le 0$ by (7.3), (7.9), and (7.10), we get from (7.1), (7.2), and (7.9) that

(7.11)

$$\begin{aligned}\dot{\widetilde{W}} &\le \theta\left\{\sum_{j=1,2}\left\{-\sum_{j=1}^{n_j} d_j[\varphi_{jls}z_{jl}^2 + \varphi_{jlm}e_{jl}^2] + \sum_{l\in J_j} d_j\gamma_j^{\chi_{jl}}\chi_{jl}^2\right\} + \frac{\dot\theta_0}{\theta_0}\sum_{j=1}^2 d_jW_j\right\}\\ &\le \theta\left\{\sum_{j=1,2}\left\{-\sum_{j=1}^{n_j}\delta_{jl}[\varphi_{jls}z_{jl}^2 + \varphi_{jlm}e_{jl}^2] + \sum_{\chi_{jl}\notin H} d_j\gamma_j^{\chi_{jl}}\chi_{jl}^2\right\} + a + \frac{\dot\theta_0}{\theta_0}\sum_{j=1}^2 d_jW_j\right\}\end{aligned}$$

along the trajectories of Σ, with δ_{jl}, δ_{1r} as in (7.4) and $\delta_{jl}, \delta_{1r} > 0$, $r \in I_1, l \in I_2$, by (7.7). Since $\sum_{j=1}^2 d_jW_j \ge \min_j\{d_jc_j\}$ whenever either $W_1 \ge c_2$ or $W_2 \ge c_1$ and, moreover, $a \equiv 0$ when $W_i \le c_i$, $i = 1, 2$, from (7.3) and (7.11) we obtain (7.5). □

8. Proof of Theorem 1.1. The proof goes as follows. For each system Σ_j, $j = 1, \dots, n$, in (4.1) we apply Theorem 5.2 to obtain a measurement feedback controller $\mathcal{C}_j$ and take the interconnection $\mathcal{C}$ of $\mathcal{C}_j$, $j = 1, \dots, n$, as measurement feedback controller for the interconnection Σ of Σ_j, $j = 1, \dots, n$, i.e., the controller (1.4) (see part A below). Theorem 1.1 follows from Lemma 6.3, once we prove that (1.1)–(1.4) is 0-GAS and UBEBS (see part B). First, we show that (1.1)–(1.4) is UBEBS. Instrumental to this and since $w \in \mathcal{L}_2^r(\mathbb{R}^{\ge})$, we prove that $\Sigma_i \circ \mathcal{C}_i$, $i = j, \dots, n$, for each $j = 1, \dots, n$ is recurrent relative to some compact trap $\mathcal{T}_j(\|w\|_2)$ (see Lemma 8.1). Finally, we show that (1.1)–(1.4) is 0-GAS (see part C). Instrumental to this, we use the traps $\mathcal{T}_j(0)$, $j = 1, \dots, n$, to define a filtered Lyapunov function $\widetilde{W}$ for (1.1)–(1.4), with $w = 0$, by repeated applications of Theorem 7.4 (see Lemma 8.2) and then we prove that $\widetilde{W}$ satisfies Lemma 6.2.

A. *Controller definition for* (1.1). Set $e_j := z_j - \sigma_j$, $j = 1, \dots, n$, and let $G(s)$ be defined as in Theorem 1.1.

Step 1. Application of Theorem 5.2 to Σ_1 in (4.1) gives $\Delta_1, h_{1s}, h_{1m} \in (0, 1]$ and $R'_{11} \ge 1$ such that for all $R_{11} \ge R'_{11}$ the controller $\mathcal{C}_1$ in (5.3), with $j = 1$, has levels $(\Delta_1, \Delta_1/h_{1s})$ and

$$W_1(z_1, \sigma_1) = \sqrt{1 + z_1^2} + \sqrt{1 + e_1^2} - 2 \tag{8.1}$$

is a smooth filtered Lyapunov function for $\Sigma_1 \circ \mathcal{C}_1$ with stability margins

$$\varphi_{1s} \sim 1/[R_{11}(1 + z_1^2)], \quad \varphi_{1m} \sim 1/[R_{11}(1 + e_1^2)] \tag{8.2}$$

and, by (5.6), smooth incremental rates

$$\gamma_1^{z_l} \sim R_{11}[\gamma_{1s}^{z_l} + \gamma_{1m}^{z_l}], \quad l = 2, \dots, n,$$
$$\gamma_1^{w} \sim R_{11}[\gamma_{1s}^{w} + \gamma_{1m}^{w}], \quad \gamma_1^{\widetilde{x}_l} \sim R_{11}[\gamma_{1s}^{\widetilde{x}_l} + \gamma_{1m}^{\widetilde{x}_l}], \quad l = 3, \dots, n+1. \tag{8.3}$$

In other words,

$$\dot{W}_1 \leq -\varphi_{1s} z_1^2 - \varphi_{1m} e_1^2 + \sum_{l=2}^{n} [\gamma_1^{z_l} z_l^2 + \gamma_1^{\widetilde{x}_{l+1}} \widetilde{x}_{l+1}^2] + \gamma_1^{w} \|w\|^2. \tag{8.4}$$

Moreover, on account of

$$G\left(\left|\sum_{l=1}^{2} s_l\right|\right) \leq \sum_{l=1}^{2} G(|s_l|) \; \forall s_l, \; l = 1, 2, \tag{8.5}$$

and by Lemma A.2 with $j = 1$, we obtain that $\widetilde{x}_2$ and $\dot{\widetilde{x}}_2$ satisfy

$$\widetilde{x}_2^2 \leq \frac{1}{R_{11}^2} \left[G^2(z_1) + G^2(e_1)\right], \tag{8.6}$$
$$\dot{\widetilde{x}}_2^2 \preceq \frac{1}{R_{11}^4} \left[\sum_{l=1}^{2} (G^2(z_l) + G^2(e_l)) + \sum_{l=3}^{n} z_l^2 + \sum_{l=3}^{n+1} \widetilde{x}_l^2 + \|w\|^2\right].$$

Step 2. Given $j = 2, \dots, n$, let

$$\varepsilon^{n-1} = \frac{R_{11}}{R'_{11}}, \quad R_{l1} = R'_{l1} \varepsilon^{n-l}, \quad l = 1, \dots, j-1, \tag{8.7}$$

and assume that

$$\widetilde{x}_l^2 \leq \frac{1}{R_{l-1,1}^2} [G^2(z_{l-1}) + G^2(e_{l-1})], \quad l = 2, \dots, j, \tag{8.8}$$

$$|\dot{\widetilde{x}}_j|^2 \preceq \frac{1}{R_{j-1,1}^4} \left[\sum_{l=j-1}^{j} (G^2(z_l) + G^2(e_l)) + \sum_{l=j+1}^{n} z_l^2 + \sum_{l=j+1}^{n+1} \widetilde{x}_l^2 + \|w\|^2\right]. \tag{8.9}$$

By application of Theorem 5.2 to Σ_j in (4.1) with the additional constraints (8.8)–(8.9) we get $\Delta_j, h_{js}, h_{jm} \in (0, 1]$ and $R'_{j1} \geq 1$ such that for all $R_{j1} \geq R'_{j1}$ the controller $\mathcal{C}_j$ (5.3) has levels $(\Delta_j, \Delta_j / h_{js})$ and

$$W_j(z_j, \sigma_j) = \sqrt{1 + z_j^2} + \sqrt{1 + e_j^2} - 2 \tag{8.10}$$

is a smooth filtered Lyapunov function for $\Sigma_j \circ \mathcal{C}_j$ with stability margins

$$\varphi_{js} \sim 1/[R_{j1}(1 + z_j^2)], \quad \varphi_{jm} \sim 1/[R_{j1}(1 + e_j^2)] \tag{8.11}$$

and, by (5.6) and (8.9), smooth incremental rates

$$\gamma_j^{z_l} \sim \frac{R_{j1}}{R_{j-1,1}^4 (1 + z_l^2)}, \quad \gamma_j^{e_l} \sim \frac{R_{j1}}{R_{j-1,1}^4 (1 + e_l^2)}, \quad l = j-1, j, \tag{8.12}$$
$$\gamma_j^{z_l} \sim R_{j1}[\gamma_{js}^{z_l} + \gamma_{jm}^{z_l} + 1], \quad l = j+1, \dots, n, \tag{8.13}$$
$$\gamma_j^{w} \sim R_{j1}[\gamma_{js}^{w} + \gamma_{jm}^{w} + 1], \quad \gamma_j^{\widetilde{x}_l} \sim R_{j1}[\gamma_{js}^{\widetilde{x}_l} + \gamma_{jm}^{\widetilde{x}_l} + 1], \quad l = j+2, \dots, n+1. \tag{8.14}$$

In other words,

$$\dot{W}_j \le -[\varphi_{js} - \gamma_j^{z_j}]z_j^2 - [\varphi_{jm} - \gamma_j^{e_j}]e_j^2 + \gamma_j^{z_{j-1}} z_{j-1}^2 + \gamma_j^{e_{j-1}} e_{j-1}^2 + \sum_{l=j+1}^{n} [\gamma_j^{z_l} z_l^2 + \gamma_j^{\widetilde{x}_{l+1}} \widetilde{x}_{l+1}^2] + \gamma_j^w \|w\|^2. \tag{8.15}$$

Set

$$R_{j1} = R'_{j1}\varepsilon^{n-j}. \tag{8.16}$$

By (8.7), (8.11), (8.12), and (8.16) we can select $\varepsilon_j \ge 1$ such that

$$\varphi_{js} - \gamma_j^{z_j} \ge \varphi_{js}/2, \quad \varphi_{jm} - \gamma_j^{e_j} \ge \varphi_{jm}/2 \tag{8.17}$$

for all $\varepsilon \ge \varepsilon_j$. Moreover, by (8.5) and Lemma A.2, $\widetilde{x}_{j+1}$ and $\dot{\widetilde{x}}_{j+1}$ satisfy

$$\widetilde{x}_l^2 \le \frac{1}{R_{l-1,1}^2}[G^2(z_{l-1}) + G^2(e_{l-1})], \quad l = 2, \dots, j+1, \tag{8.18}$$

$$|\dot{\widetilde{x}}_{j+1}|^2 \preceq \frac{1}{R_{j1}^4}\left[\sum_{l=j}^{j+1}(G^2(z_l) + G^2(e_l)) + \|w\|^2 + \sum_{l=j+2}^{n} z_l^2 + \sum_{l=j+2}^{n+1} \widetilde{x}_l^2\right]. \tag{8.19}$$

Using the constraints (8.6) and (8.18) we can express the terms $\widetilde{x}_{l+1}^2$ in (8.4) and (8.15) in terms of the exogenous inputs z_l and e_l, so that we can assume that the incremental rates are finally given for $j = 1$ by

$$\gamma_1^{z_l}, \gamma_1^{e_l} \sim R_{11}[\gamma_{1s}^{z_l} + \gamma_{1m}^{z_l} + \gamma_{1s}^{\widetilde{x}_{l+1}} + \gamma_{1m}^{\widetilde{x}_{l+1}}], \quad l = 2, \dots, n, \ \gamma_1^w \sim R_{11}[\gamma_{1s}^w + \gamma_{1m}^w] \tag{8.20}$$

and for $j = 2, \dots, n$ by

$$\gamma_j^{z_{j-1}} \sim R_{j1}/[R_{j-1,1}^4(1 + z_{j-1}^2)], \quad \gamma_j^{e_{j-1}} \sim R_{j1}/[R_{j-1,1}^4(1 + e_{j-1}^2)], \tag{8.21}$$

$$\gamma_j^{z_l}, \gamma_j^{e_l} \sim R_{j1}[\gamma_{js}^{z_l} + \gamma_{jm}^{z_l} + \gamma_{js}^{\widetilde{x}_{l+1}} + \gamma_{jm}^{\widetilde{x}_{l+1}} + 1], \quad l = j+1, \dots, n,$$

$$\gamma_j^w \sim R_{j1}[\gamma_{js}^w + \gamma_{jm}^w + 1]. \tag{8.22}$$

After performing n steps, we obtain the controller (1.4) as the interconnection of $\mathcal{C}_j$, $j = 1, \dots, n$.

B. (1.1)–(1.4) *is UBEBS.* The exogenous inputs of (1.1)–(1.4) are w. To prove that (1.1)–(1.4) is UBEBS, we need the following lemma, which is proved in the appendix. We will often use the notation $W_j(t)$ in place of $W_j(z_j(t), \sigma_j(t))$. Moreover, let $Z_j = (z_j \ \cdots \ z_n)^T$ and $S_j = (\sigma_j \ \cdots \ \sigma_n)^T$, with $Z_1 = Z$ and $S_1 = S$. Denote by $(\mathbb{R}^{\ge})^q$ the q-times vector space product $\mathbb{R}^{\ge} \times \cdots \times \mathbb{R}^{\ge}$.

LEMMA 8.1. *There exist continuous* $c_j : \mathbb{R}^+ \to \mathbb{R}^+$, $\varrho_j : (\mathbb{R}^{\ge})^{n-j+2}\mathbb{R}^{\ge} \to \mathbb{R}^{\ge}$, *nondecreasing with respect to each argument and* $\varrho_j(0, \dots, 0) = 0$, *and* $\varepsilon_j^* > 0$, $j = 1, \dots, n$, *such that for all* $\varepsilon \ge \varepsilon_j^*$, *for each* $w \in \mathcal{L}_2^r(\mathbb{R}^{\ge})$, *and* $j = 1, \dots, n$ *the interconnection* $\Sigma_i \circ \mathcal{C}_i$, $i = j, \dots, n$, *is recurrent relative to the trap* $\mathcal{T}_j(\|w\|_2) := \{(Z_j, S_j) : W_i(z_i, \sigma_i) \le \varrho_i(c_i(\varepsilon), \dots, c_n(\varepsilon), \|w\|_2), \ i = j, \dots, n\}$ *and*

$$c_1 < 1, \quad c_i \sim \prod_{l=2}^{i} \frac{R_{l1}^2}{R_{l-1,1}^4}, \quad i = 2, \dots, n. \tag{8.23}$$

Moreover, for all $t \geq 0$ along the trajectories of (1.1)–(1.4)

$$W_i(t) \leq \varrho_i(W_i(0), \dots, W_n(0), \|w\|_2), \quad i = j, \dots, n. \tag{8.24}$$

Since $W_j(z_j, \sigma_j)$, $j = 1, \dots, n$, is proper and positive definite,

$$\beta_{j1}(\|z_j, \sigma_j\|) \leq W_j(z_j, \sigma_j) \leq \beta_{j2}(\|z_j, \sigma_j\|), \quad j = 1, \dots, n, \tag{8.25}$$

for some $\beta_{j1}, \beta_{j2} \in \mathcal{K}_\infty$, $j = 1, \dots, n$, and for all z_j, σ_j. Moreover, for any continuous $\alpha : \mathbb{R}^{\geq} \times \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$ such that $\alpha(s, r)$ and $\alpha(r, s)$ are nondecreasing for each $r \geq 0$ and $\alpha(0, 0) = 0$,

$$\alpha(r, s) \leq \alpha(s, s) + \alpha(r, r) \quad \forall r, s \geq 0 \tag{8.26}$$

and $\alpha(s) := \alpha(s, s)$ is $\mathcal{K}$-class. Thus, from (8.24)–(8.26) and since $\beta_{j1}^{-1}(\varrho_j(s_j, \dots, s_{n+1}))$ is nondecreasing with respect to each argument s_i and $\beta_{j1}^{-1}(\varrho_j(0, \dots, 0)) = 0$, we have along the trajectories of (1.1)–(1.4)

$$\begin{aligned} \|z_j(t), \sigma_j(t)\| &\leq \beta_{j1}^{-1}(\varrho_j(W_j(0), \dots, W_n(0), \|w\|_2)) \\ &\leq \sum_{i=j}^{n} 2^{n+2-j} \beta_{j1}^{-1}(\varrho_j(W_i(0), \dots, W_i(0))) \\ &+ \sum_{i=j}^{n} 2^{n+2-j} \beta_{j1}^{-1}(\varrho_j(\|w\|_2, \dots, \|w\|_2)) \\ &\leq \alpha_{j1}(\|Z_j(0), S_j(0)\|) + \alpha_{j2}(\|w\|_2), \end{aligned} \tag{8.27}$$

where $\alpha_{j1}(s) := \sum_{i=j}^{n} 2^{n+2-j} \beta_{j1}^{-1}(\varrho_j(\beta_{i2}(s), \dots, \beta_{i2}(s)))$ and $\alpha_{j2}(s) := \sum_{i=j}^{n} 2^{n+2-j} \beta_{j1}^{-1}(\varrho_j(s, \dots, s))$ for $j = 1, \dots, n$. Let $\alpha_1(s) := \sum_{j=1}^{n} \alpha_{j1}(s) + s$, and $\alpha_2(s) := \sum_{j=1}^{n} \alpha_{j2}(s) + s$ and note that $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, since $\alpha_{j1}, \alpha_{j2} \in \mathcal{K}$. We obtain from (8.27) that

$$\|Z(t), S(t)\| \leq \alpha_1(\|Z(0), S(0)\|) + \alpha_2(\|w\|_2) \tag{8.28}$$

for all $t \geq 0$ along the trajectories (1.1)–(1.4). This proves that (1.1)–(1.4) is UBEBS.

C. (1.1)–(1.4) *is* 0*-GAS.* To prove that (1.1)–(1.4) is 0-GAS we construct a filtered Lyapunov function $\widetilde{W}$ for (1.1)–(1.4), with $w = 0$, and prove that it satisfies (6.8)–(6.10). To this aim, we need the following lemma, which is proved in the appendix. Let $Z_j := (z_j \ \cdots \ z_n)^T$ and $S_j = (\sigma_j \ \cdots \ \sigma_n)^T$ and set $Z = Z_1$ and $S = S_1$. Moreover, from now on we take $w = 0$.

Lemma 8.2. *Let $c_j : \mathbb{R}^+ \to \mathbb{R}^+$, $\varrho_j : (\mathbb{R}^{\geq})^{n-j+2} \to \mathbb{R}^{\geq}$, and $\varepsilon_j^* > 0$, $j = 1, \dots, n$, be as in Lemma* 8.1. *There exist $\varepsilon^{**} \geq \max_j \varepsilon_j^*$ and continuous and nondecreasing $\tau_j : \mathbb{R}^{\geq} \to [1, \infty)$, $j = 1, \dots, n$, such that*

$$\begin{aligned} \widetilde{W}(Z, S, \theta) &:= \theta W(Z, S), \\ W(Z, S) &:= W_1(z_1, \sigma_1) + \sum_{j=2}^{n} 2^{j-1} W_j(z_j, \sigma_j) \tau_1(c^{(2)}) \cdots \tau_{j-1}(c^{(j)}), \\ c^{(n)} &:= \overline{c}_n, \quad c^{(i)} := \overline{c}_i + 2\tau_i(c^{(i+1)}) c^{(i+1)}, \quad i = 2, \dots, n-1, \\ \overline{c}_i &:= \varrho_i(c_i, \dots, c_n, 0) \end{aligned} \tag{8.29}$$

is for all $\varepsilon \geq \varepsilon^{**}$ *a smooth filtered Lyapunov function for* (1.1)–(1.4) *and*

$$\dot{\widetilde{W}} \leq -\theta \sum_{i=1}^{n} [\widetilde{\varphi}_{is} z_i^2 + \widetilde{\varphi}_{im} e_i^2] \tag{8.30}$$

along the trajectories of (1.1)–(1.4)*, where*

$$\widetilde{\varphi}_{1l} \sim \varphi_{1l}, \quad \widetilde{\varphi}_{il} \sim \tau_{i-1}(c^{(i)}) \cdots \tau_{n-1}(c^{(n)}) \varphi_{il}, \quad i = 2, \ldots, n, \quad l = s, m. \tag{8.31}$$

Moreover, for each trajectory of (1.1)–(1.4) *there exists* $T > 0$ *such that*

$$\begin{aligned} &\theta(t) \geq 1 \ \forall t \geq 0; \\ &\theta(t) = 1 \ \forall t \geq T. \end{aligned} \tag{8.32}$$

First, we prove (6.8). Since

$$(1+s)^{1/2} - 1 \leq s \ \forall s \geq 0, \ \frac{r_1 + r_2}{1 + r_1 + r_2} \leq \frac{r_1}{1 + r_1} + \frac{r_2}{1 + r_2} \ \forall r_1, r_2 \geq 0, \tag{8.33}$$

it is easy to see by (8.2), (8.11), and (8.31) that

$$\tau_{i-1}(c^{(i)}) \cdots \tau_{n-1}(c^{(n)}) W_i(z_i, \sigma_i)/[1 + W_i(z_i, \sigma_i] \preceq \widetilde{\varphi}_{is}(z_i) z_i^2 + \widetilde{\varphi}_{im}(e_i) e_i^2 \tag{8.34}$$

for all z_i, σ_i and for each $i = 1, \ldots, n$. By the definition of $\widetilde{W}$ in (8.29), (8.34) and since $\tau_1(s), \ldots, \tau_{n-1}(s) \geq 1$ for all $s \geq 0$,

$$\theta \geq 1 \Rightarrow \widetilde{W}(Z, S, \theta)/[1 + \widetilde{W}(Z, S, \theta)] \preceq \theta \sum_{i=1}^{n} [\widetilde{\varphi}_{is}(z_i) z_i^2 + \widetilde{\varphi}_{im}(e_i) e_i^2]. \tag{8.35}$$

From (8.30), (8.32), with $j = 1, \ldots, n-1$, and (8.35) follows the existence of $\alpha_0 > 0$ such that $\dot{\widetilde{W}} \leq -\alpha_0 \widetilde{W}(Z, S, \theta)/[1 + \widetilde{W}(Z, S, \theta)]$ along the trajectories of (1.1)–(1.4). This proves (6.8) with $\alpha_1(s) = \alpha_0 s/[1 + s]$.

Next, we show (6.9)–(6.10). By continuity of the trajectories $(Z(t), S(t))$ of (1.1)–(1.4) with respect to $(Z(0), S(0))$ over finite intervals $[0, T]$, it follows that $\theta(0)$ is a continuous function of $(Z(0), S(0))$. Indeed, according to the proof of Lemma 8.2, $\theta = \widetilde{\theta}_1 := \theta_1 \cdots \theta_{n-1}$, with θ_i and $a_i(Z_{i+1}, S_{i+1})$, $i = 1, \ldots, n-1$, defined as in (A.54) and $\theta_i(0) := \exp^{\int_0^\infty a_i(Z_{i+1}(s), S_{i+1}(s))ds}$. Since by definition of $a_i(Z_{i+1}, S_{i+1})$ for each $(Z_{i+1,0}, S_{i+1,0}) \in \mathbb{R}^{n-i} \times \mathbb{R}^{n-i}$ and $i = 1, \ldots, n-1$ there exist $T_i^\circ \in \mathbb{R}^{\geq}$ and an open ball $\mathcal{B}(Z_{i+1,0}, S_{i+1,0})$ around $(Z_{i+1,0}, S_{i+1,0})$ such that $a_i(Z_{i+1}(s), S_{i+1}(t)) = 0$ for all $t \geq T_i^\circ$ and trajectory $(Z_{i+1}(s), S_{i+1}(t))$ ensuing from $\mathcal{B}(Z_{i+1,0}, S_{i+1,0})$, then $\theta_i(0) := \exp^{\int_0^\infty a_i(Z_{i+1}(s), S_{i+1}(s))ds} = \exp^{\int_0^{T_i^\circ} a_i(Z_{i+1}(s), S_{i+1}(s))ds}$ for all $(Z_{i+1}(0), S_{i+1}(0)) \in \mathcal{B}(Z_{i+1,0}, S_{i+1,0})$, $i = 1, \ldots, n-1$. As in the proof of Theorem 1(i) of [7], we conclude that $\theta_i(0)$ is a continuous function of $(Z_{i+1}(0), S_{i+1}(0))$ and thus, as claimed, $\theta(0)$ is a continuous function of $(Z(0), S(0))$. By this we can find a continuous and increasing function $\beta : \mathbb{R}^{\geq} \to \mathbb{R}^{+}$ such that

$$\theta(0) \leq \beta(\|Z(0), S(0)\|) \tag{8.36}$$

for all $(Z(0), S(0))$. Indeed, let $\alpha(s) = \int_s^{s+1} \max_{W(Z(0),S(0)) \leq r} \{\theta(0)\} dr$ and $\widetilde{\alpha}(s) = \alpha(s) + s$. Since $W(Z, S)$ (defined as in (8.29)) is proper and positive definite being the linear combination of proper and positive definite functions $W_j(z_j, \sigma_j)$, $\alpha(s)$ is

continuous and nondecreasing for all $s \geq 0$, and $\widetilde{\alpha}(s)$ is continuous and increasing for all $s \geq 0$. Moreover, by construction $\theta(0) \leq \alpha(W(Z(0), S(0))) \leq \widetilde{\alpha}(W(Z(0), S(0)))$ for all $Z(0), S(0)$. The desired function is $\beta(s) = \widetilde{\alpha}(\widetilde{\beta}_1(s))$, where $\widetilde{\beta}_1 \in \mathcal{K}_\infty$ is such that $W(Z, S) \leq \widetilde{\beta}_1(\|Z, S\|)$ for all (Z, S), which exists since $W(Z, S)$ is proper and positive definite. This with (8.29) and (8.36) gives (6.9) with $\alpha_2(s) = \beta(s)\widetilde{\beta}_1(s)$.

Finally, from (8.29) and (8.32), we obtain $\widetilde{W}(Z(t), S(t), \theta(t)) \geq \widetilde{\beta}_2(\|Z(t), S(t)\|)$ for all $t \geq 0$ along the trajectories of (1.1)–(1.4), with $\widetilde{\beta}_2 \in \mathcal{K}_\infty$ such that $W(Z, S) \geq \widetilde{\beta}_2(\|Z, S\|)$ for all (Z, S), which exists since $W(Z, S)$ is proper and positive definite. This implies (6.10) with $\alpha_3(s) = \widetilde{\beta}_2(s)$. By Lemma 6.2 we conclude that (1.1)–(1.4) is 0-GAS.

Appendix.

LEMMA A.1. *Let $w \in \mathbb{R}^r$, $x_h, \widetilde{x}_h \in \mathbb{R}$, $h = 1, \ldots, n+1$, $\dot{\widetilde{x}}_h \in \mathbb{R}$, $h = 1, \ldots, n$, $\widetilde{x}_1 = \dot{\widetilde{x}}_1 := 0$, $x_{n+1} = \widetilde{x}_{n+1}$, and $z_j := x_j - \widetilde{x}_j$, $j = 1, \ldots, n$, $z_{n+1} := 0$. For any continuous functions $\psi_{ji} : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^r \to \mathbb{R}$, $j = 1, \ldots, n$, $i = s, m$, satisfying (1.2) and for each $\Delta_1, \ldots, \Delta_n \in (0, 1]$ there exist $\gamma_{ji}^{\widetilde{x}_h}(\Delta_j, \ldots, \Delta_{h-1}) > 0$, $h = j+1, \ldots, n+1$, $i = s, m$, and smooth $\gamma_{ji}^{w} : \mathbb{R}^{n-j} \to \mathbb{R}^{\geq}$ and $\gamma_{ji}^{z_h} : \mathbb{R}^{n-h+1} \to \mathbb{R}^{\geq}$, $h = j+1, \ldots, n$, $i = s, m$, such that for each $j = 1, \ldots, n$*

$$\lim_{\Delta_j \to 0^+} \gamma_{js}^{\widetilde{x}_{j+1}}(\Delta_j) = 0 \tag{A.1}$$

and

$$\begin{aligned}[\psi_{js} + z_{j+1} - \dot{\widetilde{x}}_j]^2 \leq 2|\dot{\widetilde{x}}_j|^2 + \sum_{h=j+1}^{n} z_h^2 \gamma_{js}^{z_h}(z_h, \ldots, z_n) + \sum_{h=j+1}^{n+1} \widetilde{x}_h^2 \gamma_{js}^{\widetilde{x}_h}(\Delta_j, \ldots, \Delta_{h-1}) \\ + \gamma_{js}^{w}(z_{j+1}, \ldots, z_n)\|w\|^2,\end{aligned} \tag{A.2}$$

$$\psi_{jm}^2 \leq \sum_{h=j+1}^{n} z_h^2 \gamma_{jm}^{z_h}(z_h, \ldots, z_n) + \sum_{h=j+1}^{n+1} \widetilde{x}_h^2 \gamma_{jm}^{\widetilde{x}_h}(\Delta_j, \ldots, \Delta_{h-1}) + \gamma_{jm}^{w}(z_{j+1}, \ldots, z_n)\|w\|^2 \tag{A.3}$$

for all $w \in \mathbb{R}^r$, $x_h, \widetilde{x}_h \in \mathbb{R}$, $h = 1, \ldots, n+1$, $\dot{\widetilde{x}}_h \in \mathbb{R}$, $h = 1, \ldots, n$, such that $|\widetilde{x}_h| \leq \Delta_{h-1}$, $h = 2, \ldots, n+1$.

Proof. We will prove only (A.1) and (A.2) (the other inequality, (A.3), can be proved in the same way as (A.2)). Fix $j = 1, \ldots, n$. Since $\psi_{js}|_{x_i=0, i=j+1, \ldots, n+1; w=0} = 0$ for all $x_1, \ldots, x_j \in \mathbb{R}$ by (1.2) and using the first relation of (1.2), we get for all $x \in \mathbb{R}^n$, $w \in \mathbb{R}^r$, and $\widetilde{x}_h \in \mathbb{R}$, $h = j+1, \ldots, n+1$, that

$$\begin{aligned}|\psi_{js}|^2 \preceq |\psi_{js} - \psi_{js}|_{w=0}|^2 + \sum_{h=j+1}^{n} |\psi_{js}|_{x_i=\widetilde{x}_i, i=j+1, \ldots, h; w=0} - \psi_{js}|_{x_i=\widetilde{x}_i, i=j+1, \ldots, h-1; w=0}|^2 \\ + \sum_{h=j+1}^{n+1} |\psi_{js}|_{x_i=\widetilde{x}_i, i=j+1, \ldots, h; x_i=0, i=h+1, \ldots, n+1; w=0} \\ - \psi_{js}|_{x_i=\widetilde{x}_i, i=j+1, \ldots, h-1; x_i=0, i=h, \ldots, n+1; w=0}|^2 \\ \leq b_{js}(x_{j+1}, \ldots, x_{n+1})\|w\|^2 + \sum_{h=j+1}^{n} |z_h|^2 a_{jsh}(\widetilde{x}_h, \widetilde{x}_{j+1}, \ldots, \widetilde{x}_{h-1}, x_h, \ldots, x_{n+1})\end{aligned}$$

$$+\sum_{h=j+1}^{n+1}|\widetilde{x}_h|^2 a_{jsh}(0,\widetilde{x}_{j+1},\dots,\widetilde{x}_h,0,\dots,0). \tag{A.4}$$

Moreover, for any smooth function $q:\mathbb{R}^s\times\mathbb{R}^r\to\mathbb{R}$ there exist smooth $f:\mathbb{R}^s\to\mathbb{R}^+$ and $g:\mathbb{R}^r\to\mathbb{R}^+$ such that $q(x,y)\le f(x)g(y)$ for all $x\in\mathbb{R}^s$ and $y\in\mathbb{R}^r$ [10], [5]. By this, let $h=j+1,\dots,n+1$ and let $f_{jsh}:\mathbb{R}^{h-j+1}\to\mathbb{R}^+$, $g_{jsh}:\mathbb{R}^{n-h+1}\to\mathbb{R}^+$, and $\varrho_{js},\xi_{js}:\mathbb{R}^{n-h+1}\to\mathbb{R}^+$ be smooth functions such that

$$\begin{aligned}
&a_{jsh}(\widetilde{x}_h,\widetilde{x}_{j+1},\dots,\widetilde{x}_{h-1},x_h,\dots,x_{n+1})\\
&\qquad\le g_{jsh}(z_h,\dots,z_n)\max_{|\widetilde{x}_{i+1}|\le\Delta_i\in(0,1]:i=j,\dots,n} f_{jsh}(\widetilde{x}_{j+1},\dots,\widetilde{x}_n,x_{n+1}),\\
&b_{js}(x_{j+1},\dots,x_{n+1})\\
&\qquad\le \varrho_{js}(z_{j+1},\dots,z_n)\max_{|\widetilde{x}_{i+1}|\le\Delta_i\in(0,1]:i=j,\dots,n} \xi_{js}(\widetilde{x}_{j+1},\dots,\widetilde{x}_n,x_{n+1})
\end{aligned} \tag{A.5}$$

for all $\widetilde{x}_{j+1},\dots,\widetilde{x}_h,x_h,\dots,x_{n+1}$ such that $|\widetilde{x}_{i+1}|\le\Delta_i$, $i=j,\dots,n$. Moreover, let

$$\alpha_{js,j+1}(r):=\max_{|\widetilde{x}_{j+1}|\le r} a_{js,j+1}(0,\widetilde{x}_{j+1},0,\dots,0). \tag{A.6}$$

The function $\alpha_{js,j+1}:\mathbb{R}^{\ge}\to\mathbb{R}^{\ge}$ is nondecreasing for all $r\ge 0$ and continuous at $r=0$, and

$$a_{js,j+1}(0,\widetilde{x}_{j+1},0,\dots,0)\le\alpha_{js,j+1}(|\widetilde{x}_{j+1}|)\le\alpha_{js,j+1}(\Delta_j) \tag{A.7}$$

for all $\widetilde{x}_{j+1}$ such that $|\widetilde{x}_{j+1}|\le\Delta_j$. From (A.4)–(A.7) we get (A.2). Moreover, (A.1) follows by (A.7), the continuity of $\alpha_{js,j+1}(r)$ at $r=0$ and since $a_{js,j+1}(0,0,\dots,0)=0$ by (1.2). □

LEMMA A.2. *Let $w\in\mathbb{R}^r$, $x_h,\widetilde{x}_h\in\mathbb{R}$, $h=1,\dots,n+1$, $\widetilde{x}_1:=0$, $x_{n+1}=\widetilde{x}_{n+1}$, and $z_j:=x_j-\widetilde{x}_j$, $j=1,\dots,n$. For any continuous functions $\psi_{jm}:\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}^r\to\mathbb{R}$, $j=1,\dots,n-1$, satisfying (1.2), for each $j=1,\dots,n-1$, and $\Delta_1,\dots,\Delta_n\in(0,1]$*

$$G^2(e_j+\psi_{jm})\preceq G^2(e_j)+G^2(z_{j+1})+\sum_{l=j+2}^{n}z_l^2+\sum_{l=j+1}^{n+1}\widetilde{x}_l^2+\|w\|^2 \tag{A.8}$$

for all $w\in\mathbb{R}^r$, $e_j,x_h,\widetilde{x}_h\in\mathbb{R}$, $h=1,\dots,n+1$, such that $|\widetilde{x}_h|\le\Delta_{h-1}$, $h=2,\dots,n+1$.

Proof. Under assumption (1.2), (A.3) holds true by Lemma A.1. Moreover, for each continuous function $f:\mathbb{R}^q\to\mathbb{R}^{\ge}$ there exists $a>0$ such that

$$\frac{f(s)\|s\|^2}{1+f(s)\|s\|^2}\le\frac{a\|s\|^2}{1+\|s\|^2} \tag{A.9}$$

for all $s\in\mathbb{R}^q$. Indeed, pick $\delta>0$ and let $a_0>0$ be such that (A.9) holds true for all $s\in\mathbb{R}^q:\|s\|\le\delta$ with $a=a_0$. Since $f(s)(1+\|s\|^2)/[1+f(s)\|s\|^2]\le\frac{1}{\delta^2}+1$ for all $s\in\mathbb{R}^q:\|s\|\ge\delta$, then clearly (A.9) holds true with $a=\max\{a_0,\frac{1}{\delta^2}+1\}$ for all $s\in\mathbb{R}^q$. Using repeatedly $\frac{s_1}{1+s_1}\le\frac{s_2}{1+s_2}$ for all $s_2\ge s_1\ge 0$ and $\frac{\sum_l s_l}{1+\sum_l s_l}\le\sum_l\frac{s_l}{1+s_l}$ for all $s_l\ge 0$ and (A.9) with $s:=(z_{j+1}\ \cdots\ z_n\ \widetilde{x}_{j+1}\ \cdots\ \widetilde{x}_{n+1}\ w)^T$ and $f(s):=\sum_{l=j+1}^{n}\gamma_{jm}^{z_l}+\sum_{l=j+1}^{n+1}\gamma_{jm}^{\widetilde{x}_l}+\gamma_{jm}^{w}$, from (A.3) we get

$$\begin{aligned}
G^2(e_j+\psi_{jm})&\preceq G^2(e_j)+G^2(\psi_{jm})\preceq G^2(e_j)+G^2(\sqrt{f(s)}\|s\|)\\
&\preceq G^2(e_j)+G^2(\|s\|)\preceq G^2(e_j)+G^2(z_{j+1})+\sum_{l=j+2}^{n}z_l^2+\sum_{l=j+1}^{n+1}\widetilde{x}_l^2+\|w\|^2,
\end{aligned} \tag{A.10}$$

i.e., (A.8). □

LEMMA A.3. *For any continuous* $\gamma : \mathbb{R}^n \to \mathbb{R}^{\geq}$, *positive* $\varphi : \mathbb{R}^n \to \mathbb{R}^+$, *and smooth, proper, and positive definite* $W : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{\geq}$ *there exists a continuous nondecreasing function* $\tau : \mathbb{R}^{\geq} \to [1, \infty)$ *such that* $\gamma(x) \leq \tau(W(x,y))\varphi(x)$ *for all* $x, y \in \mathbb{R}^n$.

Proof. Let $\alpha \in \mathcal{K}_\infty$ be such that $W(x,y) \geq \alpha(\|x,y\|)$ for all $x, y \in \mathbb{R}^n$, which indeed exists since W is proper and positive definite. Let $\max_{0\leq\|x\|\leq r} \frac{\gamma(x)}{\varphi(x)} := \beta(r)$ and $\widetilde{\tau}(s) = \int_s^{s+1} \beta(r)dr$. The function $\widetilde{\tau} : \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$ is continuous and nondecreasing for all $s \geq 0$ and $\dfrac{\gamma(x)}{\varphi(x)} \leq \widetilde{\tau}(\|x\|) \leq \widetilde{\tau}(\|x,y\|) \leq \widetilde{\tau}(\alpha^{-1}(W(x,y))) + 1$ for all x and y. Thus, our lemma follows with $\tau(s) = \widetilde{\tau}(\alpha^{-1}(s)) + 1$. □

Proof of Lemma 8.1. First, we prove the lemma for $j = n, n-1$, and then proceed by induction. We construct a sequence of compact traps $\mathcal{T}_n^{(k)}(\|w\|_2)$ and $\mathcal{T}_{n-1}^{(k)}(\|w\|_2)$ for $\Sigma_n \circ \mathcal{C}_n$ and $\Sigma_i \circ \mathcal{C}_i$, $i = n-1, n$, respectively, parametrized by some numbers $c_n^{(k)}$ and $c_{n-1}^{(k)}$, getting smaller as k gets larger and converging to c_n and c_{n-1} satisfying (8.23). Note that since

$$\text{(A.11)} \quad \frac{(r_1+r_2+1)^2-1}{(r_1+r_2+1)^2} \leq 2\left[\frac{(r_1+1)^2-1}{(r_1+1)^2} + \frac{(r_2+1)^2-1}{(r_2+1)^2}\right] \quad \forall r_1, r_2 \geq 0$$

and $W_n(z_n, \sigma_n) = [\sqrt{1+z_n^2} - 1] + [\sqrt{1+e_n^2} - 1]$, then

$$\text{(A.12)} \qquad 2\left[\frac{z_n^2}{1+z_n^2} + \frac{e_n^2}{1+e_n^2}\right] \geq \frac{(W_n(z_n,\sigma_n)+1)^2-1}{(W_n(z_n,\sigma_n)+1)^2} \geq \frac{W_n(z_n,\sigma_n)}{1+W_n(z_n,\sigma_n)}$$

for all z_n, σ_n. Thus, from (8.7), (8.8)–(8.11), (8.15), and (8.21)–(8.22), with $j = n$, we infer the existence of $\varepsilon_n^{(0)} \geq \varepsilon_n$ and $b_n, l_n > 0$ such that for all $\varepsilon \geq \varepsilon_n^{(0)}$

$$\text{(A.13)} \quad \dot{W}_n \leq -\frac{l_n W_n}{R_{n1}(W_n+1)} + \frac{6b_n R_{n1}}{R_{n-1,1}^4} + b_n R_{n1}\|w\|^2, \quad c_n^{(0)} := \frac{24R_{n1}^2 b_n}{l_n R_{n-1,1}^4} < 1$$

along the trajectories of (1.1)–(1.4). Let $\varepsilon \geq \varepsilon_n^{(0)}$ and $s \geq t_0 \geq 0$. Note that $0 \leq a \leq 1 \Rightarrow a/[a+1] \geq a/2$ and

$$\text{(A.14)} \qquad W_n \geq c_n^{(0)} \Rightarrow -l_n W_n/[R_{n1}(W_n+1)] + 12b_n R_{n1}/R_{n-1,1}^4 \leq 0.$$

For each trajectory of (1.1)–(1.4) and $r \geq s$ such that $W_n(t) \geq c_n^{(0)}$ for all $t \in [s,r]$ along such trajectory and since $\int_s^r \|w(\lambda)\|^2 d\lambda \leq \|w\|_2^2 < \infty$, by integrating (A.13) over $[s,r]$ we obtain

$$\text{(A.15)} \quad W_n(t) \leq -\frac{6b_n R_{n1}(t-s)}{R_{n-1,1}^4} + W_n(s) + b_n R_{n1}\|w\|_2^2 \quad \forall t \in [s,r],\ W_n(s) \geq c_n^{(0)}.$$

It is easy to see that for each $s \geq t_0 \geq 0$ and trajectory of (1.1)–(1.4) such that $W_n(s) > c_n^{(0)}$ there exists $\overline{r} \geq s$ such that

$$\text{(A.16)} \qquad W_n(\overline{r}) = c_n^{(0)}.$$

Indeed, if not there would exist a trajectory of (1.1)–(1.4) such that $W_n(t) > c_n^{(0)}$ and (A.15) holds for all $t \geq s$. From (A.15) we get $W_n(\overline{r}) = c_n^{(0)}$ for $\overline{r} := s +$

$R_{n-1,1}^4/(6b_nR_{n1})][W_n(s)-c_n^{(0)}+b_n\|w\|_2^2]$, which gives a contradiction. Thus, (A.15) and (A.16) imply for each trajectory of (1.1)–(1.4) starting at $t_0\geq 0$ the existence of $T_n^{(0)}\geq t_0$ such that

$$\begin{aligned} &W_n(t)\leq \varrho_n^{(0)}(W_n(t_0),\|w\|_2):=W_n(t_0)+b_nR_{n1}\|w\|_2^2 \quad \forall t\geq t_0,\\ &W_n(t)\leq \varrho_n^{(0)}(c_n^{(0)},\|w\|_2) \quad \forall t\geq T_n^{(0)}, \end{aligned} \tag{A.17}$$

where $\varrho_n^{(0)}(\lambda_1,\cdot)$ and $\varrho_n^{(0)}(\cdot,\lambda_2)$ are nondecreasing for each $\lambda_1,\lambda_2\geq 0$ and $\varrho_n^{(0)}(0,0)=0$. In other words, $\Sigma_n\circ\mathcal{C}_n$ is recurrent relative to the trap $\mathcal{T}_n^{(0)}(\|w\|_2):=\{(z_n,\sigma_n): W_n(z_n,\sigma_n)\leq \varrho_n^{(0)}(c_n^{(0)},\|w\|_2)\}$ for all $\varepsilon\geq\varepsilon_n^{(0)}$, with $c_n^{(0)}\preceq\frac{R_{n1}^2}{R_{n-1,1}^4}$ by (A.13). Moreover, since $W_n(z_n,\sigma_n)\geq 0$ for all z_n,σ_n, from (A.13) and the definition of $c_n^{(0)}$

$$\begin{aligned} W_n(t)\geq c_n^{(0)}\ \forall t\in[s,r]&\Rightarrow \int_s^t \frac{W_n(\lambda)}{W_n(\lambda)+1}d\lambda\\ &\leq \xi_n^{(0)}(W_n(s),\|w\|_2):=\frac{2R_{n1}}{l_n}[W_n(s)+R_{n1}b_n\|w\|_2^2] \quad \forall t\in[s,r] \end{aligned} \tag{A.18}$$

along the trajectories (1.1)–(1.4), where $\xi_n^{(0)}(\lambda_1,\cdot)$ and $\xi_n^{(0)}(\cdot,\lambda_2)$ are nondecreasing for each $\lambda_1,\lambda_2\geq 0$ and $\xi_n^{(0)}(0,0)=0$. Now, we prove the existence of $\varepsilon_{n-1}^{(0)}>0$ such that for all $\varepsilon\geq\varepsilon_{n-1}^{(0)}$, $\Sigma_i\circ\mathcal{C}_i$, $i=n-1,n$, is recurrent relative to the trap $\mathcal{T}_{n-1}^{(0)}(\|w\|_2):=\{(Z_{n-1},S_{n-1}): W_{n-1}(z_{n-1},\sigma_{n-1})\leq \varrho_{n-1}^{(0)}(c_{n-1}^{(0)},c_n^{(0)},\|w\|_2), W_n(z_n,\sigma_n)\leq\varrho_n^{(0)}(c_n^{(0)},\|w\|_2)\}$ with $\varrho_{n-1}^{(0)}:(\mathbb{R}^{\geq})^3\to\mathbb{R}^{\geq}$, nondecreasing with respect to each argument and $\varrho_{n-1}^{(0)}(0,0,0)=0$, and with

$$c_{n-1}^{(0)}<1,\ c_n^{(0)}\preceq\frac{R_{n1}^2}{R_{n-1,1}^4}. \tag{A.19}$$

Let

$$\alpha_{n-1}(s):=\int_s^{s+1}\max_{W_n(z_n,\sigma_n)\leq r}\overline{\alpha}_{n-1}(z_n,\sigma_n)dr$$

with $\overline{\alpha}_{n-1}(z_n,\sigma_n):=[(\gamma_{n-1,s}^{z_n}+\gamma_{n-1,m}^{z_n}+\gamma_{n-1,s}^{\widetilde{x}_{n+1}}+\gamma_{n-1,m}^{\widetilde{x}_{n+1}}+1)(W_n(z_n,\sigma_n)+2)^2+\gamma_{n-1}^w]$. The function $\alpha_{n-1}:\mathbb{R}^{\geq}\to\mathbb{R}^{\geq}$ is continuous and nondecreasing since $\alpha_{n-1}(s)\geq\max_{W_n(z_n,\sigma_n)\leq s}\overline{\alpha}_{n-1}(z_n,\sigma_n)$ for all $s\geq 0$. Moreover, by (8.22), with $j=n-1$ and $l=n$, the nonnegativity of $W_n(z_n,\sigma_n)$, and since $z_n^2,e_n^2\leq(1+W_n(z_n,\sigma_n))^2-1$ for all z_n,σ_n,

$$\begin{aligned} \gamma_{n-1}^{z_n}z_n^2+\gamma_{n-1}^{e_n}e_n^2+\gamma_{n-1}^w\|w\|^2&\preceq R_{n-1,1}\Bigg[(\gamma_{n-1,s}^{z_n}+\gamma_{n-1,m}^{z_n}+\gamma_{n-1,s}^{\widetilde{x}_{n+1}}+\gamma_{n-1,m}^{\widetilde{x}_{n+1}}+1)\\ &\quad\cdot\frac{[(W_n+1)^2-1](2+W_n)}{W_n}+\gamma_{n-1}^w\Bigg]\left[\frac{W_n}{W_n+1}+\|w\|^2\right]\\ &\leq R_{n-1,1}\alpha_{n-1}(W_n)\left[\frac{W_n}{W_n+1}+\|w\|^2\right] \end{aligned}$$

for all z_n,σ_n,w. Note that $c_n^{(0)}<1$, $R_{n1}\geq 1$, and $W_n(z_n,\sigma_n)\leq c_n^{(0)}\Rightarrow z_n^2,e_n^2,\widetilde{x}_{n+1}^2\leq 2[(c_n^{(0)}+1)^2-1]\leq 6$ by (8.18) and the second relation of (A.13). Thus, from (8.7),

(8.10), (8.11), (8.15)–(8.17), and (8.21)–(8.22), with $j = n-1$, we derive the existence of $\varepsilon^{(0)}_{n-1} > 0$ and $b_{n-1}, l_{n-1} > 0$ such that for all $\varepsilon \geq \varepsilon^{(0)}_{n-1}$

$$c^{(0)}_{n-1} := \frac{4R^2_{n-1,1}b_{n-1}}{l_{n-1}}\left[\frac{6}{R^4_{n-2,1}} + 2n^3\alpha_{n-1}(2n)c^{(0)}_n\right] < 1 \tag{A.20}$$

and along the trajectories of (1.1)–(1.4)

$$\begin{aligned}\dot{W}_{n-1} \leq -\frac{l_{n-1}W_{n-1}}{R_{n-1,1}(W_{n-1}+1)} + 6b_{n-1}\frac{R_{n-1,1}}{R^4_{n-2,1}} \\ + b_{n-1}R_{n-1,1}\alpha_{n-1}(W_n)\left[\frac{W_n}{W_n+1} + \|w\|^2\right].\end{aligned} \tag{A.21}$$

Assume that $W_{n-1}(t) \geq c^{(0)}_{n-1}$ for all $t \in [s,r]$, $s \geq t_0$. Since $\alpha_{n-1}(a)$ is nondecreasing for all $a \geq 0$ and $c^{(0)}_n \leq 1$, by the nonnegativity of W_i and (A.17) and (A.18) we obtain for all $t \in [s,r]$

$$\begin{aligned}&\int_s^t \alpha_{n-1}(W_n(\lambda))\left[\frac{W_n(\lambda)}{W_n(\lambda)+1} + \|w(\lambda)\|^2\right]d\lambda \\ &\leq \alpha_{n-1}(2n\varrho^{(0)}_n(W_n(s),\|w\|_2))[2n^2\xi^{(0)}_n(W_n(s),\|w\|_2) + \|w\|^2_2]\end{aligned} \tag{A.22}$$

if $W_n(t) \geq c^{(0)}_n$ for all $t \in [s,r]$,

$$\alpha_{n-1}(W_n(t))\left[\frac{W_n(t)}{W_n(t)+1} + \|w(t)\|^2\right] \leq n\alpha_{n-1}(2n)[2n^2c^{(0)}_n + \|w(t)\|^2] \tag{A.23}$$

if $W_n(t) \leq c^{(0)}_n$ for all $t \in [s,r]$. Thus, using the definition of $c^{(0)}_{n-1}$ in (A.20) and by integrating (A.21) over $[s,r]$, we have that

$$\begin{aligned}&W_{n-1}(t) \geq c^{(0)}_{n-1} \quad \forall t \in [s,r] \Rightarrow W_{n-1}(t) \leq -\frac{6b_{n-1}R_{n-1,1}(t-s)}{R^4_{n-2,1}} \\ &+ \varrho^{(0)}_{n-1}(W_{n-1}(s), W_n(s), \|w\|_2)\ \forall t \in [s,r]\end{aligned} \tag{A.24}$$

with

$$\begin{aligned}\varrho^{(0)}_{n-1}(W_{n-1}(s), W_n(s), \|w\|_2) := W_{n-1}(s) + b_{n-1}R_{n-1,1}\Big\{n\alpha_{n-1}(2n)\|w\|^2_2 \\ + \alpha_{n-1}(2n\varrho^{(0)}_n(W_n(s),\|w\|_2))[2n^2\xi^{(0)}_n(W_n(s),\|w\|_2) + \|w\|^2_2]\Big\}.\end{aligned} \tag{A.25}$$

From (A.24) follows for each trajectory of (1.1)–(1.4) starting at t_0 the existence of continuous $\varrho^{(0)}_{n-1} : (\mathbb{R}^{\geq})^3 \to \mathbb{R}^{\geq}$, nondecreasing with respect to each argument and $\varrho^{(0)}_{n-1}(0,0,0) = 0$, and for each trajectory (1.1)–(1.4) the existence of $T^{(0)}_{n-1} \geq T^{(0)}_n \geq t_0$ such that

$$\begin{aligned}&W_{n-1}(t) \leq \varrho^{(0)}_{n-1}(W_{n-1}(t_0), W_n(t_0), \|w\|_2) \quad \forall t \geq t_0, \\ &W_{n-1}(t) \leq \varrho^{(0)}_{n-1}(c^{(0)}_{n-1}, c^{(0)}_n, \|w\|_2) \quad \forall t \geq T^{(0)}_{n-1}.\end{aligned} \tag{A.26}$$

Moreover, $c_n^{(0)}$ and $c_{n-1}^{(0)}$ satisfy (A.19) by (A.13) and (A.20). This proves that $\Sigma_i \circ \mathcal{C}_i$, $i = n-1, n$, is recurrent relative to the trap $\mathcal{T}_{n-1}^{(0)}(\|w\|_2)$. We also claim the existence of continuous $\xi_{n-1}^{(0)} : \mathbb{R}^{\geq} \times \mathbb{R}^{\geq} \times \mathbb{R}^{\geq} \to \mathbb{R}^{\geq}$, nondecreasing with respect to each argument and $\xi_{n-1}^{(0)}(0,0,0) = 0$, such that

$$\begin{aligned} & W_{n-1}(t) \geq c_{n-1}^{(0)} \quad \forall t \in [s,r] \\ & \Rightarrow \int_s^t \frac{W_n(\lambda)}{W_n(\lambda)+1} d\lambda \leq \xi_{n-1}^{(0)}(W_{n-1}(s), W_n(s), \|w\|_2) \quad \forall t \in [s,r] \end{aligned} \tag{A.27}$$

along the trajectories (1.1)–(1.4). Indeed, assume that $W_{n-1}(t) \geq c_{n-1}^{(0)}$ for all $t \in [s,r]$, $s \geq t_0$. Integrating (A.21) over $[s,r]$ and on account of (A.22) and (A.23), by the definition of $c_{n-1}^{(0)}$ in (A.20) and the nonnegativity of W_{n-1}, we obtain

$$\int_s^t \frac{W_n(\lambda)}{W_n(\lambda)+1} d\lambda \leq \frac{2R_{n-1,1}}{l_{n-1}} \{W_{n-1}(s) + \varrho_{n-1}^{(0)}(0, W_n(s), \|w\|_2)\} \quad \forall t \in [s,r], \tag{A.28}$$

which, upon setting $\xi_{n-1}^{(0)}(W_{n-1}(s), W_n(s), \|w\|_2)) := [2R_{n-1,1}/l_{n-1}]\{W_{n-1}(s) + \varrho_{n-1}^{(0)}(0, W_n(s), \|w\|_2)\}$, proves (A.27). Next, we prove by induction that for each $k \geq 0$ it is possible to construct smaller and smaller $c_n^{(k+1)}$ and $c_{n-1}^{(k+1)}$ and, thus, traps $\mathcal{T}_n^{(k+1)}(\|w\|_2)$ and $\mathcal{T}_{n-1}^{(k+1)}(\|w\|_2)$ for $\Sigma_n \circ \mathcal{C}_n$ and $\Sigma_i \circ \mathcal{C}_i$, $i = n-1, n$, by using $c_n^{(k)}$ and $c_{n-1}^{(k)}$. For some sufficiently large k, these $c_n^{(k+1)}$ and $c_{n-1}^{(k+1)}$ will satisfy (8.23) for $i = n-1, n$. Let $k \geq 0$.

Induction step. There exist $\varepsilon_i^{(k)} > 0$, continuous functions $c_i^{(k)} : \mathbb{R}^+ \to \mathbb{R}^+$, $\varrho_i^{(k)}, \xi_i^{(k)} : (\mathbb{R}^{\geq})^3 \to \mathbb{R}^{\geq}$, $i = n-1, n$, nondecreasing with respect to each argument and $\varrho_i^{(k)}(0,0,0) = \xi_i^{(k)}(0,0,0) = 0$, and for each trajectory of (1.1)–(1.4) starting at $t_0 \geq 0$ there exist $T_{n-1}^{(k)} \geq T_n^{(k)} \geq t_0$, such that (A.17), (A.18), (A.19), (A.26), and (A.27) hold along the trajectories of (1.1)–(1.4), with $^{(0)}$ replaced by $^{(k)}$ and for all $\varepsilon \geq \varepsilon_i^{(k)}$.

By (A.19) and since $W_{n-1}(z_{n-1}, \sigma_{n-1}) \leq c_{n-1}^{(k)} \Rightarrow z_{n-1}^2,\ e_{n-1}^2, \widetilde{x}_n^2 \leq 2[(c_{n-1}^{(k)} + 1)^2 - 1] \leq 6c_{n-1}^{(k)}$ by (8.18), from (8.7), (8.8)–(8.15), with $j = n$, and (8.21)–(8.22) we get the existence of $\varepsilon_n^{(k+1)} \geq \varepsilon_n^{(k)}$ such that for all $\varepsilon \geq \varepsilon_n^{(k+1)}$

$$\dot{W}_n \leq -\frac{l_n W_n}{R_{n1}(W_n+1)} + \frac{6b_n R_{n1} c_{n-1}^{(k)}}{R_{n-1,1}^4} + b_n R_{n1} \|w\|^2, \quad c_n^{(k+1)} := \frac{24 R_{n1}^2 b_n c_{n-1}^{(k)}}{l_n R_{n-1,1}^4} < 1, \tag{A.29}$$

$$\begin{aligned} \dot{W}_{n-1} \leq & -\frac{l_{n-1} W_{n-1}}{R_{n-1,1}(W_{n-1}+1)} + \frac{6b_{n-1} R_{n-1,1}}{R_{n-2,1}^4} \\ & + b_{n-1} R_{n-1,1} \alpha_{n-1}(W_n) \left[\frac{W_n}{W_n+1} + \|w\|^2 \right], \\ c_{n-1}^{(k+1)} := & \frac{4R_{n-1,1}^2 b_{n-1}}{l_{n-1}} \left[\frac{6}{R_{n-2,1}^4} + 2n^3 \alpha_{n-1}(2n) c_n^{(k)} \right] < 1. \end{aligned} \tag{A.30}$$

Note that in (A.13) we used the weaker bound $W_{n-1}(z_{n-1}, \sigma_{n-1}) \leq c_{n-1}^{(0)} \Rightarrow z_{n-1}^2, e_{n-1}^2, \widetilde{x}_n^2 \leq 6$ instead of $W_{n-1}(z_{n-1}, \sigma_{n-1}) \leq c_{n-1}^{(k)} \Rightarrow z_{n-1}^2, e_{n-1}^2, \widetilde{x}_n^2 \leq 6c_{n-1}^{(k)}$ as in

(A.29). Reasoning as in the case $k=0$ above, we obtain the existence of $\varepsilon_i^{(k+1)}>0$, continuous functions $\varrho_i^{(k+1)},\xi_i^{(k+1)}:(\mathbb{R}^{\geq})^{n-i+2}\to\mathbb{R}^{\geq}$, $i=n-1,n$, nondecreasing with respect to each argument and $\varrho_n^{(k+1)}(0,0,0)=0$, and for each trajectory of (1.1)–(1.4) the existence of $T_{n-1}^{(k+1)},T_n^{(k+1)}\geq T_{n-1}^{(k)}$ such that the induction step holds true with $\varepsilon_i^{(k)}$, $c_i^{(k)}$, $T_i^{(k)}$, $\varrho_i^{(k)}$, and $\xi_i^{(k)}$ replaced by $\varepsilon_i^{(k+1)}$, $c_i^{(k+1)}$, $T_i^{(k+1)}$, $\varrho_i^{(k+1)}$, and $\xi_i^{(k+1)}$, $i=n-1,n$. Thus, since, as already shown, the induction step also holds for $k=0$, it holds for all $k\geq 0$. We prove Lemma 8.1 for $j=n-1,n$ if we prove that $c_n^{(k)}$ and $c_{n-1}^{(k)}$ in (A.29) and (A.30) satisfy (8.23) for $j=n-1,n$ for sufficiently large k and ε. To this aim, we claim that there exist $\varepsilon_{n-1}^*>0$ and $k_{n-1}^*>0$ such that for all $\varepsilon\geq\varepsilon_{n-1}^*$ and $k\geq k_{n-1}^*$

$$c_i^{(k)}\sim\prod_{l=n-1}^{i}\frac{R_{l1}^2}{R_{l-1,1}^4},\quad i=n-1,n, \tag{A.31}$$

$$c_n^{(k+1)}:=\frac{24R_{n1}^2b_nc_{n-1}^{(k)}}{l_nR_{n-1,1}^4}<1,\ c_{n-1}^{(k+1)}:=\frac{4R_{n-1,1}^2b_{n-1}}{l_{n-1}}\Big[\frac{6}{R_{n-2,1}^4}+2n^3\alpha_{n-1}(2n)c_n^{(k)}\Big]<1 \tag{A.32}$$

with $c_n^{(0)}$ as in (A.13) and $c_{n-1}^{(0)}$ as in (A.20), since then Lemma 8.1 for $j=n-1,n$ follows with ε_{n-1}^*, $c_i:=c_i^{(k_{n-1}^*)}$ and $\varrho_i:=\varrho_i^{(k_{n-1}^*)}$, $j=n-1,n$. The equations (A.32), which hold for all $k\geq 0$ on account of (A.29)–(A.30), can be described by a linear discrete-time system $x(k+1)=Ax(k)+Bu(k)$ with state $x(k)=(c_n^{(k)}\ c_{n-1}^{(k)})^T$, initial condition $x(0)=(c_n^{(0)}\ c_{n-1}^{(0)})^T$, input $u(k)=24R_{n-1,1}^2b_{n-1}/[R_{n-2,1}^4l_{n-1}]$. Using induction and by (8.7) and (8.16), with $j=n$, we prove for all $k\geq 1$

$$\begin{aligned}\begin{pmatrix}c_n^{(k)}\\ c_{n-1}^{(k)}\end{pmatrix}&=A^k\begin{pmatrix}c_n^{(0)}\\ c_{n-1}^{(0)}\end{pmatrix}+[I+A+A^2+\cdots+A^{k-1}]\begin{pmatrix}0\\1\end{pmatrix}u(k)\\ &\sim\begin{pmatrix}\mathcal{O}(\varepsilon^{-(k+1)})\\ \mathcal{O}(\varepsilon^{-(k+1)})\end{pmatrix}+\begin{pmatrix}[R_{n1}^2/R_{n-1,1}^4][1+\mathcal{O}(\varepsilon^{-1})]\\ 1+\mathcal{O}(\varepsilon^{-1})\end{pmatrix}\frac{R_{n-1,1}^2}{R_{n-2,1}^4},\end{aligned} \tag{A.33}$$

where $\mathcal{O}(r)$ means infinitesimals with order greater than r and $\sim$ is meant componentwise. This proves (A.31) for some sufficiently large $\varepsilon_{n-1}^*,k_{n-1}^*$ and Lemma 8.1 for $j=n-1,n$. We complete the proof of Lemma 8.1 by induction on the number $n-j+1$, $j=1,\dots,n-1$, of systems $\Sigma_i\circ\mathcal{C}_i$ in $\Sigma_i\circ\mathcal{C}_i$, $i=j,\dots,n$. To this aim we formulate the induction step as follows, letting $j=2,\dots,n-1$.

Induction step. There exist $\varepsilon_j^*>0$, $k_j^*\geq 1$, continuous functions $\varrho_i^{(k)},\xi_i^{(k)}:(\mathbb{R}^{\geq})^{n-i+2}\to\mathbb{R}^{\geq}$, $i=j,\dots,n$, $k\geq k_j^*$, nondecreasing with respect to each argument and $\varrho_i^{(k)}(0,\dots,0)=\xi_i^{(k)}(0,\dots,0)=0$, and for each trajectory of (1.1)–(1.4) starting at $t_0\geq 0$ there exist $T_i^{(k)}\geq t_0$, $i=j,\dots,n$, $k\geq k_j^*$, such that for all $i=j,\dots,n$, $k\geq k_j^*$, and $\varepsilon\geq\varepsilon_j^*$,

$$\begin{aligned}&W_i(t)\leq\varrho_i^{(k)}(W_i(t_0),\dots,W_n(t_0),\|w\|_2)\quad\forall t\geq t_0,\\ &W_i(t)\leq\varrho_i^{(k)}(c_i^{(k)},\dots,c_n^{(k)},\|w\|_2)\quad\forall t\geq T_i^{(k)},\\ &W_i(t)\geq c_n^{(k)}\quad\forall t\in[s,r]\end{aligned} \tag{A.34}$$

$$\Rightarrow\int_s^t\frac{W_i(\lambda)}{W_i(\lambda)+1}d\lambda\leq\xi_i^{(k)}(W_i(s),\dots,W_n(s),\|w\|_2)\quad\forall t\in[s,r] \tag{A.35}$$

hold along the trajectories of (1.1)–(1.4) and

$$c_i^{(k)} \sim \prod_{l=j}^{i} \frac{R_{l1}^2}{R_{l-1,1}^4},\ i=j,\dots,n,\ c_j^{(k)} \geq c_{j+1}^{(k)} \geq \cdots \geq c_n^{(k)}, \tag{A.36}$$

$$c_n^{(k+1)} := \frac{24R_{n1}^2 b_n c_{n-1}^{(k)}}{l_n R_{n-1,1}^4} < 1,$$

$$c_i^{(k+1)} := \frac{4R_{i1}^2 b_i}{l_i}\left[\frac{6c_{i-1}^{(k)}}{R_{i-1,1}^4} + 2n^3\alpha_i(2n)c_{i+1}^{(k)}\right] < 1,\ i=j,\dots,n. \tag{A.37}$$

Let $k \geq k_j^*$ and $\varepsilon \geq \varepsilon_j^*$. We distinguish the cases $j=3,\dots,n$ and $j=2$. Assume first that $j=3,\dots,n$. Let

$$\alpha_{j-1}(s) := \int_s^{s+1} \max_{\sum_{i=j}^n W_i(z_i,\sigma_i)\leq r} \overline{\alpha}_{j-1}(Z_j,S_j)dr$$

with $\overline{\alpha}_{j-1}(Z_j,S_j) := \sum_{i=j}^n[(\gamma_{j-1,s}^{z_i} + \gamma_{j-1,m}^{z_i} + \gamma_{j-1,s}^{\widetilde{x}_{i+1}} + \gamma_{j-1,m}^{\widetilde{x}_{i+1}} + 1)(\sum_{i=j}^n W_i(z_i,\sigma_i) + 2)^2 + \gamma_{j-1}^w]$ and, thus,

$$\sum_{i=j}^n(\gamma_{j-1}^{z_i}z_i^2 + \gamma_{j-1}^{e_i}e_i^2) + \gamma_{n-1}^w\|w\|^2 \leq R_{j-1,1}\alpha_{j-1}\left(\sum_{i=j}^n W_i\right)\left[\frac{\sum_{i=j}^n W_i}{\sum_{i=j}^n W_i + 1} + \|w\|^2\right]$$

for all z_i,σ_i,w, $i=j\dots,n$. Thus, from (8.7), (8.8)–(8.11), (8.15), and (8.21)–(8.22), we derive the existence of $\varepsilon_{j-1}^{(k_j^*)} > 0$ and $b_{j-1}, l_{j-1} > 0$ such that for all $\varepsilon \geq \varepsilon_{j-1}^{(k_j^*)}$

$$c_{j-1}^{(k_j^*)} := \frac{4R_{j-1,1}^2 b_{j-1}}{l_{j-1}}\left[\frac{6}{R_{j-2,1}^4} + 2n^3\alpha_{j-1}(2n)c_j^{(k_j^*)}\right] < 1 \tag{A.38}$$

and along the trajectories of (1.1)–(1.4)

$$\begin{aligned}\dot{W}_{j-1} \leq\ & -\frac{l_{j-1}W_{j-1}}{R_{j-1,1}(W_{j-1}+1)} + 6b_{j-1}\frac{R_{j-1,1}}{R_{j-2,1}^4} \\ & + b_{j-1}R_{j-1,1}\alpha_{j-1}\left(\sum_{i=j}^n W_i\right)\left[\frac{\sum_{i=j}^n W_i}{\sum_{i=j}^n W_i + 1} + \|w\|^2\right].\end{aligned} \tag{A.39}$$

Assume that $W_{j-1}(t) \geq c_{j-1}^{(k_j^*)}$, $W_i(t) \geq c_i^{(k_j^*)}$, $i \in J \subseteq \{j,\dots,n\}$, and $W_i(t) \leq c_i^{(k_j^*)}$, $i \in \{j,\dots,n\}\backslash J$, for all $t \in [s,r]$, $s \geq t_0$. Since $\alpha_{j-1}(\sum_{i=j}^n a_i)[\sum_{i=j}^n a_i]/[\sum_{i=j}^n a_i + 1]$ is nondecreasing with respect to each $a_i \geq 0$ and on account of (8.26)

$$\frac{\sum_{i=j}^n a_i}{\sum_{i=j}^n a_i + 1}\alpha_{j-1}\left(\sum_{i=j}^n a_i\right) \leq \sum_{i=j}^n \frac{2n^2 a_i}{a_i+1}\alpha_{j-1}(2na_i) \tag{A.40}$$

for all $a_i \geq 0$, and since $c_i^{(k_j^*)} \leq 1$, $j,\dots,n$, using the induction step and the nonnegativity of W_i we obtain for all $t \in [s,r]$, $i=j\dots,n$, and for any set $J \subseteq \{j,\dots,n\}$

$$\alpha_{j-1}\left(\sum_{i=j}^{n} W_i(t)\right)\left[\frac{\sum_{i=j}^{n} W_i(t)}{\sum_{i=j}^{n} W_i(t)+1}+\|w(t)\|^2\right]$$

$$\leq 2n^2\sum_{i=j}^{n}\alpha_{j-1}(2nW_i(t))\left[\frac{W_i}{W_i+1}+\|w(t)\|^2\right], \tag{A.41}$$

$$\sum_{i\in J}\int_s^t \alpha_{j-1}(2nW_i(\lambda))\left[\frac{W_i(\lambda)}{W_i(\lambda)+1}+\|w(\lambda)\|^2\right]d\lambda$$

$$\leq \sum_{i=j}^{n}\alpha_{j-1}(2n\varrho_i^{(k_j^*)}(W_i(s),\dots,W_n(s),\|w\|_2))[\xi_i^{(k_j^*)}(W_i(s),\dots,W_n(s),\|w\|_2)+\|w\|_2^2], \tag{A.42}$$

$$\sum_{i\in\{j,\dots,n\}\setminus J}\alpha_{j-1}(2nW_i(t))\left[\frac{W_i(t)}{W_i(t)+1}+\|w(t)\|^2\right]\leq \alpha_{j-1}(2n)\sum_{i=j}^{n}[c_i^{(k_j^*)}+\|w(t)\|^2]$$

$$\leq n\alpha_{j-1}(2n)[c_j^{(k_j^*)}+\|w(t)\|^2]. \tag{A.43}$$

The last passage in (A.43) follows from being $c_j^{(k)}\geq c_{j+1}^{(k)}\geq\cdots\geq c_n^{(k)}$ in (A.37). Thus, using the definition of $c_{n-1}^{(k_j^*)}$, by integrating (A.39) over $[s,r]$

$$W_{j-1}(t)\geq c_{j-1}^{(k_j^*)}\quad\forall t\in[s,r]\Rightarrow W_{j-1}(t)\leq -\frac{6b_{n-1}R_{j-1,1}(t-s)}{R_{j-2,1}^4}$$

$$+\,\varrho_{j-1}^{(k_j^*)}(W_{j-1}(s),\dots,W_n(s),\|w\|_2)\quad\forall t\in[s,r] \tag{A.44}$$

with

$$\varrho_{j-1}^{(k_j^*)}(W_{j-1}(s),\dots,W_n(s),\|w\|_2):=W_{j-1}(s)+2n^2b_{n-1}R_{j-1,1}\Big\{n\alpha_{j-1}(2n)\|w\|_2^2$$

$$+\,\alpha_{j-1}(2n\varrho_i^{(k_j^*)}(W_i(s),\dots,W_n(s),\|w\|_2))\sum_{i=j}^{n}[\xi_i^{(k_j^*)}(W_i(s),\dots,W_n(s),\|w\|_2)+\|w\|_2^2]\Big\}.$$

From (A.44) follow the existence of $\varepsilon_{j-1}^{(k_j^*+1)}>0$ and for each trajectory of (1.1)–(1.4) starting at $t_0\geq 0$ the existence of $T_{j-1}^{(k_j^*+1)}\geq t_0$ such that (A.34) hold true for all $i=j-1,\dots,n$ with $k=k_j^*+1$. Next, we show that also (A.35) holds true for all $i=j-1,\dots,n$ with $k=k_j^*+1$. Assume that $W_{j-1}(t)\geq c_{j-1}^{(k_j^*)}$, $W_i(t)\geq c_i^{(k_j^*)}$, $i\in J\subseteq\{j,\dots,n\}$, and $W_i(t)\leq c_i^{(k_j^*)}$, $i\in\{j,\dots,n\}\setminus J$, for all $t\in[s,r]$, $s\geq t_0$. Integrating (A.39) over $[s,r]$ and using the induction step, by the definition of $c_{j-1}^{(k_j^*)}$ in (A.38) and the nonnegativity of W_{j-1}, we obtain for any set $J\subseteq\{j,\dots,n\}$

$$\int_s^t\frac{W_i(\lambda)}{W_i(\lambda)+1}d\lambda\leq\frac{2R_{j-1,1}}{l_{j-1}}\{W_{j-1}(s)+\varrho_{j-1}^{(k_j^*)}(0,W_j(s),\dots,W_n(s),\|w\|_2)\}\ \forall t\in[s,r],$$

which, upon setting $\xi_{j-1}^{(k_j^*)}(W_j(s),\dots,W_n(s),\|w\|_2):=[2R_{j-1,1}/l_{j-1}]\{W_{j-1}(s)+\varrho_{j-1}^{(k_j^*)}(0,W_j(s),\dots,W_n(s),\|w\|_2)\}$, gives (A.35) for all $i=j-1,\dots,n$ with $k=k_j^*+1$.

The case $j = 2$ can be carried out in the same way as the case $j = 3, \dots, n$ with the only difference being that

$$c_1^{(k_2^*)} := 8R_{11}^2 b_1 n^3 \alpha_1(2n) c_2^{(k_2^*)}/l_1 < 1. \tag{A.45}$$

By iterating the above arguments for $k > k_j^* + 1$ and using the bounds $W_i(z_i, \sigma_i) \le c_i^{(k)} \Rightarrow z_i^2, e_i^2, \widetilde{x}_{i+1}^2 \le 6c_i^{(k)}$, $i = j, \dots, n$, in (A.39), we obtain (A.34)–(A.35) for all $i = j-1, \dots, n$ and $k \ge k_j^*$ with $c_i^{(k)}$ defined by the following equations for $k \ge k_j^*$:

$$c_n^{(k+1)} := \frac{24R_{n1}^2 b_n c_{n-1}^{(k)}}{l_n R_{n-1,1}^4}, \quad c_i^{(k+1)} := \frac{4R_{i1}^2 b_i}{l_i}\left[\frac{6c_{i-1}^{(k)}}{R_{i-1,1}^4} + 2n^3\alpha_i(2n)c_{i+1}^{(k)}\right], \quad i = j, \dots, n, \tag{A.46}$$

and

$$\begin{aligned} c_{j-1}^{(k+1)} &:= \frac{4R_{j-1,1}^2 b_{j-1}}{l_{j-1}}\left[\frac{6}{R_{j-2,1}^4} + 2n^3\alpha_{j-1}(2n)c_j^{(k)}\right] \quad \text{if j} \ge 3, \\ c_1^{(k+1)} &:= \frac{8R_{11}^2 b_1 n^3 \alpha_1(2n) c_2^{(k)}}{l_1} \quad \text{if j} = 2, \end{aligned} \tag{A.47}$$

where the $c_i^{(k_j^*)}$, $i = j, \dots, n$, are defined (on account of the induction step) as

$$c_i^{(k_j^*)} \sim \prod_{l=j}^{i} \frac{R_{l1}^2}{R_{l-1,1}^4},\ i = j, \dots, n, \quad c_j^{(k_j^*)} \ge c_{j+1}^{(k_j^*)} \ge \cdots \ge c_n^{(k_j^*)}, \tag{A.48}$$

and $c_{j-1}^{(k_j^*)}$ as in (A.38) if $j \ge 3$ and $c_1^{(k_2^*)}$ as in (A.45) if $j = 2$. The equations (A.46) can be described by a linear discrete-time system $x(k+1) = Ax(k) + Bu(k)$ with state $x(k) = (c_n^{(k)} \ \cdots \ c_{j-1}^{(k)})^T$, initial condition $x(0) = (c_n^{(k_j^*)} \ \cdots \ c_{j-1}^{(k_j^*)})^T$, input $u(k) = 0$ if $j = 2$, and $24R_{j-1,1}^2 b_{j-1}/[l_{j-1}R_{j-2,1}^4]$ if $j \ge 3$. Thus, for all $k \ge k_j^* + 1$

$$\begin{aligned} &\begin{pmatrix} c_n^{(k)} & \cdots & c_{j-1}^{(k)} \end{pmatrix}^T = A^{k-k_j^*} \begin{pmatrix} c_n^{(k_j^*)} & \cdots & c_{j-1}^{(k_j^*)} \end{pmatrix}^T \\ &+[I + A + A^2 + \cdots + A^{k-k_j^*-1}] \begin{pmatrix} 0 & \cdots & u(k) \end{pmatrix}^T, \end{aligned}$$

where the last term is zero if $j = 2$. It is not difficult to prove that if $j \ge 3$ and for $1 \le r \le j-1$

$$[I + A + A^2 + \cdots + A^r] \begin{pmatrix} 0 \\ \vdots \\ u(k) \end{pmatrix} \sim \begin{pmatrix} 0 \\ \vdots \\ \displaystyle\prod_{h=j-1}^{r+j-1} \frac{R_{h1}^2}{R_{h-1,1}^4}\left[1 + \mathcal{O}\Big(\frac{1}{\varepsilon}\Big)\right] \\ \vdots \\ \displaystyle\frac{R_{j-1,1}^2}{R_{j-2,1}^4}\left[1 + \mathcal{O}\Big(\frac{1}{\varepsilon}\Big)\right] \end{pmatrix},$$

and for $1 \le r \le j-1$ and $s \ge 0$

$$A^{sj+r} \begin{pmatrix} 0 & \cdots & u(k) \end{pmatrix}^T \sim \varepsilon^{-2s} A^r \begin{pmatrix} 0 & \cdots & u(k) \end{pmatrix}^T, \tag{A.49}$$

where $\sim$ is meant componentwise. Moreover, for $1 \le r \le j-1$

$$A^r \begin{pmatrix} c_n^{(k_j^*)} & \cdots & c_{j-1}^{(k_j^*)} \end{pmatrix}^T$$
$$= \begin{pmatrix} \mathcal{O}\left(\varepsilon^{-(n-j+2)} \prod_{h=j}^{n} R_{h-1,1}^{-2}\right) & \cdots & \mathcal{O}\left(\varepsilon^{-2} R_{j-1,1}^{-2}\right) & \mathcal{O}\left(\varepsilon^{-1}\right) \end{pmatrix}^T$$

and for $1 \le r \le j-1$ and $s \ge 0$

$$A^{sj+r} \begin{pmatrix} c_n^{(k_j^*)} & \cdots & c_{j-1}^{(k_j^*)} \end{pmatrix}^T \sim \varepsilon^{-2s} A^r \begin{pmatrix} c_n^{(k_j^*)} & \cdots & c_{j-1}^{(k_j^*)} \end{pmatrix}^T. \tag{A.50}$$

From this follows the existence of $\varepsilon_{j-1}^* \ge \varepsilon_j^*$ and $k_{j-1}^* \ge k_j^*$ such that the induction step (A.34)–(A.35) and (A.37) hold true with j replaced by $j-1$, and $c_i^{(k)}$, $i = j-1, \dots, n$, satisfy (A.36) if $j \ge 3$; otherwise $c_1^{(k)} < 1$ if $j = 2$. Thus, since (A.34)–(A.35) and (A.37) hold true for $j = n-1$ with (A.31), they hold true for all $j = 1, \dots, n$. Moreover, $c_i^{(k)}$, $i = 2, \dots, n$, satisfy (A.36) and $c_1^{(k)} < 1$ for all $\varepsilon \ge \varepsilon_1^*$ and $k \ge k_1^*$. This also concludes the proof of Lemma 8.1 with $c_j := c_j^{(k_1^*)}$ and $\varrho_j := \varrho_j^{(k_1^*)}$, $j = 1, \dots, n$. □

Proof of Lemma 8.2. Let ε_i^*, ϱ_i, and c_i, $i = 1, \dots, n$, be as in Lemma 8.1. Moreover, let

$$\overline{c}_i(\varepsilon) := \varrho_i(c_i(\varepsilon), \dots, c_n(\varepsilon), 0),\ \mathcal{T}_j(0) := \{(Z_j, S_j) : W_i(z_i, \sigma_i) \le \overline{c}_i(\varepsilon),\ i = j, \dots, n\}$$

and set $w = 0$ (we are proving internal stability). Define recursively a filtered Lyapunov function for (1.1)–(1.4) by using Theorem 7.4. First, find a filtered Lyapunov function for $\Sigma_i \circ \mathcal{C}_i$, $i = n-1, n$; then proceed by induction on the number $n-j+1$, $j = 1, \dots, n-1$, of systems $\Sigma_i \circ \mathcal{C}_i$ in $\Sigma_i \circ \mathcal{C}_i$, $i = j, \dots, n$. By (8.15) and (8.17), with $j = n, n-1$, $\Sigma_n \circ \mathcal{C}_n$ has for all $\varepsilon \ge \max_i \varepsilon_i^*$ filtered Lyapunov function W_n, stability margins $\varphi_{ns}/2$, $\varphi_{nm}/2$, and incremental rates $\gamma_n^{z_{n-1}}$, $\gamma_n^{e_{n-1}}$ (given in (8.21)), while $\Sigma_{n-1} \circ \mathcal{C}_{n-1}$ has for all $\varepsilon \ge \max_i \varepsilon_i^*$ Lyapunov function W_{n-1}, stability margins $\varphi_{n-1,s}/2$, $\varphi_{n-1,m}/2$, and incremental rates $\gamma_{n-1}^{z_{n-2}}$, $\gamma_{n-1}^{e_{n-2}}$, $\gamma_{n-1}^{z_n}$, $\gamma_{n-1}^{e_n}$ (given in (8.21)–(8.22)). Choose $\varepsilon_{n-1}^{**} \ge \max_i \varepsilon_i^*$ such that for all $\varepsilon \ge \varepsilon_{n-1}^{**}$ the following hold:

(1) $\varphi_{ns}/2$ and $\varphi_{nm}/2$ locally saturate $\gamma_{n-1}^{z_n}$ and $\gamma_{n-1}^{e_n}$, respectively, with levels $(\overline{c}_n, \tau_{n-1}\ (\overline{c}_n))$, where $\tau_{n-1} : \mathbb{R}^{\ge} \to [1, \infty)$ is a continuous nondecreasing function such that $\gamma_{n-1}^{z_n} \le \tau_{n-1}(W_n)\varphi_{ns}/2$ and $\gamma_{n-1}^{e_n} \le \tau_{n-1}(W_n)\varphi_{nm}/2$ for all z_n, σ_n: by (8.22), with $j = n-1$ and $l = n$, and Lemma A.3 this can be done by taking $\tau_{n-1}(s) \sim R_{n-1,1} R_{n1} \widetilde{\tau}_{n-1}(s)$, with $\widetilde{\tau}_{n-1} : \mathbb{R}^{\ge} \to [1, \infty)$ a continuous nondecreasing function such that

$$[(1 + W_n(z_n, \sigma_n))^2 - 1][\gamma_{n-1,s}^{z_n} + \gamma_{n-1,m}^{z_n} + \gamma_{n-1,s}^{\widetilde{x}_{n+1}} + \gamma_{n-1,m}^{\widetilde{x}_{n+1}} + 1] \le \widetilde{\tau}_{n-1}(W_n(z_n, \sigma_n))$$

for all z_n, σ_n.

(2) $\varphi_{n-1,s}/2$ and $\varphi_{n-1,m}/2$ saturate $\gamma_n^{z_{n-1}}$ and $\gamma_n^{e_{n-1}}$ with level $1/[3\tau_{n-1}(\overline{c}_n)]$: this follows since $\widetilde{\tau}_{n-1}(\overline{c}_n) \preceq 1$ (by continuity of $\widetilde{\tau}_{n-1}(s)$ and $\overline{c}_n \preceq 1$) and from (8.21) with $j = n$, by (8.7) and (8.16) with $j = n$.

(3) $\Sigma_n \circ \mathcal{C}_n$ is recurrent relative to the trap $\mathcal{T}_n(0)$, and $\Sigma_{n-1} \circ \mathcal{C}_{n-1}$ is recurrent relative to the trap $\{(z_{n-1}, \sigma_{n-1}) : W_{n-1}(z_{n-1}, \sigma_{n-1}) \le \overline{c}_{n-1}\}$: this follows by Lemma 8.1.

Application of Theorem 7.4 to $\Sigma_i \circ \mathcal{C}_i$, $i = n-1, n$, with $c_1 \to \overline{c}_{n-1}, c_2 \to \overline{c}_n, \kappa_2 \to 1/[3\tau_{n-1}(\overline{c}_n)], \kappa_1 \to \tau_{n-1}(\overline{c}_n)$, $d_1 \to 1$, $d_2 \to 2\tau_{n-1}(\overline{c}_{n-1})$, $\theta_1, \theta_2 \to 1$, $\theta_0 \to \theta_{n-1}$,

$\tau_2(s) \to \kappa_n$, and $\tau_1(s) \to \tau_{n-1}(s)$, gives that

$$\begin{aligned}
&\widetilde{W}^{(n-1)}(Z_{n-1}, S_{n-1}, \theta_{n-1}) = \theta_{n-1} W^{(n-1)}(Z_{n-1}, S_{n-1}),\\
&W^{(n-1)}(Z_{n-1}, S_{n-1}) = W_{n-1}(z_{n-1}, \sigma_{n-1}) + 2\tau_{n-1}(\overline{c}_n) W_n(z_n, \sigma_n),\\
&\dot{\theta}_{n-1} = -[a_{n-1}(z_n, \sigma_n)/\min\{\overline{c}_{n-1}, 2\overline{c}_n \tau_{n-1}(\overline{c}_n)\}]\theta_{n-1},\ \theta_{n-1}(0) = e^{\int_0^\infty a_{n-1}(\tau)d\tau},
\end{aligned}$$

with $a_{n-1}(z_n, \sigma_n) = \max\{\tau_{n-1}(W_n(z_n, \sigma_n)) - \tau_{n-1}(\overline{c}_n), 0\}[\varphi_{ns} z_n^2/2 + \varphi_{nm} e_n^2/2]$, is for all $\varepsilon \geq \varepsilon_n^{**}$ a smooth filtered Lyapunov function for the interconnection $\Sigma_j \circ \mathcal{C}_j$, $j = n-1, n$, with

$$\dot{\widetilde{W}}^{(n-1)} \leq \theta_{n-1} \left\{ -\sum_{i=n-1}^{n} [\varphi_{is}^{(n-1)} z_i^2 + \varphi_{im}^{(n-1)} e_i^2] + \gamma_{n-1}^{z_{n-2}} z_{n-2}^2 + \gamma_{n-1}^{e_{n-2}} e_{n-2}^2 \right\} \tag{A.51}$$

along the trajectories of $\Sigma_j \circ \mathcal{C}_j$, $j = n-1, n$, where $\varphi_{n-1,l}^{(n-1)} \sim \varphi_{n-1,l}$ and $\varphi_{nl}^{(n-1)} \sim \tau_{n-1}(\overline{c}_n)\varphi_{nl}$, $l = s, m$. Moreover, the interconnection $\Sigma_j \circ \mathcal{C}_j$, $j = n-1, n$, is recurrent relative to the trap $\{(Z_{n-1}, S_{n-1}) : W^{(n-1)}(Z_{n-1}, S_{n-1}) \leq c^{(n-1)} := \overline{c}_{n-1} + 2\tau_{n-1}(\overline{c}_n)\}$ and for each trajectory of $\Sigma_j \circ \mathcal{C}_j$, $j = n-1, n$, there exists $T_{n-1} > 0$ such that

$$\theta_{n-1}(t) \geq 1 \quad \forall t \geq 0, \qquad \theta_{n-1}(t) = 1 \quad \forall t \geq T_{n-1}. \tag{A.52}$$

For each $j = 2, \ldots, n-1$ and $i = j, \ldots, n-1$ let

$$\begin{aligned}
&\varphi_{nl}^{(n)} \sim \varphi_{nl},\ \varphi_{il}^{(i)} \sim \varphi_{il},\ \varphi_{hl}^{(i)} \sim \tau_i(c^{(i+1)}) \cdots \tau_{n-1}(c^{(n)})\varphi_{hl}, \quad h = i+1, \ldots, n; l = s, m,\\
&\widetilde{W}^{(n)}(Z_n, S_n) = W_n(z_n, \sigma_n), \quad c^{(n)} = \overline{c}_n, \quad c^{(i)} = \overline{c}_i + 2\tau_i(c^{(i+1)})c^{(i+1)},
\end{aligned} \tag{A.53}$$

and

$$\begin{aligned}
&\widetilde{W}^{(j)}(Z_j, S_j, \widetilde{\theta}_j) = \widetilde{\theta}_j W^{(j)}(Z_j, S_j), \quad \widetilde{\theta}_j = \theta_j \cdots \theta_{n-1},\\
&W^{(i)}(Z_i, S_i) = W_i(z_i, \sigma_i) + 2\tau_i(c^{(i+1)}) W^{(i+1)}(Z_{i+1}, S_{i+1}),\\
&\dot{\theta}_i = -[a_i(Z_{i+1}, S_{i+1})/\min\{\overline{c}_i, 2c^{(i+1)}\tau_i(c^{(i+1)})\}]\theta_i, \quad \theta_i(0) = e^{\int_0^\infty a_i(\tau)d\tau},\\
&i = j, \ldots, n-1,
\end{aligned} \tag{A.54}$$

where $a_i(Z_{i+1}, S_{i+1}) = \max\{\tau_i(W^{(i+1)}(Z_{i+1}, S_{i+1})) - \tau_i(c^{(i+1)}), 0\} \sum_{l=i+1}^{n} [\varphi_{ls}^{(l+1)} z_l^2 + \varphi_{lm}^{(l+1)} e_l^2]$ and $\tau_i : \mathbb{R}^{\geq} \to [1, \infty)$, $i = j, \ldots, n-1$, are continuous nondecreasing functions.

Induction hypothesis. Let $j = 2, \ldots, n-1$. Assume the existence of $\varepsilon_j^{**} \geq \max_i \varepsilon_i^*$ such that for all $\varepsilon_j \geq \varepsilon_j^{**}$

$$\tau_i(s) \sim R_{i1} R_{i+1,1} \widetilde{\tau}_i(s), \quad i = j, \ldots, n-1, \tag{A.55}$$

with $\widetilde{\tau}_i, \tau_i : \mathbb{R}^{\geq} \to [1, \infty)$ continuous nondecreasing functions such that $\gamma_i^{z_{i+1}} \leq \tau_i(W^{(i+1)})\varphi_{i+1,s}^{(i)}$, $\gamma_i^{e_{i+1}} \leq \tau_i(W^{(i+1)})\varphi_{i+1,m}^{(i)}$ and

$$[(1 + W_{i+1}(z_{i+1}, \sigma_{i+1}))^2 - 1][\gamma_{is}^{z_{i+1}} + \gamma_{im}^{z_{i+1}} + \gamma_{is}^{\widetilde{x}_{i+2}} + \gamma_{im}^{\widetilde{x}_{i+2}} + 1] \leq \widetilde{\tau}_i(W^{(i+1)}(Z_{i+1}, S_{i+1}))$$

for all Z_{i+1}, S_{i+1}. Assume also that for all $\varepsilon \geq \varepsilon_j^{**}$

$$\dot{\widetilde{W}}^{(j)} \leq \widetilde{\theta}_j \left\{ -\sum_{i=j}^{n} [\varphi_{is}^{(j)} z_i^2 + \varphi_{im}^{(j)} e_i^2] + \gamma_j^{z_{j-1}} z_{j-1}^2 + \gamma_j^{e_{j-1}} e_{j-1}^2 \right\} \tag{A.56}$$

along the trajectories of $\Sigma_i \circ \mathcal{C}_i$, $i = j, \dots, n$, and, moreover, the interconnection $\Sigma_j \circ \mathcal{C}_j$, $i = j, \dots, n$, is recurrent relative to the trap $\{(Z_j, S_j) : W^{(j)}(Z_j, S_j) \le c^{(j)}\}$ and for each trajectory of $\Sigma_i \circ \mathcal{C}_i$, $i = j-1, \dots, n$, there exists $T_j > 0$ such that

$$\theta_i(t) \ge 1 \quad \forall t \ge 0, \quad \theta_i(t) = 1 \quad \forall t \ge T_j, \quad i = j, \dots, n-1. \tag{A.57}$$

Note that by (8.15) and (8.17), with j replaced by $j-1$, for all $\varepsilon \ge \varepsilon_j^{**}$

$$\begin{aligned}\dot W_{j-1} \le & -[\varphi_{j-1,s} z_{j-1}^2/2 + \varphi_{j-1,m} e_{j-1}^2/2] + \gamma_{j-1}^{z_{j-2}} z_{j-2}^2 + \gamma_{j-1}^{e_{j-2}} e_{j-2}^2 \\ & + \sum_{l=j}^{n} [\gamma_{j-1}^{z_l} z_l^2 + \gamma_{j-1}^{e_l} e_l^2]\end{aligned}$$

along the trajectories of $\Sigma_i \circ \mathcal{C}_i$, $i = j-1, \dots, n$. Choose $\varepsilon_{j-1}^{**} \ge \varepsilon_j^{**}$ such that for all $\varepsilon \ge \varepsilon_{j-1}^{**}$ the following hold:

(1) $\varphi_{is}^{(j)}/2$ and $\varphi_{im}^{(j)}/2$, $i = j, \dots, n$, locally saturate $\gamma_{j-1}^{z_i}$ and $\gamma_{j-1}^{e_i}$, respectively, with levels $(c^{(j)}, \tau_{j-1}(c^{(j)}))$, where $\tau_{j-1} : \mathbb{R}^{\ge} \to [1, \infty)$ is a continuous nondecreasing function such that $\gamma_{j-1}^{z_j} \le \tau_{j-1}(W^{(j)})\varphi_{js}/2$ and $\gamma_{j-1}^{e_j} \le \tau_{j-1}(W^{(j)})\varphi_{jm}/2$ for all Z_j, S_j: by Lemma A.3 and since $\tau_i(s) \ge 1$ for all s, $i = j, \dots, n$, this can be done by taking $\tau_{j-1}(s) \sim R_{j-1,1} R_{j1} \widetilde\tau_{j-1}(s)$, with $\widetilde\tau_{j-1} : \mathbb{R}^{\ge} \to [1, \infty)$ a continuous nondecreasing function such that for all Z_j, S_j

$$[(1 + W_j(z_j, \sigma_j))^2 - 1][\gamma_{j-1,s}^{z_j} + \gamma_{j-1,m}^{z_j} + \gamma_{j-1,s}^{\widetilde x_{j+1}} + \gamma_{j-1,m}^{\widetilde x_{j+1}} + 1] \le \widetilde\tau_{j-1}(W^{(j)}(Z_j, S_j)). \tag{A.58}$$

(2) $\varphi_{j-1,s}/2$ and $\varphi_{j-1,m}/2$ saturate $\gamma_j^{z_{j-1}}$ and $\gamma_j^{e_{j-1}}$ with level $1/[3\tau_{j-1}(c^{(j)})]$: this follows from (A.55), (8.21), and (A.58) and since $\widetilde\tau_{j-1} \preceq 1$.

(3) $\Sigma_i \circ \mathcal{C}_i$, $i = j, \dots, n$, is recurrent relative to the trap $\{(Z_j, S_j) : W^{(j)}(Z_j, S_j) \le c_i^{(j)}\}$, and $\Sigma_{j-1} \circ \mathcal{C}_{j-1}$ is recurrent relative to the trap $\{(z_{j-1}, \sigma_{j-1}) : W_{j-1}(z_{j-1}, \sigma_{j-1}) \le \bar c_{j-1}\}$: this follows from Lemma 8.1 and the induction step.

By application of Theorem 7.4 to $\Sigma_i \circ \mathcal{C}_i$, $i = j-1, \dots, n$, we prove that there exists $\varepsilon_{j-1}^{**} \ge \varepsilon_{j-1}^{*}$ such that (A.53)–(A.57) hold for all $\varepsilon \ge \varepsilon_{j-1}^{**}$ with j replaced by $j-1$, the interconnection $\Sigma_j \circ \mathcal{C}_j$, $i = j-1, \dots, n$, is recurrent relative to the trap $\{(Z_{j-1}, S_{j-1}) : W^{(j-1)}(Z_{j-1}, S_{j-1}) \le c^{(j-1)}\}$, and for each trajectory of $\Sigma_i \circ \mathcal{C}_i$, $i = j-1, \dots, n$, there exist $T_{j-1} > 0$ such that $\theta_{j-1}(t) \ge 1$ for all $t \ge 0$ and $\theta_{j-1}(t) = 1$ for all $t \ge T_{j-1}$. Thus, since the induction step holds true for $j = n-1$, it holds true for all $j = 1, \dots, n-1$. This completes the proof of Lemma 8.2 with $\varepsilon^{**} = \varepsilon_1^{**}$, $\widetilde W = \widetilde W^{(1)}$, $\theta = \widetilde\theta_1$, and $\widetilde\varphi_{il} = \varphi_{il}^{(1)}$. □

Acknowledgment. The author wishes to thank the reviewers for their suggestions for improving the presentation of the paper.

REFERENCES

[1] D. Angeli, E. D. Sontag, and Y. Wang, *Further equivalences and semiglobal versions of integral input to state stability*, Dynam. Control, 10 (2000), pp. 127–149.

[2] S. Battilotti, *Robust stabilization of nonlinear systems with pointwise norm bounded uncertainty: A control Lyapunov function approach*, IEEE Trans. Automat. Control, 44 (1999), pp. 1–15.

[3] S. Battilotti, *Lyapunov design of global measurement feedback controllers for nonlinear systems*, in Proceedings of the 5th Annual IFAC Symposium on Nonlinear Control Systems, St. Petersburg, Russia, 2001.

[4] S. Battilotti, *A separation result for systems with feedback constraints*, Systems Control Lett., 55 (2006), pp. 369–375.
[5] Z. Chen and J. Huang, *Dissipativity, stabilization, and regulation of cascade-connected systems*, IEEE Trans. Automat. Control, 49 (2004), pp. 635–650.
[6] R. A. Freeman and L. Praly, *Integrator backstepping for bounded controls and rates*, IEEE Trans. Automat. Control, 43 (1998), pp. 258–262.
[7] M. Jankovic, R. Sepulchre, and P. Kokotovic, *Constructive Lyapunov stabilization of nonlinear cascade systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 1723–1735.
[8] Z. P. Jiang and Y. Mareels, *A small gain control method for nonlinear cascaded systems with dynamic uncertainty*, IEEE Trans. Automat. Control, 42 (1999), pp. 292–308.
[9] H. K. Khalil, *Nonlinear Systems*, 1st ed., Macmillan, New York, 1992.
[10] W. Lin and C. Qian, *Adaptive control of nonlinearly parametrized systems: A nonsmooth feedback framework*, IEEE Trans. Automat. Control, 47 (2002), pp. 757–774.
[11] F. Mazenc and A. Iggidr, *Backstepping with bounded feedbacks for systems in feedback form*, in Proceedings of the 5th Annual IFAC Symposium on Nonlinear Control Systems, St. Petersburg, Russia, 2001.
[12] F. Mazenc and L. Praly, *Adding integrations, saturated controls, and stabilization for feed forward systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 1559–1578.
[13] R. Sepulchre, M. Jankovic, and P. Kokotovic, *Constructive Nonlinear Control*, Springer-Verlag, Berlin, 1996.
[14] A. R. Teel, *On $\mathcal{L}_p$ performance induced by feedback with multiple saturations*, ESIAM Control Optim. Calc. Var., 1 (1997), pp. 225–240.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 116–142

ADAPTIVE SPACE-TIME FINITE ELEMENT METHODS FOR PARABOLIC OPTIMIZATION PROBLEMS*

DOMINIK MEIDNER† AND BORIS VEXLER‡

Abstract. In this paper we derive a posteriori error estimates for space-time finite element discretizations of parabolic optimization problems. The provided error estimates assess the discretization error with respect to a given quantity of interest and separate the influences of different parts of the discretization (time, space, and control discretization). This allows us to set up an efficient adaptive algorithm which successively improves the accuracy of the computed solution by construction of locally refined meshes for time and space discretizations.

Key words. parabolic equations, optimal control, parameter identification, a posteriori error estimation, mesh refinement

AMS subject classifications. 65N30, 49K20, 65M50, 35K55

DOI. 10.1137/060648994

1. Introduction.

In this paper we develop an adaptive algorithm for efficient solution of time-dependent optimization problems governed by parabolic partial differential equations. The optimization problems are formulated in a general setting including optimal control as well as parameter identification problems. Both, time and space discretization of the state equation are based on the finite element method as proposed, e.g., in [10, 11]. In [2] we have shown that this type of discretization allows for a natural translation of the optimality conditions from the continuous to the discrete level. This gives rise to exact computation of the derivatives required in the optimization algorithms on the discrete level.

The main goal of this paper is to derive a posteriori error estimates which assess the error between the solution of the continuous and the discrete optimization problem with respect to a given quantity of interest. This quantity of interest may coincide with the cost functional or express another goal for the computation. In order to set up an efficient adaptive algorithm we will separate the influences of the time and space discretizations on the error in the quantity of interest. This allows us to balance different types of errors and successively to improve the accuracy by construction of locally refined meshes for time and space discretizations.

The use of adaptive techniques based on a posteriori error estimation is well accepted in the context of finite element discretization of partial differential equations; see, e.g., [9, 28, 3]. In the past several years the application of these techniques has also been investigated for optimization problems governed by partial differential equations. Energy-type error estimators for the error in the state, control, and adjoint variable are developed in [20, 21] in the context of distributed elliptic optimal control problems subject to pointwise control constraints. Recently, these techniques were also applied in the context of optimal control problems governed by linear parabolic equations; see [19]. In a recent preprint [24] an anisotropic error estimate is derived

*Received by the editors January 4, 2006; accepted for publication (in revised form) October 25, 2006; published electronically March 22, 2007.
http://www.siam.org/journals/sicon/46-1/64899.html

†Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, INF 294, 69120 Heidelberg, Germany (dominik.meidner@iwr.uni-heidelberg.de).

‡Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenberger Straße 69, 4040 Linz, Austria (boris.vexler@oeaw.ac.at).

for the error due to the space discretization of an optimal control problem governed by the linear heat equation.

However, in many applications, the error in global norms does not provide useful error bounds for the error in the quantity of physical interest. In [1, 3] a general concept for a posteriori estimation of the discretization error with respect to the cost functional in the context of optimal control problems is presented. In papers [4, 5], this approach is extended to the estimation of the discretization error with respect to an arbitrary functional depending on both the control and state variables, i.e., with respect to a quantity of interest. This allows, among other things, an efficient treatment of parameter identification and model calibration problems.

The main contribution of this paper is the extension of these approaches to optimization problems governed by parabolic partial differential equations.

In this paper, we consider optimization problems under constraints of (nonlinear) parabolic differential equations

$$\begin{aligned} \partial_t u + A(q,u) &= f \\ u(0) &= u_0(q). \end{aligned} \tag{1.1}$$

Here, the state variable is denoted by u and the control variable by q. Both, the differential operator A and the initial condition u_0 may depend on q. This allows a simultaneous treatment of both optimal control and parameter identification problems. For optimal control problems, the operator A is typically given by

$$A(q,u) = \bar{A}(u) - B(q),$$

with a (nonlinear) operator $\bar{A}$ and a (usually linear) control operator B. In parameter identification problems, the variable q denotes the unknown parameters to be determined and may enter the operator A in a nonlinear way. The case of initial control is included via the q-dependent initial condition $u_0(q)$.

The target of the optimization is to minimize a given cost functional $J(q,u)$ subject to the state equation (1.1).

For the numerical solution of this optimization problem the state variable has to be discretized in space and in time. Moreover, if the control (parameter) space is infinite dimensional, it has to be discretized too. For fixed time, space, and control discretizations this leads to a finite dimensional optimization problem. We introduce σ as a general discretization parameter including the space, time, and control discretizations and denote the solution of the discrete problem by (q_σ, u_σ). For this discrete solution we derive an a posteriori error estimate with respect to the cost functional J of the following form:

$$J(q,u) - J(q_\sigma, u_\sigma) \approx \eta_k^J + \eta_h^J + \eta_d^J. \tag{1.2}$$

Here, η_k^J, η_h^J, and η_d^J denote the error estimators, which can be evaluated from the computed discrete solution; η_k^J assesses the error due to the time discretization, η_h^J due to the space discretization, and η_d^J due to the discretization of the control space. The structure of the error estimate (1.2) allows for equilibration of different discretization errors within an adaptive refinement algorithm to be described in the following discussion.

For many optimization problems the quantity of physical interest coincides with the cost functional, which explains the choice of the error measure (1.2). However, in the case of parameter identification or model calibration problems, the cost functional is only an instrument for the estimation of the unknown parameters. Therefore, the

value of the cost functional in the optimum and the corresponding discretization error are of secondary importance. This motivates error estimation with respect to a given functional I depending on the state and control (parameter) variables. In this paper we extend the corresponding results from [4, 5, 29] to parabolic problems and derive an a posteriori error estimator of the form

$$I(q,u) - I(q_\sigma, u_\sigma) \approx \eta_k^I + \eta_h^I + \eta_d^I,$$

where again η_k^I and η_h^I estimate the temporal and spatial discretization errors and η_d^I estimates the discretization error due to the discretization of the control space.

In section 5.2 we will describe an adaptive algorithm based on these error estimators. Within this algorithm the time, space, and control discretizations are separately refined for efficient reduction of the total error equilibrating different types of the error. This local refinement relies on the computable representation of the error estimators as a sum of local contributions (error indicators), see the discussion in section 5.1.

To the authors' knowledge, this is the first paper describing the a posteriori error estimation for optimization problems governed by parabolic differential equations including the separation of different types of the discretization error.

The outline of the paper is as follows: In the next section we describe necessary optimality conditions for the problem under consideration and sketch the Newton-type optimization algorithm on the continuous level. This algorithm will be applied on the discrete level for fixed discretizations within an adaptive refinement procedure. In section 3 we present the space-time finite element discretization of the optimization problem. Section 4 is devoted to the derivation of the error estimators in a general setting. In section 5 we discuss numerical evaluation of these error estimators and the adaptive algorithm in details. In the last section we present two numerical examples illustrating the behavior of the proposed methods. The first example deals with boundary control of the heat equation, whereas the second one is concerned with the identification of Arrhenius parameters in a simplified gaseous combustion model by means of point measurements of the concentrations.

2. Optimization. The optimization problems considered in this paper are formulated in the following abstract setting: Let Q be a Hilbert space for the controls (parameters) with scalar product $(\cdot,\cdot)_Q$. Moreover, let V and H be Hilbert spaces, which build together with the dual space V^* of V a Gel'fand triple $V \hookrightarrow H \hookrightarrow V^*$. The duality pairing between the Hilbert spaces V and its dual V^* is denoted by $\langle\cdot,\cdot\rangle_{V^*\times V}$, and the scalar product in H is denoted by $(\cdot,\cdot)_H$. A typical choice for these spaces could be

$$V = \left\{ v \in H^1(\Omega) \,\middle|\, v\big|_{\partial\Omega_{\mathrm{D}}} = 0 \right\} \text{ and } H = L^2(\Omega), \tag{2.1}$$

where $\partial\Omega_{\mathrm{D}}$ denotes the part of the boundary of Ω with prescribed Dirichlet boundary conditions.

For a time interval $(0,T)$ we introduce the Hilbert space $X := W(0,T)$ defined as

$$W(0,T) = \left\{ v \,\middle|\, v \in L^2((0,T),V) \text{ and } \partial_t v \in L^2((0,T),V^*) \right\}. \tag{2.2}$$

It is well known that the space X is continuously embedded in $C([0,T],H)$; see, e.g., [8]. Furthermore, we use the inner product of $L^2((0,T),H)$ given by

$$(u,v) := (u,v)_{L^2((0,T),H)} = \int_0^T (u(t),v(t))_H \, dt \tag{2.3}$$

for setting up the weak formulation of the state equation.

By means of the spatial semilinear form $\bar{a}\colon Q\times V\times V \to \mathbb{R}$ defined for a differential operator $A\colon Q \times V \to V^*$ by

$$\bar{a}(q,\bar{u})(\bar{\varphi}) := \langle A(q,\bar{u}),\bar{\varphi}\rangle_{V^*\times V},$$

we can define the semilinear form $a(\cdot,\cdot)(\cdot)$ on $Q \times X \times X$ as

$$a(q,u)(\varphi) := \int_0^T \bar{a}(q,u(t))(\varphi(t))\,dt$$

which is assumed to be three times Gâteaux differentiable and linear in the third argument.

Remark 2.1. If the control variable q depends on time, this has to be incorporated by an obvious modification of the definitions of the semilinear forms.

After these preliminaries, we pose the *state equation* in a weak form: Find for given control $q \in Q$ the *state variable* $u \in X$ such that

$$\begin{aligned}(\partial_t u,\varphi) + a(q,u)(\varphi) &= (f,\varphi) \quad \forall \varphi \in X,\\ u(0) &= u_0(q),\end{aligned} \tag{2.4}$$

where $f \in L^2((0,T),V^*)$ represents the right-hand side of the state equation and $u_0\colon Q \to H$ denotes a three times Gâteaux differentiable mapping describing parameter-dependent initial conditions. The usage of the inner product $(\cdot,\cdot)$ defined in (2.3) for stating the formulation (2.4) is possible since the inner product on H is an equivalent representation of the duality pairing of V and V^* due to the properties of the Gel'fand triple.

Remark 2.2. There are several sets of assumptions on the nonlinearity in $\bar{a}(\cdot,\cdot)(\cdot)$ and its dependence on the control variable q allowing the state equation (2.4) to be well-posed. Typical examples are different semilinear equations, where the form $\bar{a}(\cdot,\cdot)(\cdot)$ consists of a linear elliptic part and a nonlinear term depending on u and ∇u. Due to the fact that the development of the proposed adaptive algorithm does not depend on the particular structure of the nonlinearity in $\bar{a}$, we do not specify a set of assumptions on it but assume that the state equation (2.4) possesses a unique solution $u = S(q) \in X$ for each $q \in Q$.

The cost functional $J\colon Q \times X \to \mathbb{R}$ is defined using two three times Gâteaux differentiable functionals $J_1\colon V \to \mathbb{R}$ and $J_2\colon H \to \mathbb{R}$ by

$$J(q,u) = \int_0^T J_1(u)\,dt + J_2(u(T)) + \frac{\alpha}{2}\|q-\bar{q}\|_Q^2, \tag{2.5}$$

where the regularization (or cost) term is added which involves $\alpha \geq 0$ and a reference parameter $\bar{q} \in Q$.

The corresponding optimization problem is formulated as follows:

$$\text{Minimize } J(q,u) \text{ subject to the state equation (2.4), } (q,u) \in Q \times X. \tag{2.6}$$

The question of existence and uniqueness of solutions to such optimization problems is discussed, e.g., in [18, 13, 27]. Throughout the paper, we assume problem (2.6) to admit a (locally) unique solution. Moreover, we assume the existence of a neighborhood $W \subset Q\times X$ of the optimal solution, such that the linearized form $\bar{a}'_u(q,u(t))(\cdot,\cdot)$ considered as a linear operator

$$\bar{a}'_u(q,u(t))\colon V \to V^*$$

is an isomorphism for all $(q,u) \in W$ and almost all $t \in (0,T)$. This assumption will allow all considered linearized and adjoint problems to be well-posed.

Provided the existence of a solution operator $S\colon Q \to X$ for the state equation (2.4) (see Remark 2.2), we can define the reduced cost functional $j\colon Q \to \mathbb{R}$ by $j(q) = J(q, S(q))$. This definition allows us to reformulate problem (2.6) as an unconstrained optimization problem:

$$\text{Minimize } j(q), \quad q \in Q. \tag{2.7}$$

We assume the solution operator S to be two times differentiable; see, e.g., [27] for a discussion of this issue.

For the reduced optimization problem (2.7) we apply Newton's method to reach a control q which satisfies the first order necessary optimality condition

$$j'(q)(\tau q) = 0 \quad \forall \tau q \in Q.$$

Starting with an initial guess q^0, the next Newton iterate is obtained by $q^{i+1} = q^i + \delta q$, where the update $\delta q \in Q$ is the solution of the linear problem:

$$j''(q)(\delta q, \tau q) = -j'(q)(\tau q) \quad \forall \tau q \in Q. \tag{2.8}$$

Thus, we need suitable expressions for the first and second derivatives of the reduced cost functional j. To this end, we introduce the Lagrangian $\mathcal{L}\colon Q \times X \times X \to \mathbb{R}$, defined as

$$\mathcal{L}(q,u,z) = J(q,u) + (f - \partial_t u, z) - a(q,u)(z) - (u(0) - u_0(q), z(0))_H. \tag{2.9}$$

With its aid, we obtain the following standard representation of the first derivative $j'(q)(\tau q)$.

THEOREM 2.1.

- *If for given $q \in Q$ the state $u \in X$ fulfills the state equation*

$$\mathcal{L}'_z(q,u,z)(\varphi) = 0 \quad \forall \varphi \in X,$$

  with $(q,u) \in W \subset Q \times X$,
- *and if additionally $z \in X$ is chosen as a solution of the adjoint state equation*

$$\mathcal{L}'_u(q,u,z)(\varphi) = 0 \quad \forall \varphi \in X,$$

then the following expression of the first derivative of the reduced cost functional holds:

$$\begin{aligned} j'(q)(\tau q) &= \mathcal{L}'_q(q,u,z)(\tau q) \\ &= \alpha(q - \bar{q}, \tau q)_Q - a'_q(q,u)(\tau q, z) + (u'_0(q)(\tau q), z(0))_H. \end{aligned}$$

Remark 2.3. The optimality system of the considered optimization problem (2.6) is given by the derivatives of the Lagrangian used in Theorem 2.1 above:

$$\begin{aligned} &\mathcal{L}'_z(q,u,z)(\varphi) = 0 \quad \forall \varphi \in X && \text{(State equation)}, \\ &\mathcal{L}'_u(q,u,z)(\varphi) = 0 \quad \forall \varphi \in X && \text{(Adjoint state equation)}, \\ &\mathcal{L}'_q(q,u,z)(\psi) = 0 \quad \forall \psi \in Q && \text{(Gradient equation)}. \end{aligned} \tag{2.10}$$

For the explicit formulation of the dual equation in this setting, see, e.g., [2].

In the same manner one can gain representations of the second derivatives of j in terms of the Lagrangian; see, e.g., [2] where two different kinds of expressions are discussed: Either one can build up the whole Hessian and solve the system (2.8) by an arbitrary linear solver, or one can just compute matrix-vector products of the Hessian times a given vector and use this to solve (2.8) by the conjugate gradient method.

The presented Newton's method will be used to solve discrete optimization problems arising from discretizing the states and the controls as, e.g., shown in the following section. In practical realizations, Newton's method has to be combined with some globalization techniques such as line search or trust region to enlarge its area of convergence; see, e.g., [23, 7].

Remark 2.4. The solution u of the underlying state equation is typically required in the whole time interval for the computation of the adjoint solution z. If all data are stored, the storage grows linearly with respect to the number of time intervals in the time discretization. For reducing the required memory one can apply checkpointing techniques; see, e.g., [15, 14]. In [2] we analyze such a strategy in the context of space-time finite element discretization of parabolic optimization problems.

3. Discretization. In this section, we discuss the discretization of the optimization problem (2.6). To this end, we use Galerkin finite element methods in space and time to discretize the state equation. This allows us to give a natural computable representation of the discrete gradient and Hessian in the same manner as shown in section 2 for the continuous problem. The use of exact discrete derivatives is important for the convergence of the optimization algorithms. Moreover, our systematic approach to a posteriori error estimation relies on using the Galerkin-type discretizations.

The first of the following subsections is devoted to semidiscretization in time by *continuous Galerkin* (cG) and *discontinuous Galerkin* (dG) methods. Section 3.2 deals with the space discretization of the semidiscrete problems arising from time discretization. For the numerical analysis of these schemes we refer to [10].

The discretization of the control space Q is kept rather abstract by choosing a finite dimensional subspace $Q_d \subset Q$. A possible concretion of this choice is shown in the numerical examples in section 6. For the variational discretization concept, where the control variable is not discretized explicitly, we refer to [16]; for a superconvergence based discretization of the control variable, see [22].

3.1. Time discretization of the states. To define a semidiscretization in time, let us partition the time interval $[0,T]$ as

$$[0,T] = \{0\} \cup I_1 \cup I_2 \cup \cdots \cup I_M$$

with subintervals $I_m = (t_{m-1}, t_m]$ of size k_m and time points

$$0 = t_0 < t_1 < \cdots < t_{M-1} < t_M = T.$$

We define the discretization parameter k as a piecewise constant function by setting $k\big|_{I_m} = k_m$ for $m = 1, \ldots, M$.

By means of the subintervals I_m, we define for $r \in \mathbb{N}_0$ two semidiscrete spaces X_k^r and $\widetilde{X}_k^r$:

$$X_k^r = \left\{ v_k \in C([0,T], H) \,\middle|\, v_k\big|_{I_m} \in \mathcal{P}^r(I_m, V) \right\} \subset X,$$
$$\widetilde{X}_k^r = \left\{ v_k \in L^2((0,T), V) \,\middle|\, v_k\big|_{I_m} \in \mathcal{P}^r(I_m, V) \text{ and } v_k(0) \in H \right\}.$$

Here, $\mathcal{P}^r(I_m, V)$ denotes the space of polynomials up to order r defined on I_m with values in V. Thus, X_k^r consists of piecewise polynomials which are continuous in time and will be used as trial space in the cG method, whereas the functions in $\widetilde{X}_k^r$ may have discontinuities at the edges of the subintervals I_m. This space will be used in what follows as test space in the cG method and as trial and test space in the dG method.

3.1.1. Continuous Galerkin methods. Using the semidiscrete spaces defined above, the cG(r) formulation of the state equation can be directly stated as follows: Find for given control $q_k \in Q$ a state $u_k \in X_k^r$ such that

$$\begin{aligned} (\partial_t u_k, \varphi) + a(q_k, u_k)(\varphi) &= (f, \varphi) \quad \forall \varphi \in \widetilde{X}_k^{r-1}, \\ u_k(0) &= u_0(q_k). \end{aligned} \tag{3.1}$$

Remark 3.1. This equation is assumed to posses a unique solution for each $q \in Q$, cf. Remark 2.2. In special cases the existence and uniqueness can be shown by separation of variables and by using the fact that $\widetilde{X}_k^r$ is finite dimensional with respect to time.

The corresponding semidiscretized optimization problem reads

(3.2) Minimize $J(q_k, u_k)$ subject to the state equation (3.1), $(q_k, u_k) \in Q \times X_k^r$.

Since the state equation semidiscretized by the cG(r) method has the same form as in the continuous setting, the corresponding Lagrangian is analogically defined on $Q \times X_k^r \times \widetilde{X}_k^{r-1}$ as

$$\mathcal{L}(q_k, u_k, z_k) = J(q_k, u_k) + (f - \partial_t u_k, z_k) - a(q_k, u_k)(z_k) - (u_k(0) - u_0(q_k), z_k(0))_H.$$

3.1.2. Discontinuous Galerkin methods. To define the dG(r) discretization we employ the following definition for functions $v_k \in \widetilde{X}_k^r$:

$$v_{k,m}^+ := \lim_{t \to 0^+} v_k(t_m + t), \quad v_{k,m}^- := \lim_{t \to 0^+} v_k(t_m - t) = v_k(t_m), \quad [v_k]_m := v_{k,m}^+ - v_{k,m}^-.$$

Then, the dG(r) semidiscretization of the state equation (2.4) reads as follows: Find for given control $q_k \in Q$ a state $u_k \in \widetilde{X}_k^r$ such that

$$\begin{aligned} \sum_{m=1}^{M} \int_{I_m} (\partial_t u_k, \varphi)_H \, dt + a(q_k, u_k)(\varphi) + \sum_{m=0}^{M-1} ([u_k]_m, \varphi_m^+)_H &= (f, \varphi) \quad \forall \varphi \in \widetilde{X}_k^r, \\ u_{k,0}^- &= u_0(q_k). \end{aligned} \tag{3.3}$$

This equation is assumed to be well-posed, cf. Remark 3.1.

The semidiscrete optimization problem for the dG(r) time discretization has the form

(3.4) Minimize $J(q_k, u_k)$ subject to the state equation (3.3), $(q_k, u_k) \in Q \times \widetilde{X}_k^r$.

Then we pose the Lagrangian $\widetilde{\mathcal{L}}\colon Q \times \widetilde{X}_k^r \times \widetilde{X}_k^r \to \mathbb{R}$ associated with the dG(r) time discretization for the state equation as

$$\begin{aligned} \widetilde{\mathcal{L}}(q_k, u_k, z_k) = J(q_k, u_k) + (f, z_k) - \sum_{m=1}^{M} \int_{I_m} (\partial_t u_k, z_k)_H \, dt \\ - a(q_k, u_k)(z_k) - \sum_{m=0}^{M-1} ([u_k]_m, z_{k,m}^+)_H - (u_{k,0}^- - u_0(q_k), z_{k,0}^-)_H. \end{aligned}$$

3.2. Space discretization of the states. In this subsection, we first describe the finite element discretization in space. To this end, we consider two- or three-dimensional shape-regular meshes; see, e.g., [6]. A mesh consists of quadrilateral or hexahedral cells K, which constitute a nonoverlapping cover of the computational domain $\Omega \subset \mathbb{R}^n$, $n \in \{2, 3\}$. The corresponding mesh is denoted by $\mathcal{T}_h = \{K\}$, where we define the discretization parameter h as a cellwise constant function by setting $h\big|_K = h_K$ with the diameter h_K of the cell K.

On the mesh $\mathcal{T}_h$ we construct a conform finite element space $V_h \subset V$ in a standard way:

$$V_h^s = \left\{ v \in V \,\middle|\, v\big|_K \in \mathcal{Q}^s(K) \text{ for } K \in \mathcal{T}_h \right\}.$$

Here, $\mathcal{Q}^s(K)$ consists of shape functions obtained via bi- or trilinear transformations of polynomials in $\widehat{\mathcal{Q}^s}(\widehat{K})$ defined on the reference cell $\widehat{K} = (0,1)^n$.

To obtain the fully discretized versions of the time discretized state equations (3.1) and (3.3), we utilize the space-time finite element spaces

$$X_{k,h}^{r,s} = \left\{ v_{kh} \in C([0,T], V_h^s) \,\middle|\, v_{kh}\big|_{I_m} \in \mathcal{P}^r(I_m, V_h^s) \right\} \subset X_k^r$$

and

$$\widetilde{X}_{k,h}^{r,s} = \left\{ v_{kh} \in L^2((0.T), V_h^s) \,\middle|\, v_{kh}\big|_{I_m} \in \mathcal{P}^r(I_m, V_h^s) \text{ and } v_{kh}(0) \in V_h^s \right\} \subset \widetilde{X}_k^r.$$

Remark 3.2. By the above definition of the discrete spaces $X_{k,h}^{r,s}$ and $\widetilde{X}_{k,h}^{r,s}$, we have assumed that the spatial discretization is fixed for all time intervals. However, in many application problems the use of different meshes $\mathcal{T}_h^m$ for each of the subintervals I_m will lead to more efficient adaptive discretizations. The consideration of such dynamically changing meshes can be included in the formulation of the dG(r) schemes in a natural way. The corresponding formulation of the cG(r) method is more involved due to the continuity requirement in the trial space. The treatment of dynamic meshes for the forward simulation of parabolic problems within an adaptive algorithm is discussed in [26]. It will be analyzed in a forthcoming paper in the context of parabolic optimization problems.

Then, the so-called cG(s)cG(r) discretization of the state equation (2.4) can be stated as follows: Find for given control $q_{kh} \in Q$ a state $u_{kh} \in X_{k,h}^{r,s}$ such that

$$(3.5)\quad (\partial_t u_{kh}, \varphi) + a(q_{kh}, u_{kh})(\varphi) + (u_{kh}(0), \varphi(0))_H = (f, \varphi) + (u_0(q_{kh}), \varphi(0))_H \quad \forall \varphi \in \widetilde{X}_{k,h}^{r-1,s}.$$

The cG(s)dG(r) discretization has the following form: Find for given control $q_{kh} \in Q$ a state $u_{kh} \in \widetilde{X}_{k,h}^{r,s}$ such that

$$(3.6)\quad \sum_{m=1}^{M} \int_{I_m} (\partial_t u_{kh}, \varphi)_H \, dt + a(q_{kh}, u_{kh})(\varphi) + \sum_{m=0}^{M-1} ([u_{kh}]_m, \varphi_m^+)_H + (u_{kh,0}^-, \varphi_0^-)_H = (f, \varphi) + (u_0(q_{kh}), \varphi_0^-)_H \quad \forall \varphi \in \widetilde{X}_{k,h}^{r,s}.$$

These fully discretized state equations are assumed to posses unique solutions for each $q_{kh} \in Q$; see Remark 3.1.

Thus, the optimization problems with fully discretized states are given by

(3.7)
Minimize $J(q_{kh}, u_{kh})$ subject to the state equation (3.5), $(q_{kh}, u_{kh}) \in Q \times X_{k,h}^{r,s}$,

for the cG(s)cG(r) discretization and by

(3.8)
Minimize $J(q_{kh}, u_{kh})$ subject to the state equation (3.6), $(q_{kh}, u_{kh}) \in Q \times \widetilde{X}_{k,h}^{r,s}$,

for the cG(s)dG(r) discretization of the state space.

The definition of the Lagrangians $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ for fully discretized states can be directly transferred from the formulations for semidiscretization in time just by restriction of the state spaces X_k^r and $\widetilde{X}_k^r$ to the subspaces $X_{k,h}^{r,s}$ and $\widetilde{X}_{k,h}^{r,s}$, respectively. With the aid of these Lagrangians, the derivatives of the reduced functionals $j_k(q_k) = J(q_k, S_k(q_k))$ and $j_{kh}(q_{kh}) = J(q_{kh}, S_{kh}(q_{kh}))$ on the different discretization levels can be expressed in the same manner as described on the continuous level in Theorem 2.1. Thus, we obtain exact derivatives of the reduced cost functional on the discrete level; see [2] for details.

Remark 3.3. The dG(r) and cG(r) schemes are known to be time discretization schemes of order $r+1$. The cG(r) schemes lead to a A-stable discretization whereas the dG(r) schemes are even strongly A-stable.

Remark 3.4. Due to the fact that the test space is discontinuous in time for both dG(r) and cG(r) discretization, these methods (although globally formulated) can be interpreted as time-stepping schemes. To illustrate this fact, we present the time-stepping scheme for the low order cG(s)dG(0) method: For the state equation we obtain with the abbreviations $U_0 := u_{kh}(0)$ and $U_m := u_{hk}\big|_{I_m}$ for $m = 1, \dots, M$ the following time-stepping formulation:

- $m = 0$:

$$(U_0, \varphi)_H = (u_0(q), \varphi)_H \quad \forall \varphi \in V_h^s,$$

- $m = 1, \dots, M$:

$$(U_m, \varphi)_H + k_m \bar{a}(q, U_m)(\varphi) = (U_{m-1}, \varphi)_H + \int_{I_m} (f(t), \varphi)_H \, dt \quad \forall \varphi \in V_h^s.$$

This scheme is a variant of the implicit Euler scheme. If the time integrals are approximated by the box rule, then the resulting scheme is equivalent to the implicit Euler method. However, a better approximation of these time integrals leads to a scheme which allows for better error estimates with respect to the required smoothness of the solution and has advantages in the case of long time integration ($T \gg 1$); see, e.g., [12].

The exact computation of the derivatives on the discrete level mentioned above is not disturbed even by the numerical integration. This can be shown by computing the schemes for the auxiliary equations by means of the inner product based on the underlying quadrature rule (e.g., the box rule or the trapezoidal rule).

3.3. Discretization of the controls. As proposed in the beginning of the current section, the discretization of the control space Q is kept rather abstract. It is done by choosing a finite dimensional subspace $Q_d \subset Q$. Then, the formulation of the state equation, the optimization problems, and the Lagrangians defined on the fully discretized state space can be directly transferred to the level with fully discretized control and state spaces by replacing Q by Q_d. The full discrete solutions will be indicated by the subscript σ which collects the discretization indices k, h, and d.

4. Derivation of the a posteriori error estimator. In this section, we will establish a posteriori error estimators for the error arising due to the discretization of the control and state spaces in terms of the cost functional J and an arbitrary quantity of interest I.

For this, we first recall a modification of an abstract result from [3] which we will later use to establish the desired a posteriori error estimators.

PROPOSITION 4.1. *Let Y be a function space and L a three times Gâteaux differentiable functional on Y. We seek a stationary point y_1 of L on a subspace $Y_1 \subset Y$, i.e.,*

$$L'(y_1)(\hat{y}_1) = 0 \quad \forall \hat{y}_1 \in Y_1. \tag{4.1}$$

This equation is approximated by a Galerkin method using a subspace $Y_2 \subset Y$. The approximative problem seeks $y_2 \in Y_2$ satisfying

$$L'(y_2)(\hat{y}_2) = 0 \quad \forall \hat{y}_2 \in Y_2. \tag{4.2}$$

If the continuous solution fulfills additionally

$$L'(y_1)(\hat{y}_2) = 0 \quad \forall \hat{y}_2 \in Y_2, \tag{4.3}$$

then we have for arbitrary $\hat{y}_2 \in Y_2$ the error representation

$$L(y_1) - L(y_2) = \frac{1}{2} L'(y_2)(y_1 - \hat{y}_2) + \mathcal{R}, \tag{4.4}$$

where the remainder term $\mathcal{R}$ is given with $e := y_1 - y_2$ by

$$\mathcal{R} = \frac{1}{2} \int_0^1 L'''(y_2 + se)(e, e, e) \cdot s \cdot (s-1)\, ds.$$

Proof. Even if the assumptions are weakened compared to the variant in [3], the proof presented there can be transferred directly. ∎

Remark 4.1. Usually this proposition is formulated for the case $Y_1 = Y$; then condition (4.3) is automatically fulfilled.

In what follows, we present the derivation of an error estimator for the fully discrete optimization problem in the case of dG time discretization only. The cG time discretization can be treated in a similar way.

4.1. Error estimator for the cost functional. In what follows, we use the abstract result of Proposition 4.1 for derivation of error estimators in terms of the cost functional J:

$$J(q, u) - J(q_\sigma, u_\sigma).$$

Here, $(q, u) \in Q \times X$ denotes the continuous optimal solution of (2.6), and $(q_\sigma, u_\sigma) = (q_{khd}, u_{khd}) \in Q_d \times \widetilde{X}_{k,h}^{r,s}$ is the optimal solution of the full discretized problem.

To separate the influences of the different discretizations on the discretization error we are interested in, we split

$$\begin{aligned} J(q, u) - J(q_\sigma, u_\sigma) = \quad & J(q, u) - J(q_k, u_k) \\ & + J(q_k, u_k) - J(q_{kh}, u_{kh}) \\ & + J(q_{kh}, u_{kh}) - J(q_\sigma, u_\sigma), \end{aligned}$$

where $(q_k, u_k) \in Q \times \widetilde{X}_k^r$ is the solution of the time discretized problem (3.4) and $(q_{kh}, u_{kh}) \in Q \times \widetilde{X}_{k,h}^{r,s}$ is the solution of the time and space discretized problem (3.8) with still undiscretized control space Q.

THEOREM 4.1. *Let (q,u,z), (q_k,u_k,z_k), (q_{kh},u_{kh},z_{kh}), and $(q_\sigma,u_\sigma,z_\sigma)$ be stationary points of $\mathcal{L}$, resp., $\widetilde{\mathcal{L}}$ on the different levels of discretization, i.e.,*

$$\begin{aligned}
\mathcal{L}'(q,u,z)(\hat{q},\hat{u},\hat{z}) = \widetilde{\mathcal{L}}'(q,u,z)(\hat{q},\hat{u},\hat{z}) = 0 \quad & \forall (\hat{q},\hat{u},\hat{z}) \in Q \times X \times X,\\
\widetilde{\mathcal{L}}'(q_k,u_k,z_k)(\hat{q}_k,\hat{u}_k,\hat{z}_k) = 0 \quad & \forall (\hat{q}_k,\hat{u}_k,\hat{z}_k) \in Q \times \widetilde{X}_k^r \times \widetilde{X}_k^r,\\
\widetilde{\mathcal{L}}'(q_{kh},u_{kh},z_{kh})(\hat{q}_{kh},\hat{u}_{kh},\hat{z}_{kh}) = 0 \quad & \forall (\hat{q}_{kh},\hat{u}_{kh},\hat{z}_{kh}) \in Q \times \widetilde{X}_{k,h}^{r,s} \times \widetilde{X}_{k,h}^{r,s},\\
\widetilde{\mathcal{L}}'(q_\sigma,u_\sigma,z_\sigma)(\hat{q}_\sigma,\hat{u}_\sigma,\hat{z}_\sigma) = 0 \quad & \forall (\hat{q}_\sigma,\hat{u}_\sigma,\hat{z}_\sigma) \in Q_d \times \widetilde{X}_{k,h}^{r,s} \times \widetilde{X}_{k,h}^{r,s}.
\end{aligned}$$

Then there holds for the errors with respect to the cost functional due to the time, space, and control discretizations

$$\begin{aligned}
J(q,u) - J(q_k,u_k) &= \frac{1}{2}\widetilde{\mathcal{L}}'(q_k,u_k,z_k)(q-\hat{q}_k, u-\hat{u}_k, z-\hat{z}_k) + \mathcal{R}_k,\\
J(q_k,u_k) - J(q_{kh},u_{kh}) &= \frac{1}{2}\widetilde{\mathcal{L}}'(q_{kh},u_{kh},z_{kh})(q_k-\hat{q}_{kh}, u_k-\hat{u}_{kh}, z_k-\hat{z}_{kh}) + \mathcal{R}_h,\\
J(q_{kh},u_{kh}) - J(q_\sigma,u_\sigma) &= \frac{1}{2}\widetilde{\mathcal{L}}'(q_\sigma,u_\sigma,z_\sigma)(q_{kh}-\hat{q}_\sigma, u_{kh}-\hat{u}_\sigma, z_{kh}-\hat{z}_\sigma) + \mathcal{R}_d.
\end{aligned}$$

Here, $(\hat{q}_k,\hat{u}_k,\hat{z}_k) \in Q \times \widetilde{X}_k^r \times \widetilde{X}_k^r$, $(\hat{q}_{kh},\hat{u}_{kh},\hat{z}_{kh}) \in Q \times \widetilde{X}_{k,h}^{r,s} \times \widetilde{X}_{k,h}^{r,s}$, and $(\hat{q}_\sigma,\hat{u}_\sigma,\hat{z}_\sigma) \in Q_d \times \widetilde{X}_{k,h}^{r,s} \times \widetilde{X}_{k,h}^{r,s}$ can be chosen arbitrarily, and the remainder terms $\mathcal{R}_k$, $\mathcal{R}_h$, and $\mathcal{R}_d$ have the same form as given in Proposition 4.1 *for $L = \widetilde{\mathcal{L}}$.*

Proof. Since all the used solution pairs are optimal solutions of the optimization problem on different discretizations levels, we obtain for arbitrary $z \in X$, $z_k \in \widetilde{X}_k^r$, and $z_{kh}, z_\sigma \in \widetilde{X}_{k,h}^{r,s}$

$$J(q,u) - J(q_k,u_k) = \widetilde{\mathcal{L}}(q,u,z) - \widetilde{\mathcal{L}}(q_k,u_k,z_k), \tag{4.5a}$$

$$J(q_k,u_k) - J(q_{kh},u_{kh}) = \widetilde{\mathcal{L}}(q_k,u_k,z_k) - \widetilde{\mathcal{L}}(q_{kh},u_{kh},z_{kh}), \tag{4.5b}$$

$$J(q_{kh},u_{kh}) - J(q_\sigma,u_\sigma) = \widetilde{\mathcal{L}}(q_{kh},u_{kh},z_{kh}) - \widetilde{\mathcal{L}}(q_\sigma,u_\sigma,z_\sigma), \tag{4.5c}$$

whereas the identity

$$J(q,u) = \mathcal{L}(q,u,z) = \widetilde{\mathcal{L}}(q,u,z)$$

follows from the fact that the $u \in X$ is continuous, and thus the additional jump terms in $\widetilde{\mathcal{L}}$ compared to $\mathcal{L}$ vanish.

To apply the abstract error identity (4.4) on the three right-hand sides in (4.5), we choose the spaces Y_1 and Y_2 of Proposition 4.1 as

$$\begin{aligned}
&\text{for (4.5a)}: && Y_1 = Q \times X \times X, && Y_2 = Q \times \widetilde{X}_k^r \times \widetilde{X}_k^r,\\
&\text{for (4.5b)}: && Y_1 = Q \times \widetilde{X}_k^r \times \widetilde{X}_k^r, && Y_2 = Q \times \widetilde{X}_{k,h}^{r,s} \times \widetilde{X}_{k,h}^{r,s},\\
&\text{for (4.5c)}: && Y_1 = Q \times \widetilde{X}_{k,h}^{r,s} \times \widetilde{X}_{k,h}^{r,s}, && Y_2 = Q_d \times \widetilde{X}_{k,h}^{r,s} \times \widetilde{X}_{k,h}^{r,s}.
\end{aligned}$$

Hence, for the second and third pairing we have $Y_2 \subset Y_1$, since we have $\widetilde{X}_{k,h}^{r,s} \subset \widetilde{X}_k^r$ and $Q_d \subset Q$. Thus we can choose $Y = Y_1$ in these cases. For the choice of the spaces

for (4.5a), we have to take into account the fact that $\widetilde{X}_k^r \not\subset X$. Thus, we choose $Y = Y_1 + Y_2$ and have to ensure condition (4.3):

$$\widetilde{\mathcal{L}}'(q,u,z)(\hat{q},\hat{u},\hat{z}) = 0 \quad \forall (\hat{q},\hat{u},\hat{z}) \in Q \times \widetilde{X}_k^r \times \widetilde{X}_k^r.$$

Since the solutions $u \in X$ and $z \in X$ are continuous in time with respect to H, the additional jump terms in $\widetilde{\mathcal{L}}$ compared to $\mathcal{L}$ vanish, and we may prove equivalently

$$\begin{aligned}
\mathcal{L}_z'(q,u,z)(\hat{z}) &= 0 \quad \forall \hat{z} \in \widetilde{X}_k^r,\\
\mathcal{L}_u'(q,u,z)(\hat{u}) &= 0 \quad \forall \hat{u} \in \widetilde{X}_k^r,\\
\mathcal{L}_q'(q,u,z)(\hat{q}) &= 0 \quad \forall \hat{q} \in Q.
\end{aligned}$$

We demonstrate the details of the construction for the adjoint state equation

$$\mathcal{L}_u'(q,u,z)(\hat{u}) = 0 \quad \forall \hat{u} \in \widetilde{X}_k^r$$

which we can write after integration by parts in time as

$$\begin{aligned}
-\sum_{m=1}^{M} \int_{I_m} (\hat{u}, \partial_t z)_H \, dt + a_u'(q,u)(\hat{u},z)& \\
+ (\hat{u}_M^-, z(T))_H = \int_I J_1'(u)(\hat{u})\, dt + J_2'(u(T))(\hat{u}_M^-) \quad \forall \hat{u} \in \widetilde{X}_k^r.&
\end{aligned}$$

Since the continuous adjoint solution z fulfills

$$(\varphi, z(T))_H = J_2'(u(T))(\varphi) \quad \forall \varphi \in H,$$

the terms containing $\hat{u}_M^- \in V \subset H$ cancel out, and we have to ensure

$$-\sum_{m=1}^{M} \int_{I_m} (\hat{u}, \partial_t z)_H \, dt + a_u'(q,u)(\hat{u},z) = \int_I J_1'(u)(\hat{u})\, dt \quad \forall \hat{u} \in \widetilde{X}_k^r.$$

Since we have that X is dense in $L^2((0,T),V)$ in regards to the $L^2((0,T),V)$ norm and due to $\widetilde{X}_k^r \subset L^2((0,T),V)$, we obtain then directly the stated condition

$$\mathcal{L}_u'(q,u,z)(\hat{u}) = 0 \quad \forall \hat{u} \in \widetilde{X}_k^r.$$

The remaining derivatives of $\mathcal{L}$ can be treated in a similar matter. The assertion of the theorem follows then by application of Proposition 4.1. ∎

By means of the residuals of the three equations building the optimality system (2.10),

$$\begin{aligned}
\tilde{\rho}^u(q,u)(\varphi) &:= \widetilde{\mathcal{L}}_z'(q,u,z)(\varphi),\\
\tilde{\rho}^z(q,u,z)(\varphi) &:= \widetilde{\mathcal{L}}_u'(q,u,z)(\varphi),\\
\tilde{\rho}^q(q,u,z)(\varphi) &:= \widetilde{\mathcal{L}}_q'(q,u,z)(\varphi),
\end{aligned}$$

the statement of Theorem 4.1 can be rewritten as

(4.6a)
$$J(q,u) - J(q_k,u_k) \approx \frac{1}{2}\Big(\tilde{\rho}^u(q_k,u_k)(z-\hat{z}_k) + \tilde{\rho}^z(q_k,u_k,z_k)(u-\hat{u}_k)\Big),$$

(4.6b)
$$J(q_k,u_k) - J(q_{kh},u_{kh}) \approx \frac{1}{2}\Big(\tilde{\rho}^u(q_{kh},u_{kh})(z_k-\hat{z}_{kh}) + \tilde{\rho}^z(q_{kh},u_{kh},z_{kh})(u_k-\hat{u}_{kh})\Big),$$

(4.6c)
$$J(q_{kh},u_{kh}) - J(q_\sigma,u_\sigma) \approx \frac{1}{2}\tilde{\rho}^q(q_\sigma,u_\sigma,z_\sigma)(q_{kh}-\hat{q}_\sigma).$$

Here, we employed the fact that the terms

$$\begin{aligned} &\tilde{\rho}^q(q_k,u_k,z_k)(q-\hat{q}_k), \qquad &&\tilde{\rho}^q(q_{kh},u_{kh},z_{kh})(q_k-\hat{q}_{kh}),\\ &\tilde{\rho}^u(q_\sigma,u_\sigma)(z_{kh}-\hat{z}_\sigma), \qquad &&\tilde{\rho}^z(q_\sigma,u_\sigma,z_\sigma)(u_{kh}-\hat{u}_\sigma) \end{aligned}$$

are zero for the choice

$$\begin{aligned} &\hat{q}_k = q \in Q, \qquad &&\hat{q}_{kh} = q_k \in Q,\\ &\hat{z}_\sigma = z_{kh} \in \widetilde{X}^{r,s}_{k,h}, \qquad &&\hat{u}_\sigma = u_{kh} \in \widetilde{X}^{r,s}_{k,h}. \end{aligned}$$

This is possible since for the errors $J(q,u)-J(q_k,u_k)$ and $J(q_k,u_k)-J(q_{kh},u_{kh})$ only the state space is discretized, and for $J(q_{kh},u_{kh}) - J(q_\sigma,u_\sigma)$ we keep the discrete state space while discretizing the control space Q.

4.2. Error estimator for an arbitrary functional. We now tend toward an error estimation of the different types of discretization errors in terms of a given functional $I\colon Q \times X \to \mathbb{R}$ describing the quantity of interest. This will be done using solutions of some auxiliary problems. In order to ensure the solvability of these problems we assume that the semidiscrete and the full discrete optimal solutions (q_k,u_k), (q_{kh},u_{kh}), and (q_σ,u_σ) are in the neighborhood $W \subset Q \times X$ of the optimal solution (q,u) introduced in section 2.

We define exterior Lagrangians $\mathcal{M}\colon [Q\times X\times X]^2 \to \mathbb{R}$ and $\widetilde{\mathcal{M}}\colon [Q\times \widetilde{X}^r_k\times \widetilde{X}^r_k]^2 \to \mathbb{R}$ as

$$\mathcal{M}(\xi,\chi) = I(q,u) + \mathcal{L}'(\xi)(\chi),$$

with $\xi = (q,u,z), \chi = (p,v,y)$, and

$$\widetilde{\mathcal{M}}(\xi_k,\chi_k) = I(q_k,u_k) + \widetilde{\mathcal{L}}'(\xi_k)(\chi_k),$$

with $\xi_k = (q_k,u_k,z_k), \chi_k = (p_k,v_k,y_k)$.

Now we are in a similar setting to that in the preceding subsection: We split the total discretization error with respect to I as

$$\begin{aligned} I(q,u) - I(q_\sigma,u_\sigma) = \quad & I(q,u) - I(q_k,u_k)\\ &+ I(q_k,u_k) - I(q_{kh},u_{kh})\\ &+ I(q_{kh},u_{kh}) - I(q_\sigma,u_\sigma) \end{aligned}$$

and obtain the following theorem.

THEOREM 4.2. *Let (ξ,χ), (ξ_k,χ_k), (ξ_{kh},χ_{kh}), and (ξ_σ,χ_σ) be stationary points of $\mathcal{M}$, resp., $\widetilde{\mathcal{M}}$ on the different levels of discretization, i.e.,*

$$\begin{aligned}
\mathcal{M}'(\xi,\chi)(\hat\xi,\hat\chi) = \widetilde{\mathcal{M}}'(\xi,\chi)(\hat\xi,\hat\chi) = 0 \quad & \forall(\hat\xi,\hat\chi)\in[Q\times X\times X]^2,\\
\widetilde{\mathcal{M}}'(\xi_k,\chi_k)(\hat\xi_k,\hat\chi_k) = 0 \quad & \forall(\hat\xi_k,\hat\chi_k)\in[Q\times \widetilde{X}_k^r\times \widetilde{X}_k^r]^2,\\
\widetilde{\mathcal{M}}'(\xi_{kh},\chi_{kh})(\hat\xi_{kh},\hat\chi_{kh}) = 0 \quad & \forall(\hat\xi_{kh},\hat\chi_{kh})\in[Q\times \widetilde{X}_{k,h}^{r,s}\times \widetilde{X}_{k,h}^{r,s}]^2,\\
\widetilde{\mathcal{M}}'(\xi_\sigma,\chi_\sigma)(\hat\xi_\sigma,\hat\chi_\sigma) = 0 \quad & \forall(\hat\xi_\sigma,\hat\chi_\sigma)\in[Q_d\times \widetilde{X}_{k,h}^{r,s}\times \widetilde{X}_{k,h}^{r,s}]^2.
\end{aligned}$$

Then there holds for the errors with respect to the quantity of interest due to the time, space, and control discretizations

$$\begin{aligned}
I(q,u)-I(q_k,u_k) &= \frac{1}{2}\widetilde{\mathcal{M}}'(\xi_k,\chi_k)(\xi-\hat\xi_k,\chi-\hat\chi_k)+\mathcal{R}_k,\\
I(q_k,u_k)-I(q_{kh},u_{kh}) &= \frac{1}{2}\widetilde{\mathcal{M}}'(\xi_{kh},\chi_{kh})(\xi_k-\hat\xi_{kh},\chi_k-\hat\chi_{kh})+\mathcal{R}_h,\\
I(q_{kh},u_{kh})-I(q_\sigma,u_\sigma) &= \frac{1}{2}\widetilde{\mathcal{M}}'(\xi_\sigma,\chi_\sigma)(\xi_{kh}-\hat\xi_\sigma,\chi_{kh}-\hat\chi_\sigma)+\mathcal{R}_d.
\end{aligned}$$

Here, $(\hat\xi_k,\hat\chi_k)\in[Q\times\widetilde{X}_k^r\times\widetilde{X}_k^r]^2$, $(\hat\xi_{kh},\hat\chi_{kh})\in[Q\times\widetilde{X}_{k,h}^{r,s}\times\widetilde{X}_{k,h}^{r,s}]^2$, and $(\hat\xi_\sigma,\hat\chi_\sigma)\in[Q_d\times\widetilde{X}_{k,h}^{r,s}\times\widetilde{X}_{k,h}^{r,s}]^2$ can be chosen arbitrarily, and the remainder terms $\mathcal{R}_k$, $\mathcal{R}_h$, and $\mathcal{R}_d$ have the same form as given in Proposition 4.1 for $L=\widetilde{\mathcal{M}}$.

Proof. Due to the optimality of the solution pairings on the different discretization levels, we have the representations

$$I(q,u)-I(q_k,u_k) = \widetilde{\mathcal{M}}(\xi,\chi)-\widetilde{\mathcal{M}}(\xi_k,\chi_k), \tag{4.7a}$$

$$I(q_k,u_k)-I(q_{kh},u_{kh}) = \widetilde{\mathcal{M}}(\xi_k,\chi_k)-\widetilde{\mathcal{M}}(\xi_{kh},\chi_{kh}), \tag{4.7b}$$

$$I(q_{kh},u_{kh})-I(q_\sigma,u_\sigma) = \widetilde{\mathcal{M}}(\xi_{kh},\chi_{kh})-\widetilde{\mathcal{M}}(\xi_\sigma,\chi_\sigma), \tag{4.7c}$$

where the identity

$$I(q,u)=\mathcal{M}(\xi,\chi)=\widetilde{\mathcal{M}}(\xi,\chi)$$

again follows from the fact that the $u\in X$ is continuous and thus the additional jump terms in $\widetilde{\mathcal{M}}$ compared to $\mathcal{M}$ vanish.

Similar to the proof of Theorem 4.1, we choose the spaces Y_1 and Y_2 for application of Proposition 4.1 as

$$\begin{aligned}
&\text{for (4.7a)}: && Y_1=[Q\times X\times X]^2, && Y_2=[Q\times\widetilde{X}_k^r\times\widetilde{X}_k^r]^2,\\
&\text{for (4.7b)}: && Y_1=[Q\times\widetilde{X}_k^r\times\widetilde{X}_k^r]^2, && Y_2=[Q\times\widetilde{X}_{k,h}^{r,s}\times\widetilde{X}_{k,h}^{r,s}]^2,\\
&\text{for (4.7c)}: && Y_1=[Q\times\widetilde{X}_{k,h}^{r,s}\times\widetilde{X}_{k,h}^{r,s}]^2, && Y_2=[Q_d\times\widetilde{X}_{k,h}^{r,s}\times\widetilde{X}_{k,h}^{r,s}]^2,
\end{aligned}$$

and we end up with the stated error representations. ∎

To apply Theorem 4.2 for instance to $I(q_{kh},u_{kh})-I(q_\sigma,u_\sigma)$, we have to require that

$$\widetilde{\mathcal{M}}'(\xi_\sigma,\chi_\sigma)(\hat\xi_\sigma,\hat\chi_\sigma)=0 \quad \forall(\hat\xi_\sigma,\hat\chi_\sigma)\in[\widetilde{X}_{k,h}^{r,s}\times\widetilde{X}_{k,h}^{r,s}\times Q_d]^2.$$

For solving this system, we have to consider the concrete form of $\widetilde{\mathcal{M}}'$:

$$\begin{aligned}
&\widetilde{\mathcal{M}}'(\xi_\sigma,\chi_\sigma)(\delta\xi_\sigma,\delta\chi_\sigma) =\\
&\qquad I_q'(q_\sigma,u_\sigma)(\delta q_\sigma)+I_u'(q_\sigma,u_\sigma)(\delta u_\sigma)+\widetilde{\mathcal{L}}'(\xi_\sigma)(\delta\chi_\sigma)+\widetilde{\mathcal{L}}''(\xi_\sigma)(\chi_\sigma,\delta\xi_\sigma).
\end{aligned}$$

Since $\xi_\sigma = (q_\sigma, u_\sigma, z_\sigma)$ is the solution of the discrete optimization problem, it fulfills already $\widetilde{\mathcal{L}}'(\xi_\sigma)(\delta\chi_\sigma) = 0$. Thus, the solution triple $\chi_\sigma = (p_\sigma, v_\sigma, y_\sigma) \in Q_d \times \widetilde{X}_{k,h}^{r,s} \times \widetilde{X}_{k,h}^{r,s}$ has to fulfill

$$\begin{aligned}(4.8)\quad \widetilde{\mathcal{L}}''(\xi_\sigma)(\chi_\sigma, \delta\xi_\sigma) = \\ - I_q'(q_\sigma, u_\sigma)(\delta q_\sigma) - I_u'(q_\sigma, u_\sigma)(\delta u_\sigma) \quad \forall \delta\xi_\sigma \in Q_d \times \widetilde{X}_{k,h}^{r,s} \times \widetilde{X}_{k,h}^{r,s}.\end{aligned}$$

Solving this system of equations is—apart from a different right-hand side—equivalent to the execution of one step of a (reduced) SQP-type method.

After splitting $y_\sigma = y_\sigma^{(0)} + y_\sigma^{(1)}$, where $y_\sigma^{(0)} \in \widetilde{X}_{k,h}^{r,s}$ is the solution of

$$\widetilde{\mathcal{L}}_{zu}''(\xi_\sigma)(y_\sigma^{(0)}, \varphi) = -I_u'(q_\sigma, u_\sigma)(\varphi) \quad \forall \varphi \in \widetilde{X}_{k,h}^{r,s},$$

we can rewrite system (4.8) in terms of the full discrete reduced Hessian $j_\sigma''(q)$ as

$$j_\sigma''(q_\sigma)(p_\sigma, \delta q_\sigma) = -I_q'(q_\sigma, u_\sigma)(\delta q_\sigma) - \mathcal{L}_{zq}''(\xi_\sigma)(y_\sigma^{(0)}, \delta q_\sigma) \quad \forall \delta q_\sigma \in Q_d,$$

where $j_\sigma''(q_\sigma)(p_\sigma, \delta q_\sigma)$ can be expressed as

$$\widetilde{\mathcal{L}}_{qq}''(\xi_\sigma)(p_\sigma, \delta q_\sigma) + \widetilde{\mathcal{L}}_{uq}''(\xi_\sigma)(v_\sigma, \delta q_\sigma) + \widetilde{\mathcal{L}}_{zq}''(\xi_\sigma)(y_\sigma^{(1)}, \delta q_\sigma).$$

The computation of $j_\sigma''(q_\sigma)(p_\sigma, \cdot)$ requires here the solution of the two auxiliary equations for $v_\sigma \in \widetilde{X}_{k,h}^{r,s}$ and $y_\sigma^{(1)} \in \widetilde{X}_{k,h}^{r,s}$:

$$\begin{aligned}\widetilde{\mathcal{L}}_{uz}''(\xi_\sigma)(v_\sigma, \varphi) &= -\widetilde{\mathcal{L}}_{qz}''(\xi_\sigma)(p_\sigma, \varphi) \quad \forall \varphi \in \widetilde{X}_{k,h}^{r,s}, \\ \widetilde{\mathcal{L}}_{zu}''(\xi_\sigma)(y_\sigma^{(1)}, \varphi) &= -\widetilde{\mathcal{L}}_{qu}''(\xi_\sigma)(p_\sigma, \varphi) - \widetilde{\mathcal{L}}_{uu}''(\xi_\sigma)(v_\sigma, \varphi) \quad \forall \varphi \in \widetilde{X}_{k,h}^{r,s}.\end{aligned}$$

By means of the residuals of the presented equations for p, v, and y, i.e.,

$$\begin{aligned}\tilde{\rho}^v(\xi, p, v)(\varphi) &:= \widetilde{\mathcal{L}}_{uz}''(\xi)(v, \varphi) + \widetilde{\mathcal{L}}_{qz}''(\xi)(p, \varphi), \\ \tilde{\rho}^y(\xi, p, v, y)(\varphi) &:= \widetilde{\mathcal{L}}_{zu}''(\xi)(y, \varphi) + \widetilde{\mathcal{L}}_{qu}''(\xi)(p, \varphi) + \widetilde{\mathcal{L}}_{uu}''(\xi)(v, \varphi) + I_u'(q, u)(\varphi), \\ \tilde{\rho}^p(\xi, p, v, y)(\varphi) &:= \widetilde{\mathcal{L}}_{qq}''(\xi)(p, \varphi) + \widetilde{\mathcal{L}}_{uq}''(\xi)(v, \varphi) + \widetilde{\mathcal{L}}_{zq}''(\xi)(y, \varphi) + I_q'(q, u)(\varphi),\end{aligned}$$

and the already defined residuals $\tilde{\rho}^u$, $\tilde{\rho}^z$, and $\tilde{\rho}^q$, the result of Theorem 4.2 can be expressed as

$$\begin{aligned}I(q, u) - I(q_k, u_k) \approx \frac{1}{2}\Big(&\tilde{\rho}^u(q_k, u_k)(y - \hat{y}_k) + \tilde{\rho}^z(q_k, u_k, z_k)(v - \hat{v}_k) \\ &+ \tilde{\rho}^v(\xi_k, p_k, v_k)(z - \hat{z}_k) + \tilde{\rho}^y(\xi_k, p_k, v_k, y_k)(u - \hat{u}_k)\Big), \\ I(q_k, u_k) - I(q_{kh}, u_{kh}) \approx \frac{1}{2}\Big(&\tilde{\rho}^u(q_{kh}, u_{kh})(y_k - \hat{y}_{kh}) + \tilde{\rho}^z(q_{kh}, u_{kh}, z_{kh})(v_k - \hat{v}_{kh}) \\ &+ \tilde{\rho}^v(\xi_{kh}, p_{kh}, v_{kh})(z_k - \hat{z}_{kh}) \\ &+ \tilde{\rho}^y(\xi_{kh}, p_{kh}, v_{kh}, y_{kh})(u_k - \hat{u}_{kh})\Big), \\ I(q_{kh}, u_{kh}) - I(q_\sigma, u_\sigma) \approx \frac{1}{2}\Big(&\tilde{\rho}^q(q_\sigma, u_\sigma, z_\sigma)(p_{kh} - \hat{p}_\sigma) + \tilde{\rho}^p(\xi_\sigma, p_\sigma, v_\sigma, y_\sigma)(q_{kh} - \hat{q}_\sigma)\Big).\end{aligned}$$

As for the estimator for the error in the cost functional, we employed here the fact that the terms

$$\begin{aligned}
\tilde{\rho}^q(q_k,u_k,z_k)(p-\hat{p}_k), &\qquad \tilde{\rho}^p(\xi_k,p_k,v_k,y_k)(q-\hat{q}_k),\\
\tilde{\rho}^q(q_{kh},u_{kh},z_{kh})(p_k-\hat{p}_{kh}), &\qquad \tilde{\rho}^p(\xi_{kh},p_{kh},v_{kh},y_{kh})(q_k-\hat{q}_{kh}),\\
\tilde{\rho}^u(q_\sigma,u_\sigma)(y_{kh}-\hat{y}_\sigma), &\qquad \tilde{\rho}^z(q_\sigma,u_\sigma,z_\sigma)(v_{kh}-\hat{v}_\sigma),\\
\tilde{\rho}^v(\xi_\sigma,p_\sigma,v_\sigma)(z_{kh}-\hat{z}_\sigma), &\qquad \tilde{\rho}^y(\xi_\sigma,p_\sigma,v_\sigma,y_\sigma)(u_{kh}-\hat{u}_\sigma)
\end{aligned}$$

vanish if $\hat{p}_k$, $\hat{q}_k$, $\hat{p}_{kh}$, $\hat{q}_{kh}$, $\hat{y}_\sigma$, $\hat{v}_\sigma$, $\hat{z}_\sigma$, $\hat{u}_\sigma$ are chosen appropriately.

Remark 4.2. As already mentioned in the introduction of this section, we obtain almost identical results for the time discretization by the cG method to those presented here. The difference simply consists in the tilde on the variables. The arguments of the proofs are exactly the same.

Remark 4.3. For the error estimation with respect to the cost functional no additional equations have to be solved. The error estimation with respect to a given quantity of interest requires the computation of the auxiliary variables p_σ, v_σ, y_σ. The additional numerical effort is similar to the execution of one step of the SQP or Newton's method.

5. Numerical realization.

5.1. Evaluation of the error estimators. In this subsection, we concretize the a posteriori error estimator developed in the previous section for the cG(1)cG(1) and cG(1)dG(0) space-time discretizations on quadrilateral meshes in two space dimensions. That is, we consider the combination of cG(1) or dG(0) time discretization with piecewise bilinear finite elements for the space discretization. As in the previous section, we will present only the concrete expressions for the dG time discretization; the cG discretization can be treated in exactly the same manner.

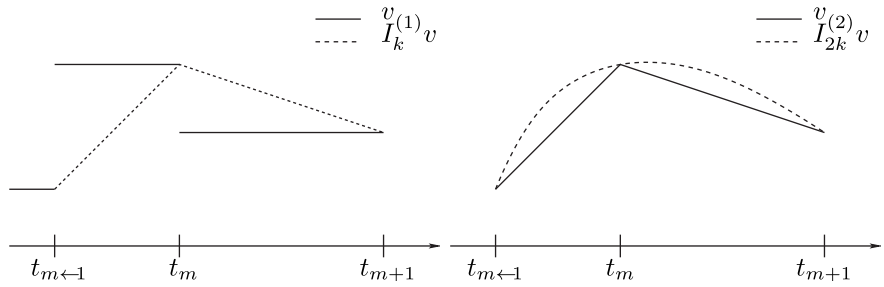
The error estimates presented in the previous section involve interpolation errors of the time, space, and the control discretizations. We approximate these errors using interpolations in higher order finite element spaces. To this end, we introduce linear operators Π_h, Π_k, and Π_d, which will map the computed solutions to the approximations of the interpolation errors:

$$\begin{aligned}
z-\hat{z}_k &\approx \Pi_k z_k, & u-\hat{u}_k &\approx \Pi_k u_k,\\
z_k-\hat{z}_{kh} &\approx \Pi_h z_{kh}, & u_k-\hat{u}_{kh} &\approx \Pi_h u_{kh},\\
q_{kh}-\hat{q}_\sigma &\approx \Pi_d q_\sigma,
\end{aligned}$$

$$\begin{aligned}
y-\hat{y}_k &\approx \Pi_k y_k, & v-\hat{v}_k &\approx \Pi_k v_k,\\
y_k-\hat{y}_{kh} &\approx \Pi_h y_{kh}, & v_k-\hat{v}_{kh} &\approx \Pi_h v_{kh},\\
p_{kh}-\hat{p}_\sigma &\approx \Pi_d p_\sigma.
\end{aligned}$$

For the case of cG(1)cG(1) and cG(1)dG(0) discretizations of the state space considered here, the operators are chosen depending on the test and trial space as

$$\begin{aligned}
\Pi_k &= I_k^{(1)} - \mathrm{id} \quad \text{with} \quad I_k^{(1)}\colon \widetilde{X}_k^0 \to X_k^1,\\
\Pi_k &= I_{2k}^{(2)} - \mathrm{id} \quad \text{with} \quad I_{2k}^{(2)}\colon X_k^1 \to X_{2k}^2,\\
\Pi_h &= I_{2h}^{(2)} - \mathrm{id} \quad \text{with} \quad I_{2h}^{(2)}\colon \begin{cases} X_{k,h}^{1,1} \to X_{k,2h}^{1,2}\\ \widetilde{X}_{k,h}^{0,1} \to \widetilde{X}_{k,2h}^{0,2}. \end{cases}
\end{aligned}$$



(a) *Piecewise linear interpolation of a piecewise constant function.*

(b) *Piecewise quadratic interpolation of a piecewise linear function.*

FIG. 5.1. *Temporal interpolation.*

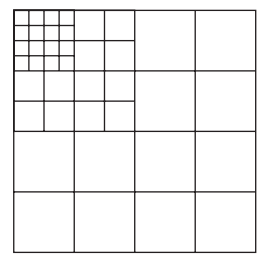


FIG. 5.2. *Patched mesh.*

The action of the piecewise linear and piecewise quadratic interpolation operators $I_k^{(1)}$ and $I_{2k}^{(2)}$ in time is depicted in Figure 5.1. The piecewise biquadratic spatial interpolation $I_{2h}^{(2)}$ can be easily computed if the underlying mesh provides a patch structure. That is, one can always combine four adjacent cells to a macrocell on which the biquadratic interpolation can be defined. An example of such a patched mesh is shown in Figure 5.2.

The choice of Π_d depends on the discretization of the control space Q. If the finite dimensional subspaces Q_d are constructed similar to the discrete state spaces, one can directly choose for Π_d a modification of the operators Π_k and Π_h defined above. If, e.g., the controls q depend only on time and the discretization is done with piecewise constant polynomials, we can choose $\Pi_d = I_d^{(1)} - \mathrm{id}$. If the control space Q is already finite dimensional, which is usually the case in the context of parameter estimation, it is possible to choose $\Pi_d = 0$, and thus, the estimator for the error $J(q_{kh}, u_{kh}) - J(q_\sigma, u_\sigma)$ is zero—as well as this discretization error itself.

In order to make the error representations from the previous section computable, we replace the residuals linearized on the solution of semidiscretized problems by the linearization at full discrete solutions.

We finally obtain the following computable a posteriori error estimator for the cost functional J:

$$J(q,u) - J(q_\sigma, u_\sigma) \approx \eta_k^J + \eta_h^J + \eta_d^J,$$

with

$$\eta_k^J := \frac{1}{2}\Big(\tilde{\rho}^u(q_\sigma,u_\sigma)(\Pi_k z_\sigma) + \tilde{\rho}^z(q_\sigma,u_\sigma,z_\sigma)(\Pi_k u_\sigma)\Big),$$
$$\eta_h^J := \frac{1}{2}\Big(\tilde{\rho}^u(q_\sigma,u_\sigma)(\Pi_h z_\sigma) + \tilde{\rho}^z(q_\sigma,u_\sigma,z_\sigma)(\Pi_h u_\sigma)\Big),$$
$$\eta_d^J := \frac{1}{2}\tilde{\rho}^q(q_\sigma,u_\sigma,z_\sigma)(\Pi_d q_\sigma).$$

For the quantity of interest I the error estimator is given by

$$I(q,u) - I(q_\sigma,u_\sigma) \approx \eta_k^I + \eta_h^I + \eta_d^I,$$

with

$$\begin{aligned}\eta_k^I := \frac{1}{2}\Big(&\tilde{\rho}^u(q_\sigma,u_\sigma)(\Pi_k y_\sigma) + \tilde{\rho}^z(q_\sigma,u_\sigma,z_\sigma)(\Pi_k v_\sigma)\\ &+ \tilde{\rho}^v(\xi_\sigma,v_\sigma,p_\sigma)(\Pi_k z_\sigma) + \tilde{\rho}^y(\xi_\sigma,v_\sigma,y_\sigma,p_\sigma)(\Pi_k u_\sigma)\Big),\\ \eta_h^I := \frac{1}{2}\Big(&\tilde{\rho}^u(q_\sigma,u_\sigma)(\Pi_h y_\sigma) + \tilde{\rho}^z(q_\sigma,u_\sigma,z_\sigma)(\Pi_h v_\sigma)\\ &+ \tilde{\rho}^v(\xi_\sigma,v_\sigma,p_\sigma)(\Pi_h z_\sigma) + \tilde{\rho}^y(\xi_\sigma,v_\sigma,y_\sigma,p_\sigma)(\Pi_h u_\sigma)\Big),\\ \eta_d^I := \frac{1}{2}\Big(&\tilde{\rho}^q(q_\sigma,u_\sigma,z_\sigma)(\Pi_d p_\sigma) + \tilde{\rho}^p(\xi_\sigma,v_\sigma,y_\sigma,p_\sigma)(\Pi_d q_\sigma)\Big).\end{aligned}$$

To give an impression of the terms that have to be evaluated for the error estimators, we present for the implicit Euler variant of the cG(1)dG(0) discretization the explicit form of the state residuals $\tilde{\rho}^u(q_\sigma,u_\sigma)(\Pi_k z_\sigma)$ and $\tilde{\rho}^u(q_\sigma,u_\sigma)(\Pi_h z_\sigma)$ and the adjoint state residuals $\tilde{\rho}^z(q_\sigma,u_\sigma,z_\sigma)(\Pi_k u_\sigma)$ and $\tilde{\rho}^z(q_\sigma,u_\sigma,z_\sigma)(\Pi_h u_\sigma)$. For simplicity of notation, we assume here q to be independent on time. Since we evaluate the arising integrals over time for the residuals weighted with z_σ or u_σ by the right endpoint rule and for the residuals weighted with $I_k^{(1)} z_\sigma$ or $I_k^{(1)} u_\sigma$ by the trapezoidal rule, we have to ensure the right-hand side f to be continuous in time, i.e., $f \in C([0,T],H)$. Then we obtain with the abbreviations $U_0 := u_\sigma(0)$, $U_m := u_\sigma\big|_{I_m}$, $Z_0 := z_\sigma(0)$, and $Z_m = z_\sigma\big|_{I_m}$ the following parts of the error estimators:

$$\begin{aligned}\tilde{\rho}^u(q_\sigma,u_\sigma)(\Pi_k z_\sigma) = \sum_{m=1}^{M}\Big\{&(U_m - U_{m-1}, Z_m - Z_{m-1})_H\\ &+ \frac{k_m}{2}\bar{a}(q_\sigma,U_m)(Z_m - Z_{m-1})\\ &+ \frac{k_m}{2}(f(t_{m-1}),Z_{m-1})_H - \frac{k_m}{2}(f(t_m),Z_m)_H\Big\},\end{aligned}$$

$$\begin{aligned}\tilde{\rho}^z(q_\sigma,u_\sigma,z_\sigma)(\Pi_k u_\sigma) = \sum_{m=1}^{M}\Big\{&\frac{k_m}{2}\bar{a}'_u(q_\sigma,U_m)(U_m,Z_m)\\ &- \frac{k_m}{2}\bar{a}'_u(q_\sigma,U_{m-1})(U_{m-1},Z_m)\\ &+ \frac{k_m}{2}J_1'(U_{m-1})(U_{m-1}) - \frac{k_m}{2}J_1'(U_m)(U_m)\Big\},\end{aligned}$$

$$\begin{aligned}\tilde{\rho}^u(q_\sigma,u_\sigma)(\Pi_h z_\sigma) = \sum_{m=1}^{M} \Big\{ & k_m(f(t_m), I_{2h}^{(2)} Z_m - Z_m)_H \\ & - k_m \bar{a}(q_\sigma, U_m)(I_{2h}^{(2)} Z_m - Z_m) \\ & - (U_m - U_{m-1}, I_{2h}^{(2)} Z_m - Z_m)_H \Big\} \\ & - (U_0 - u_0(q_\sigma), I_{2h}^{(2)} Z_0 - Z_0)_H,\end{aligned}$$

$$\begin{aligned}\tilde{\rho}^z(q_\sigma,u_\sigma,z_\sigma)(\Pi_h u_\sigma) = \sum_{m=1}^{M} \Big\{ & k_m J_1'(U_m)(I_{2h}^{(2)} U_m - U_m) \\ & - k_m \bar{a}_u'(q_\sigma, U_m)(I_{2h}^{(2)} U_m - U_m, Z_m) \\ & + (I_{2h}^{(2)} U_{m-1} - U_{m-1}, Z_m - Z_{m-1})_H \Big\} \\ & + J_2'(U_M)(I_{2h}^{(2)} U_M - U_M) - (I_{2h}^{(2)} U_M - U_M, Z_M)_H.\end{aligned}$$

For the cG(1)cG(1) discretization the terms that have to be evaluated are very similar and the evaluation can be treated as presented here for the cG(1)dG(0) discretization. The presented a posteriori error estimators are directed towards two aims: assessment of the discretization error and improvement of the accuracy by local refinement. For the second aim the information provided by the error estimator has to be localized to cellwise or nodewise contributions (local error indicators). For details of the localization procedure we refer, e.g., to [3].

5.2. Adaptive algorithm. The goal of the adaption of the different types of discretizations has to be the equilibrated reduction of the corresponding discretization errors. If a given tolerance (TOL) has to be reached, this can be done by refining each discretization as long as the value of this part of the error estimator is greater than $\frac{1}{3}$TOL. We want to present here a strategy which will equilibrate the different discretization errors even if no tolerance is given.

The aim of the equilibration algorithm presented in what follows is to obtain discretization such that

$$|\eta_k| \approx |\eta_h| \approx |\eta_d|$$

and to keep this property during the further refinement. Here, the estimators η_i denote the estimators η_i^J for the cost functional J or η_i^I for the quantity of interest I.

For doing this equilibration, we choose an "equilibration factor" $e \approx 1-5$ and propose the following strategy: We compute a permutation (a,b,c) of the discretization indices (k,h,d) such that

$$|\eta_a| \geq |\eta_b| \geq |\eta_c|,$$

and we define the relations

$$\gamma_{ab} := \left|\frac{\eta_a}{\eta_b}\right| \geq 1, \qquad \gamma_{bc} := \left|\frac{\eta_b}{\eta_c}\right| \geq 1.$$

Then we decide by means of Table 5.1 in every repetition of the adaptive refinement algorithm given by Algorithm 5.1 which discretization shall be refined. For every discretization to be adapted we select by means of the local error indicators the cells for refinement. For this purpose there are several strategies available; see, e.g., [3].

TABLE 5.1
Equilibration strategy.

Relation between the estimators	Discretizations to be refined
$\gamma_{ab} \leq e$ and $\gamma_{bc} \leq e$	a, b, and c
$\gamma_{bc} > e$	a and b
else ($\gamma_{ab} > e$ and $\gamma_{bc} \leq e$)	a

ALGORITHM 5.1 (ADAPTIVE REFINEMENT ALGORITHM).

Choose an initial triple of discretizations $\mathcal{T}_{\sigma_0}$, $\sigma_0 = (k_0, h_0, d_0)$ for the space-time discretization of the states and an appropriate discretization of the controls, and set $n = 0$.
loop
  Compute the optimal solution pair $(q_{\sigma_n}, u_{\sigma_n})$.
  Evaluate the a posteriori error estimators η_{k_n}, η_{h_n}, and η_{d_n}.
  if $\eta_{k_n} + \eta_{h_n} + \eta_{d_n} \leq TOL$ ***then***
    break
  else
    Determine the discretization(s) to be refined by means of Table 5.1.
  end if
  Refine $\mathcal{T}_{\sigma_n} \to \mathcal{T}_{\sigma_{n+1}}$ depending on the size of η_{k_n}, η_{h_n}, and η_{d_n} to equilibrate the three discretization errors.
  Increment n.
end loop

6. Numerical examples. This section is devoted to the numerical validation of the theoretical results presented in the previous sections. This will be done by means of an optimal control problem with time-dependent boundary control (see section 6.1) and a parameter estimation problem (see section 6.2).

6.1. Example 1: Neumann boundary control problem. We consider the linear parabolic state equation on the two-dimensional unit square $\Omega := (0,1)^2$ (see Figure 6.1) with final time $T = 1$ given by

$$\begin{aligned} \partial_t u - \nu \Delta u + u &= f && \text{in } \Omega \times (0,T), \\ \partial_n u(x,t) &= 0 && \text{on } \Gamma_0 \times (0,T), \\ \partial_n u(x,t) &= q_i(t) && \text{on } \Gamma_i \times (0,T), i = 1,2, \\ u(x,0) &= 0 && \text{on } \Omega. \end{aligned} \tag{6.1}$$

The control $q = (q_1, q_2)$ acts as a purely time-dependent boundary control of Neumann type on the two parts of the boundary denoted by Γ_1 and Γ_2. Thus, the control space Q is chosen as $[L^2(0,T)]^2$, and the spaces V and H used in the definition of the state space X are set to $V = H^1(\Omega)$ and $H = L^2(\Omega)$.

As the cost functional J to be minimized subject to the state equation, we choose the functional

$$J(q,u) := \frac{1}{2} \int_0^T \int_\Omega (u(x,t) - 1)^2 \, dx \, dt + \frac{\alpha}{2} \int_0^T \{q_1^2(t) + q_2^2(t)\} \, dt$$

of the tracking type endowed with a $L^2(0,T)$-regularization.

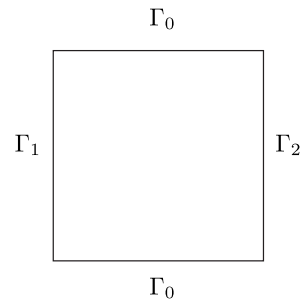


FIG. 6.1. *Example* 1: *Computational domain* Ω.

For the computations, the right-hand side of f is chosen as

$$f(x,t) = 10t \exp\left(1 - \frac{1}{1 - 100\|x - \tilde{x}\|^2}\right), \quad \tilde{x} = \left(\frac{2}{3}, \frac{1}{2}\right),$$

and the parameters α and ν are set to

$$\alpha = 0.1, \quad \nu = 0.1.$$

The discretization of the state space is done here via the cG(1)cG(1) space-time Galerkin method which is a variant of the Crank–Nicolson scheme. Consequently, the state is discretized in time by piecewise linear polynomials and the adjoint state by piecewise constant polynomials. The controls are discretized using piecewise constant polynomials on a partition of the time interval $(0,T)$ which has to be at most as fine as the time discretization of the states.

Remark 6.1. If the discretization of the control is chosen such that the gradient equation

$$\int_{\Gamma_i} z(x,t)\,dx + \alpha q_i(t) = 0, \qquad i = 1,2, \quad t \in (0,T),$$

can be fulfilled pointwise on the discrete level, the residual ρ^q of this equation as well as the error due to discretization of the control space vanish; cf. (4.6c). Thus, it is only reasonable to discretize the controls at most as fine as the adjoint state.

In Table 6.1 we show the development of the discretization error and the a posteriori error estimators during an adaptive run with local refinement of all three types of discretizations. Here, M denotes the number of time steps, N denotes the number of nodes in the spatial mesh, and $\dim Q_d$ is the number of degrees of freedom for the discretization of the control. The effectivity index given in the last column of this table is defined as usual by

$$I_{\text{eff}} := \frac{J(q,u) - J(q_\sigma, u_\sigma)}{\eta_k^J + \eta_h^J + \eta_d^J}.$$

The table also demonstrates the desired equilibration of the different discretization errors and the sufficient quality of the error estimators. Here and in what follows, the "exact" values $J(q,u)$ and $I(q,u)$ are obtained approximatively by extrapolation of the values of these functionals computated on a sequence of fine discretizations.

A comparison of the error $J(q,u) - J(q_\sigma, u_\sigma)$ for the different refinement strategies is depicted in Figure 6.2:

TABLE 6.1
Example 1: *Local refinement with equilibration.*

M	N	$\dim Q_d$	η_k^J	η_h^J	η_d^J	$\eta_k^J + \eta_h^J + \eta_d^J$	$J(q,u) - J(q_\sigma, u_\sigma)$	I_{eff}
64	25	16	$-9.7{\cdot}10^{-05}$	$2.0{\cdot}10^{-03}$	$-8.5{\cdot}10^{-04}$	$1.088 \cdot 10^{-03}$	$-2.567 \cdot 10^{-04}$	-0.2360
64	81	20	$-1.1{\cdot}10^{-04}$	$-1.0{\cdot}10^{-03}$	$-3.2{\cdot}10^{-04}$	$-1.543 \cdot 10^{-03}$	$-7.818 \cdot 10^{-04}$	0.5065
64	289	20	$-1.3{\cdot}10^{-04}$	$-4.8{\cdot}10^{-04}$	$-3.2{\cdot}10^{-04}$	$-9.458 \cdot 10^{-04}$	$-8.009 \cdot 10^{-04}$	0.8468
74	813	32	$-4.7{\cdot}10^{-05}$	$-2.2{\cdot}10^{-05}$	$-1.3{\cdot}10^{-04}$	$-2.058 \cdot 10^{-04}$	$-2.116 \cdot 10^{-04}$	1.0285
74	813	48	$-4.8{\cdot}10^{-05}$	$-2.2{\cdot}10^{-05}$	$-7.7{\cdot}10^{-05}$	$-1.476 \cdot 10^{-04}$	$-1.493 \cdot 10^{-04}$	1.0109
87	2317	76	$-2.7{\cdot}10^{-05}$	$1.1{\cdot}10^{-05}$	$-2.9{\cdot}10^{-05}$	$-4.516 \cdot 10^{-05}$	$-4.559 \cdot 10^{-05}$	1.0094
104	8213	128	$-1.8{\cdot}10^{-05}$	$2.7{\cdot}10^{-06}$	$-1.3{\cdot}10^{-05}$	$-2.931 \cdot 10^{-05}$	$-2.842 \cdot 10^{-05}$	0.9696
208	8213	128	$-4.3{\cdot}10^{-06}$	$2.7{\cdot}10^{-06}$	$-1.5{\cdot}10^{-05}$	$-1.674 \cdot 10^{-05}$	$-1.661 \cdot 10^{-05}$	0.9923
208	8213	192	$-4.2{\cdot}10^{-06}$	$2.7{\cdot}10^{-06}$	$-7.0{\cdot}10^{-06}$	$-8.573 \cdot 10^{-06}$	$-8.335 \cdot 10^{-06}$	0.9722

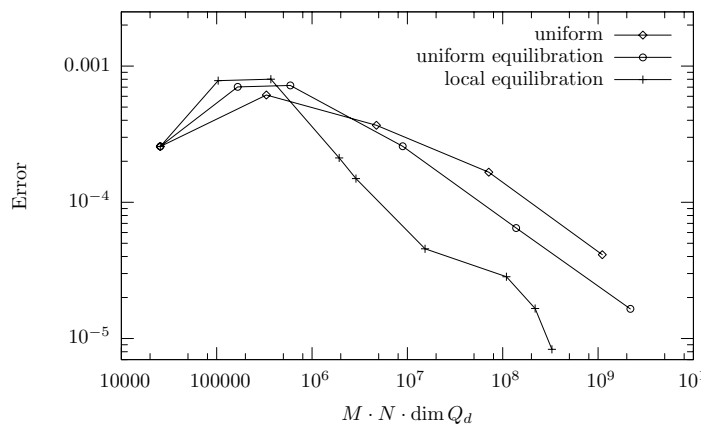


FIG. 6.2. *Example* 1: *Comparison of different refinement strategies.*

- "Uniform": Here, we apply uniform refinement of all discretizations after each run of the optimization loop.
- "Uniform equilibration": Here, we also allow for only uniform refinements but use the error estimators within the equilibration strategy (Table 5.1) to decide which discretizations have to be refined.
- "Local equilibration": Here, we combine local refinement of all discretizations with the proposed equilibration strategy.

It shows, e.g., that to reach a discretization error of $4 \cdot 10^{-5}$ the uniform refinement needs about 70 times the number of degrees of freedom the fully adaptive refinement needs.

In Table 6.2 we present the numerical justification for splitting the total discretization error in three parts regarding the discretization of time, space, and control: The table demonstrates the independence of each part of the error estimator on the refinement of the other parts. This feature is especially important to reach a equilibration of the discretization errors by applying the adaptive refinement algorithm.

6.2. Example 2: Parameter estimation. The state equation for the following example is taken from [17]. It describes the major part of gaseous combustion under the low Mach number hypothesis. Under this assumption, the motion of the fluid becomes independent from temperature and species concentration. Hence, one can solve the temperature and the species equation alone specifying any solenoidal velocity field.

TABLE 6.2
Example 1: *Independence of one part of the error estimator on the refinement of the other parts.*

M	N	$\dim Q_d$	η_k^J	η_h^J	η_d^J
256	289	16		$-4.9104\cdot10^{-04}$	$-8.6152\cdot10^{-04}$
512	289	16		$-4.9110\cdot10^{-04}$	$-8.6232\cdot10^{-04}$
1024	289	16	—	$-4.9111\cdot10^{-04}$	$-8.6251\cdot10^{-04}$
2048	289	16		$-4.9111\cdot10^{-04}$	$-8.6256\cdot10^{-04}$
4096	289	16		$-4.9112\cdot10^{-04}$	$-8.6258\cdot10^{-04}$
1024	25	16	$-3.8360\cdot10^{-07}$		$-8.7015\cdot10^{-04}$
1024	81	16	$-4.3463\cdot10^{-07}$		$-8.5900\cdot10^{-04}$
1024	289	16	$-4.5039\cdot10^{-07}$	—	$-8.6251\cdot10^{-04}$
1024	1089	16	$-4.5529\cdot10^{-07}$		$-8.6398\cdot10^{-04}$
1024	4225	16	$-4.6096\cdot10^{-07}$		$-8.6432\cdot10^{-04}$
4096	289	16	$-2.8171\cdot10^{-08}$	$-4.9112\cdot10^{-04}$	
4096	289	32	$-3.0332\cdot10^{-08}$	$-4.8826\cdot10^{-04}$	
4096	289	64	$-3.1317\cdot10^{-08}$	$-4.8688\cdot10^{-04}$	—
4096	289	128	$-3.1704\cdot10^{-08}$	$-4.8651\cdot10^{-04}$	
4096	289	256	$-3.1828\cdot10^{-08}$	$-4.8642\cdot10^{-04}$	

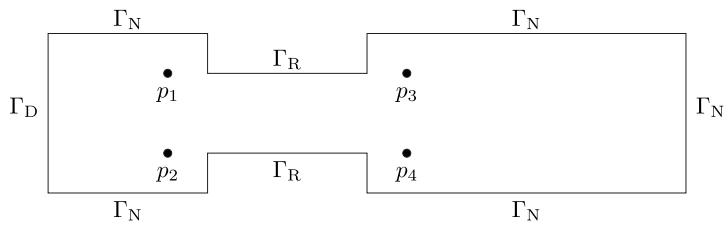


FIG. 6.3. *Example* 2: *Computational domain* Ω *and measurement points* p_i.

Introducing the dimensionless temperature $\theta = \frac{T-T_{\text{unburnt}}}{T_{\text{burnt}}-T_{\text{unburnt}}}$, denoting by Y the species concentration, and assuming constant diffusion coefficients yields

$$\begin{aligned} \partial_t\theta - \Delta\theta &= \omega(Y,\theta) && \text{in } \Omega\times(0,T),\\ \partial_t Y - \frac{1}{\mathrm{Le}}\Delta Y &= -\omega(Y,\theta) && \text{in } \Omega\times(0,T), \end{aligned} \tag{6.2}$$

where the Lewis number Le is the ratio of diffusivity of heat and diffusivity of mass. We use a simple one-species reaction mechanism governed by an Arrhenius law

$$\omega(Y,\theta) = \frac{\beta^2}{2\mathrm{Le}} Y \mathrm{e}^{\frac{\beta(\theta-1)}{1+\alpha(\theta-1)}}$$

in which an approximation for large activation energy has been employed.

Here, we consider a freely propagating laminar flame described by (6.2) and its response to a heat absorbing obstacle, a set of cooled parallel rods with rectangular cross section (cf. Figure 6.3). Thus, the boundary conditions are chosen as

$$\begin{aligned} \theta &= 1 && \text{on } \Gamma_D\times(0,T),\\ Y &= 0 && \text{on } \Gamma_D\times(0,T),\\ \partial_n\theta &= 0 && \text{on } \Gamma_N\times(0,T),\\ \partial_n Y &= 0 && \text{on } \Gamma_N\times(0,T),\\ \partial_n\theta &= -k\theta && \text{on } \Gamma_R\times(0,T),\\ \partial_n Y &= 0 && \text{on } \Gamma_R\times(0,T), \end{aligned}$$

where the heat absorption is modeled by Robin boundary conditions on Γ_R.

The initial condition is the analytical solution of a one-dimensional right-traveling flame in the limit $\beta \to \infty$ located left of the obstacle:

$$\theta(0,x) = \begin{cases} 1 & \text{for } x_1 \le \tilde{x}_1 \\ \mathrm{e}^{\tilde{x}_1 - x_1} & \text{for } x_1 > \tilde{x}_1 \end{cases} \qquad \text{on } \Omega,$$

$$Y(0,x) = \begin{cases} 0 & \text{for } x_1 \le \tilde{x}_1 \\ 1 - \mathrm{e}^{\mathrm{Le}(\tilde{x}_1 - x_1)} & \text{for } x_1 > \tilde{x}_1 \end{cases} \qquad \text{on } \Omega.$$

For the computations, the occurring parameters are set to

$$\mathrm{Le} = 1, \qquad \beta = 10, \qquad k = 0.1, \qquad \tilde{x}_1 = 9,$$

whereas the parameter α occurring in the Arrhenius law will be the objective of the parameter estimation.

To use the same notations as in the theoretical parts of this article, we define the pair of solution components $u := (\theta, Y) \in \hat{u} + X^2$ and denote the parameter α to be estimated by $q \in Q := \mathbb{R}$. For definition of the state space X we use the spaces V and H as given by (2.1). The function $\hat{u}$ is defined to fulfill the prescribed Dirichlet data as $\hat{u}\big|_{\Gamma_{\mathrm{D}}} = (1, 0)$.

The unknown parameter α is estimated here using information from pointwise measurements of θ and Y at four measurement points $p_i \in \Omega$ $(i = 1, \ldots, 4)$ at final time $T = 60$. This parameter identification problem can be formulated as a cost functional of least squares type:

$$J(q,u) = \frac{1}{2} \sum_{i=1}^{4} \Big\{ (\theta(p_i, T) - \tilde{\theta}_i)^2 + (Y(p_i, T) - \tilde{Y}_i)^2 \Big\}.$$

The values of artificial measurements $\tilde{\theta}_i$ and $\tilde{Y}_i$ $(i = 1, \ldots, 4)$ are obtained from a reference solution computed on fine discretizations.

The consideration of point measurements does not fulfill the assumption on the cost functional in (2.5), since the point evaluation is not bounded as a functional on $H = L^2(\Omega)$. Therefore, the point functionals here may be understood as regularized functionals defined on $L^2(\Omega)$. For an a priori error estimate of elliptic parameter identification problems with pointwise measurements, we refer to [25].

For this type of parameter estimation problem one is usually not interested in reducing the discretization error measured in terms of the cost functional. The focus is rather on reducing the error in the parameter q to be estimated. Hence, we use the quantity of interest I given by

$$I(q,u) = q$$

and apply the techniques presented in section 4.2 for estimating the discretization error with respect to I. Since the control space Q in this application is given as $Q = \mathbb{R}$, it is not necessary to discretize Q. Thus, there is no discretization error due to the Q-discretization and the a posteriori error estimator consists only of η_k^I and η_h^I.

TABLE 6.3
Example 2: *Local refinement with equilibration.*

M	N	η_k^I	η_h^I	$\eta_k^I + \eta_h^I$	$I(q,u) - I(q_{kh}, u_{kh})$	I_{eff}
512	269	$-8.4\cdot10^{-03}$	$4.3\cdot10^{-02}$	$3.551 \cdot 10^{-02}$	$-2.859 \cdot 10^{-02}$	-0.8051
512	685	$-9.0\cdot10^{-03}$	$5.2\cdot10^{-03}$	$-3.778 \cdot 10^{-03}$	$-4.854 \cdot 10^{-02}$	12.8480
690	1871	$-3.7\cdot10^{-03}$	$-1.4\cdot10^{-02}$	$-1.860 \cdot 10^{-02}$	$-3.028 \cdot 10^{-02}$	1.6280
968	5611	$-2.9\cdot10^{-03}$	$-6.3\cdot10^{-03}$	$-9.292 \cdot 10^{-03}$	$-1.104 \cdot 10^{-02}$	1.1885
1036	14433	$-2.7\cdot10^{-03}$	$-2.3\cdot10^{-03}$	$-5.118 \cdot 10^{-03}$	$-5.441 \cdot 10^{-03}$	1.0630
1044	43979	$-2.7\cdot10^{-03}$	$-8.3\cdot10^{-04}$	$-3.613 \cdot 10^{-03}$	$-3.588 \cdot 10^{-03}$	0.9932

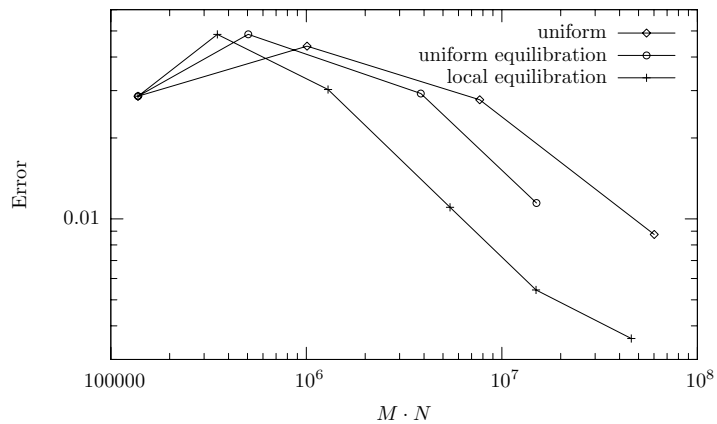


FIG. 6.4. *Example* 2: *Comparison of different refinement strategies.*

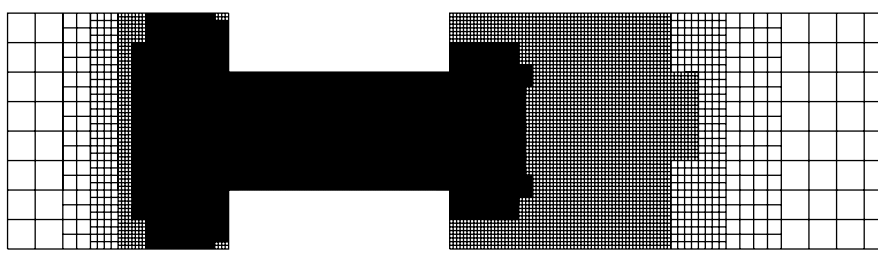


FIG. 6.5. *Example* 2: *Local refined mesh.*

The results of a computation with equilibrated adaption of the space and time discretization using cG(1)dG(0) are shown in Table 6.3. The discretization parameters M and N as well as the effectivity index I_{eff} are defined as in section 6.1.

Similar to section 6.1, we compare in Figure 6.4 the fully adaptive refinement with equilibration and uniform refinements with and without equilibration. By local refinement of all involved discretizations we reduce the necessary degrees of freedom to reach a total error of 10^{-2} by a factor of 11 compared to a uniform refinement without equilibration.

Finally, we present in the Figures 6.5 and 6.6 a typical locally refined spatial mesh and a distribution of the time step size obtained by the space-time-adaptive refinement.

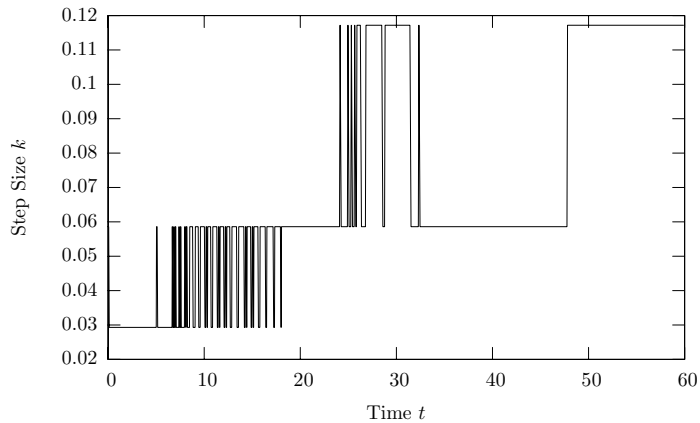


FIG. 6.6. *Example* 2: *Visualization of the adaptively determined time step size* k.

REFERENCES

[1] R. BECKER, H. KAPP, AND R. RANNACHER, *Adaptive finite element methods for optimal control of partial differential equations: Basic concepts*, SIAM J. Control Optim., 39 (2000), pp. 113–132.

[2] R. BECKER, D. MEIDNER, AND B. VEXLER, *Efficient numerical solution of parabolic optimization problems by finite element methods*, Optim. Methods Softw., to appear.

[3] R. BECKER AND R. RANNACHER, *An Optimal Control Approach to A-Posteriori Error Estimation*, Acta Numer. 2001, Arieh Iserles, ed., Cambridge University Press, London, 2001, pp. 1–102.

[4] R. BECKER AND B. VEXLER, *A posteriori error estimation for finite element discretizations of parameter identification problems*, Numer. Math., 96 (2004), pp. 435–459.

[5] R. BECKER AND B. VEXLER, *Mesh refinement and numerical sensitivity analysis for parameter calibration of partial differential equations*, J. Comput. Phys., 206 (2005), pp. 95–110.

[6] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[7] A. R. CONN, N. GOULD, AND PH. L. TOINT, *Trust-Region Methods*, MPS/SIAM Ser. Optim., SIAM, Philadelphia, 2000.

[8] R. DAUTRAY AND J.-L. LIONS, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 5, Springer-Verlag, Berlin, 1992.

[9] K. ERIKSSON, D. ESTEP, P. HANSBO, AND C. JOHNSON, *Introduction to Adaptive Methods for Differential Equations*, Acta Numer. 1995, Arieh Iserles, ed., Cambridge University Press, London, 1995, pp. 105–158.

[10] K. ERIKSSON, D. ESTEP, P. HANSBO, AND C. JOHNSON, *Computational differential equations*, Cambridge University Press, Cambridge, 1996.

[11] K. ERIKSSON, C. JOHNSON, AND V. THOMÉE, *Time discretization of parabolic problems by the discontinuous Galerkin method*, RAIRO Modelisation Math. Anal. Numer., 19 (1985), pp. 611–643.

[12] K. ERIKSSON AND C. JOHNSON, *Adaptive finite element methods for parabolic problems V: Long-time integration*, SIAM J. Numer. Anal., 32 (1995), pp. 1750–1763.

[13] A. V. FURSIKOV, *Optimal Control of Distributed Systems: Theory and Applications*, Transl. Math. Monogr. 187, AMS, Providence, 1999.

[14] A. GRIEWANK AND A. WALTHER, *Revolve: An implementation of checkpointing for the reverse or adjoint mode of computational differentiation*, ACM Trans. Math. Software, 26 (2000), pp. 19–45.

[15] A. GRIEWANK, *Achieving logarithmic growth of temporal and spatial complexity in reverse automatic differentiation*, Optim. Methods Softw., 1 (1992), pp. 35–54.

[16] M. HINZE, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–61.

[17] J. LANG, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems. Theory, Algorithm, and Applications*, Lecture Notes in Earth Sci. 16, Springer-Verlag, Berlin, 1999.

[18] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Grundlehren Math. Wiss. 170, Springer-Verlag, Berlin, 1971.
[19] W. Liu, H. Ma, T. Tang, and N. Yan, *A posteriori error estimates for discontinuous galerkin time-stepping method for optimal control problems governed by parabolic equations*, SIAM J. Numer. Anal., 42 (2004), pp. 1032–1061.
[20] W. Liu and N. Yan, *A posteriori error estimates for distributed convex optimal control problems*, Adv. Comput. Math, 15 (2001), pp. 285–309.
[21] W. Liu and N. Yan, *A posteriori error estimates for control problems governed by nonlinear elliptic equations*, Appl. Numer. Math., 47 (2003), pp. 173–187.
[22] C. Meyer and A. Rösch, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim., 43 (2004), pp. 970–985.
[23] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1999.
[24] M. Picasso, *Anisotropic A Posteriori Error Estimates for an Optimal Control Problem Governed by the Heat Equation*, Internat. J. Numer. Methods PDE, 22 (2006), pp. 1314–1336.
[25] R. Rannacher and B. Vexler, *A priori error estimates for the finite element discretization of elliptic parameter identification problems with pointwise measurements*, SIAM J. Control Optim., 44 (2005), pp. 1844–1863.
[26] M. Schmich and B. Vexler, *Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations*, SIAM J. Sci. Comput., submitted.
[27] F. Tröltzsch, *Optimale Steuerung Partieller Differentialgleichungen*, Friedr. Vieweg & Sohn Verlag, Wiesbaden, 2005.
[28] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley/Teubner, New York/Stuttgart, 1996.
[29] B. Vexler, *Adaptive Finite Elements for Parameter Identification Problems*, Ph.D. thesis, Institut für Angewandte Mathematik, Universität Heidelberg, 2004.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 143–155

EXTENDED KRONECKER SUMMATION FOR CLUSTER TREATMENT OF LTI SYSTEMS WITH MULTIPLE DELAYS*

ALI FUAT ERGENC†, NEJAT OLGAC†, AND HASSAN FAZELINIA†

Abstract. A new procedure is presented for determining the kernel and the offspring hypersurfaces for general linear time invariant (LTI) dynamics with multiple delays. These hypersurfaces, as they have very recently been introduced in a concept paper [R. Sipahi and N. Olgac, *Automatica*, 41 (2005), pp. 1413–1422], form the basis of the overriding paradigm which is called the cluster treatment of characteristic roots (CTCR). In fact, these two sets of hypersurfaces exhaustively represent the locations in the domain of the delays where the system possesses at least one pair of imaginary characteristic roots. To determine the kernel and offspring we use the extraordinary features of the "extended Kronecker summation" operation in this paper. The end result is that the infinite-dimensional problem reduces to a finite-dimensional one (and preferably into an eigenvalue problem). Following the procedure described in this paper, we are able to shorten the computational time considerably in determining these hypersurfaces. We demonstrate these concepts via some example case studies. One of the examples treats a 3-delay system. For this case another interesting perspective, called the "building block," is also utilized to display the kernel in three-dimensional space in the domain of "spectral delays."

Key words. linear time-delayed systems, Kronecker sum, multiple delays, stability, robust stability

AMS subject classifications. 15A15, 15A09, 15A23

DOI. 10.1137/06065180X

1. Introduction and the problem statement.

We consider linear time invariant, retarded multiple time-delayed systems (LTI-MTDS), the general form of which is given as

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \sum_{j=1}^{p} \mathbf{B}_j \mathbf{x}(t-\tau_j), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A}$, $\mathbf{B}_j$, $j = 1 \dots p$, are all constant matrices in $\mathbb{R}^{n\times n}$ and the vector of time delays $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_p) \in \mathbb{R}^{p+}$ of which the elements are rationally independent from each other. As a note of formalism we use boldface capital notation for vector and matrix quantities in the text. We refer to the right (and left) half open complex plane as $\mathbb{C}^+$ ($\mathbb{C}^-$), while $\mathbb{C}^0$ is used to indicate the imaginary axis. Therefore $\mathbb{C}^+ \bigcup \mathbb{C}^- \bigcup \mathbb{C}^0 = \mathbb{C}$ represents the entire complex plane.

The characteristic equation of the system in (1) is

$$\begin{aligned} CE(s, \tau_1, \dots, \tau_p) &= \det\left[s\mathbf{I} - \mathbf{A} - \sum_{j=1}^{p} \mathbf{B}_j e^{-\tau_j s} \right] \\ &= \mathbf{A_0}(s) + \mathbf{A}_{p+1}(s, \tau_1, \dots, \tau_p) + \sum_{j=1}^{p} e^{-n_j \tau_j s} \mathbf{A}_j \left(s, \tau_1, \dots, \tau_{j-1}, \tau_{j+1}, \dots, \tau_p\right), \end{aligned} \tag{2}$$

*Received by the editors February 9, 2006; accepted for publication (in revised form) September 25, 2006; published electronically March 30, 2007. This research was supported in part by awards from DoE (DE-FG02-04ER25656), NSF (CMS-0439980), and NSF (DMI-0522910).
http://www.siam.org/journals/sicon/46-1/65180.html

†Department of Mechanical Engineering, University of Connecticut, Storrs, CT 06269-3139 (ergenc@engr.uconn.edu, olgac@engr.uconn.edu, h_fazeli@engr.uconn.edu).

where $A_0(s)$ is an nth degree polynomial in s and $\mathbf{A}_j$'s ($j = 1 \dots p$) are quasi polynomials in s and all the delays except τ_j. n_j is the highest order of commensuracy of delay τ_j in the dynamics ($n_j \leq n$). $\mathbf{A}_j$ contains s terms with the highest degree of $n-1$ and they are the factors multiplying the representative exponential of the highest commensuracy of τ_j, i.e., $e^{-n_j \tau_j s}$. Since the system is "retarded" the s^n term appears only in $\mathbf{A}_0(s)$ which is free of delays. $\mathbf{A}_{p+1}$ is another quasi polynomial which contains all the remaining terms with lower commensuracy levels (in τ_j) than n_j, $j = 1 \dots p$.

The stability robustness of this general class of systems has been studied for over four decades, resulting in some respectable volumes of literature [11, 13, 7, 15, 18]. One of the mainstream research foci has been the stability assessment of these systems for a given delay vector, $\boldsymbol{\tau}$. The determination of the robustness of such systems against uncertainties in delay and other parameters (i.e., uncertain $\boldsymbol{\tau}$, $\mathbf{A}$, and $\mathbf{B}_j$) are also widely investigated [5, 10, 17, 16, 20, 21]. The class, with delay uncertainty only, is declared to be an *N-P* hard (nondeterministic-polynomial time hard) problem [27]. Many further investigations appeared later on some simplified forms of the problem given here [25, 12, 4]. A very recent paradigm which is introduced by the authors, cluster treatment of characteristic roots (CTCR), brought a practical and numerically efficient procedure for the problem when there are only two delays ($p = 2$) [23, 24, 22]. In fact, this procedure produces a unique stability robustness tableau in the domain of uncertain delays, $\boldsymbol{\tau} \in \mathbb{R}^{p+}$. The numerical efficiency is comfortably demonstrated for cases $n = 3$, $p = 2$ [23], still respecting the difficulties attributed to the *N-P* hardness of the problem. That is, the CTCR method solves the stability robustness problem completely in the delay space, however, with nondeterministic polynomial time hard numerical complexity still remaining. The most critical step in CTCR is the *exhaustive* determination of the stability switching trajectories in the delay space. The primary contributions of the present paper are on this issue.

The key novelties introduced by the CTCR paradigm are the concept of "*kernel* and *offspring* hypersurfaces" and their intriguing characteristics, which were unrecognized earlier. Leaving the details to later segments of the paper, we simply describe the *kernel* hypersurface as the creating loci of *all* the points in $\boldsymbol{\tau} \in \mathbb{R}^{p+}$ space, which render at least one pair of imaginary characteristic roots ($\pm\omega i$) of (2) (or a root at the origin, $s = 0$). Let us denote a generic point on the *kernel* with τ_{ker}. This *kernel* point has an important descriptor: there is no $\boldsymbol{\tau} \in \mathbb{R}^{p+}$, which generates the same imaginary root $\pm\omega i$ and is closer to the origin than $\boldsymbol{\tau}_{\text{ker}}$. This is equivalent to stating that the set $\{\boldsymbol{\tau} | CE(\boldsymbol{\tau}, \omega i) = 0, CE(\boldsymbol{\tau}_{\text{ker}}, \omega i) = 0, |\boldsymbol{\tau}| < |\boldsymbol{\tau}_{\text{ker}}|\}$ is an empty set. When the inequality condition is reversed, we find the *offspring* hypersurfaces of the system on which τ's generate the same imaginary root, $\pm\omega i$. That is, the points on the offspring satisfy the following property: $\{\boldsymbol{\tau} | CE(\boldsymbol{\tau}, \omega i) = 0, CE(\boldsymbol{\tau}_{\text{ker}}, \omega i) = 0, |\boldsymbol{\tau}| > |\boldsymbol{\tau}_{\text{ker}}|\}$. We will give a summary of the outstanding features of the *kernel* and the *offspring* hypersurfaces later in the text for the completeness of the presentation.

A critical and comforting observation under the CTCR paradigm is the claim that (2) can possess imaginary characteristic roots only at some manageably small number of *kernel* hypersurface segments in the domain of the delays. And this number is shown to be upper bounded by n^2. Indeed this finite number of hypersurface segments constitutes the "*kernel* hypersurface set" as explained in section 2.

The text is structured as follows: Section 2 reviews the concept of "*kernel*" and "*offspring*" hypersurfaces under the umbrella of the CTCR paradigm. Section 3 states the two fundamental propositions as the foundation of the new paradigm and presents

the steps of the CTCR procedure for the stability robustness assessment of LTI-MDTS. In the same section, we take advantage of an intriguing procedure called the "Kronecker sum" of matrices. We show that it provides considerable computational advantage over the alternative procedure called the "Rekasius substitution," which was utilized in earlier pursuit of CTCR. This point constitutes the main contribution of the paper. We also present a different perspective on the problem, using the domain of "spectral delays" as a companion computational effort. It is referred to as "building block" representation. Section 4 contains example case studies including one with three delays, $p = 3$.

2. Review of CTCR paradigm. We present an overview of the CTCR paradigm borrowing from [23, 24, 22]. The underlying philosophy can be expressed via the following interlinked observations: (i) The system (1) is infinite-dimensional. That is, it has infinitely many characteristic roots in the finite complex plane, $\mathbb{C}$. (ii) Its stability is guaranteed if there exist no characteristic roots in the open right half plane, $\mathbb{C}^+$. (iii) It is impossible to track all of the infinitely many roots. (iv) One has to focus *only* on the occurrence of the imaginary root crossings at $\pm\omega i$, which can be encountered only at some special settings of $\boldsymbol{\tau} \in \mathbb{R}^{p+}$ [11, 13, 18]. (v) These points show continuous variations resulting in continuously varying imaginary roots. Those surfaces are the only locations where the stability switching can take place: let us call them the "switching hypersurfaces." (vi) One must determine, then, *exhaustively* all such hypersurfaces in $\boldsymbol{\tau} \in \mathbb{R}^{p+}$ space. (vii) However, there is still a countably infinite number of these hypersurfaces [11, 13]. (viii) One has to introduce a feature-based discipline, what we call "the clustering" operation, to those crossings in order to bring the analysis to a manageable size. Clearly, this route can be taken only if there is such a discipline in the root formation.

CTCR achieves precisely this objective, determining the two extraordinary "clustering features" of LTI-MDTS as described in detail later. The deployment of these features on the single delay and two-delay cases has already been reported in earlier investigations [23, 24, 19]. This document presents the first generalized treatment for systems with nth order-p delay treatment and offers an example with three delays.

We state a series of relevant postulates and propositions first. As discussed above, the imaginary characteristic roots, $\pm\omega i$, play a critical role. Let us define a set, which exhaustively contains *all* such frequencies ω for the entire parameter space $\boldsymbol{\tau} \in \mathbb{R}^{p+}$:

$$\Omega = \left\{\omega \,|CE\left(s = \omega i, \boldsymbol{\tau}\right) = 0, \boldsymbol{\tau} \in \mathbb{R}^{p+}, \omega \in \mathbb{R}\right\}. \tag{3}$$

It is intuitively obvious that, except degeneracies (such as a standing root at $s = 0$), Ω contains a continuum of ω values, not discrete [11, 13, 18]. That is, if there is a $\langle \boldsymbol{\tau}, \omega \rangle$ correspondence, $\boldsymbol{\tau} + \boldsymbol{\varepsilon}$ should result in another imaginary root, $\langle \boldsymbol{\tau} + \boldsymbol{\varepsilon}, \omega + \varepsilon_\omega \rangle$. Clearly, $\langle \bullet, \bullet \rangle$ notation implies a causal relation that the p members of the first argument result in the second argument as the imaginary root. The in-depth analysis of existence/uniqueness of $\boldsymbol{\varepsilon}$ and ε_ω correspondence is beyond the scope of this presentation, and therefore it is suppressed.

POSTULATE 1. *The stability posture of the system in* (1) *is determined by the number of open right half plane characteristic roots of* (2)*, which we call NU (abbreviated from* ***N****umber of* ***U****nstable roots). This number is naturally a function of the delays, which are the only parameters in* (1)*, i.e., NU* $(\tau_1, \ldots, \tau_p)$*. Wherever NU* $= 0$*, the system is labeled as "stable." For any change in NU one has to pass through a point on the switching hypersurfaces,* $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)$*, for which a characteristic root* $\pm\omega i$ *exists. This is a direct result of the "root continuity" argument in the parametric*

space $\boldsymbol{\tau} \in \mathbb{R}^{p+}$ [11, 13]. *Obviously* $\Delta NU = \pm 2$ *if* $\omega \neq 0$, *and* $\Delta NU = \pm 1$ *if* $\omega = 0$ *is a simple root crossing. As a critical side note, if* $s = 0$ *is a simple root, it is a stationary one independent of* $\boldsymbol{\tau}$. *Therefore when there is a crossing at* $s = 0$, *it has to appear in the form of a multiple root. That is, at least* $CE(\boldsymbol{\tau}, s)|_{s=0} = 0$ *and* $\frac{d}{ds} CE(\boldsymbol{\tau}, s)|_{s=0} = 0$ *must be satisfied jointly for the same delay set* $\boldsymbol{\tau}$. *As stated above we represent such occurrences of* $\boldsymbol{\tau}$ *yielding* $\pm\omega i$ *as a characteristic root, with the notation* $\langle \boldsymbol{\tau}, \omega \rangle$.

As per El'sgol'ts's D-subdivison principle [8] one can state that a region in $\boldsymbol{\tau} \in \mathbb{R}^{p+}$ space where NU is constant has to be enclosed by hypersurfaces which belong to either one of the following two general classes:

H_1: a hypersurface on which $\langle \boldsymbol{\tau}, \omega \rangle$ occurrence is encountered.

H_2: a hypersurface defined by one or more of the delays being zero, where $\langle \boldsymbol{\tau}, \omega \rangle$ correspondence does not necessarily occur.

We iterate some more on these classes next.

POSTULATE 2. *The* H_2 *class of hypersurfaces is clearly arising from the hard bounds of* $\tau_j \geq 0$, $j = 1 \ldots p$, *posed by the problem statement, and these hypersurfaces are uniquely defined. The* H_1 *class, however, represents a countably infinite number of hypersurfaces possessing the feature*

$$\left\langle \tau_1 \pm \frac{2\pi}{\omega} j_1, \tau_2 \pm \frac{2\pi}{\omega} j_2, \ldots, \tau_p \pm \frac{2\pi}{\omega} j_p, \omega \right\rangle, \quad j_k = 0, 1, \ldots \quad k = 1, \ldots, p. \tag{4}$$

This equation indicates that if $s = \omega i$ occurred at a point $(\tau_1, \ldots, \tau_p)$, infinitely many equidistance grid points in $\boldsymbol{\tau} \in \mathbb{R}^{p+}$ (with the grid size $\frac{2\pi}{\omega}$) would also result in the same characteristic root. Small perturbations on $(\tau_1, \ldots, \tau_p)$ would yield small perturbations on the resultant imaginary root as per the root continuity argument [8]. They sprout in a set of countably infinite number of hypersurfaces which are pointwise equidistant as per (4).

As such the boundaries of the closed regions mentioned above would duplicate themselves as infinitely many hypersurfaces. In order to successfully complete the stability robustness analysis one has to determine *all* of the hypersurfaces where $\langle \boldsymbol{\tau}, \omega \rangle$ correspondence occurs and *all* of these closed regions. This is an impossible task to undertake, unless a well-structured approach is followed. This line of rationale is precisely what prompts the CTCR paradigm.

DEFINITION 1 (kernel hypersurfaces). *Those hypersurfaces which consist of all the points in* $\boldsymbol{\tau} \in \mathbb{R}^{p+}$ *as per* (4) *complying with* $\langle \boldsymbol{\tau}, \omega \rangle$ *correspondence except with the constraint that*

$$0 \leq \tau_k < \frac{2\pi}{\omega} \quad \forall k = 1, \ldots, p \tag{5}$$

are called the kernel hypersurfaces. *This constraint implies that the points on the* kernel hypersurface *exhibit the smallest positive member for each one of its p elements. Notice that there are* ∞^p *(p-dimensional infinite) candidate points in* $\boldsymbol{\tau} = \mathbb{R}^{p+}$ *defined by* (4) *resulting in the same imaginary root,* ωi. *All of these points are represented by a single point on the* kernel hypersurface. *And similar unique points on the* kernel hypersurface *(call it the "*kernel *points") exist for all possible* $\omega \in \Omega$. *Consequently, this hypersurface formation is unique for a given system* (1). *The following notation encapsulates the complete definition of the* kernel hypersurfaces*:*

$$\wp_{\text{ker}} = \textit{kernel hypersurfaces} = \left\{ \boldsymbol{\tau} | \langle \boldsymbol{\tau}, \omega \rangle, \boldsymbol{\tau} \in \mathbb{R}^{p+}, \omega \in \Omega, 0 \leq \tau_k < \frac{2\pi}{\omega} \ \forall k = 1, 2, \ldots, p \right\}. \tag{6}$$

DEFINITION 2 (offspring hypersurfaces). *Those* ∞^p *hypersurfaces which are obtained from the* kernel hypersurface *by a pointwise nonlinear transformation given in* (4) *are called the* offspring. *This definition simply utilizes the fact that a point on the* kernel *will result in* ∞^p *(p-dimensional infinity)* offspring. *The complete formalism for* offspring hypersurfaces *can be given as*

$$\wp_{\text{off}} = \textit{offspring hypersurfaces} = \left\{ \boldsymbol{\tau} | \, \langle \boldsymbol{\tau}, \omega \rangle, \boldsymbol{\tau} \in \mathbb{R}^{p+}, \omega \in \Omega \right\} \setminus \textit{kernel}. \tag{7}$$

3. Determination of the kernel and the offspring hypersurfaces. As per the earlier discussions for the stability robustness of the system we need to determine *all* the kernel and offspring hypersurfaces exhaustively. That amounts to determining the complete set of $\langle \boldsymbol{\tau}, \omega \rangle$ correspondence for the entire $\boldsymbol{\tau} = \mathbb{R}^{p+}$ domain. This mission is computationally very demanding. In our earlier research we utilized a holographic mapping procedure called the Rekasius substitution for this purpose [23]. An alternate procedure is studied here: the extended Kronecker sum method. It is based on the properties of the Kronecker summation of matrices as described for single delay systems in [6, 26]. The treatment prescribed here is an extended version of the process to multiple rationally independent delay cases, and therefore the name "extended Kronecker sum method." We first describe the Kronecker summation operation for clarity and state the main theorem of the paper. A Lyapunov function-based study, which is conducted by another researcher, independently from the authors, has recently been presented at an international conference [14]. It results in the similar outcome of the Kronecker summation method here. Separately, we also recognize that, from the numerical perspective, this process identically coincides with a new concept called the *building block*, which offers a very interesting yet simple representation of the kernel and offspring hypersurfaces [22, 9].

Kronecker summation of two matrices and its properties. In matrix algebra [2, 3] the Kronecker sum of two square matrices $\mathbf{M}_1$ $(n_1 \times n_1)$ and $\mathbf{M}_2$ $(n_2 \times n_2)$ is defined as

$$\mathbf{M}_1 \oplus \mathbf{M}_2 = \mathbf{M}_1 \otimes \mathbf{I}_{n_2} + \mathbf{I}_{n_1} \otimes \mathbf{M}_2, \quad \text{where } \mathbf{M}_1 \in \mathbb{R}^{n_1 \times n_1}, \mathbf{M}_2 \in \mathbb{R}^{n_2 \times n_2}.$$

Here $\oplus$ denotes the *Kronecker summation* and $\otimes$ the *Kronecker product* operations. The most critical feature of the Kronecker summation of $\mathbf{M}_1$ and $\mathbf{M}_2$ is that this new square matrix

$$\mathbf{M}_1 \oplus \mathbf{M}_2 \in \mathbb{R}^{(n_1 \cdot n_2) \times (n_1 \cdot n_2)}$$

has $n_1 \cdot n_2$ eigenvalues which are indeed pairwise combinatoric summations of the n_1 eigenvalues of $\mathbf{M}_1$ and n_2 eigenvalues of $\mathbf{M}_2$. That is, the Kronecker sum operation, in fact, induces the "eigenvalue addition" character to the matrices. We take advantage of this feature as discussed next in a definition and the highlight theorem.

DEFINITION 3. *We define the auxiliary characteristic equation (ACE) of the system in* (1)*, with* $z_j = e^{-\tau_j s}$*:*

$$ACE(\mathbf{z}) = \det \left[\mathbf{A} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{A} + \sum_{j=1}^{p} (\mathbf{B}_j \otimes \mathbf{I} z_j + \mathbf{I} \otimes \mathbf{B}_j z_j^{-1}) \right] = 0. \tag{8}$$

THEOREM 1. *For the system given in* (1) *the following findings are equivalent:*

(a) *A vector of p-dimensional unitary complex numbers* $\mathbf{z} = \{z_j\} \in \mathbb{C}_u^p$, $|z_j| = 1$ *for all* $j = 1 \dots p$ *satisfies ACE.* $\mathbb{C}_u^p$ *is the complete set of such complex numbers.*

(b) *There exists at least one pair of imaginary characteristic roots,* $\pm\omega i$, *of* (2).

(c) *There exists a corresponding delay vector* $\boldsymbol{\tau} = \{\tau_j\} \in \wp_{\text{ker}} \bigcup \wp_{\text{off}}$, *where* $\langle \boldsymbol{\tau}, \omega \rangle$ *holds.*

Proof of Theorem 1. The Laplace transformation of the LTI-MTDS equation (1) is

$$sX(s) = \mathbf{A}X(s) + \sum_{j=1}^{p} \mathbf{B}_j X(s) z_j, \tag{9}$$

where the $z_j = e^{-\tau_j s}$, $j = 1 \dots p$, represents p unitary complex numbers for $s = \omega i$. The following equation is directly obtained from (9):

$$(s\mathbf{I} - \Delta(\mathbf{z}(s)))X(s) = 0, \tag{10}$$

$$\Delta(\mathbf{z}(s)) = \mathbf{A} + \sum_{j=1}^{p} \mathbf{B}_j z_j \in \mathbb{R}^{n \times n}, \quad \text{where } \mathbf{z} = \{z_j\} \in \mathbb{C}_u^p.$$

In order to find nontrivial solution of $X(s)$ the matrix $(s\mathbf{I} - \Delta(\mathbf{z}(s)))$ should be singular, in other words, $\det \Delta(\mathbf{z}(s)) = 0$, or

$$\det \left[s\mathbf{I} - \mathbf{A} - \sum_{j=1}^{p} \mathbf{B}_j z_j \right] = 0. \tag{11}$$

Due to the fact that $\mathbf{A}$, $\mathbf{B}_j$, $j = 1 \dots p$, are all constant matrices in $\mathbb{R}^{n \times n}$, the complex conjugates of s and (indirectly) of z_j also satisfy (11):

$$\det \left[s^*\mathbf{I} - \mathbf{A} - \sum_{j=1}^{p} \mathbf{B}_j z_j^* \right] = 0. \tag{12}$$

For a point on the $\wp_{\text{ker}} \bigcup \wp_{\text{off}}$ hypersurface set $\boldsymbol{\tau} = (\tau_1, \dots, \tau_p) \in \wp_{\text{ker}} \bigcup \wp_{\text{off}}$, $s = \omega i$ is an element of the spectrum. Therefore $s^* = -\omega i$, $z_j = \left\{e^{-\tau_j \omega i}\right\}$ and $z_j^* = \left\{e^{\tau_j \omega i}\right\} = z_j^{-1}$, $j = 1 \dots p$. Since the sum of the two certain eigenvalues of $\Delta(s, \mathbf{z})$ and $\Delta(s^*, \mathbf{z}^*)$ is zero, then the Kronecker sum of these two matrices must be singular when such $\langle \boldsymbol{\tau}, \omega \rangle$ correspondence occurs. That is,

$$\det \left[\left(\mathbf{A} + \sum_{j=1}^{p} \mathbf{B}_j z_j \right) \oplus \left(\mathbf{A} + \sum_{j=1}^{p} \mathbf{B}_j z_j^* \right) \right] = 0. \tag{13}$$

Note that in (13) all the s and $\boldsymbol{\tau}$ terms are incorporated into z_j, where z_j represents the unitary complex numbers, i.e., $|z_j| = 1$. The task of determining $\wp_{\text{ker}} \bigcup \wp_{\text{off}}$ in $\boldsymbol{\tau} \in \mathbb{R}^{p+}$ space is now reduced to evaluating the solution set of $\mathbf{z} \in \mathbb{C}_u^p$ which satisfies (13). Using the Kronecker summation definition, (13) can be rewritten as

$$ACE(\mathbf{z}) = \det \left[\mathbf{A} \otimes \mathbf{I} + \sum_{j=1}^{p} (\mathbf{B}_j \otimes \mathbf{I}) z_j + \mathbf{I} \otimes \mathbf{A} + \sum_{j=1}^{p} (\mathbf{I} \otimes \mathbf{B}_j z_j^*) \right] = 0. \tag{14}$$

Equation (14) is in fact a multinomial in terms of the n components of $\mathbf{z}$, and the highest degree of any one of these components is n^2 in this multinomial. We denote the complete solution set for (14) by $\mathbf{Z}$. That is,

$$\mathbf{Z} = \{\mathbf{z} \,|ACE\,(\mathbf{z}) = 0,\ \mathbf{z} \in \mathbb{C}_u^p\}, \quad \mathbf{Z} \in \mathbb{C}_u^p. \tag{15}$$

Inversely it is trivial to prove that for every solution $\mathbf{z} \in \mathbb{C}_u^p$ of (15) there exists at least one imaginary characteristic root, $s = \pm\omega i$. □

In other words, (8) is both a necessary and sufficient condition for a point $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p) \in \mathbb{R}^{p+}$ to be on either the kernel or the offspring hypersurfaces. Since this equation is completely free from the delays, and only a function of $\mathbf{z}$, the procedure is now considerably simplified to find $\mathbf{z} \in \mathbb{C}_u^p$ solutions of (8) exhaustively. To determine the imaginary characteristic roots of (2) one simply plugs such a $\mathbf{z}$ into (11) and solves for s. These roots reveal the crossing frequencies we are interested in and form the earlier-mentioned set in (3), i.e.,

$$\Omega = \{\omega \,|CE\,(s = \omega i, \mathbf{z}) = 0, \mathbf{z} \in \mathbf{Z}\}\,. \tag{16}$$

One then uses the individual components of s to determine the respective delays which are

$$\tau_{jk} = \frac{\arg(z_j) \mp 2k\pi}{\omega}, \quad j = 1 \ldots p, \quad k = 0, 1, 2, \ldots, \tag{17}$$

where τ_{jk} implies the jth delay value for various k values. One of these delays forms the $\tau_{j\,\mathrm{ker}}$, via the feature described in (5). The remaining delays form an infinitely many equidistant grid in $\mathbb{R}^{p+}$ for the same crossing root, ωi, as also expressed in (4).

Definition 4 (the building blocks). *The complex vector $\mathbf{z} \in \mathbb{C}_u^p$ can be identified exactly by a p-dimensional real vector $\boldsymbol{\tau}\omega \in \mathbb{R}^p$ (i.e., the fundamental phase angle of each component) as per* (17). *The complete solution set $\mathbf{Z}$ of* (15) *corresponds to some hypersurfaces in p-dimensional space of $\boldsymbol{\tau}\omega$. These hypersurfaces are called the "building blocks (BBs)."*

Due to the constraint on $\tau_j\omega \in [0, 2\pi]$ for the fundamental phase angle (also see the discussion below on $\tau_{\mathrm{ker,j}}$), however, the complete BB set can now be confined to a p-dimensional generalized cube of size 2π. This encapsulated BB set on the new domain has recently been studied by the authors' group in [22, 9]. There are some interesting features of BBs which are proven in [9]. The interested reader is referred to that document.

Definition 5 (spectral delay space). *We name the p-dimensional space of $\boldsymbol{\tau}\omega \in \mathbb{R}^p$ the spectral delay space (SDS), as it is the domain of BB representation.*

Determination of the Kernel and the Offspring. The explicit representation $ACE(\mathbf{z}) = 0$ of (14) is, in fact, a class of holographic mappings from $\boldsymbol{\tau} \in \mathbb{R}^{p+}$ space to $\mathbf{z} \in \mathbb{C}_u^p$ space. Every single $\boldsymbol{\tau}$ which renders an imaginary root for (2) creates a corresponding $\mathbf{z}$ which is unique, and it satisfies (14). Inversely, however, for each $\mathbf{z}$ that satisfies (14) one can find infinitely many equidistant $\boldsymbol{\tau}$ points but only one single imaginary root, ωi. Thus $\mathbf{z} \leftrightarrow \boldsymbol{\tau}$ mapping is of "holographic" class [1].

Consequently, if one obtains the complete solution set $\mathbf{Z}$ as given in (15) it would be sufficient to create the kernel and the offspring hypersurfaces we seek. For this, one has to create the transition

$$\mathbf{z} \rightarrow \omega \rightarrow \boldsymbol{\tau}$$

as described in the proof of Theorem 1. The hypersurfaces created in $\boldsymbol{\tau} \in \mathbb{R}^{p+}$ form the complete set of hypersurfaces $\wp_{\text{ker}} \bigcup \wp_{\text{off}}$. The *kernel* is simply identified using the pointwise feature, which selects the minimum positive delay set for any given ω root crossing frequency (or a solution vector $\mathbf{z}$ of (15)). In mathematical formalism the *kernel hypersurfaces* are made of points defined by

$$\tau_{\text{ker}} = \{\min \tau_{jk}\}, \quad j = 1 \dots p, \quad k = 0, 1, 2, \dots, \quad 0 \leq \tau_{\text{ker,j}} < \frac{2\pi}{\omega}, \tag{18}$$

as indicated without details in section 1. Notice that the min operation in (18) is componentwise applied to all p elements of the vector. And the remaining τ_{jk} values from (17) will create the points on the *offspring hypersurfaces,* and these points correspond to the same ω. In short, $\mathbf{z} \rightarrow \boldsymbol{\tau}_{\text{ker}}$ correspondence is one-to-one; one point in the BB corresponds to one point on the kernel hypersurface.

Corollary 1. *The number of kernel hypersurface segments is upperbounded by* n^2.

Proof. The proof is simply by recognizing the fact that the number of BBs is upperbounded by n^2. If we fix $z_1, \dots, z_{p-1}$ in (14), there can be at most n^2 unitary solutions for z_p. This fact implies that there can be a maximum of n^2 layers of the BBs. One-to-one correspondence from these BBs to the kernel hypersurfaces results in the conclusion that kernel hypersurfaces can have at most n^2 segments. ∎

The numerical procedure with Kronecker summation for obtaining the *kernel* and the *offspring* can be performed very efficiently. We will demonstrate this capability in some example case studies next, with a companion treatment of the BB concept and the SDS.

4. Example case studies.

Example 1. A case is borrowed from [23], with $n = 3$ and $p = 2$:

$$\mathbf{A} = \begin{pmatrix} -1 & 13.5 & -1 \\ -3 & -1 & -2 \\ -2 & -1 & -4 \end{pmatrix}, \quad \mathbf{B}_1 = \begin{pmatrix} -5.9 & 0 & 0 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}, \quad \mathbf{B}_2 = \begin{pmatrix} 0 & 7.1 & -70.3 \\ 0 & -1 & 5 \\ 0 & 0 & 6 \end{pmatrix}. \tag{19}$$

The corresponding characteristic equation is

$$\begin{aligned} CE(s, \tau_1, \tau_2) = {} & s^3 + 6s^2 + 45.5s + 111 + (96.9s + 18.3)e^{-(\tau_1+\tau_2)s} + (-5s^2 - 121.3s + \\ & 20.1)e^{-\tau_2 s} + (5.9s^2 + 4.5s - 42.2)e^{-\tau_1 s} + (-6s - 203.4)e^{-2\tau_2 s} + 119.4e^{-(\tau_1+2\tau_2)s} = 0. \end{aligned} \tag{20}$$

This quasi polynomial contains all types of complex formations of transcendental terms, such as individual delays, their commensurate forms (e.g., $e^{-2\tau_2 s}$), cross-talk terms (e.g., $e^{-(\tau_1+\tau_2)s}$), and both cross-talk with commensuracy (e.g., $e^{-(\tau_1+2\tau_2)s}$). As it is accepted in the delay differential equations literature, such a formation presents a very complex problem.

Using Theorem 1 *ACE* is obtained as follows:

(21)

$$
\begin{aligned}
ACE(z_1,z_2) = &\left(\begin{array}{l} 1.7022z_2^6 + 9.0713z_2^5 + 40.7183z_2^4 - 21.1282z_2^3 - 40.1780z_2^2 \\ -123.5525z_2 - 34.0495 + 4.3008z_2^{-1} + 50.1953z_2^{-2} + 7.2436z_2^{-3} \\ +8.6351z_2^{-4} \end{array}\right) z_1^6 \\
+ &\left(\begin{array}{l} -9.2124z_2^6 - 34.3982z_2^5 - 164.0727z_2^4 + 309.8357z_2^3 + 308.8355z_2^2 + 340.0104z_2 \\ -93.5932 - 455.9817z_2^{-1} - 2.4843z_2^{-2} - 85.8515z_2^{-3} + 17.8037z_2^{-4} \\ -7.3179z_2^{-5} \end{array}\right) z_1^5 \\
+ &\left(\begin{array}{l} 16.2587z_2^6 + 19.9951z_2^5 + 204.5044z_2^4 - 883.9491z_2^3 - 429.7956z_2^2 + 26.9372z_2 \\ +763.3490 + 964.9333z_2^{-1} - 78.4191z_2^{-2} - 63.6748z_2^{-3} - 22.3465z_2^{-4} \\ -37.63910z_2^{-5} - 1.4883z_2^{-6} \end{array}\right) z_1^4 \\
+ &\left(\begin{array}{l} -8.5227z_2^6 + 45.9059z_2^5 - 69.8016z_2^4 + 773.7642z_2^3 + 75.7380z_2^2 - 631.0328z_2 \\ -1340.0621 - 631.03287z_2^{-1} + 75.7380z_2^{-2} + 773.7642z_2^{-3} - 69.8016z_2^{-4} \\ +45.9059z_2^{-5} - 8.5227z_2^{-6} \end{array}\right) z_1^3 \\
+ &\left(\begin{array}{l} -1.4883z_2^6 - 37.63910z_2^5 - 22.3465z_2^4 - 63.6748z_2^3 - 78.4191z_2^2 + 964.9333z_2 \\ +763.3490 + 26.9372z_2^{-1} - 429.7956z_2^{-2} - 883.9491z_2^{-3} + 204.5044z_2^{-4} \\ +19.9951z_2^{-5} + 16.2587z_2^{-6} \end{array}\right) z_1^2 \\
+ &\left(\begin{array}{l} -7.3179z_2^5 + 17.8037z_2^4 - 85.8515z_2^3 - 2.4843z_2^2 - 455.9817z_2 - 93.5932 \\ +340.0104z_2^{-1} + 308.8355z_2^{-2} + 309.8357z_2^{-3} - 164.0727z_2^{-4} - 34.3982z_2^{-5} \\ -9.2124z_2^{-6} \end{array}\right) z_1^1 \\
+ &\left(\begin{array}{l} 8.6351z_2^4 + 7.2436z_2^3 + 50.1953z_2^2 + 4.3008z_2 - 34.0495 - 123.5525z_2^{-1} \\ -40.1780z_2^{-2} - 21.1282z_2^{-3} + 40.7183z_2^{-4} + 9.0713z_2^{-5} + 1.7022z_2^{-6} \end{array}\right).
\end{aligned}
$$

For a numerical algorithm to obtain the $(z_1, z_2) \in \mathbb{C}_u^2$ exhaustively, the following steps are followed:

1. Select z_2 using a secondary parameter, θ, as $z_2 = e^{i\theta}$, $\theta \in [0, 2\pi]$; start with $\theta = 0$.
2. With this value of z_2, solve the roots of (21). Those roots with unity magnitude will form the corresponding z_1 values. Obviously, there can be a maximum of 6 such solutions which is less than $n^2 = 9$.
3. Substitute z_1 (for $e^{-\tau_1 \omega i}$) and z_2 (for $e^{-\tau_2 \omega i}$) into (20) and solve for ω, the imaginary spectrum of the system.
4. Then increase θ by a desired resolution, $\Delta\theta$, and repeat steps 1–3.
5. As θ reaches 2π a complete set of imaginary spectra and the corresponding kernel points $(\tau_1, \tau_2) \in \wp_{\text{ker}}$ are obtained using (17). The offspring $\wp_{\text{off}}$ is trivially generated using (17) again.

Notice that ACE is a self-inversive polynomial with interspersed zeros on the unit circle. Although not used here, this feature may be of some computational value in future studies.

Further deployment of Postulates 1 and 2 of CTCR results in the stability robustness tableau of Figure 1 against delay uncertainties. The stable regions are shaded, and the number of unstable roots, NU, is shown sparingly on the figure. The stability outlook matches with that of [23] precisely.

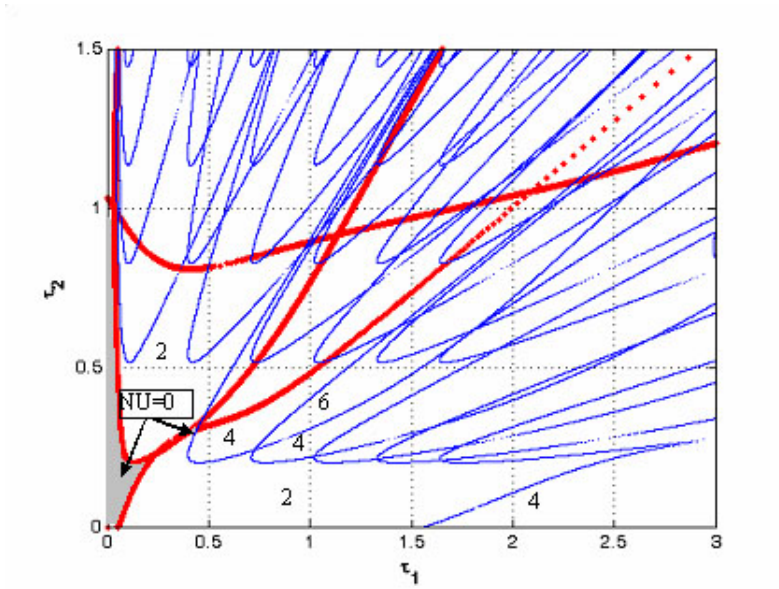


FIG. 1. *Kernel (red, thick) and offspring hyperplanes of Example* 1.

Two main objectives of this example are the following:

1. To demonstrate that the extended Kronecker summation method coincides with the earlier Rekasius-based determination of $\wp_{\text{ker}}$ and $\wp_{\text{off}}$. Clearly both of these procedures are serving for the initial step of the CTCR paradigm.
2. To show substantial computational improvement by utilizing the Kronecker summation method. The computation of kernel and offspring hypersurfaces takes 0.7 seconds on a PC with Pentium-Centrino 2.13 GHz and 1 GB RAM, as opposed to 19.2 sec as stated in [23] which is based on Rekasius substitution. We make a qualified remark that such CPU times can vary considerably for various code writing styles. Nevertheless, we wish to give the reader a comparative perspective.

Example 2. We take a challenging case with three delays $(n = 2, p = 3)$ next:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -8 & -3 \end{pmatrix}, \quad \mathbf{B}_1 = \begin{pmatrix} 0 & 0 \\ -1 & -3 \end{pmatrix}, \quad \mathbf{B}_2 = \begin{pmatrix} 0 & 0 \\ -8 & 1 \end{pmatrix}, \quad \mathbf{B}_3 = \begin{pmatrix} 0 & 0 \\ -5 & 0 \end{pmatrix}$$

with the corresponding characteristic equation

$$CE(s, \tau_1, \tau_2, \tau_3) = s^2 + 3s + 8 + (3s + 1)e^{-\tau_1 s} + (-s + 8)e^{-\tau_2 s} + 5e^{-\tau_3 s} = 0.$$

For this system the $ACE(z_1, z_2, z_3)$ is trivial to obtain from (14), but it is not given here to conserve space. For this example case study, we display in Figure 2 the BB in the SDS, $(\tau_1\omega, \tau_2\omega, \tau_3\omega)$. Notice that for each $(z_1, z_2, z_3) \in \mathbb{C}_u^3$ solution of ACE one obtains a $\langle(\tau_1, \tau_2, \tau_3), \omega\rangle$ correspondence. The relation between z_j and $\tau_j\omega$, $j = 1, 2, 3$, is trivial:

$$\tau_j\omega = \arg(z_1), \quad 0 < \arg(z_1) < 2\pi.$$

The BB represents the complete kernel hypersurface set in the SDS. Every point on the BB has an ω root crossing frequency (which are not shown here), and the BB

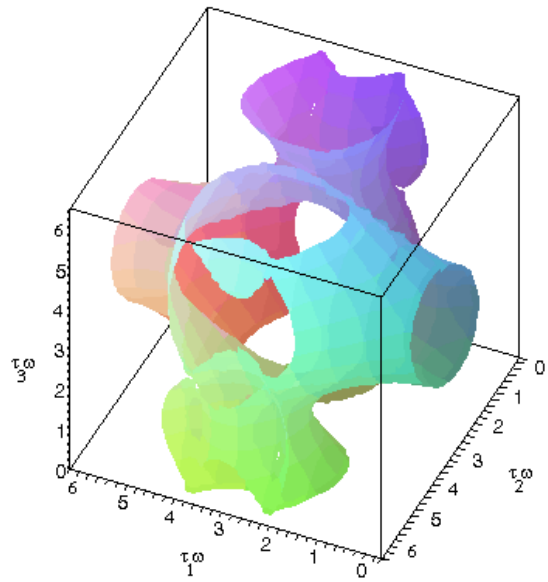


FIG. 2. *BB representation of the system in Example* 2.

traces the hypersurfaces with the following property:

$$\{\omega | \boldsymbol{\tau}\omega \in BB\} = \Omega.$$

That is, the BB represents the complete kernel. For easier visualization, selected cross-sections of Figure 2 are given in Figure 3 for $\tau_3\omega = 0, \frac{\pi}{2}, \pi$, and $\frac{3\pi}{2}$. The offspring form by simple shift operations (by 2π) in any of the three SDS coordinates in accordance with (17); see Figure 3 (blue). The BB in Figure 2 can be transformed into the (τ_1, τ_2, τ_3) delay space to create the ultimate kernel and offspring as shown in Figure 4 for various levels of τ_3. The cross-sections of hypersurfaces with the $\tau_3 =$ constant planes (the kernel and offspring hypercurves) are shown in red and black, respectively. The stable regions are shaded. Notice several small regions of stability (such as $\tau_1 = 1.5, \tau_2 = 0.2, \tau_3 = 2$ or $\tau_1 = 2, \tau_2 = 0.4, \tau_3 = 2.5$) which are easily detected by the CTCR procedure. Computationally this operation is quite efficient. Just to give an idea to the reader, each one of the frames in Figure 4 is obtained within 0.8 sec CPU time. Therefore, it is only a matter of computational capacity to create a sufficiently dense cross-section of the kernel and the offspring even in the case of three independent delays. Another study of the authors' group, on a simpler dynamics with three delays (without cross-talk), was also presented at a recent conference [28].

5. Conclusions. This paper presents the stability robustness of LTI systems with multiple delays against uncertainties in delays. The main contribution is in the utilization of some intriguing properties of the Kronecker summation of matrices. One of them (the eigenvalue summation feature) yields a numerically efficient process for determining the stability switching hypersurfaces, which are called the kernel and the offspring hypersurfaces. This effort constitutes the initial step for the umbrella paradigm, the CTCR, which, in turn, resolves the stability robustness of LTI systems with multiple delays against delay uncertainties. Numerical efficiency improvement is the main benefit of using the Kronecker summation property. A companion perspective, the BB representation, is also given for a three-delay example case study.

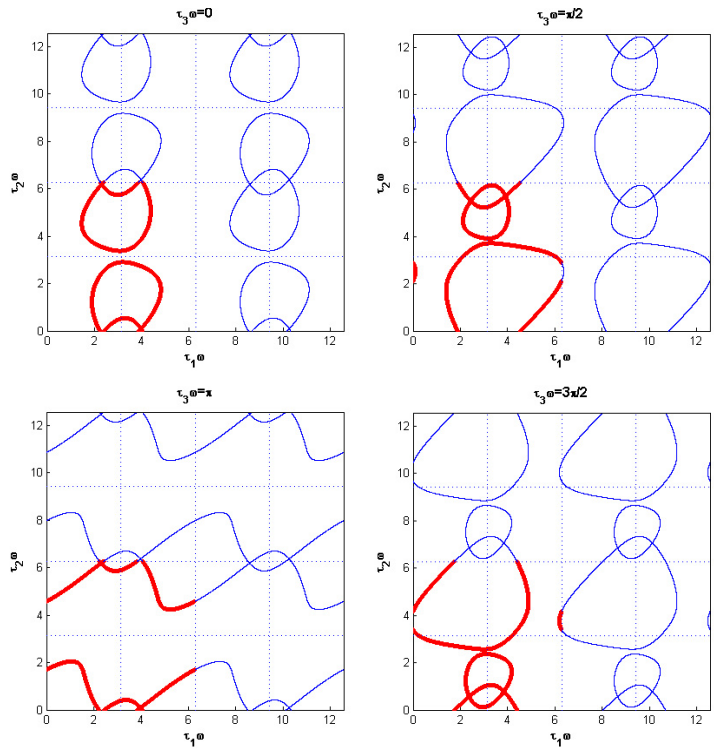


FIG. 3. *Cross-sectional outlook of the BB (red, thick) and the offspring (blue, thin) for various values of* $\tau_3\omega$.

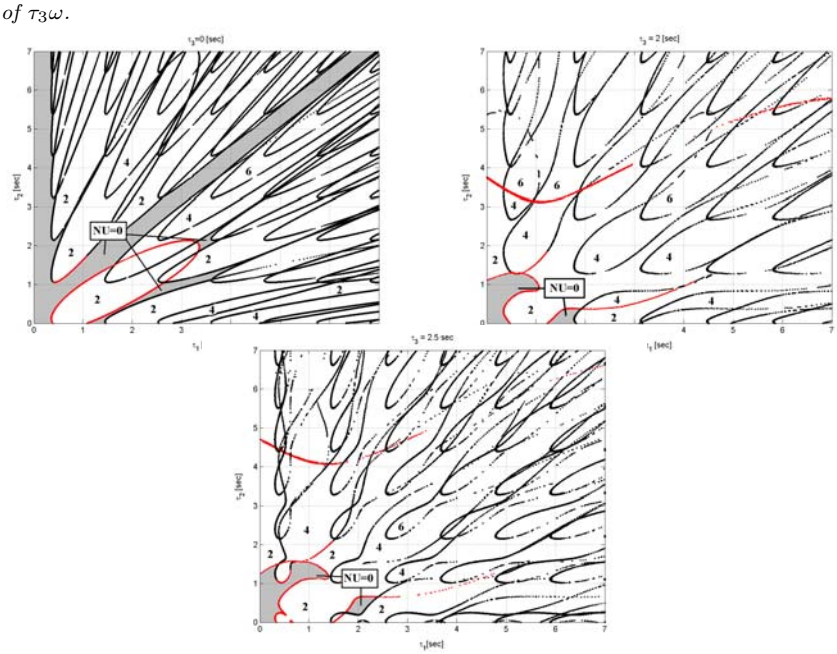


FIG. 4. *Cross-sections of kernel (red, thick) and offspring hypersurfaces (black, thin) for various* τ_3 *values. Stable regions* ($NU = 0$) *are shaded.*

REFERENCES

[1] D. H. Barnhart, N. A. Halliwell, and J. M. Coupland, *Holographic velocimetry with object conjugate reconstruction (OCR): Simultaneous velocity mapping in fluid and solid mechanics*, in Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 2089–2104.

[2] D. S. Bernstein, *Matrix Mathematics*, Princeton University Press, Princeton, NJ, 2005.

[3] J. W. Brewer, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits and Systems, 25 (1978), pp. 772–781.

[4] S. A. Campbell, *Stability and bifurcation of a simple neural network with multiple time delays*, Fields Inst. Commun., 21 (1999), pp. 65–79.

[5] H. Chapellat and S. P. Bhattacharyya, *A generalization of Kharitonov's theorem: Robust stability of interval plants*, IEEE Trans. Automat. Control, 34 (1989), pp. 306–311.

[6] J. Chen, G. Gu, and C. N. Nett, *A new method for computing delay margins for stability of linear delay systems*, Systems Control Lett., 26 (1995), pp. 107–117.

[7] K. L. Cooke and P. van den Driessche, *On zeros of some transcendental equations*, Funkcial. Ekvac., 29 (1986), pp. 77–90.

[8] L. E. El'sgol'ts and S. B. Norkin, *Introduction to the Theory and Application of Differential Equations with Deviating Arguments*, Academic Press, New York, 1973.

[9] H. Fazelinia, R. Sipahi, and N. Olgac, *Stability analysis of multiple time delayed systems using "building block" concept*, IEEE Trans. Automat. Control, to appear.

[10] M. Fu, A. W. Olbrot, and M. P. Polis, *Robust stability for time-delayed systems: The edge theorem and graphical tests*, IEEE Trans. Automat. Control, 34 (1989), pp. 813–819.

[11] J. K. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[12] J. K. Hale and W. Huang, *Global geometry of the stable regions for two delay differential equations*, J. Math. Anal. Appl., 178 (1993), pp. 344–362.

[13] J. K. Hale and S. M. Verduyn Lunel, *An Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[14] E. Jarlebring, *Computing the stability region in delay-space of a TDS using polynomial eigenproblems*, in Proceedings of the Sixth IFAC Workshop on Time-Delay Systems, L'Aquila, Italy, 2006.

[15] E. W. Kamen, *Linear systems with commensurate time delays: Stability and stabilization independent of delay*, IEEE Trans. Automat. Control, 25 (1982), pp. 367–375.

[16] V. L. Kharitonov, *Robust stability analysis of time delay systems: A survey*, Annu. Rev. Control, 23 (1999), pp. 185–196.

[17] V. L. Kharitonov and A. P. Zhabko, *Robust stability of time-delay systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 2388–2397.

[18] R. Nussbaum, *Differential delay equations with two time lags*, Mem. Amer. Math. Soc., 16 (1978).

[19] N. Olgac and R. Sipahi, *An exact method for the stability analysis of time delayed LTI systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 793–797.

[20] J. P. Richard, *Time-delay systems: An overview of some recent advances and open problems*, Automatica, 39 (2003), pp. 1667–1694.

[21] G. J. Silva, A. Datta, and S. P. Bhattacharyya, *PI stabilization of first-order systems with time delay*, Automatica, 37 (2001), pp. 2025–2031.

[22] R. Sipahi, H. Fazelinia, and N. Olgac, *Generalization of Cluster Treatment of Characteristic Roots for Robust Stability of Multiple Time Delayed Systems*, http://www.engr.uconn.edu/alarm/reports.html (2005).

[23] R. Sipahi and N. Olgac, *Complete stability robustness of third-order LTI multiple time-delay systems*, Automatica, 41 (2005), pp. 1413–1422.

[24] R. Sipahi and N. Olgac, *A unique methodology for the stability robustness of multiple time delay systems*, Systems Control Lett., 55 (2006), pp. 819–825.

[25] G. Stepan, *Retarded Dynamical Systems: Stability and Characteristic Function*, Longman Scientific and Technical, Harlow, John Wiley and Sons, New York, 1989.

[26] J.-H. Su, *The asymptotic stability of linear autonomous systems with commensurate time delays*, IEEE Trans. Automat. Control, 40 (1995), pp. 1114–1117.

[27] O. Toker and H. Ozbay, *Complexity issues in robust stability of linear delay-differential systems*, Math. Control Signals Systems, 9 (1996), pp. 386–400.

[28] R. Sipahi and N. Olgac, *Stability map of systems with three independent delays*, in Proceedings of the American Control Conference, Minneapolis, MN, 2006.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 156–175

MEAN-VARIANCE PORTFOLIO SELECTION UNDER PARTIAL INFORMATION*

JIE XIONG† AND XUN YU ZHOU‡

Abstract. This paper is concerned with a continuous-time mean-variance portfolio selection problem in a (possibly incomplete) market with multiple stocks and a bond. Only the past price movements of the stocks and the bond are the information available to the investors. A separation principle is shown to hold in this setting. Efficient strategies based on the aforementioned partial information are derived, which involve the optimal filter of the stock appreciation rate processes. The main methodological contribution of the paper is to employ the particle system representation to develop analytical and numerical approaches in obtaining the filter as well as solving the related backward stochastic differential equation.

Key words. mean-variance portfolio selection, continuous time, partial information, nonlinear filtering, backward stochastic differential equation, particle system representation

AMS subject classifications. 91B28, 93C41, 93E11

DOI. 10.1137/050641132

1. Introduction.

In the Nobel Prize winning work [19], Markowitz proposed the mean-variance portfolio selection model for a single investment period, where an agent seeks to minimize the risk of his investment, measured by the variance of his return, subject to a given mean return. The dynamic extension of the Markowitz model, especially in continuous time, has been studied extensively in recent years; see, e.g., Li and Ng [15], Zhao and Ziemba [30], Zhou and Li [31], Lim [16], Bielecki et al. [2], and Xia [25]. (In particular, refer to Steinbach [23] and Bielecki et al. [2] for elaborative discussions on the history of the mean-variance model.) In many of these works, explicit, analytic forms of efficient portfolios have been obtained. However, in all these works it is assumed that the driving Brownian motions are completely observable by an investor, which in reality is more of an exception than a rule. Practically, the investor can observe only the stock prices (including past and present) on which he will base his investment decisions. This leads to the so-called partially observed portfolio selection problem, and this paper aims to solve the problem in the realm of mean variance. An important finding in this paper is that the separation principle (for separating filtering and optimization) turns out to hold in the mean-variance setting, which in turn greatly simplifies the problem. Another main contribution of the paper is to employ the particle system representation, which has been developed quite recently for solving stochastic partial differential equations (SPDEs), to develop analytical and numerical approaches in obtaining the filter as well as solving the related backward stochastic differential equation (BSDE).

*Received by the editors September 24, 2005; accepted for publication (in revised form) October 22, 2006; published electronically March 30, 2007.

http://www.siam.org/journals/sicon/46-1/64113.html

†Department of Mathematics, University of Tennessee, Knoxville, TN 37996-1300, and Department of Mathematics, Hebei Normal University, Shijiazhuang 050016, People's Republic of China (jxiong@math.utk.edu). The research of this author was partially supported by NSA grant H98230-05-1-0043. Most of this work was done when this author was visiting the Chinese University of Hong Kong, the hospitality and financial support of which are gratefully acknowledged.

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk). The research of this author was supported by RGC grant CUHK4175/03E and the Croucher Senior Research Fellowship.

Asset allocation and asset pricing based on partial information under various setups have been studied extensively in the financial economics literature; see, for example, Lakner [14], Brennan and Xia [3], Xia [26], Rogers [22], Nagai and Peng [20], and Yang and Xiong [27]. Detemple [4], Dothan and Feldman [5], and Gennotte [7] established a separation principle. However, all these works are predominantly done within the expected utility framework. (Refer to [2, 23, 30] for discussions on crucial differences between the utility and mean-variance models.) Pham [21] considered a mean-variance hedging problem for a general semimartingale model and proved a separation principle for a diffusion model (though it is not a multidimensional geometric Brownian motion as in the present paper). Although in theory a mean-variance problem a là Markowitz can be formulated as a mean-variance hedging problem, there are subtleties that must be considered such as the feasibility and the determination of the Lagrange multiplier (see [2, 16]). Moreover, the analysis in [21] is rather involved due to the martingale method employed, whereas here we will give a very direct, clean, and short proof for a separation principle.

The rest of this article is organized as follows. In section 2, we formulate the mean-variance portfolio selection model under partial information. In section 3, we derive the innovation process associated with the filtering problem, which leads to the separation principle. Section 4 studies the optimal filter in details for two cases. Section 5 is devoted to the optimization part as well as the final solution to the partially observed mean-variance problem. A numerical solution to a related BSDE is presented, which is of independent interest.

2. The model. We consider a market consisting of d stocks and a bond whose prices are stochastic processes $S_i(t)$, $i = 0, 1, \dots, d$, governed by the following SDEs:

$$\begin{cases} dS_i(t) = S_i(t)\left(\mu_i(t)dt + \sum_{j=1}^m \tilde{\sigma}_{ij}(t)d\tilde{W}_j(t)\right), & i = 1, 2, \dots, d, \\ dS_0(t) = S_0(t)\mu_0(t)dt, \quad t \geq 0, \end{cases}$$

where $\tilde{W} := (\tilde{W}_1, \dots, \tilde{W}_m)^*$ is a standard Brownian motion defined on a filtered complete probability space $(\Omega, \mathcal{F}, P; \{\mathcal{F}_t\}_{t\geq 0})$; $\mu_i(t)$, $i = 1, 2, \dots, d$, are the appreciation rate processes of the stocks; $\mu_0(t)$ is the interest rate process; and the $d \times m$ matrix valued process $\tilde{\Sigma}(t) := (\tilde{\sigma}_{ij}(t))$ is the volatility process. Here and throughout the paper A^* denotes the transpose of a matrix A.

Let

$$\mathcal{G}_t := \sigma(S_i(s) : \ s \leq t, \ i = 0, 1, 2, \dots, d), \ \ t \geq 0.$$

In our model $\mathcal{G}_t$, rather than $\mathcal{F}_t^{\tilde{W}}$ (the filtration generated by $\tilde{W}$), is the only information available to the investors at time t.

By Itô's formula, we have

$$d\log S_i(t) = \left(\mu_i(t) - \frac{1}{2}a_{ii}(t)\right)dt + \sum_{j=1}^m \tilde{\sigma}_{ij}(t)d\tilde{W}_j(t), \ \ i = 1, 2, \dots, d, \tag{2.1}$$

where

$$a_{ij}(t) := \sum_{k=1}^m \tilde{\sigma}_{ik}(t)\tilde{\sigma}_{jk}(t), \ \ i, j = 1, 2, \dots, d.$$

The following assumptions on the market coefficients will be in force throughout this paper.

Assumption (*ND*). For any $t \geq 0$, the $d \times d$ matrix $A(t) := (a_{ij}(t))$ is of full rank a.s. (almost surely).

Assumption (*BC*). There exists a finite constant C such that $\forall\ t \geq 0$, $\forall i, j$,

$$|\tilde{\sigma}_{ij}(t)| \leq C \qquad \text{a.s.}$$

Assumption (*IC*). $\mathbb{E}\int_0^T (\mu_0(t)^2 + |\mu(t)|^2)dt < \infty$.

Remark 2.1. In this article, we allow $d < m$ as long as condition (ND) is satisfied; in other words, the market itself is allowed to be incomplete. It is interesting to note that, unlike the full information case, the incompleteness of the market does not impose essential difficulty in the partial information case. This can be explained as follows. In the classical model of incomplete market (with full information), there are vast amounts of information available. Namely, one has to seek optimal portfolios in the class of all $\mathcal{F}_t^{\tilde{W}}$-adapted portfolios. When $m > d$, the number of available stocks is fewer than that of the (independent) random factors and hence, some of the market risks cannot be completely eliminated by composing an appropriate stock portfolio. As a result, portfolio selection problems become harder than the case with a complete market because some contingent claims cannot be replicated. In our current setup, the available information comes *only* from the stocks themselves, and any other information is not observable anyway. Therefore, the model is essentially complete, as also will be evident in what follows, although the market is indeed incomplete in the conventional sense.

It is easy to show that the quadratic covariation process between $\log S_i(t)$ and $\log S_j(t)$ is given by $\int_0^t a_{ij}(s)ds$. Therefore, the matrix valued process $(a_{ij}(t))$ is $\mathcal{G}_t$-adapted. Let $\Sigma(t) \equiv (\sigma_{ij}(t))$ be the square root of $A(t)$. Then, $\sigma_{ij}(t)$ is $\mathcal{G}_t$-adapted, i.e., it is completely observable. As we shall see in (3.5) below, the stock price $S_i(t)$ satisfies an equivalent SDE which depends on $\sigma_{ij}(t)$ instead of $\tilde{\sigma}_{ij}(t)$. Moreover, $\mu_0(t) = \frac{d}{dt}\log S_0(t)$ is also $\mathcal{G}_t$-adapted. Therefore, we do not need to consider the filtering problem for the stochastic interest rate and volatility processes.

However, the stochastic process $\mu(t) := (\mu_1(t), \ldots, \mu_d(t))^*$ is not necessarily $\mathcal{G}_t$-adapted and hence, its value is not available to the investors. Note that $\mu(t)$, being a very general process, does not need to be even $\mathcal{F}_t^{\tilde{W}}$-adapted.

Denote by $L^2_{\mathcal{G}}(0,T;\mathbb{R}^n)$ the set of $\mathbb{R}^n$-valued, $\mathcal{G}_t$-adapted processes $f(t)$ with $\mathbb{E}\int_0^T |f(t)|^2 dt < \infty$. (Similar notation $L^2_{\mathcal{H}}(0,T;\mathbb{R}^n)$ can be defined for any filtration $\mathcal{H}_t$.) $L^2_{\mathcal{G}}(0,T;\mathbb{R}^n)$ becomes a Hilbert space endowed with the norm $\| f \|_{L^2_{\mathcal{G}}(0,T;\mathbb{R}^n)} := (\mathbb{E}\int_0^T |f(t)|^2 dt)^{\frac{1}{2}}$.

We now define the class of admissible portfolios (investment strategies).

Definition 2.2. *A d-dimensional process $u(t) \equiv (u_1(t), \ldots, u_d(t))^*$ is an* admissible portfolio *if $u(t) \in L^2_{\mathcal{G}}(0,T;\mathbb{R}^d)$.*

In the preceding definition, $u_i(t)$ represents the worth (dollar amount) of an agent's wealth in the ith stock, $i = 1, 2, \ldots, d$. It is well known that under a so-called self-financed portfolio, the wealth process of an agent, starting with an initial wealth x_0, satisfies the following *wealth equation* (see, e.g., [11]):

$$\begin{cases} dx(t) = \big(\mu_0(t)x(t) + \sum_{i=1}^d (\mu_i(t) - \mu_0(t))u_i(t)\big)dt \\ \qquad\qquad + \sum_{i=1}^d \sum_{j=1}^m \tilde{\sigma}_{ij}(t)u_i(t)d\tilde{W}_j(t), \quad t \geq 0, \\ x(0) = x_0. \end{cases} \tag{2.2}$$

The partially observed mean-variance portfolio selection model is formulated as the following optimization model:

$$\text{Minimize} \quad \mathrm{Var}(x(T)) = \mathbb{E}(x(T) - \mathbb{E}x(T))^2$$

$$\text{subject to} \quad \begin{cases} u(t) \text{ is self-financed and admissible,} \\ (x(t), u(t)) \text{ satisfies (2.2) with initial wealth } x_0, \\ \mathbb{E}x(T) = z, \end{cases} \tag{2.3}$$

where $x_0, z \in \mathbb{R}$ are given constants.

3. Separation principle. In this section, we consider the filtering problem associated with our model (2.3) and establish a separation principle. Specifically, we define the innovation process for the filtering problem. Based on this process, we will derive a $\mathcal{G}_t$-adapted representation for the wealth process corresponding to any self-financed admissible portfolio.

THEOREM 3.1. *Under any self-financed admissible portfolio $u(t)$, the corresponding wealth process $x(t)$ satisfies the following SDE:*

$$\begin{cases} dx(t) = \left(x(t)\mu_0(t) + \sum_{i=1}^d (\bar{\mu}_i(t) - \mu_0(t))u_i(t)\right) dt + \sum_{i,j=1}^d \sigma_{ij}(t)u_i(t)d\nu_j(t), \ t \geq 0, \\ x(0) = x_0, \end{cases} \tag{3.1}$$

where $\bar{\mu}_i(t) := \mathbb{E}(\mu_i(t)|\mathcal{G}_t)$ is the optimal filter of $\mu_i(t)$, and the innovation process *$\nu(t) \equiv (\nu_1(t), \ldots, \nu_d(t))^*$ is a d-dimensional Brownian motion given by*

$$d\nu(t) := \Sigma(t)^{-1} d\log S(t) - \Sigma(t)^{-1} \left(\bar{\mu}(t) - \frac{1}{2}\tilde{A}(t)\right) dt, \tag{3.2}$$

where

$$S(t) := (S_1(t), \ldots, S_d(t))^*, \quad \log S(t) := (\log S_1(t), \ldots, \log S_d(t))^*,$$
$$\bar{\mu}(t) := (\bar{\mu}_1(t), \ldots, \bar{\mu}_d(t))^*, \quad \text{and } \tilde{A}(t) := (a_{11}(t), \ldots, a_{dd}(t))^*.$$

Proof. From (2.1), we see that

$$\log S_i(t) - \log S_i(0) - \int_0^t \left(\mu_i(s) - \frac{1}{2}a_{ii}(s)\right) ds = \sum_{j=1}^m \int_0^t \tilde{\sigma}_{ij}(s) d\tilde{W}_j(s), \quad i = 1, 2, \ldots, d,$$

are martingales with a quadratic covariation process $\int_0^t A(s)ds = \int_0^t \Sigma(s)^2 ds$. By the martingale representation theorem, there exists a standard Brownian motion $W \equiv (W_1, \ldots, W_d)$ on $(\Omega, \mathcal{F}, P)$ such that

$$\sum_{j=1}^m \tilde{\sigma}_{ij}(t) d\tilde{W}_j(t) = \sum_{j=1}^d \sigma_{ij}(t) dW_j(t), \qquad i = 1, \ \ldots, \ d. \tag{3.3}$$

Thus,

$$d\log S_i(t) = \left(\mu_i(t) - \frac{1}{2}a_{ii}(t)\right) dt + \sum_{j=1}^d \sigma_{ij}(t) dW_j(t), \qquad i = 1, \ \ldots, \ d. \tag{3.4}$$

Equivalently, the stock prices satisfy the following modified SDE:

$$dS_i(t) = S_i(t)\left(\mu_i(t)dt + \sum_{j=1}^{d}\sigma_{ij}(t)dW_j(t)\right), \qquad i = 1, \ldots, d. \tag{3.5}$$

Note that $\Sigma(t)$ is invertible. Let $\tilde{S}(t)$ be defined by

$$d\tilde{S}(t) := \Sigma(t)^{-1} d\log S(t).$$

We can write the observation equation (3.4) in the classical form (cf. (8.1.1) in [10]):

$$\tilde{S}(t) = \tilde{S}(0) + \int_0^t \Sigma(s)^{-1}\left(\mu(s) - \frac{1}{2}\tilde{A}(s)\right)ds + W(t) \tag{3.6}$$

with the observation σ-field $\mathcal{G}_t$. By Theorem 8.1.3 and Remark 8.1.1 in Kallianpur [10], $(\nu(t), \mathcal{G}_t)$ is a d-dimensional Brownian motion such that $\sigma(\nu(u) - \nu(s) : u \geq s \geq t)$ is independent of $\mathcal{G}_t$.

By (3.6) and (3.2), we get

$$\Sigma(t)dW(t) = \Sigma(t)d\nu(t) + (\bar{\mu}(t) - \mu(t))dt. \tag{3.7}$$

The desired form of wealth equation (3.1) then follows from (2.2), (3.3), and (3.7). □

Remark 3.2. A notorious difficulty in tackling general stochastic optimization problems with partial information is that one usually cannot separate the filtering and optimization, except for some very rare situations. The significance of Theorem 3.1 is that for the specific mean-variance portfolio selection problem, the separation principle happens to hold: one can simply replace the appreciation rate with its filter in the wealth equation and then solve the resulting optimization problem as in the complete information case.

4. Filtering. In this section, we study the filtering problem (for the appreciation rate process) by considering two cases associated with the volatility processes. The aim is to study the optimal filter $U(t)$ given by

$$\langle U(t), f\rangle := \mathbb{E}(f(\mu(t))|\mathcal{G}_t) \qquad \forall\, f \in C_b(\mathbb{R}^d).$$

4.1. Case 1: Nonrandom $\tilde{\Sigma}$. In this subsection we consider the case when the original volatility process $\tilde{\Sigma}(t)$ is a deterministic matrix valued function of t and $\mu(t)$ a d-dimensional Markov process (with a generator L) independent of $\tilde{W}$. By the definitions of $\Sigma(t)$ and W, where W is defined via (3.3), it is clear that $\Sigma(t)$ is also nonrandom and $\mu(t)$ is independent of W. Then the filtering problem becomes a classical one with the signal $\mu(t)$ and observation $\tilde{S}(t)$ given by (3.6). In this case, $\mathcal{G}_t = \mathcal{F}_t^{\tilde{S}}$.

Remark 4.1. By Theorem 8.3.1 in [10], every square integrable $\mathcal{F}_t^{\tilde{S}}$-martingale (and hence, in the present case every $\mathcal{G}_t$-martingale) Y_t can be represented as

$$Y_t = \mathbb{E}(Y_0) + \int_0^t \Phi(s)^* d\nu(s), \tag{4.1}$$

where $\Phi(s) \in L^2_{\mathcal{F}^{\tilde{S}}}(0, T; \mathbb{R}^d)$. This fact will be useful in Proposition 5.5 below in establishing an optimal portfolio since $\mathbb{H} = AC(\mathcal{G})$ in this case (cf. Definition 5.1). Namely, the market is complete in the sense of Definition 5.2.

In Kurtz and Xiong [12, 13], a large class of SPDEs, with the filtering equations as a special case, and their numerical solutions are studied based on a technique called the particle system representation. In this subsection, we demonstrate that the particle system representation itself can be used to derive the filtering equations directly. Note that the generator L of the appreciation rate process $\mu(t)$ is not required to be a second-order differential operator. In fact, even the continuity of $\mu(t)$ in t need not to be assumed. It is also worth mentioning that the results in this and the next subsections are not covered by those of [12, 13].

Introduce the following integrability condition:

$$\int_0^T \left| \left(\mu(s)^* - \frac{1}{2}\tilde{A}(s)^* \right) \Sigma(s)^{-1} \right|^2 ds < \infty \qquad \text{a.s.} \tag{4.2}$$

Applying Girsanov's formula to (3.6) and noting that μ and W are independent, under the probability $\tilde{P}$ defined below, we get that $\tilde{S}(t)$ is a Brownian motion independent of $\mu(t)$, where $dP = M(T)d\tilde{P}$ with

$$M(t) := \exp\biggl(\int_0^t \left(\mu(s)^* - \frac{1}{2}\tilde{A}(s)^* \right) \Sigma(s)^{-1} d\tilde{S}(s) - \frac{1}{2} \int_0^t \left| \left(\mu(s)^* - \frac{1}{2}\tilde{A}(s)^* \right) \Sigma(s)^{-1} \right|^2 ds \biggr).$$

By the Kallianpur–Striebel formula, we can represent the optimal filter $U(t)$ as

$$\langle U(t), f \rangle = \frac{\langle V(t), f \rangle}{\langle V(t), 1 \rangle}, \tag{4.3}$$

where for $f \in C_b(\mathbb{R}^d)$,

$$\langle V(t), f \rangle := \tilde{\mathbb{E}} \left(M(t) f(\mu(t)) | \mathcal{G}_t \right)$$

is the unnormalized filter. Here $\tilde{\mathbb{E}}$ refers to the expectation under the new probability measure $\tilde{P}$.

LEMMA 4.2. *Suppose that $\mu(t)$ is a d-dimensional Markov process (with generator L) independent of W satisfying* (4.2). *Then, $V(t)$ is represented as*

$$\langle V(t), f \rangle = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} M^i(t) f(\mu^i(t)), \tag{4.4}$$

where $\mu^1(t)$, $\mu^2(t)$, ... are independent copies of $\mu(t)$ and

$$M^i(t) := \exp\biggl(\int_0^t \left(\mu^i(s)^* - \frac{1}{2}\tilde{A}(s)^* \right) \Sigma(s)^{-1} d\tilde{S}(s) - \frac{1}{2} \int_0^t \left| \left(\mu^i(s)^* - \frac{1}{2}\tilde{A}(s)^* \right) \Sigma(s)^{-1} \right|^2 ds \biggr). \tag{4.5}$$

Proof. Note that

$$dM^i(t) = M^i(t) \left(\mu^i(t)^* - \frac{1}{2}\tilde{A}(t)^* \right) \Sigma(t)^{-1} d\tilde{S}(t). \tag{4.6}$$

Denote by $\mathcal{X}$ the collection of all those processes $\mu(t)$ satisfying (4.2) and by $\mathcal{Y}$ the collection of all measurable $\mathbb{R}^{2d}$-valued random vectors. Then $\mathcal{X}$ and $\mathcal{Y}$ are measurable spaces. For $\mu^i \in \mathcal{X}$, the SDE (4.6) has a unique strong solution. Therefore, for each fixed $t \in [0,T]$ there is a measurable functional $F_t : C([0,T],\mathbb{R}^d) \times \mathcal{X} \to \mathcal{Y}$ such that $(\mu^i(t), M^i(t)) = F_t(\tilde{S}, \mu^i)$. As a consequence, under the conditional probability $\tilde{P}(\cdot|\mathcal{F}_t^{\tilde{S}})$, $(\mu^i(t), M^i(t))$ is completely determined by μ^i, $i = 1, 2, \ldots$. Since $\tilde{S}, \mu^1, \mu^2, \ldots$ are independent, the strong law of large numbers yields

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} M^i(t) f(\mu^i(t)) = \tilde{\mathbb{E}}\left(M(t) f(\mu(t)) | \mathcal{F}_t^{\tilde{S}} \right). \tag{4.7}$$

Since $\mathcal{F}_t^{\tilde{S}} = \mathcal{G}_t$, we obtain (4.4). □

Now we derive via Itô's formula and (4.4)–(4.5) an SPDE for the unnormalized filter $V(t)$. Let $\mu^i(t)$, $i = 1, 2, \ldots$, be as in Lemma 4.2. It is well known (see the standard textbooks of Ethier and Kurtz [6] or Stroock and Varadhan [24]) that there are independent martingales $N_f^i(t)$ such that

$$df(\mu^i(t)) = Lf(\mu^i(t))dt + dN_f^i(t), \quad f \in D(L), \tag{4.8}$$

where $D(L)$ is the domain of L.

Applying Itô's formula to (4.6) and (4.8), we get

$$\begin{aligned} d(M^i(t) f(\mu^i(t))) &= M^i(t) \left(Lf(\mu^i(t))dt + dN_f^i(t) \right) \\ &\quad + M^i(t) f(\mu^i(t)) \left(\left(\mu^i(t)^* - \frac{1}{2}\tilde{A}(t)^* \right) \Sigma(t)^{-1} d\tilde{S}(t) \right). \end{aligned}$$

Taking an average for $i = 1, 2, \ldots, k$, letting $k \to \infty$, and applying Lemma 4.2, we see that $V(t)$ satisfies the following Zakai equation:

$$d\langle V(t), f\rangle = \langle V(t), Lf\rangle\, dt + \langle V(t), G_t f\rangle\, d\tilde{S}(t), \tag{4.9}$$

where

$$G_t f(\mu) := f(\mu) \left(\mu^* - \frac{1}{2}\tilde{A}(t)^* \right) \Sigma(t)^{-1}.$$

Making use of (4.3), by Itô's formula, we have

$$\begin{aligned} d\langle U(t), f\rangle &= \langle U(t), Lf\rangle\, dt \\ &\quad + \left(\langle U(t), G_t f\rangle - \langle U(t), G_t 1\rangle \langle U(t), f\rangle \right) \left(d\tilde{S}(t) - \langle U(t), G_t 1\rangle\, dt \right). \end{aligned}$$

Note that

$$\begin{aligned} \langle U(t), G_t 1\rangle &= \mathbb{E}\left(\left(\mu(t)^* - \frac{1}{2}\tilde{A}(t)^* \right) \Sigma(t)^{-1} \middle| \mathcal{G}_t \right) \\ &= \left(\bar{\mu}(t)^* - \frac{1}{2}\tilde{A}(t)^* \right) \Sigma(t)^{-1}. \end{aligned}$$

Hence (3.2) can be rewritten as

$$d\tilde{S}(t) - \langle U(t), G_t 1\rangle\, dt = d\nu(t).$$

Therefore, $U(t)$ satisfies the following Fujisaki–Kallianpur–Kunita (FKK) equation:

$$d\langle U(t), f\rangle = \langle U(t), Lf\rangle\, dt + (\langle U_t, G_t f\rangle - \langle U(t), G_t 1\rangle \langle U(t), f\rangle)\, d\nu(t). \tag{4.10}$$

Remark 4.3. Since L is not necessarily a second-order differential operator, which was used heavily in [12] in proving that an SPDE of the form (4.10) has a unique strong solution, we cannot establish such a property for a general L by the same argument in [12]. However, by Theorem 9.1 in [1], under suitable conditions, the solution to (4.10) is indeed unique.

Next, let us discuss the numerical implementation of the preceding filter. Recall that $\mu^1(t)$, $\mu^2(t)$, ... are independent copies of $\mu(t)$. For $\delta > 0$, let $\eta_\delta(t) := j\delta$ for $j\delta \le t < (j+1)\delta$, $j = 0, 1, \ldots$. We approximate $M^i(t)$ by the Euler scheme:

$$M^{\delta,i}(t) := \exp\Bigg(\int_0^t \left(\mu^i(\eta_\delta(s))^* - \frac{1}{2}\tilde{A}(\eta_\delta(s))^*\right)\Sigma(s)^{-1} d\tilde{S}(s) - \frac{1}{2}\int_0^t \left|\left(\mu^i(\eta_\delta(s))^* - \frac{1}{2}\tilde{A}(\eta_\delta(s))^*\right)\Sigma(s)^{-1}\right|^2 ds\Bigg).$$

Define

$$\left\langle V^{n,\delta}(t), f\right\rangle := \frac{1}{n}\sum_{i=1}^n M^{\delta,i}(t) f(\mu^i(t)) \text{ and } \langle V^n(t), f\rangle := \frac{1}{n}\sum_{i=1}^n M^i(t) f(\mu^i(t)).$$

We then combine both approximations by using $\bar{V}^n := V^{n,1/n^{1/2\alpha}}$ to approximate the unnormalized filter, where $\alpha > 0$ is given in (4.11) below.

Although we did not assume the sample path continuity of $\mu(t)$, we need some kind of continuity in the sense of moments. We assume that $|\Sigma(t)^{-1}| \le K$ and with some $\alpha > 0$,

$$\mathbb{E}\left(|\mu(t) - \mu(s)|^2\right) \le |t - s|^\alpha. \tag{4.11}$$

For example, if $\mu(t)$ is a compound Poisson process, then (4.11) holds with $\alpha = 1$.

Let $\bar{\mathbb{R}}^d$ be the one-point compactification of $\mathbb{R}^d$. Then $C_b(\bar{\mathbb{R}}^d)$ is a separable Banach space. Let $\mathcal{M}_F(\bar{\mathbb{R}}^d)$ be the space of finite Borel measures on $\bar{\mathbb{R}}^d$, and let d be a distance defined on $\mathcal{M}_F(\bar{\mathbb{R}}^d)$ whose topology coincides with the weak convergence topology. More precisely, let $\{f_k\} \subset C_b^1(\bar{\mathbb{R}}^d)$ be a dense subset of $C_b(\bar{\mathbb{R}}^d)$. We define

$$d(\nu_1, \nu_2) := \sum_{k=1}^\infty \frac{|\langle \nu_1 - \nu_2, f_k\rangle|}{2^k \|f_k\|_L}, \quad \nu_1, \nu_2 \in \mathcal{M}_F(\bar{\mathbb{R}}^d),$$

where

$$\|f\|_L := \sup_{x\in\bar{\mathbb{R}}^d} |f(x)| + \sup_{x,y\in\bar{\mathbb{R}}^d} \frac{|f(x) - f(y)|}{|x - y|}, \quad f \in C_b^1(\bar{\mathbb{R}}^d).$$

THEOREM 4.4. *Suppose that* $|\Sigma(t)^{-1}| \le K$ *and* (4.11) *holds. Then, for each fixed* t, *there exists* $M > 0$, *such that* $\forall\, n$,

$$\tilde{\mathbb{E}}\left(d(\bar{V}^n(t), V(t))\right) \le \frac{M}{\sqrt{n}}. \tag{4.12}$$

Proof. By the conditional independence, it is easy to show that $\forall\ f \in C_b(\bar{\mathbb{R}}^d)$,

$$\tilde{\mathbb{E}}\left(\langle V^n(t) - V(t), f\rangle^2\right) \le \frac{c_1(T)^2\|f\|_\infty^2}{n}.$$

Next we note that

$$\begin{aligned}
&\tilde{\mathbb{E}}\left\{\left|\log M^{\delta,i}(t) - \log M^i(t)\right|^2\right\}\\
&\le 2\int_0^T \mathbb{E}\left|\left(\mu^i(\eta_\delta(s))^* - \frac{1}{2}\tilde{A}(\eta_\delta(s))^*\right)\Sigma(s)^{-1} - \left(\mu^i(s)^* - \frac{1}{2}\tilde{A}(s)^*\right)\Sigma(s)^{-1}\right|^2 ds\\
&\quad + T\int_0^T \mathbb{E}\left|\left|\left(\mu^i(\eta_\delta(s))^* - \frac{1}{2}\tilde{A}(\eta_\delta(s))^*\right)\Sigma(s)^{-1}\right|^2 - \left|\left(\mu^i(s)^* - \frac{1}{2}\tilde{A}(s)^*\right)\Sigma(s)^{-1}\right|^2\right|^2 ds\\
&\le c_2(T)\delta^\alpha.
\end{aligned}$$

Let $\tilde{d}$ be the Waserstein metric on $\mathcal{M}_F(\bar{\mathbb{R}}^d)$, namely,

$$\tilde{d}(v_1, v_2) := \inf\left\{|\langle v_1 - v_2, f\rangle| :\ \ |f(x)| \le 1\ \forall x,\ \ |f(x) - f(y)| \le |x - y|\ \forall x, y\right\}.$$

Then

$$\begin{aligned}
&\tilde{d}(V^{n,\delta}(t), V^n(t))\\
&\le \frac{1}{n}\sum_{i=1}^n \left(M^{\delta,i}(t) \vee M^i(t)\right)\left(|\mu^{\delta,i}(t) - \mu^i(t)| + |\log M^{\delta,i}(t) - \log M^i(t)|\right).
\end{aligned}$$

Thus

$$\tilde{\mathbb{E}}\tilde{d}(V^{n,\delta}(t), V^n(t)) \le c_3(T)\delta^\alpha.$$

It is clear that $d \le \tilde{d}$. So

$$\begin{aligned}
\tilde{\mathbb{E}}\left(d(\bar{V}^n(t), V(t))\right) &\le \tilde{\mathbb{E}}\tilde{d}(V^{n,1/n^{1/2\alpha}}, V^n(t)) + \tilde{\mathbb{E}}d(V^n(t), V(t))\\
&\le c_2(T)(n^{-1/2\alpha})^\alpha + \frac{c_1(T)}{\sqrt{n}} = \frac{c_3(T)}{\sqrt{n}}. \qquad \square
\end{aligned}$$

4.2. Case 2: Random $\tilde{\Sigma}$. In this subsection we discuss a case when the volatility is a random process with the following structure: $\sigma(t)$ is a function of $\mu(t)$ plus a white noise, namely,

$$d\sigma_{ij}(t) = h^{ij}(\mu(t))dt + dB^{ij}(t), \qquad 1 \le i \le j \le d, \tag{4.13}$$

where $B^{ij}(t)$, $1 \le i \le j \le d$, are independent Brownian motions. Note that the independence assumption is imposed for ease of presentation only. In fact, the arguments below remain valid if we replace the $\frac{d(d+1)}{2}$-dimensional Brownian motion $(B^{ij}(t),$ $1 \le i \le j \le d)$ with a linear transformation of this process.

In this case, we have a classical filtering problem with observations $\tilde{S}(t)$ and $\Sigma(t) \equiv (\sigma_{ij}(t))$ given by (3.6) and (4.13), respectively. Then

$$\tilde{B}_{ij}(t) := \sigma_{ij}(t) - \int_0^t \left\langle U(s), h^{ij}\right\rangle ds$$

are Brownian motions adapted to $\mathcal{G}_t$ and are independent of $\nu(t)$. Namely, $\tilde{\nu}(t) :=$ $(\tilde{B}_{ij}(t), \nu(t); 1 \le i \le j \le d)$ forms the $\left(\frac{d(d+1)}{2} + d\right)$-dimensional innovation process for

the filtering problem. The FKK equation can be derived similarly to the arguments as in subsection 4.1 leading to (4.10). Namely, we need only replace $\nu(t)$ and $\big(\mu(t) - \frac{1}{2}\tilde{A}(t)\big)^*\Sigma(t)^{-1}$ with $\tilde{\nu}(t)$ and $\big(h_{ij}(\mu(t)),\ \big(\mu(t) - \frac{1}{2}\tilde{A}(t)\big)^*\Sigma(t)^{-1},\ 1 \le i \le j \le d\big)$, respectively. The numerical scheme can also be given by employing a similar method from subsection 4.1. We leave the details to the interested reader.

Remark 4.5. The condition (4.1) does not hold for the present model. In fact, $\tilde{B}_{ij}(t)$ is a $\mathcal{G}_t$-martingale independent of $\nu(t)$; hence, it cannot be represented as the stochastic integrals with respect to $\nu(t)$.

5. Optimization. In this section, we derive the optimal strategy of the partially observed mean-variance problem (2.3) in three steps. First, we derive from (3.1) a constraint on the terminal wealth $x(T)$. Then, we solve a static optimization problem under this constraint to find the best terminal wealth, $x^*(T)$. After that, we show that there is a portfolio such that $x^*(T)$ is its terminal wealth. Finally, we give a numerical scheme in solving the BSDE involved in deriving the optimal portfolio and prove the convergence of the proposed scheme.

5.1. The optimal value of $x(T)$. Let

$$\rho(t) := \exp\bigg(-\int_0^t \mu_0(s)ds - \sum_{i,j=1}^d \int_0^t \sigma_{ij}^{-1}(s)(\bar{\mu}_i(s) - \mu_0(s))d\nu_j(s) - \frac{1}{2}\sum_{j=1}^d \int_0^t \bigg|\sum_{i=1}^d \sigma_{ij}^{-1}(s)(\bar{\mu}_i(s) - \mu_0(s))\bigg|^2 ds\bigg),$$

where, with an abuse of notation, $\sigma_{ij}^{-1}(s)$ denotes the ijth element of $\Sigma(s)^{-1}$. Denote

$$\theta_j(t) := \sum_{i=1}^d \sigma_{ij}^{-1}(t)(\bar{\mu}_i(t) - \mu_0(t)). \tag{5.1}$$

By Itô's formula, we get

$$d\rho(t) = -\rho(t)\mu_0(t)dt - \sum_{j=1}^d \rho(t)\theta_j(t)d\nu_j(t), \quad \rho(0) = 1. \tag{5.2}$$

Applying Itô's formula to (5.2) and (3.1), we have

$$d(x(t)\rho(t)) = \rho(t)\sum_{i,j=1}^d \sigma_{ij}(t)u_i(t)d\nu_j(t) - x(t)\sum_{j=1}^d \rho(t)\theta_j(t)d\nu_j(t). \tag{5.3}$$

Therefore, $x(t)\rho(t)$ is a $\mathcal{G}_t$-martingale and hence,

$$x(t) = \rho(t)^{-1}\mathbb{E}(\rho(T)x(T)|\mathcal{G}_t).$$

In particular, taking $t = 0$ we have

$$\mathbb{E}(\rho(T)x(T)) = x(0) = x_0. \tag{5.4}$$

Definition 5.1. *A contingent claim $v \in \mathbb{H} := L^2(\Omega, \mathcal{G}_T, P)$ is called* attainable *if there is $\Phi(s) \in L^2_{\mathcal{G}}(0, T; \mathbb{R}^d)$ such that*

$$v\rho(T) = \mathbb{E}(v\rho(T)) + \int_0^T \Phi(s)^* d\nu(s).$$

Denote the collection of all attainable contingent claims by $AC(\mathcal{G})$. It is easy to see that $AC(\mathcal{G})$ is a subspace of $\mathbb{H}$. Denote by $\mathbb{H}_0$ the closure of $AC(\mathcal{G})$.

DEFINITION 5.2. *The market is complete if* $AC(\mathcal{G}) = \mathbb{H}$.

Remark 5.3. If $\Sigma(t)$ is nonrandom as in subsection 4.1, then the market is complete.

Now we seek

$$\min_{v\in\mathbb{H}_0} \mathbb{E}(v-z)^2 \tag{5.5}$$

subject to constraints

$$\mathbb{E}v = z \quad \text{and} \quad \mathbb{E}(\rho(T)v) = x_0. \tag{5.6}$$

THEOREM 5.4. *Let* α *and* β *be the orthogonal projections on* $\mathbb{H}_0$ *of* 1 *and* $\rho(T)$*, respectively. Then the optimal solution to the optimization problem* (5.5) *under constraint* (5.6) *is given by*

$$v = \frac{(z\langle\beta,\beta\rangle_{\mathbb{H}} - x_0\langle\alpha,\beta\rangle_{\mathbb{H}})\,\alpha + (-z\langle\alpha,\beta\rangle_{\mathbb{H}} + x_0\langle\alpha,\alpha\rangle_{\mathbb{H}})\,\beta}{\langle\alpha,\alpha\rangle_{\mathbb{H}}\langle\beta,\beta\rangle_{\mathbb{H}} - \langle\alpha,\beta\rangle_{\mathbb{H}}^2}. \tag{5.7}$$

Proof. Note that

$$\mathbb{E}(v-z)^2 = \mathbb{E}(v-z\alpha)^2 + z^2\mathbb{E}(1-\alpha)^2.$$

So, the optimization problem becomes

$$\min_{v\in\mathbb{H}_0} \|v-z\alpha\|_{\mathbb{H}}^2$$

subject to constraints

$$\langle v,\alpha\rangle_{\mathbb{H}} = z \quad \text{and} \quad \langle v,\beta\rangle_{\mathbb{H}} = x_0. \tag{5.8}$$

Using Lagrange multipliers, we define

$$f(v,\lambda_1,\lambda_2) := \|v-z\alpha\|_{\mathbb{H}}^2 - 2\lambda_1(\langle v,\alpha\rangle_{\mathbb{H}} - z) - 2\lambda_2(\langle v,\beta\rangle_{\mathbb{H}} - x_0),\ (v,\lambda_1,\lambda_2)\in\mathbb{H}\times\mathbb{R}^2.$$

Taking the Fréchet derivative and setting it to be zero, we have

$$2(v-z\alpha) - 2\lambda_1\alpha - 2\lambda_2\beta = 0.$$

This implies

$$v = z\alpha + \lambda_1\alpha + \lambda_2\beta.$$

Plugging the above into the constraints (5.8), we obtain the values of λ_1 and λ_2, which lead to (5.7). ∎

5.2. Replicate v. In this subsection, we seek the wealth process $x(t)$ which satisfies (3.1) and $x(T) = v$, where $v \in \mathbb{H}_0$ is given by (5.7). Namely, we seek a solution to the following BSDE:

$$\begin{cases} dx(t) = \big(x(t)\mu_0(t) + \sum_{j=1}^d(\bar{\mu}_j(t) - \mu_0(t))u_j(t)\big)dt + \sum_{i,j=1}^d \sigma_{ij}(t)u_i(t)d\nu_j(t),\ 0\le t\le T, \\ x(T) = v. \end{cases} \tag{5.9}$$

Let

$$Z_j(t) := \sum_{i=1}^{d} \sigma_{ij}(t) u_i(t).$$

Then

$$u_i(t) = \sum_{j=1}^{d} \sigma_{ij}^{-1}(t) Z_j(t) \tag{5.10}$$

and (5.9) becomes

(5.11)
$$\begin{cases} dx(t) = \left(x(t)\mu_0(t) + \sum_{j=1}^{d} \theta_j(t) Z_j(t) \right) dt + \sum_{j=1}^{d} Z_j(t) d\nu_j(t), \ 0 \le t \le T, \\ x(T) = v. \end{cases}$$

If $\mathcal{F}_t^{\nu} = \mathcal{G}_t$, then (5.11) is the usual BSDE whose solution exists, assuming that $\mu_0(t)$ and $\theta_j(t)$ are essentially bounded (cf. [28]). However, it is well known that, in general, $\mathcal{F}_t^{\nu} \neq \mathcal{G}_t$.

Now we prove the existence of a unique square integrable solution for the BSDE (5.11), under the following additional condition.

Assumption (*UB*). For some $\delta > 0$, $A(t) \ge \delta I$ a.s., almost every $t \ge 0$, and $\mu_0(t)$ and $\mu(t)$ are essentially bounded.

Under this assumption, $\theta_j(t)$ is also essentially bounded.

PROPOSITION 5.5. *If $v \in \mathbb{H}_0$, then* (5.11) *has a unique $\mathcal{G}_t$-adapted, square integrable solution* $(x(t), Z_j(t),\ j = 1, 2, \ldots, d)$.

Proof. If $(x(t), Z_j(t),\ j = 1, 2, \ldots, d)$ is a $\mathcal{G}_t$-adapted, square integrable solution to (5.11), as in (5.3), we have that

$$x(t)\rho(t) = x_0 + \sum_{j=1}^{d} \int_0^t \rho(s) \left(\sum_{i=1}^{d} \sigma_{ij}(s) u_i(s) - x(s)\theta_j(s) \right) d\nu_j(s)$$

is a $\mathcal{G}_t$-local martingale. Hence, there is an increasing sequence of $\mathcal{G}_t$-stopping times $\{\tau_n\}$ with $\tau_n \to T$ as $n \to \infty$ such that for each n,

$$x(t \wedge \tau_n)\rho(t \wedge \tau_n) = \mathbb{E}(x(T \wedge \tau_n)\rho(T \wedge \tau_n) | \mathcal{G}_t).$$

For any fixed $t \in [0, T]$,

$$x(t \wedge \tau_n)\rho(t \wedge \tau_n) \le \sup_{0 \le s \le T} x(s) \sup_{0 \le s \le T} \rho(s),$$

whereas the right-hand side of the above is a square integrable random variable by virtue of the Cauchy–Schwarz inequality and the standard L^2 estimation on the supernorm of the solutions to SDEs. Hence, we obtain by the dominated convergence theorem that

$$x(t)\rho(t) = \mathbb{E}(x(T)\rho(T) | \mathcal{G}_t).$$

Namely,

$$x(t) = \rho(t)^{-1} \mathbb{E}(v\rho(T) | \mathcal{G}_t). \tag{5.12}$$

This implies the uniqueness of the solution.

To prove the existence we first assume that $v \in AC(\mathcal{G})$. We show that $x(t)$ given by (5.12) is a solution to (5.11). As $v \in AC(\mathcal{G})$, $v\rho(T)$ is square integrable in view of Definition 5.1, and we have the representation

$$\mathbb{E}(v\rho(T)|\mathcal{G}_t) = \mathbb{E}(v\rho(T)) + \sum_{j=1}^{d} \int_0^t \Phi^j(s) d\nu_j(s), \tag{5.13}$$

where each Φ^j is square integrable. Define

$$Z_j(t) := x(t)\theta_j(t) + \rho(t)^{-1}\Phi^j(t), \quad 0 \le t \le T. \tag{5.14}$$

By Itô's formula, it is easy to show that $(x(t), Z(t)) \equiv (x(t), Z_j(t),\ j = 1, 2, \ldots, d)$ satisfies (5.11). Moreover, a stopping time argument exactly as in [28, pp. 352–353] establishes the square integrability of $(x(t), Z(t))$.

Next, let $v \in \mathbb{H}_0$. Then there is a sequence $\{v_n\} \subset AC(\mathcal{G})$ such that $v_n \to v$ in $\mathbb{H}$. By the above proof there is a unique square integrable solution $(x_n(t), Z_n(t))$ to (5.11) with $x_n(T) = v_n$. Moreover,

$$\sup_n \mathbb{E} \int_0^T (|x_n(t)|^2 + |Z_n(t)|^2) dt \le K \sup_n \mathbb{E}|v_n|^2 < +\infty;$$

see page 349, Theorem 2.2 in [28]. This implies that $(x_n(t), Z_n(t))$ is a bounded sequence in $L^2_{\mathcal{G}}(0, T; \mathbb{R}^{d+1})$. Hence there is a subsequence (still denoted as $(x_n(t), Z_n(t))$), along with $(x(t), Z(t)) \in L^2_{\mathcal{G}}(0, T; \mathbb{R}^{d+1})$, so that

$$(x_n(t), Z_n(t)) \to (x(t), Z(t)) \quad \text{weakly in } L^2_{\mathcal{G}}(0, T; \mathbb{R}^{d+1}).$$

By Mazur's theorem there is a sequence $(\tilde{x}_n(t), \tilde{Z}_n(t))$, each element of which is a convex combination of those in $\{(x_n(t), Z_n(t))\}$, so that

$$(\tilde{x}_n(t), \tilde{Z}_n(t)) \to (x(t), Z(t)) \quad \text{(strongly) in } L^2_{\mathcal{G}}(0, T; \mathbb{R}^{d+1}).$$

Since (5.11) is a linear equation, by a standard technique we conclude that $(x(t), Z(t))$ is a square integrable solution to (5.11). □

Summarizing, we get the following.

THEOREM 5.6. *For every z, there exists an optimal portfolio to the mean-variance problem* (2.3)*, which is a self-financed, admissible portfolio replicating v given by* (5.7)*. Moreover, this optimal portfolio $u(t)$ is given by* (5.10) *and the corresponding optimal wealth process is $x(t)$, where $(x(t), Z_j(t),\ j = 1, 2, \ldots, d)$ is the square integrable solution to* (5.11)*.*

So solving our partially observed mean-variance problem boils down to solving the BSDE (5.11). Numerical solutions to some classes of nonlinear BSDEs have been developed lately [29, 8]. However, in those works the drift coefficients of the BSDEs are assumed to be deterministic functions. In our model, the coefficients $\mu_0(t)$ and $\theta_j(t)$ are random in general. To the best of our knowledge, solving such a BSDE numerically remains open. In the next subsection, we shall give a numerical solution to this BSDE based upon the constructive proof of the last theorem.

5.3. Numerical solution. In this subsection, we assume that the market is complete (see Definition 5.2) and seek a numerical solution to (5.11). By virtue of the constructive proof of Proposition 5.5 the solution is given by (5.12) and (5.14). We now propose a numerical scheme to approximate (5.12) and (5.14).

To start with, note that as the market is complete, $\alpha = 1$ and $\beta = \rho(T)$. By (5.7), we get

$$v = z + \frac{x_0 - z\mathbb{E}\rho(T)}{\mathrm{Var}(\rho(T))}(\rho(T) - \mathbb{E}\rho(T)). \tag{5.15}$$

As in the proof of Proposition 5.5, the key to solving (5.11) is the martingale representation of the $\mathcal{G}_t$-martingale $\mathbb{E}(\rho(T)v|\mathcal{G}_t)$. We will establish a particle representation for this martingale.

Let (θ^1, ν^1), (θ^2, ν^2), ... be independent copies of (θ, ν) which appeared in (5.2). Now we define $\rho^i(t, t')$, $t, t' \geq 0$, in two steps. First, for $t \leq t'$, let $\rho^i(t, t') := \rho(t)$ which is given by (5.2). Second, for $t > t'$, let $\rho^i(t, t')$ be given by (5.2) with (b, ν) replaced by (b^i, ν^i):

$$d\rho^i(t, t') = -\rho^i(t, t')\mu_0(t)dt - \sum_{j=1}^{d} \rho^i(t, t')\theta_j^i(t)d\nu_j^i(t), \quad \rho^i(t', t') = \rho(t'). \tag{5.16}$$

Let $v^i(T, t)$ be given by (5.15) with $\rho(T)$ replaced by $\rho^i(T, t)$:

$$v^i(T, t) = z + \frac{x_0 - z\mathbb{E}\rho(T)}{\mathrm{Var}(\rho(T))}(\rho^i(T, t) - \mathbb{E}\rho(T)). \tag{5.17}$$

THEOREM 5.7. *Let v be given by* (5.15). *Then*

$$\mathbb{E}(\rho(T)v|\mathcal{G}_t) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \rho^i(T, t)v^i(T, t). \tag{5.18}$$

Proof. Note that the SDE (5.16) has a unique strong solution. Therefore, there exists a measurable functional $F_{t,T}$ such that

$$(\rho^i(T, t), v^i(T, t)) = F_{t,T}(\theta|_{[0,t]}, \nu|_{[0,t]}, \theta^i|_{[t,T]}, \nu^i|_{[t,T]}).$$

Note that $(\bar{\mu}(t), \nu(t))$ are $\mathcal{G}_t$-measurable. Further, as discussed in section 2, $\sigma_{ij}^{-1}(t)$ and $\mu_0(t)$ are also $\mathcal{G}_t$-measurable. Thus, $\theta(t)$ defined by (5.1) is $\mathcal{G}_t$-measurable and we have

$$(\rho^i(T, t), v^i) = G_{t,T}(S|_{[0,t]}, \theta^i|_{[t,T]}, \nu^i|_{[t,T]})$$

for a measurable functional $G_{t,T}$. By the independence of $S|_{[0,t]}$, $(\theta^i|_{[t,T]}, \nu^i|_{[t,T]})$, $i = 1, 2, \ldots$, it follows from the strong law of large numbers under the conditional probability (given $\mathcal{G}_t$) that (5.18) holds. ∎

For notational simplicity, we now assume $d = T = 1$. Denote the process in (5.18) by $N(t)$. By (5.13), we have

$$\langle N, \nu \rangle_t = \int_0^t \Phi(s) ds$$

which can be approximated by (cf. Jacod and Shiryaev [9, Theorem 1.4.47])

$$\sum_{k=1}^{m} (N(t_k) - N(t_{k-1}))(\nu(t_k) - \nu(t_{k-1}).$$

Based on the above, we now approximate Φ by piecewise constants, i.e., approximate Φ on $(\frac{k}{n}, \frac{k+1}{n})$ by

$$\begin{aligned} \Phi^n\left(\frac{k}{n}\right) := n \sum_{j=1}^{m} & \left(N\left(\frac{k-1}{n} + \frac{j}{mn}\right) - N\left(\frac{k-1}{n} + \frac{j-1}{mn}\right)\right) \\ & \times \left(\nu\left(\frac{k-1}{n} + \frac{j}{mn}\right) - \nu\left(\frac{k-1}{n} + \frac{j-1}{mn}\right)\right), \\ & n = 1, 2, \ldots, \end{aligned} \tag{5.19}$$

where $m = m_n$ is to be chosen later.

LEMMA 5.8. *For $1 < p < 2$, we have*

$$\begin{aligned} & \mathbb{E}\left| \Phi^n\left(\frac{k}{n}\right) - n \int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi(s) ds \right|^p \\ & \quad \le c\left(m^{\epsilon - p} + n^{\frac{p}{2}} m^{-\frac{p}{2}}\right) \left(\mathbb{E} \int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi(t)^2 dt \right)^{\frac{p}{2}}, \end{aligned} \tag{5.20}$$

for any $\epsilon > 0$, where c depends only on p.

Proof. Let

$$\pi(t) := \frac{k-1}{n} + \frac{j-1}{mn}, \qquad \text{for } \frac{k-1}{n} + \frac{j-1}{mn} \le t < \frac{k-1}{n} + \frac{j}{mn}.$$

Apply Itô's formula, we have

$$\begin{aligned} & \Phi^n\left(\frac{k}{n}\right) - n \int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi(s) ds \\ & \quad = n \int_{\frac{k-1}{n}}^{\frac{k}{n}} \left\{ \Phi(t)(\nu(t) - \nu(\pi(t))) + (N(t) - N(\pi(t))) \right\} d\nu(t). \end{aligned}$$

For $p \in (1,2)$, by Doob's inequality, we have

$$\begin{aligned}
&\mathbb{E}\left(\left|\Phi^n\left(\frac{k}{n}\right) - n\int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi(s)ds\right|^p\right) \\
&\quad \le c_p n^p \mathbb{E}\left(\int_{\frac{k-1}{n}}^{\frac{k}{n}} \left\{\Phi(t)^2(\nu(t)-\nu(\pi(t)))^2 + (N(t)-N(\pi(t)))^2\right\} dt\right)^{\frac{p}{2}} \\
&\quad \le c_p n^p \mathbb{E}\left(\left(\int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi(t)^2 dt\right)^{\frac{p}{2}} \sup_{\frac{k-1}{n}\le t\le \frac{k}{n}} (\nu(t)-\nu(\pi(t)))^p\right) \\
&\qquad + c_p n^p \mathbb{E}\left(\int_{\frac{k-1}{n}}^{\frac{k}{n}} (N(t)-N(\pi(t)))^2 dt\right)^{\frac{p}{2}} \\
&\quad \le c_p n^p \left(\mathbb{E}\int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi(t)^2 dt\right)^{\frac{p}{2}} \left(\mathbb{E}\sup_{\frac{k-1}{n}\le t\le \frac{k}{n}} (\nu(t)-\nu(\pi(t)))^{\frac{2p}{2-p}}\right)^{\frac{2-p}{2}} \\
&\qquad + c_p n^p \left(\mathbb{E}\int_{\frac{k-1}{n}}^{\frac{k}{n}} (N(t)-N(\pi(t)))^2 dt\right)^{\frac{p}{2}}.
\end{aligned}$$

Note that for independent, identically distributed (i.i.d.) normal random variables ξ_i with mean 0 and variance σ^2, we have

$$\mathbb{E}\sup_{1\le i\le m} \xi_i^p \le \left(\mathbb{E}\sup_{1\le i\le m} \xi_i^{p/\epsilon}\right)^{\epsilon} \le \left(\mathbb{E}\sum_{i=1}^m \xi_i^{p/\epsilon}\right)^{\epsilon} \le cm^{\epsilon}\sigma^p$$

for any $\epsilon > 0$. Thus

$$\left(\mathbb{E}\sup_{\frac{k-1}{n}\le t\le \frac{k}{n}} (\nu(t)-\nu(\pi(t)))^{\frac{2p}{2-p}}\right)^{\frac{2-p}{2}} \le \frac{cm^{\epsilon}}{(nm)^p}.$$

On the other hand,

$$\begin{aligned}
\mathbb{E}\int_{\frac{k-1}{n}}^{\frac{k}{n}} (N(t)-N(\pi(t)))^2 dt &= \mathbb{E}\int_{\frac{k-1}{n}}^{\frac{k}{n}} \int_{\pi(t)}^t \Phi(s)^2 ds dt \\
&= \mathbb{E}\sum_{j=1}^m \int_{\frac{k-1}{n}+\frac{j-1}{mn}}^{\frac{k-1}{n}+\frac{j}{mn}} \int_{\frac{k-1}{n}+\frac{j-1}{mn}}^t \Phi(s)^2 ds dt \\
&\le \frac{1}{nm}\mathbb{E}\int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi(t)^2 dt.
\end{aligned}$$

Equation (5.20) then follows easily. □

Next we need to approximate $N(k\delta)$, $k = 0, 1, \ldots, mn$, where $\delta = \frac{1}{mn}$. To this end, we need to approximate $\rho^i(t,t')$ by time-discretization. Recall that $\rho^i(t,t')$ is given by (5.2) for $t \le t'$ and (5.16) for $t > t'$.

For $j \le k$, let

$$\begin{aligned}
\rho^{i,\delta}(j\delta, k\delta) := \rho^i((j-1)\delta, k\delta) &- \rho^i((j-1)\delta, k\delta)\mu_0((j-1)\delta)\delta \\
&- \rho^i((j-1)\delta, k\delta)\theta((j-1)\delta)(\nu(j\delta) - \nu((j-1)\delta)),
\end{aligned}$$

with $\rho^{i,\delta}(0,k\delta) := 1$. For $j > k$, let

$$\begin{aligned}\rho^{i,\delta}(j\delta,k\delta) := \rho^i((j-1)\delta,k\delta) - \rho^i((j-1)\delta,k\delta)\mu_0((j-1)\delta)\delta \\ -\rho^i((j-1)\delta,k\delta)\theta((j-1)\delta)(\nu^i(j\delta)-\nu^i((j-1)\delta)).\end{aligned}$$

Let

$$v^{i,\delta}(T,k\delta) := z + \frac{x_0 - z\mathbb{E}\rho(T)}{\mathrm{Var}(\rho(T))}(\rho^{i,\delta}(T,k\delta) - \mathbb{E}\rho(T)),$$

and

$$N^{n,\delta}(k\delta) := \frac{1}{n}\sum_{i=1}^{n}\rho^{i,\delta}(T,k\delta)v^{i,\delta}(T,k\delta). \tag{5.21}$$

Now, in view of (5.14), we define

$$Z^n(t) := \rho^n\left(\frac{[nt]}{n}\right)^{-1}\left(\theta^n\left(\frac{[nt]}{n}\right)N^{n,\delta}\left(\frac{[nt]}{n}\right) + \Phi^{n,\delta}\left(\frac{[nt]}{n}\right)\right), \tag{5.22}$$

where $\delta = \frac{1}{m_n n}$ (with m_n suitably chosen), $\Phi^{n,\delta}$ is defined as in (5.19) for Φ^n with N replaced by $N^{n,\delta}$, the approximate θ^n and ρ^n of θ and ρ can be defined by the same method as in section 4.1 such that $\forall\ \beta > 1$,

$$\mathbb{E}\sup_{0\le t\le T}\left|\rho^n\left(\frac{[nt]}{n}\right)^{-1}\theta^n\left(\frac{[nt]}{n}\right)N^{n,\delta}\left(\frac{[nt]}{n}\right) - \rho(t)^{-1}\theta(t)N(t)\right|^{\beta} \to 0, \tag{5.23}$$

and

$$\mathbb{E}\sup_{0\le t\le T}\left|\rho^n\left(\frac{[nt]}{n}\right)^{-1} - \rho(t)^{-1}\right|^{\beta} \to 0. \tag{5.24}$$

Finally, we define the *approximate portfolio* by

$$u^n(t) := \Sigma(t)^{-1}Z^n(t).$$

Since we do not know the continuity of $\Phi(s)$, we cannot obtain $\Phi^n(s) \to \Phi(s)$ (cf. (5.23)). Therefore, it is not clear whether $u^n(t) \to u(t)$. However, we have the convergence of their corresponding terminal wealths.

THEOREM 5.9. *Suppose that* $\mathbb{E}(|v\rho(T)|^2) < \infty$. *Then*

$$\mathbb{E}|x^n(T) - x(T)| \to 0 \quad \text{as } n \to \infty. \tag{5.25}$$

Proof. For simplicity of notation, we assume $\mu_0(t) \equiv 0$. Then by (5.11), we have

$$\begin{aligned}\mathbb{E}|x^n(T) - x(T)| \le c\mathbb{E}\int_0^T |\theta^n(t)Z^n(t) - \theta(t)Z(t)|^2dt \\ +c\mathbb{E}\left(\int_0^T |Z^n(t) - Z(t)|^2dt\right)^{1/2}.\end{aligned}$$

We estimate only the second term (the first can be evaluated similarly). Note that by (5.22) and (5.19),

$$|Z^n(t) - Z(t)| \le \left| \rho^n \left(\frac{[nt]}{n} \right)^{-1} \theta^n \left(\frac{[nt]}{n} \right) N^{n,\delta} \left(\frac{[nt]}{n} \right) - \rho(t)^{-1} \theta(t) N(t) \right| + \left| \rho^n \left(\frac{[nt]}{n} \right)^{-1} \Phi^{n,\delta} \left(\frac{[nt]}{n} \right) - \rho(t)^{-1} \Phi(t) \right|.$$

By (5.23), the first term on the right-hand side of the above inequality tends to 0. Note that

$$\begin{aligned}
&\mathbb{E}\left[\left(\int_0^T \left| \rho^n \left(\frac{[nt]}{n} \right)^{-1} \Phi^{n,\delta} \left(\frac{[nt]}{n} \right) - \rho(t)^{-1} \Phi(t) \right|^2 dt \right)^{\frac{1}{2}} \right] \\
&\quad \le \mathbb{E}\left[\left(\int_0^T \rho^n \left(\frac{[nt]}{n} \right)^{-2} \left| \Phi^{n,\delta} \left(\frac{[nt]}{n} \right) - \Phi^n \left(\frac{[nt]}{n} \right) \right|^2 dt \right)^{\frac{1}{2}} \right] \\
&\qquad + \mathbb{E}\left[\left(\int_0^T \left| \rho^n \left(\frac{[nt]}{n} \right)^{-2} - \rho(t)^{-2} \right| \Phi^n \left(\frac{[nt]}{n} \right)^2 dt \right)^{\frac{1}{2}} \right] \\
&\qquad + \mathbb{E}\left[\left(\frac{1}{n} \sum_{k=1}^n \left| \Phi^n \left(\frac{k}{n} \right) - \int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi(s) ds \right|^2 \sup_{0 \le t \le T} \rho(t)^{-2} \right)^{\frac{1}{2}} \right] \\
&\qquad + \mathbb{E}\left[\left(\sum_{k=1}^n \int_{\frac{k-1}{n}}^{\frac{k}{n}} \left| \Phi(t) - n \int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi(s) ds \right|^2 \sup_{0 \le t \le T} \rho(t)^{-2} \right)^{\frac{1}{2}} \right].
\end{aligned}$$

The convergence of the first term follows from similar arguments in subsection 4.1, that of the second term from (5.24), and that of the fourth term from the same arguments as in the proof of Lebesgue's continuity theorem; namely, first approximate Φ by uniformly continuous functions and then prove the conclusion for such functions. Finally, the third term is dominated by

$$\begin{aligned}
&\mathbb{E} \max_{1 \le k \le n} \left| \Phi^n \left(\frac{k}{n} \right) - \int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi(s) ds \right| \sup_{0 \le t \le T} \rho(t)^{-1} \\
&\quad \le \left(\mathbb{E} \sum_{k=1}^n \left| \Phi^n \left(\frac{k}{n} \right) - \int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi(s) ds \right|^p \right)^{\frac{1}{p}} \left(\mathbb{E} \sup_{0 \le t \le T} \rho(t)^{-\frac{p}{p-1}} \right)^{\frac{p-1}{p}} \\
&\quad \le cn \left(m^{\epsilon - p} + n^{\frac{p}{2}} m^{-p} \right)
\end{aligned}$$

which converges to 0 if we take $m = n^\beta$ with $\beta > \max(\frac{1+\epsilon}{p}, \frac{1}{2})$. □

Remark 5.10. It follows from (5.25) that $\mathbb{E}x^n(T) \to \mathbb{E}x(T)$. However, it is not clear whether $\mathrm{Var}(x^n(T)) \to \mathrm{Var}(x(T))$. To achieve this, we need a higher moment (in p) in (5.20) for which we require $\int_0^T \Phi(s)^2 ds$ to have a higher moment. To be

precise, if we assume that $v\rho(T)$ has a moment of order $p' > 2$, then

$$\mathbb{E}\left[\left(\int_0^T \Phi(s)^2 ds\right)^{p'/2}\right] < \infty.$$

The proof of (5.20) can be adapted for $2 < p < p'$ and, hence, (5.25) can be strengthened to

$$\mathbb{E}\left(|x^n(T) - x(T)|^2\right) \to 0.$$

In this case we get the convergence of both $\mathbb{E}x^n(T)$ and $\mathrm{Var}(x^n(T))$.

Remark 5.11. If the market is not complete, the numerical approximation of this section remains valid if we replace 1 and $\rho^i(T,t)$ with their projections α and β^i on $\mathbb{H}_0$ (i.e., use the formula (5.7) instead of (5.15)). However, it remains *open* how to calculate α and β^i numerically.

Acknowledgments. The authors thank the associate editor and two anonymous referees for constructive comments that have led to an improved version.

REFERENCES

[1] A. G. Bhatt, G. Kallianpur, and R. L. Karandikar, *Uniqueness and robustness of solution of measure-valued equations of nonlinear filtering*, Ann. Probab., 23 (1995), pp. 1895–1938.
[2] T. R. Bielecki, H. Jin, S. R. Pliska, and X. Y. Zhou, *Continuous-time mean-variance portfolio selection with bankruptcy prohibition*, Math. Finance, 15 (2005), pp. 213–244.
[3] M. J. Brennan and Y. Xia, *Assessing asset pricing anomalies*, Rev. Finan. Stud., 14 (2001), pp. 905–942.
[4] J. B. Detemple, *Asset pricing in a production economy with incomplete information*, J. Finance, 41 (1986), pp. 383–391.
[5] M. U. Dothan and D. Feldman, *Equilibrium interest rates and multiperiod bonds in a partially observable economy*, J. Finance, 41 (1986), pp. 369–382.
[6] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1985.
[7] G. Gennotte, *Optimal portfolio choice under incomplete information*, J. Finance, 41 (1986), pp. 733–746.
[8] E. Gobet, J.-P. Lemor, and X. Warin, *A regression-based Monte Carlo method to solve backward stochastic differential equations*, Ann. Appl. Probab., 15 (2004), pp. 2172–2202.
[9] J. Jacod and A. N. Shiryaev, *Limit Theorems for Stochastic Processes*, Springer-Verlag, New York, 1987.
[10] G. Kallianpur, *Stochastic Filtering Theory*, Springer-Verlag, New York, 1980.
[11] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.
[12] T. Kurtz and J. Xiong, *Particle representations for a class of nonlinear SPDEs*, Stochastic Process. Appl., 83 (1999), pp. 103–126.
[13] T. Kurtz and J. Xiong, *Numerical solutions for a class of SPDEs with application to filtering*, in Stochastics in Finite and Infinite Dimension: In Honor of Gopinath Kallianpur, Trends Math., T. Hida, R. Karandikar, H. Kunita, B. Rajput, S. Watanabe, and J. Xiong, eds., Birkhäuser Boston, Boston, MA, 2000, pp. 233–258.
[14] P. Lakner, *Optimal trading strategy for an investor: The case of partial information*, Stochastic Process. Appl., 76 (1998), pp. 77–97.
[15] D. Li and W. L. Ng, *Optimal dynamic portfolio selection: Multi-period mean-variance formulation*, Math. Finance, 10 (2000), pp. 387–406.
[16] A. E. B. Lim, *Quadratic hedging and mean-variance portfolio selection with random parameters in an incomplete market*, Math. Oper. Res., 29 (2004), pp. 132–161.
[17] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes,* I, *General Theory*, 2nd ed., Springer, New York, 2001.
[18] J. Ma, P. Protter, and J. Yong, *Solving forward-backward stochastic differential equations explicitly—a four step scheme*, Probab. Theory Related Fields, 98 (1994), pp. 339–359.

[19] H. M. Markowitz, *Portfolio selection*, J. Finance, 7 (1952), pp. 77–91.
[20] H. Nagai and S. Peng, *Risk-sensitive dynamic portfolio optimization with partial information on infinite time horizon*, Ann. Appl. Probab., 12 (2002), pp. 173–195.
[21] H. Pham, *Mean-variance hedging for partially observed drift processes*, Int. J. Theor. Appl. Finance, 4 (2001), pp. 263–284.
[22] L. C. G. Rogers, *The relaxed investor and parameter uncertainty*, Finance Stoch., 5 (2001), pp. 131–154.
[23] M. C. Steinbach, *Markowitz revisited: Mean-variance models in financial portfolio analysis*, SIAM Rev., 43 (2001), pp. 31–85.
[24] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, 1979.
[25] J. Xia, *Mean-variance portfolio choice: Quadratic partial hedging*, Math. Finance, 15 (2005), pp. 533–538.
[26] Y. Xia, *Learning about predictability: The effects of parameter uncertainty on dynamic asset allocation*, J. Finance, 56 (2001), pp. 205–246.
[27] Z. J. Yang and J. Xiong, *Maximizing the expected utility from terminal wealth with partial information and the valuation of information*, submitted.
[28] J. Yong and X. Y. Zhou, *Stochastic Control: Hamiltonian Systems and HJB Equations*, Springer, New York, 1999.
[29] J. Zhang, *A numerical scheme for BSDEs*, Ann. Appl. Probab., 14 (2004), pp. 459–488.
[30] Y. G. Zhao and W. T. Ziemba, *Mean-Variance versus Expected Utility in Dynamic Investment Analysis*, working paper, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, BC, Canada, 2000.
[31] X. Y. Zhou and D. Li, *Continuous-time mean-variance portfolio selection: A stochastic LQ framework*, Appl. Math. Optim., 42 (2000), pp. 19–33.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 176–209

PARTIALLY OBSERVED INVENTORY SYSTEMS: THE CASE OF ZERO-BALANCE WALK*

ALAIN BENSOUSSAN†, METIN ÇAKANYILDIRIM‡, AND SURESH P. SETHI§

Abstract. In many inventory control contexts, inventory levels are only partially (i.e., not fully) observed. This may be due to nonobservation of demand, spoilage, misplacement, or theft of inventory. We study a partially observed inventory system where the demand is not observed, inventory level is noticed when it reaches zero, the unmet demand is lost, and replenishment orders must be decided so as to minimize the total discounted costs over an infinite horizon. This problem has an infinite-dimensional state space, and for it we establish the existence of a feedback policy when single-period costs are bounded or when the discount factor is sufficiently small. We also provide an approximately optimal feedback policy that uses a finite state representation.

Key words. stochastic inventory problem, partial observations, the Zakai equation, lost sales

AMS subject classifications. 90B05, 93E20, 93C41, 90C39

DOI. 10.1137/040620321

1. Introduction.

Inventory control is among the most important topics in operations research because of large investments in inventory and their effect on the profitability of the firms. In 1999, for example, the investment into the inventory by U.S. businesses alone amounted to 1.1 trillion dollars [31]. Because of the importance of inventory control decisions, there has resulted an extensive literature on the topic [3, 31]. For the motivation of our research, one of the critical assumptions in the vast inventory literature, dating back to at least the Harris lot size model of 1913 [17], has been that the level of inventory at any given time is fully observed. Some of the most celebrated results, such as the optimality of the base stock policy [1], have been obtained under the assumption of full observation. Yet the inventory level is often not fully observed in practice, as elaborated below. In such cases, most of the well-known inventory policies are not even admissible, let alone optimal.

The study of systems with partially observed inventories is important in many real-life situations. We shall introduce some of the possible instances where inventories can only be partially observed by the inventory manager (IM).

Transaction errors. Unintentional mistakes happen from time to time during inventory transactions. Some of these transactions are inventory counting, receiving, checking out at the cash register, etc. An example is checking out at a grocery store. If a customer buys two types of different soups each at the same price, the sales clerk often scans only one soup type twice. A similar example with different types of yogurts can be found in Raman, DeHoratius, and Ton [26]. In such cases, the recorded inventory levels of the items involved will differ from their actual levels.

*Received by the editors December 6, 2004; accepted for publication (in revised form) August 20, 2006; published electronically April 6, 2007. This material is based upon work supported by National Science Foundation grant 0509278.

http://www.siam.org/journals/sicon/46-1/62032.html

†International Center for Decision and Risk Analysis, School of Management, P.O. Box 830688, SM 30, University of Texas at Dallas, Richardson, TX 75083-0688 (bensoussan@utdallas.edu).

‡School of Management, P.O. Box 830688, SM 30, University of Texas at Dallas, Richardson, TX 75083-0688 (metin@utdallas.edu).

§Center for Intelligent Supply Networks, School of Management, P.O. Box 830688, SM 30, University of Texas at Dallas, Richardson, TX 75083-0688 (sethi@utdallas.edu).

When stock-keeping units are discrete, it may be possible to eliminate counting errors. On the other hand, when they are not discrete, such as oil in a refinery, exact measurements are difficult to obtain. While measurement errors cause inventory to be not fully observed, it is often the mistakes in reporting transactions that lead to partial observation of inventories. Raman, DeHoratius, and Ton [26] report a retailer who has inaccurate inventory records for 65% of its stock-keeping units. It is roughly estimated that the retailer loses 10% of its current profit due to these inaccuracies. They go on to say: "[this particular retailer] is not an isolated case; this [inaccuracy] problem is common at other retailers." Common use of modern information technology tends to reduce transaction errors. However, as pointed out by Axsäter [3], deployment of big-ticket computer technology is not always economically feasible.

Misplaced inventory. When a part of the inventory on hand is misplaced, it is not available to meet a demand until it is found. Often the misplaced inventories are not immediately found, and thus they remain unobserved to the IM. This causes the total inventory that is available to meet the demand to become partially observed. Misplaced inventory can be quite large and have a significant impact on the bottom line. It is reported in [26] that customers of a "leading retailer" cannot find 16% of the items in the stores because those items are misplaced. Misplacement of the items reduces the profit by roughly 25% at this retailer.

Misplacement is more likely when the location of items in storage is altered dynamically. According to [3], "It is easier to keep the records accurate if the items have fixed locations. On the other hand, this can lead to inefficient space utilization. By dynamically locating items, the same item can be stored in more than one location." The recent trends in supply chain management such as crossdocking (see, e.g., p. 412 of [13]) also cause dynamic locations.

Misplaced inventories are eventually discovered either by inspection or by chance. When the misplaced items are placed in their proper shelves, they become available once again to meet customer demands. Thus, misplacements and their recoveries can cause the actual inventory to be, respectively, less and more than the recorded inventory.

Spoilage. Products can naturally lose their properties while they are held in the inventory [25]. Examples with limited lifetime are drugs, chemicals, and food products. If the lifetime is limited and not immediately observed, then the actual inventory is less than the recorded inventory, and it is partially observed.

If the lifetime is deterministic as in the case of drugs, an implementation of RFID (radio frequency identification) tags called SMC (smart medicine cabinet) can be used to track the expired drugs [28]. Thus, the SMC can make drug spoilage fully observed. However, investments into technology such as SMC must be justified with an economic analysis, which requires the evaluation of the optimal cost under partial observations [9]. As an example of random lifetime, consider the number of batteries in a Sears retail store. Only when these batteries are inspected (say by measuring their voltage, one by one) does the inventory level of fully functioning batteries become known. When the spoiled inventory is observed immediately, the associated model (e.g., [24]), in spite of being challenging to work with, has full observations.

In retail stores, customers can cause damage to products, making them unsuitable for sale. Some examples are tearing of a package to try on the contained cloth item, wearing down a shoe by trying it on and walking, erasing software on computers on demonstration, spilling food on clothes, and scratching a car during a test drive. So long as the damages are not detected by the IM, the actual inventory is not fully observed.

Product quality and yield. When the product quality is low or a production process has a low yield [29], the actual inventory is not known. Receipts at a warehouse can include products that are defective or that do not conform to quality standards. It is very often the case that nonconformance of a product is not immediately observed by the IM. Receipts are usually added to the inventory at the warehouse without full inspection. As a result, the inventory on record may consist of both nondefective products (available to meet customer demands) and defective products (not fit for sale). Since the defective products are not immediately observed, the actual (nondefective) inventory becomes partially observed.

If production lead times are long, an IM may have to place a particular order before observing the yields from previous orders, so that the production of the particular order is completed by a given due date. Thus, partial observability of inventory can be caused by due dates and long production lead times as well as process yields.

Theft. The items in the inventory can be stolen by thieves who violently break into the inventory storage, by the warehouse employees who calmly pilfer, or by the customers who shoplift. Since violent break-ins are generally investigated, they are usually observed and therefore not relevant for our study. We focus more on continuous pilferage or shoplifting, because they are not always observed without inventory inspections. Instances of theft at furniture retailers and at food wholesalers have been documented in [12, 23]. Axsäter [3] says: "... thefts may be a major problem. Apart from the loss in value, thefts ... also lead to inaccurate inventory records." Thus, the IM relying on inventory records ends up overestimating the available inventory until a stockout occurs. In this case, there are shortage costs in addition to costs of reordering, expediting, and re-receiving items to replace the stolen units. Typically, costs of the expedited items to urgently meet backlogged demands are much more than their regular costs.

When there is no physical inventory, i.e., the inventory is zero or negative, then none of the following would happen: transaction errors, misplaced inventories, spoilage, inventory level uncertainty due to yield and quality, or theft. Most companies pay utmost attention to an item when its inventory reaches zero. At these companies, employees walk around the shelves to identify the stocked-out items and verify the inventory levels for those items. This process is implemented at the office supplies store Staples and is called "zero-balance walk" in [16, 26]. Thus, a model based on a zero-balance walk process can be built by assuming that the inventory levels are fully observed when they are zero. It is the purpose of this paper to formulate and analyze a zero-balance walk model. This paper is part of a greater effort to build a comprehensive theory of inventory control under partial observations. In related works, we study some of the other inventory models with partial observations [6, 7, 8].

There have been a few studies of partial observations in the inventory control context. In these studies, partial observations are about demands rather than inventories. Among these, a common assumption is that of unobserved demands in the periods when lost sales occur. That is, the demand is observed fully when it is smaller than the available inventory. Otherwise, only the event that it is larger than the inventory is observed. When the underlying demand distribution is not known but estimated from the demand observations, partial demand observations limit the data available for estimation. This is called estimation with censored (demand) data. Ding, Puterman, and Bisi [14] and Lu, Song, and Zhu [22] have a multiperiod newsvendor model with censored demand. They assume the leftover inventory in a period to be salvaged entirely so that every period starts with zero inventory. This assumption decouples the periods from each other as far as the inventory evolution is concerned. However,

the periods are still coupled together by the current estimate of the demand distribution. The demand distribution is updated in a Bayesian fashion in each period with that period's demand or the observation of the lost sales event. Thus, there is an evolution equation that maps one period's demand distribution to the next period's. This evolution is affected by the choice of the order quantity. Before [14], Lariviere and Porteus [21] treated a similar problem for the restricted case of exponential demand distributions with gamma conjugate priors.

Treharne and Sox [27] study a periodic review inventory model with Markov modulated demand. A simple example of such a demand occurs when there are two demand states—High and Low—of a Markov chain and there are two demand distributions, one for each state. Unlike the Markovian demand cases (treated, e.g., in [10, 11]), it is not observed whether the demand state is High or Low. Instead, probabilities are used to represent the event that the demand state is High or Low. These probabilities along with the current level of inventory constitute the state of the system. The probabilities are updated in each period in accordance with that period's demand. Neither the current level of inventory nor the order size affects the probability updates. Evolution of the probabilities that capture partial observability is totally independent of the order quantities. The evolution equations can be written down in the first period, and they will include the random demands in the forthcoming periods. To make the discussion simple, here we mention only two demand states, but Treharne and Sox consider finitely many demand states. Consequently, they have a finite-dimensional state for their system.

The models we have described above make simplifying assumptions to end up with an easily workable setup. Ding, Puterman, and Bisi assume that the leftover inventory is salvaged every period, while Treharne and Sox have updates of the probabilities which are independent of the controls. Thus, they can capture only a limited amount of the dynamics associated with partial observations. Besides, they do not consider the issue of existence of optimal policies. Consequently, they do not require the methodology developed in this paper. Without such a methodology, however, inventory models with partial observations will remain largely unexplored.

A main reason for why the analysis of inventory problems under partial observations has been neglected lies in its mathematical difficulty. Whereas one works with a finite-dimensional state space in the full observation case, one usually has to deal with an infinite-dimensional state space in the partial observation setting. More specifically, the inventory level at a given time is no longer a system state in $\Re^n$; it must now be represented by its conditional probability given some limited information available at that time. Thus, the analysis takes place in the space of probability distributions. This is, of course, inevitable, and simplifies only in particular situations when, for instance, the separation principle applies; see [5] for an example.

Concerning controls of dynamic systems in general, a great step forward was achieved in the applied mathematics and engineering control literature, when the Zakai equation [30] was discovered. Prior to that, the evolution of the conditional probability had been studied with the highly nonlinear Kushner equation [18]. The Zakai equation uses a transformation that changes the Kushner equation into a pair of linear equations. This transformation corresponds to the concept of "change of measure" [15]. While it does not remove the infinite dimensionality, the linearity has permitted a number of important control problems with partial observations to be solved [5]. Of course, there remain numerical difficulties due to the infinite dimensionality of the state. Nevertheless, a sound theory is available.

The key idea in going from the Kushner equation to the Zakai equation is in in-

troducing unnormalized conditional probabilities in place of conditional probabilities. This linearizes the state equation, and the problem becomes much simpler to study. Ideas of this kind have not been introduced yet in the context of solving partial observation control problems in management. While the standard Zakai setup cannot be directly applied to inventory problems, we show that unnormalized conditional probabilities can be introduced and are indeed quite appropriate.

In the next section, we formulate the problem with normalized probabilities and switch to unnormalized probabilities. In sections 3 and 4, we examine the existence and uniqueness of the solution under the assumptions of bounded costs, bounded order quantities, and small discount rates. An asymptotically optimal control scheme is provided in section 5. Section 6 includes a brief conclusion and directions for future research.

2. The zero-balance walk model. We study a periodic review inventory problem with partially observed inventory levels. In our model, the inventory levels are not automatically observed by the IM who decides on order quantities. The order of events in any given period t is as follows: The IM observes the event when the inventory level falls to zero, but he does not observe the inventory level when it is positive. The manager determines how much to order and the order is delivered instantaneously. Next the customer demand occurs, but it is not observed by the IM unless the inventory level drops to zero. In each period, the IM incurs inventory related costs, but he does not observe these costs immediately. Lastly the state defining the inventory level is updated for the next period.

In classical inventory settings, the inventory level I_t at the beginning of period t is observed, and is used to determine the order quantity q_t in period t. Each period t has a random demand D_t defined on the probability space $(\Omega, \mathcal{F}, P)$. The demand is met, to the extent possible, from the on-hand stock $I_t + q_t$. We suppose that the demand that is not immediately met from the on-hand stock is lost. Then the evolution of inventory dynamics is given as follows:

$$I_{t+1} = (I_t + q_t - D_t)^+ \quad \text{for } t \geq 1. \tag{1}$$

We assume demand D_t to be independently and identically distributed (i.i.d.) random variables with the same distribution as D, where the density and the cumulative distribution function of D are denoted by f and F, respectively. Let $\bar{F} = 1 - F$.

When the demand is met entirely, inventory holding costs apply to the remaining inventory. Otherwise, there are lost sales costs. It is well known that a base stock (or an order up to S) policy is optimal for this setting. We investigate the validity of the optimality of a base stock policy, or lack of it, for the zero-balance walk model.

In the zero-balance walk model, the inventory levels are partially observed by the IM as follows:

$$I_1 \text{ is either 0 or its distribution is known.} \tag{2}$$

In general, the IM does not observe the demand or the inventory level. However, looking at empty shelves and concluding $I_t = 0$ does not take much effort, and constitutes a free observation. Thus, we allow I_t to be observed only when the inventory shelf is empty, i.e., $[I_t = 0]$. To study such partial observations of the inventory levels, we introduce a signal (message) random variable

$$z_t := \mathbb{1}_{I_t=0}, \quad t \geq 1. \tag{3}$$

The signal z_t is a discrete-time Markov chain with the state space $\{0,1\}$: 1 means an empty shelf and 0 means a nonempty shelf.

When the inventory levels are fully observed, the order q_t is adapted to the sigma field $\mathcal{F}_t := \sigma(\{I_j : 1 \le j \le t\})$ generated by the inventory levels observed by period t. Note that the demand observations up to and including the beginning of period t also generate the same field, i.e., $\mathcal{F}_t = \sigma(\{I_1, D_j : 1 \le j \le t-1\})$. With our partial observations model, q_t is adapted to $\mathcal{Z}_t := \sigma(\{z_j : 1 \le j \le t\})$. Clearly $\mathcal{Z}_t \subset \mathcal{F}_t$, so our partial observations model must decide on order quantities on the basis of less than full information.

Given a stationary cost function $c(I_t, q_t)$ that depends on the beginning inventory level I_t and the order size q_t in period t, and with $\tilde{q}$ defining the admissible sequence of actions $\tilde{q} = \{q_1, q_2, \dots\}$, the total discounted cost is defined by

$$J(\zeta, \pi, \tilde{q}) := \mathrm{E} \sum_{t=1}^{\infty} \alpha^t c(I_t, q_t), \tag{4}$$

where $\alpha < 1$ is the discount factor. The initial conditions are a pair $(\zeta, \pi(x))$, where ζ is 1 or 0. If ζ is 1, then $I_1 = 0$. If ζ is 0, then $I_1 > 0$ and $\pi(\cdot)$ is the probability distribution of I_1. We look for an admissible control $\tilde{q} = \{q_1, q_2, \dots\}$, with q_t adapted to $\mathcal{Z}_t$, $t \ge 1$, such that $J(\zeta, \pi, \tilde{q})$ is minimized.

Special cases. To make the form of the single-period cost $c(I, q)$ concrete, we can consider $c(I, q) = c_1 q + hI + b\mathrm{E}[(D - I - q)^+]$, which is often used in the inventory control literature [11]. The cost parameters c_1, h, and b can be interpreted as the cost of purchasing an item, the cost of holding an item in the inventory charged at the beginning of a period, and the opportunity cost of not selling an item when there is demand for it. Since $b\mathrm{E}[(D - I - q)^+] \le b\mathrm{E}[D]$, $c(I, q)$ is of linear growth in I and q. Another example includes a nonzero fixed cost of ordering. These observations will inspire an assumption on the bounds of the general single-period cost $c(I, q)$ in section 3.

2.1. Evolution of state probabilities. We now develop the conditional probability density $\pi_t(.)$ of I_t given $\mathcal{Z}_{t-1}$ and $I_t > 0$. By definition,

$$\int_0^x \pi_t(y)dy = \mathrm{P}(I_t \le x | \mathcal{Z}_{t-1}, I_t > 0).$$

Since the event $[I_t = 0]$ is observable, conditional probabilities are needed only when $I_t > 0$.

For any real and bounded test function $\varphi(.)$, we can use the conditional Bayes theorem (e.g., [15]) to obtain

$$\int_0^\infty \varphi(x)\pi_t(x)dx = \mathrm{E}[\varphi(I_t)|\mathcal{Z}_{t-1}, I_t > 0] = \frac{\mathrm{E}[\varphi(I_t)\mathbb{1}_{I_t>0}|\mathcal{Z}_{t-1}]}{\mathrm{E}[\mathbb{1}_{I_t>0}|\mathcal{Z}_{t-1}]} = \frac{\mathrm{E}[\varphi(I_t)\mathbb{1}_{I_t>0}|\mathcal{Z}_{t-1}]}{\mathrm{P}(I_t > 0|\mathcal{Z}_{t-1})}. \tag{5}$$

In order to obtain a recursive expression for π_t in terms of π_{t-1}, we begin with expressing $\mathrm{E}(\varphi(I_t)|\mathcal{Z}_t)$ in terms of conditional expectations with respect to $\mathcal{Z}_{t-1}$ in the next lemma.

Lemma 1.

$$\begin{aligned} E(\varphi(I_t)|\mathcal{Z}_t) &= \mathbb{1}_{I_t=0}\varphi(0) + \mathbb{1}_{I_t>0}\frac{E(\varphi(I_t)\mathbb{1}_{I_t>0}|\mathcal{Z}_{t-1})}{P(I_t > 0|\mathcal{Z}_{t-1})} \\ &= \mathbb{1}_{I_t=0}\varphi(0) + \mathbb{1}_{I_t>0}E(\varphi(I_t)|\mathcal{Z}_{t-1}, I_t > 0). \end{aligned} \tag{6}$$

Proof. Beginning with the left-hand side of (6), we have

$$\mathrm{E}(\varphi(I_t)|\mathcal{Z}_t) = \mathrm{E}[\varphi(I_t)(\mathbb{1}_{I_t=0} + \mathbb{1}_{I_t>0})|\mathcal{Z}_t)] = \varphi(0)\mathbb{1}_{I_t=0} + \mathrm{E}[\varphi(I_t)\mathbb{1}_{I_t>0}|\mathcal{Z}_t]. \tag{7}$$

Now take the last term in (7) and obtain

$$\begin{aligned}\mathrm{E}(\varphi(I_t)\mathbb{1}_{I_t>0}|\mathcal{Z}_t) &= \mathbb{1}_{I_t>0}\mathrm{E}(\varphi(I_t)|\mathcal{Z}_t)\\ &= \mathbb{1}_{I_t>0}\psi(z_1,\dots,z_{t-1},z_t)\\ &= \mathbb{1}_{I_t>0}\psi(z_1,\dots,z_{t-1},0),\end{aligned} \tag{8}$$

where the first equality follows from $\mathcal{Z}_t$-measurability of $\mathbb{1}_{I_t>0}$. The second equality merely expresses $\mathrm{E}(\varphi(I_t)|\mathcal{Z}_t)$ as $\psi(z_1,\dots,z_{t-1},z_t)$ for some measurable function ψ. The last equality follows from the fact that $I_t > 0 \Leftrightarrow z_t = 0$.

We now take the expectation of (8) with respect to $\mathcal{Z}_{t-1}$. Since $\mathcal{Z}_{t-1} \subseteq \mathcal{Z}_t$ and since $\psi(z_1,\dots,z_{t-1},0)$ is $\mathcal{Z}_{t-1}$-measurable, we obtain

$$\mathrm{E}[\varphi(I_t)\mathbb{1}_{I_t>0}|\mathcal{Z}_{t-1}] = \psi(z_1,\dots,z_{t-1},0)\mathrm{E}[\mathbb{1}_{I_t>0}|\mathcal{Z}_{t-1}] = \psi(z_1,\dots,z_{t-1},0)\mathrm{P}(I_t>0|\mathcal{Z}_{t-1}) \tag{9}$$

or

$$\psi(z_1,\dots,z_{t-1},0) = \frac{\mathrm{E}(\varphi(I_t)\mathbb{1}_{I_t>0}|\mathcal{Z}_{t-1})}{\mathrm{P}(I_t>0|\mathcal{Z}_{t-1})}. \tag{10}$$

Using (10) in (8) and substituting into (7) the resulting expression for $\mathrm{E}[\varphi(I_t)\mathbb{1}_{I_t>0}|\mathcal{Z}_t]$, we obtain the first equality in (6). Using (5) gives the second equality. □

Instead of the conditional expectations in Lemma 1, the left-hand side in (6) can also be expressed by using the conditional density function π_t. Using (5) on the right-hand side of (6) gives

$$\mathrm{E}(\varphi(I_t)|\mathcal{Z}_t) = \mathbb{1}_{I_t=0}\varphi(0) + \mathbb{1}_{I_t>0}\int_0^\infty \varphi(z)\pi_t(z)dz. \tag{11}$$

The density π_t is obtained by setting (6) and (11) to be equal. For $I_t = 0$, this equality yields $\pi_t = \delta$, which is the Dirac delta function taking the value of zero everywhere except at 0, where it is infinite. For the more interesting case of $I_t > 0$, the next lemma molds (6) into a convenient form to set (11) equal to (6) and solve for π_t.

Lemma 2.

$$\begin{aligned}E(\varphi(I_t)|\mathcal{Z}_t)\mathbb{1}_{I_t>0} &= \mathbb{1}_{I_{t-1}=0}\frac{\int_0^\infty \varphi(z)f(q_{t-1}-z)\mathbb{1}_{q_{t-1}\geq z}dz}{F(q_{t-1})}\\ &\quad + \mathbb{1}_{I_{t-1}>0}\frac{\int_0^\infty \varphi(z)\int_{(z-q_{t-1})^+}^\infty f(y+q_{t-1}-z)\pi_{t-1}(y)dydz}{\int_0^\infty F(y+q_{t-1})\pi_{t-1}(y)dy}.\end{aligned} \tag{12}$$

Proof. Consider the numerator in the second term on the right-hand side of (6). We see that

$$
\begin{aligned}
\mathrm{E}(\varphi(I_t)\mathbb{1}_{I_t>0}|\mathcal{Z}_{t-1}) &= \mathrm{E}(\varphi(I_{t-1}+q_{t-1}-D_{t-1})\mathbb{1}_{I_{t-1}+q_{t-1}-D_{t-1}>0}|\mathcal{Z}_{t-1}) \\
&= \mathrm{E}\left(\mathrm{E}(\varphi(I_{t-1}+q_{t-1}-D_{t-1})\mathbb{1}_{I_{t-1}+q_{t-1}-D_{t-1}>0}|\mathcal{Z}_{t-1},I_{t-1})|\mathcal{Z}_{t-1}\right) \\
&\qquad \text{because } \mathcal{Z}_{t-1}=\sigma(\{z_1,\dots,z_{t-1}\})\subseteq\sigma(\{z_1,\dots,z_{t-1},I_{t-1}\}) \\
&= \mathrm{E}\left(\int_0^{\infty}\varphi(I_{t-1}+q_{t-1}-y)\mathbb{1}_{I_{t-1}+q_{t-1}-y>0}f(y)dy|\mathcal{Z}_{t-1}\right) \\
&= \mathrm{E}\left(\int_0^{q_{t-1}+I_{t-1}}\varphi(I_{t-1}+q_{t-1}-y)f(y)dy|\mathcal{Z}_{t-1}\right) \\
&= \mathrm{E}\left(\int_0^{q_{t-1}+I_{t-1}}\varphi(x)f(I_{t-1}+q_{t-1}-x)dx|\mathcal{Z}_{t-1}\right) \\
&\qquad \text{set } x:=I_{t-1}+q_{t-1}-y \\
&= \mathrm{E}\left(\int_0^{\infty}\varphi(x)f(I_{t-1}+q_{t-1}-x)\mathbb{1}_{I_{t-1}+q_{t-1}-x\geq 0}dx|\mathcal{Z}_{t-1}\right) \\
&= \int_0^{\infty}\varphi(x)\mathrm{E}\left(f(I_{t-1}+q_{t-1}-x)\mathbb{1}_{I_{t-1}+q_{t-1}-x\geq 0}|\mathcal{Z}_{t-1}\right)dx. \qquad (13)
\end{aligned}
$$

Use (11) with the time index $t-1$ instead of t and replace $\varphi(I_{t-1})$ with $f(I_{t-1}+q_{t-1}-x)\mathbb{1}_{I_{t-1}+q_{t-1}-x\geq 0}$ to obtain

$$
\begin{aligned}
&\mathrm{E}(f(I_{t-1}+q_{t-1}-x)\mathbb{1}_{I_{t-1}+q_{t-1}-x\geq 0}|\mathcal{Z}_{t-1}) \\
&\quad = \mathbb{1}_{I_{t-1}=0}f(q_{t-1}-x)\mathbb{1}_{q_{t-1}-x\geq 0} \\
&\quad\quad + \mathbb{1}_{I_{t-1}>0}\int_0^{\infty}f(y+q_{t-1}-x)\mathbb{1}_{y+q_{t-1}-x\geq 0}\pi_{t-1}(y)dy. \qquad (14)
\end{aligned}
$$

Inserting (14) into (13), we obtain

$$
\begin{aligned}
\mathrm{E}(\varphi(I_t)\mathbb{1}_{I_t>0}|\mathcal{Z}_{t-1}) &= \mathbb{1}_{I_{t-1}=0}\int_0^{\infty}\varphi(x)f(q_{t-1}-x)\mathbb{1}_{x\leq q_{t-1}}dx \\
&\quad + \mathbb{1}_{I_{t-1}>0}\int_0^{\infty}\varphi(x)\left(\int_{(x-q_{t-1})^+}^{\infty}f(y+q_{t-1}-x)\pi_{t-1}(y)dy\right)dx.
\end{aligned}
$$

Now consider the denominator in the second term on the right-hand side of (6) to obtain

$$
\begin{aligned}
\mathrm{P}(I_t>0|\mathcal{Z}_{t-1}) &= \mathrm{E}(\mathbb{1}_{I_{t-1}+q_{t-1}-D_{t-1}>0}|\mathcal{Z}_{t-1}) \\
&= \mathrm{E}\{\mathrm{E}(\mathbb{1}_{I_{t-1}+q_{t-1}-D_{t-1}>0}|\mathcal{Z}_{t-1},I_{t-1})|\mathcal{Z}_{t-1}\} \\
&= \mathrm{E}\{F(I_{t-1}+q_{t-1})|\mathcal{Z}_{t-1}\} \\
&= \mathbb{1}_{I_{t-1}=0}F(q_{t-1})+\mathbb{1}_{I_{t-1}>0}\int_0^{\infty}F(y+q_{t-1})\pi_{t-1}(y)dy.
\end{aligned}
$$

Inserting the numerator and the denominator into (6) yields the desired result. ∎

Having obtained the conditional expectation in Lemma 2, we go back to the conditional probability π_t as defined in (11) for $I_t>0$. Setting the second term on the right-hand side of (11) equal to (12), we have

$$
\pi_t(x)=\mathbb{1}_{I_{t-1}=0}\left\{\frac{f(q_{t-1}-x)\mathbb{1}_{x\leq q_{t-1}}}{F(q_{t-1})}\right\}+\mathbb{1}_{I_{t-1}>0}\left\{\frac{\int_{(x-q_{t-1})^+}^{\infty}f(y+q_{t-1}-x)\pi_{t-1}(y)dy}{\int_0^{\infty}F(y+q_{t-1})\pi_{t-1}(y)dy}\right\}. \qquad (15)
$$

This expression specializes to the conditional probabilities stated in the next theorem.

THEOREM 1. *The conditional probability π_t can be expressed recursively as follows:*

$$\pi_t(x) = \left\{ \begin{array}{ll} \mathbb{1}_{x \le q_{t-1}} \dfrac{f(q_{t-1}-x)}{F(q_{t-1})} & \text{if } I_{t-1}=0 \\ \dfrac{\int_{(x-q_{t-1})^+}^{\infty} \pi_{t-1}(y) f(y+q_{t-1}-x)dy}{\int_0^{\infty} \pi_{t-1}(y)F(y+q_{t-1})dy} & \text{if } I_{t-1}>0 \end{array} \right\}. \tag{16}$$

Note that the denominators in (16) are $\mathrm{P}(D_{t-1} < I_{t-1}+q_{t-1})$, which is $\mathrm{P}(I_t > 0)$.

When $I_t > 0$, π_t is an absolutely continuous p.d.f. (probability density function). Note that the recursive equations for $I_{t-1} > 0$ and $I_{t-1} = 0$ coincide for $\pi_{t-1} = \delta$, so the equation for $I_{t-1} > 0$ applies even when $I_{t-1} = 0$. Since the largest value of I_t is $I_{t-1} + q_{t-1}$, π_t has a support of $[0, \sum_{i=1}^{t-1} q_i]$. If $I_{t'} = 0$ for some $t' < t$, then the support is $[0, \sum_{i=t'}^{t-1} q_i]$. Since π_1, f, and F are all given, the evolution of π_t can be controlled only by $\tilde{q} = \{q_1, q_2, \dots\}$.

The conditional probability evolves according to a highly nonlinear equation,

$$\begin{aligned} \pi_t(x) &= z_{t-1}\frac{f(q_{t-1}-x)\mathbb{1}_{x\le q_{t-1}}}{F(q_{t-1})} \\ &\quad + (1-z_{t-1})\frac{\int_{(x-q_{t-1})^+}^{\infty} f(y+q_{t-1}-x)\pi_{t-1}(y)dy}{\int_0^{\infty} F(q_{t-1}+y)\pi_{t-1}(y)dy}, \qquad t \ge 2, \\ \pi_1(x) &= \pi(x), \end{aligned} \tag{17}$$

which corresponds to the Kushner equation [18] in our inventory context.

We can linearize (17) as follows. Set

$$p_t(x) := \lambda_t \pi_t(x), \tag{18}$$

where λ_t is a *weighting factor* to be defined shortly. On account of this weighting, $p_t(x)$ can be viewed as unnormalized probability. Furthermore, it evolves according to the linear equation

$$\begin{aligned} p_t(x) &= z_{t-1} f(q_{t-1}-x)\mathbb{1}_{x\le q_{t-1}} + (1-z_{t-1})\int_{(x-q_{t-1})^+}^{\infty} f(y+q_{t-1}-x)p_{t-1}(y)dy, \\ p_1(x) &= \pi(x). \end{aligned} \tag{19}$$

This equation corresponds to the Zakai equation for systems with diffusions in [30, 5]. By integrating both sides of (18),

$$\begin{aligned} \lambda_t &= \int_0^{\infty} p_t(x)dx \\ &\overset{(19)}{=} z_{t-1}F(q_{t-1}) + (1-z_{t-1})\int_0^{\infty} F(q_{t-1}+y)p_{t-1}(y)dy \\ &\overset{(18)}{=} z_{t-1}F(q_{t-1}) + (1-z_{t-1})\lambda_{t-1}\int_0^{\infty} F(q_{t-1}+y)\pi_{t-1}(y)dy. \end{aligned}$$

The last equation defines λ_t recursively starting with $\lambda_1 = 1$. However, note that λ_t depends on π_{t-1} on the right-hand side. The normalized probabilities can easily be

computed from the unnormalized probabilities as follows:

$$\pi_t(x) = \frac{p_t(x)}{\int_0^\infty p_t(x)dx}. \tag{20}$$

These equations can be written in the operator form in the space

$$\mathcal{H} := \left\{ p \in L^1(\Re^+) \ : \ \int_0^\infty x|p(x)|dx < \infty \right\},$$

where $L^1(\Re^+)$ is the space of integrable functions whose domain is the set of nonnegative real numbers. If we define regular addition and multiplication by a scalar on $\mathcal{H}$ and include negative valued functions in $\mathcal{H}$, then $\mathcal{H}$ becomes a subspace of $L^1(\Re^+)$. Working with the subspace $\mathcal{H}$ is convenient for some of our arguments. However, we are ultimately interested in unnormalized probabilities, which are nonnegative. For them, we will specify an appropriate subset of $\mathcal{H}$ in section 2.2.

Let us equip the subspace $\mathcal{H}$ with the norm

$$||p|| = \int_0^\infty |p(x)|dx + \int_0^\infty x|p(x)|dx. \tag{21}$$

The dual space of $\mathcal{H}$ is denoted by $\mathcal{H}_*$, and it is the space of functions ϕ with linear growth, i.e.,

$$\mathcal{H}_* = \left\{ \phi \ : \ \sup_{x>0} \frac{|\phi(x)|}{1+x} < \infty \right\}.$$

Furthermore, we have the inner product

$$\langle p, \phi \rangle = \int_0^\infty p(x)\phi(x)dx \quad \text{for } p \in \mathcal{H},\ \phi \in \mathcal{H}_*.$$

For any scalar $q > 0$ and $p \in \mathcal{H}$, we define the linear operator ρ as

$$\rho(q,p)(x) = \int_{(x-q)^+}^\infty f(y+q-x)p(y)dy,$$

which is established in section 2.2 to be from $\mathcal{H}$ to $\mathcal{H}$. For the Dirac delta function $\delta \notin \mathcal{H}$, we define $\rho(q,\delta)(x) = f(q-x)\mathbb{1}_{x \le q}$. This gives us $\rho(0,\delta)(x) = 0$ almost everywhere in $\Re^+$. Define the nonlinear operator θ as

$$\theta(q,p)(x) = \frac{\rho(q,p)(x)}{\langle \rho(q,p), 1 \rangle}. \tag{22}$$

With these notations, we can write (17) and (19) in the operator form:

$$\pi_t(x) = z_{t-1}\theta(q_{t-1},\delta)(x) \ + \ (1-z_{t-1})\theta(q_{t-1},\pi_{t-1})(x), \tag{23}$$

$$p_t(x) = z_{t-1}\rho(q_{t-1},\delta)(x) \ + \ (1-z_{t-1})\rho(q_{t-1},p_{t-1})(x), \tag{24}$$

with the initial conditions

$$\pi_1 = p_1 = \pi. \tag{25}$$

Once again, we emphasize that (24) is a linear equation, while (23) is nonlinear.

2.2. Properties of the operators. In preparation to obtain some required operator properties, we need some identities and inequalities. By changing the order of integration, we obtain

$$\begin{aligned}\int_0^\infty |\rho(q,p)(x)|dx &\le \int_0^\infty \int_{(x-q)^+}^\infty f(y+q-x)|p(y)|dydx \\ &= \int_0^\infty \int_0^{y+q} f(y+q-x)|p(y)|dxdy \\ &= \int_0^\infty |p(y)| \int_0^{y+q} f(y+q-x)dxdy = \int_0^\infty |p(y)|F(y+q)dy. \end{aligned} \tag{26}$$

Using similar operations, we see that

$$\begin{aligned}\int_0^\infty x|\rho(q,p)(x)|dx &\le \int_0^\infty \int_{(x-q)^+}^\infty xf(y+q-x)|p(y)|dydx \\ &= \int_0^\infty \int_0^{y+q} xf(y+q-x)|p(y)|dxdy \\ &= \int_0^\infty p(y) \int_0^{y+q} (y+q-z)f(z)dzdy \\ &= \int_0^\infty (y+q)|p(y)|F(y+q)dy. \end{aligned} \tag{27}$$

Furthermore,

$$\begin{aligned}\int_0^\infty |\theta(q,p)(x)|dx &= \int_0^\infty \left|\frac{\rho(q,p)(x)}{\int_0^\infty \rho(q,p)(x)dx}\right| dx = \frac{\int_0^\infty |\rho(q,p)(x)|dx}{|\int_0^\infty \rho(q,p)(x)dx|} \\ &\overset{(26)}{\le} \frac{\int_0^\infty F(y+q)|p(y)|dy}{|\int_0^\infty F(y+q)p(y)dy|} \end{aligned} \tag{28}$$

and

$$\begin{aligned}\int_0^\infty x|\theta(q,p)(x)|dx &= \int_0^\infty \left|\frac{x\rho(q,p)(x)}{\int_0^\infty \rho(q,p)(x)dx}\right| dx = \frac{\int_0^\infty x|\rho(q,p)(x)|dx}{|\int_0^\infty \rho(q,p)(x)dx|} \\ &\overset{(26,27)}{\le} \frac{\int_0^\infty (y+q)|p(y)|F(y+q)dy}{|\int_0^\infty p(y)F(y+q)dy|} \\ &= q + \frac{\int_0^\infty y|p(y)|F(y+q)dy}{|\int_0^\infty p(y)F(y+q)dy|}. \end{aligned} \tag{29}$$

Next we derive some properties for nonnegative p. For this, we define

$$\mathcal{H}^+ := \{p \in \mathcal{H} : p \ge 0\}.$$

Note that $\mathcal{H}^+$ is not a subspace, as it does not include $-p$ for any $p > 0$.

Properties.

- Operator $\theta(q,p)$ is well defined if $\langle \rho(q,p), 1\rangle > 0$, i.e.,

$$\langle \rho(q,p), 1\rangle = \int_0^\infty p(y)F(y+q)dy \ne 0. \tag{30}$$

This property is satisfied if $p \neq 0$, $p \in \mathcal{H}^+$, and $F(y) > 0$ for all $y > 0$. Then, $\int_0^\infty p(y)F(y+q)dy > 0$. Otherwise, $p(y)F(y+q) = 0$ is a.e. satisfied, which along with $F(y) > 0$ for all $y > 0$ implies $p(y) = 0$ a.e. This contradicts $p \neq 0$.

Moreover, the operator ρ preserves the nonzero property: $p \neq 0 \implies \rho(q,p) \neq 0$. We establish the contrapositive of this statement. If $\rho(q,p) = 0$, then $\langle \rho(q,p), 1\rangle = 0$, which is possible only under $p = 0$.

Furthermore, note that the equality in (30) specializes to $\langle \rho(q,\delta), 1\rangle = F(q)$.

- Operator $\theta(q,p)(x)$ yields a valid p.d.f. if $p \in \mathcal{H}^+$. Clearly, $\theta(q,p)(x) \geq 0$ and $\int_0^\infty \theta(q,p)(x)dx = 1$.
- Operator $\rho(q,p)$ is a linear operator from $L^1(\Re^+)$ to $L^1(\Re^+)$ and also from $\mathcal{H}$ to $\mathcal{H}$. Moreover, $\mathcal{H}^+$ is closed under the operator ρ: $\rho(q,p) \in \mathcal{H}^+$ when $p \in \mathcal{H}^+$, because

$$\begin{aligned} ||\rho(q,p)|| &= \int_0^\infty |\rho(q,p)(x)|dx + \int_0^\infty x|\rho(q,p)(x)|dx \\ &\overset{(26,27)}{\leq} \int_0^\infty F(y+q)|p(y)|dy + \int_0^\infty (y+q)F(y+q)|p(y)|dy \\ &\leq \int_0^\infty |p(y)|dy + \int_0^\infty y|p(y)|dy + q\int_0^\infty |p(y)|dy < \infty, \end{aligned}$$

where the last less-than-or-equal-to relation is due to $F(y+q) \leq 1$.

- Operator $\theta(q,p)$ maps each element p, satisfying $\langle \rho(q,p), 1\rangle \neq 0$, from $L^1(\Re^+)$ to $L^1(\Re^+)$ and also from $\mathcal{H}$ to $\mathcal{H}$. Moreover, $\mathcal{H}^+$ is closed under the operator θ: $\theta(q,p) \in \mathcal{H}^+$ when $p \in \mathcal{H}^+$, because

$$\begin{aligned} ||\theta(q,p)|| &= \int_0^\infty |\theta(q,p)(x)|dx + \int_0^\infty x|\theta(q,p)(x)|dx \\ &\overset{(28,29)}{\leq} \frac{\int_0^\infty F(y+q)|p(y)|dy}{|\int_0^\infty F(y+q)p(y)dy|} + q + \frac{\int_0^\infty y|p(y)|F(y+q)dy}{|\int_0^\infty p(y)F(y+q)dy|} \\ &\leq \frac{\int_0^\infty |p(y)|dy}{|\int_0^\infty F(y)p(y)dy|} + q + \frac{\int_0^\infty y|p(y)|dy}{|\int_0^\infty p(y)F(y)dy|}. \end{aligned}$$

Because of (30), the denominator is positive. Consequently, the right-hand side is finite and $\theta(q,p) \in \mathcal{H}^+$.

- Operator $\rho(q,p)$ when $q = 0$ is a contraction mapping: when no order is made, the inventory distribution shifts to the left. This follows from

$$\begin{aligned} ||\rho(0,p)|| &= \left\| \int_x^\infty f(y-x)p(y)dy \right\| \\ &\leq \int_0^\infty \int_x^\infty f(y-x)|p(y)|dydx + \int_0^\infty x\int_x^\infty f(y-x)|p(y)|dydx \\ &= \int_0^\infty |p(y)| \int_0^y f(y-x)dxdy + \int_0^\infty |p(y)| \int_0^y xf(y-x)dxdy \\ &\leq \int_0^\infty |p(y)|F(y)dy + \int_0^\infty |p(y)|yF(y)dy \\ &\leq \int_0^\infty |p(y)|dy + \int_0^\infty |p(y)|ydy \qquad \text{(since } F \leq 1\text{)} \\ &= ||p||. \end{aligned} \tag{31}$$

- Operator θ is homogenous of degree 0 in p because

$$\langle \theta(q,p),1\rangle = \int_0^\infty \frac{\rho(q,p)}{\langle \rho(q,p),1\rangle} dx = 1 \quad \text{and}$$

$$\theta(q,\lambda p) = \theta(q,p) \quad \text{for each constant } \lambda \in \Re.$$

2.3. The Bellman equation. We write $p_t(\tilde{q})$ and $\pi_t(\tilde{q})$ to emphasize the dependence of the states p_t or π_t on the control policy. We assume that $c(I_t, q_t)$ has linear growth in I_t for every fixed q_t, i.e., $c(.,q_t) \in \mathcal{H}_*$. The cost function can be written as follows:

$$\begin{aligned} J(\zeta,\pi,\tilde{q}) &= \sum_{t=1}^{\infty} \alpha^t \mathrm{E}[\mathrm{E}[c(I_t,q_t)|\mathcal{Z}_t]] \\ &= \sum_{t=1}^{\infty} \alpha^t \mathrm{E}\{z_t c(0,q_t) + (1-z_t)\langle c(I_t,q_t), \pi_t(\tilde{q})\rangle\}, \end{aligned}$$

where $\pi_t(\tilde{q})$ is the solution of (17). Recall that the initial conditions $\zeta_1 = \zeta \in \{0,1\}$ and $\pi_1 = \pi$ are given. In what follows, we study only the discounted infinite horizon costs, so the time index t is suppressed. We define the value function

$$V(\zeta,\pi) := \inf_{\tilde{q}} J(\zeta,\pi,\tilde{q}).$$

Looking one period ahead from period one,

$$V(\zeta,\pi) = \inf_q \{\zeta c(0,q) + (1-\zeta)\langle c(.,q),\pi(\cdot)\rangle + \alpha \mathrm{E}[V(\zeta_2,\pi_2)|\zeta,\pi]\},$$

where

$$\begin{aligned} \mathrm{E}[V(\zeta_2,\pi_2)|\zeta,\pi] &= \mathrm{E}[V(\zeta_2, \zeta\theta(q,\delta) + (1-\zeta)\theta(q,\pi))|\zeta,\pi] \\ &= \mathrm{P}(I_2 = 0|\zeta) V(1, \zeta\theta(q,\delta) + (1-\zeta)\theta(q,\pi)) \\ &\quad + \mathrm{P}(I_2 > 0|\zeta) V(0, \zeta\theta(q,\delta) + (1-\zeta)\theta(q,\pi)). \end{aligned}$$

Then $V(\zeta,\pi)$ can be written more explicitly, depending on ζ, as follows:

$$\begin{aligned} V(0,\pi) = \inf_q \Bigg\{ &\langle c(.,q),\pi(.)\rangle + \alpha V(1,\theta(q,\pi)) \int_0^\infty \bar{F}(y+q)\pi(y)dy \\ &+ \alpha V(0,\theta(q,\pi)) \int_0^\infty F(y+q)\pi(y)dy \Bigg\}, \\ V(1,\pi) = \inf_q \left\{ c(0,q) + \alpha V(1,\theta(q,\delta))\bar{F}(q) + \alpha V(0,\theta(q,\delta))F(q) \right\}. \end{aligned}$$

If we write $v := V(1,\pi)$ which, in fact, is not dependent on π, and $V(\pi) := V(0,\pi)$, then we obtain the following system:

$$\begin{aligned} V(\pi) = \inf_q \Bigg\{ &\langle c(.,q),\pi(.)\rangle \\ &+ \alpha v \int_0^\infty \bar{F}(y+q)\pi(y)dy + \alpha V(\theta(q,\pi)) \int_0^\infty F(y+q)\pi(y)dy \Bigg\}, \end{aligned} \tag{32}$$

$$v = \inf_q \left\{ c(0,q) + \alpha v \bar{F}(q) + \alpha V(\theta(q,\delta))F(q) \right\}. \tag{33}$$

A direct study of the system in (32)–(33) is not very easy. The matters simplify considerably when working with the unnormalized probability $p \in \mathcal{H}^+$. The unnormalized probability evolves in accordance with the linear operator ρ. To make ideas concrete, we define a new value function $Z(\cdot)$ as follows:

$$Z(p) := V\left(\frac{p}{\lambda}\right)\lambda, \quad \lambda := \int_0^\infty p(x)dx.$$

It follows from (32) that

(34)

$$\begin{aligned} Z(p) &= \lambda \inf_q \left\{ \langle c(.,q), p(.)/\lambda\rangle \ + \alpha v \int_0^\infty \bar{F}(y+q)(p(y)/\lambda)dy \right. \\ &\qquad \left. + \alpha V(\theta(q,p/\lambda)) \int_0^\infty F(y+q)(p(y)/\lambda)dy \right\} \\ &= \inf_q \left\{ \langle c(.,q), p(.)\rangle \ + \alpha v \int_0^\infty \bar{F}(y+q)p(y)dy + \alpha V(\theta(q,p)) \int_0^\infty F(y+q)p(y)dy \right\}, \end{aligned}$$

where we use the fact that θ is a homogenous operator of degree 0. Now consider the term $V(\theta(q,\pi))$ on the right-hand side. Recall that the λ value corresponding to $\rho(q,p)$ is $\langle \rho(q,p), 1\rangle$. Thus,

$$\begin{aligned} Z(\rho(q,p)) &= \left\{ \int_0^\infty \rho(q,p)(x)dx \right\} \left\{ V\left(\frac{\rho(q,p)}{\langle \rho(q,p), 1\rangle} \right) \right\} \\ &= \left\{ \int_0^\infty F(y+q)p(y)dy \right\} \{V\left(\theta(q,p)\right)\}. \end{aligned}$$

This equality can be used to eliminate $V(\theta(q,p))$ from (34). It can be specialized for $p = \delta$ to eliminate $V(\theta(q,\delta))$ from the expression for v. Eventually, we obtain the following new system of equations:

(35)

$$Z(p) = \inf_q \left\{ \langle c(.,q), p(.)\rangle \ + \alpha v \int_0^\infty \bar{F}(y+q)p(y)dy + \alpha Z(\rho(q,p)) \right\} \quad \text{for all } p \in \mathcal{H}^+,$$

(36)

$$v = \inf_q \left\{ c(0,q) + \alpha v \bar{F}(q) + \alpha Z(\rho(q,\delta)) \right\}.$$

Recall that for the Dirac delta function $\delta \notin \mathcal{H}^+ \subseteq \mathcal{H}$, we have defined $\rho(q,\delta)(x) = f(q-x)\mathbb{1}_{x\le q}$ and $\rho(0,\delta) = 0$ almost everywhere.

From (35) and (36), it follows that

$$Z(\mu p) = \mu Z(p) \quad \text{for every } \mu > 0. \tag{37}$$

Thus, $Z(0) = 0$.

Unlike the operator θ, ρ is a linear operator. Thus, it is easier to study the system in (35)–(36) than that in (32)–(33). The linearity facilitates our arguments dealing with the existence of an optimal feedback control and our discussion when it is finite. Furthermore, it helps in studying finite approximations of the infinite-dimensional state space as well as in building associated approximate solutions to (35)–(36).

We conclude this section by remarking that this problem and the methodology developed here have not appeared before in the inventory control literature.

3. Existence of a solution to the Bellman equation.

3.1. Bounded costs. For the existence result, we bound the single-period cost. To include the special cases given in section 2, we consider positive constants c, c_0, c_1, c_2, and h such that

$$c_2 + cq < c(I,q) \leq c_0 + c_1 q + hI \quad \text{for } I \geq 0. \tag{38}$$

To include the special cases, it is sufficient to set $c_0 \geq c(0,0)$, where $c(0,0)$ represents the maximum expected cost of lost sales that can be incurred in a period. Let $a_0 := \max\{c_0/(1-\alpha), h\}$.

In the subsequent analysis, we shall assume $c(I,q)$ to be continuous in I. Furthermore, we shall assume continuity in q for convenience in exposition. In the case when there is a fixed cost of ordering, $c(I,q)$ will be discontinuous in q at $q=0$. However, the affected proofs can be easily extended to handle this case.

To accommodate our unnormalized conditional probabilities, we define the functional space

$$\mathcal{B} := \left\{ \phi(p) : \mathcal{H}^+ \to \Re : \sup_{p \in \mathcal{H}^+} \frac{|\phi(p)|}{||p||} < \infty \right\} \tag{39}$$

equipped with the norm

$$||\phi||_{\mathcal{B}} := \sup_{p \in \mathcal{H}^+} \frac{|\phi(p)|}{||p||}, \tag{40}$$

where $||p||$ still refers to the norm that we initially defined in $\mathcal{H} \supseteq \mathcal{H}^+$. For any $\phi \in \mathcal{B}$, we must have $\phi(0) = 0$.

We will often speak of a pair (v, Z), which is made of a scalar $v \in \Re$ and an element $Z \in \mathcal{B}$. As such, (v, Z) can be considered as an element in the functional space $\Re \times \mathcal{B}$. We suppress the argument p in $Z(p)$ when Z is considered as a functional. Thus, when the existence, uniqueness, or continuity of Z is under consideration, we only write Z in what follows. For example, a solution of (35)–(36) will be a pair of the form (v, Z). We will write $(v_1, Z_1) \leq (v_2, Z_2)$ to mean $(v_1, Z_1(p)) \leq (v_2, Z_2(p))$ for any given $p \in \mathcal{H}^+$.

We search for a solution (v, Z) of (35)–(36) in $\Re \times \mathcal{B}$. A property of this solution is presented next.

LEMMA 3. *Each solution (v, Z) of (35)–(36) in $\Re \times \mathcal{B}$ satisfies $||Z||_{\mathcal{B}} \leq a_0/(1-\alpha)$.*

Proof. For any given $p \in \mathcal{H}^+$, by setting $q = 0$, we obtain

$$Z(p) \leq \left\{ \langle c(.,0), p(.) \rangle + \alpha v \int_0^\infty \bar{F}(y) p(y) dy + \alpha Z(\rho(0,p)) \right\}. \tag{41}$$

From (38) and the fact that $\bar{F} \leq 1$, we can write

$$\begin{aligned} Z(p) &\leq c_0 \int_0^\infty p(x)dx + h \int_0^\infty x p(x)dx + \alpha v \int_0^\infty p(x)dx + \alpha Z(\rho(0,p)) \qquad (42) \\ &\overset{(40)}{\leq} (c_0 + \alpha v) \int_0^\infty p(x)dx + h \int_0^\infty x p(x)dx + \alpha ||Z||_{\mathcal{B}} ||\rho(0,p)|| \\ &\overset{(31)}{\leq} (c_0 + \alpha v) \int_0^\infty p(x)dx + h \int_0^\infty x p(x)dx + \alpha ||Z||_{\mathcal{B}} ||p||. \qquad (43) \end{aligned}$$

Now use the Bellman equation for v along with $\rho(0,\delta)=0$ and $Z(0)=0$ to obtain the second inequality:

$$0 \le v \le c(0,0) + \alpha v \le c_0 + \alpha v.$$

This implies $v \le c_0/(1-\alpha)$. Then $c_0 + \alpha v \le c_0 + \alpha c_0/(1-\alpha) = c_0/(1-\alpha)$, which can be inserted into the upper bound for $Z(p)$ above to obtain

$$Z(p) \le (c_0/(1-\alpha)) \int_0^\infty p(x)dx + h \int_0^\infty xp(x)dx + \alpha ||Z||_{\mathcal{B}} ||p||$$

and, in turn,

$$Z(p) \le a_0||p|| + \alpha||Z||_{\mathcal{B}}||p||.$$

By dividing both sides by $||p||$ and taking the supremum over $p \in \mathcal{H}^+$, we obtain

$$||Z||_{\mathcal{B}} \le a_0 + \alpha||Z||_{\mathcal{B}},$$

which implies $||Z||_{\mathcal{B}} \le a_0/(1-\alpha)$. □

Define the function $G : \Re \times \mathcal{H} \to \Re$, given (v, Z), as

$$G(q,p;v,Z) := \langle c(.,q), p(.)\rangle + \alpha v \int_0^\infty \bar{F}(y+q)p(y)dy + \alpha Z(\rho(q,p)).$$

For $p = \delta$,

$$G(q,\delta;v,Z) = c(0,q) + \alpha v\bar{F}(q) + \alpha Z(\rho(q,\delta)).$$

Define the map $T : \Re \times \mathcal{B} \to \Re \times \mathcal{B}$ as

$$T\begin{pmatrix} v \\ Z(p) \end{pmatrix} := \begin{pmatrix} \inf_q G(q,\delta;v,Z) \\ \inf_q G(q,p;v,Z) \end{pmatrix}. \tag{44}$$

Define Z_0 as the value function that solves the Bellman equations when $q = 0$. Then it must solve

$$\langle c(.,0), p(.)\rangle + \alpha v_0 \int_0^\infty \bar{F}(y)p(y)dy + \alpha Z_0(\rho(0,p)) = Z_0(p), \tag{45}$$

where $v_0 := Z_0(p=\delta)$. By (36) and $Z(0)=0$, we have $v_0 = c_0 + \alpha v_0$. The existence and uniqueness of the functional Z_0 are established in the next lemma.

LEMMA 4. *Z_0 exists and is uniquely defined.*

Proof. First, we look for a solution of (45). Consider a linear and bounded map $U : \mathcal{H} \to \Re$. We can locally define the norm of U as

$$||U||_{\mathcal{B}^-} = \sup_{p\in\mathcal{H}} |U(p)|/||p||. \tag{46}$$

This norm is defined on the functional space $\mathcal{B}^-$, which can be constructed like $\mathcal{B}$ but with $\mathcal{H}$ instead of $\mathcal{H}^+$ in (39). In comparison to $\phi \in \mathcal{B}$ and Z_0, U has a larger domain that includes negative p. Here it is more convenient to work in the subspace $\mathcal{H}$ instead of $\mathcal{H}^+$.

Define

$$\tau(U)(p) := \langle c(.,0), p(.)\rangle + \alpha v_0 \int_0^\infty \bar{F}(y)p(y)dy + \alpha U(\rho(0,p)).$$

The map τ is an affine function of U and it is linear in p. For linear maps $U_1, U_2 \in \mathcal{B}^-$, we have

$$\begin{aligned} |\tau(U_1)(p) - \tau(U_2)(p)| &= \alpha|U_1(\rho(0,p)) - U_2(\rho(0,p))| \\ &= \alpha|(U_1 - U_2)(\rho(0,p))| \\ &\leq \alpha||\rho(0,p)|| \cdot ||U_1 - U_2||_{\mathcal{B}^-} \\ &\overset{(31)}{\leq} \alpha||p|| \cdot ||U_1 - U_2||_{\mathcal{B}^-}. \end{aligned}$$

This equality above follows from the linearity of U_1 and U_2. Using the inequality above, we can deduce that

$$\sup_{p\in\mathcal{H}} \frac{|\tau(U_1)(p) - \tau(U_2)(p)|}{||p||} \leq \alpha||U_1 - U_2||_{\mathcal{B}^-}.$$

The left-hand side above is the norm $||\tau(U_1) - \tau(U_2)||$ of $\tau(U_1) - \tau(U_2)$, so we arrive at

$$||\tau(U_1) - \tau(U_2)|| \leq \alpha||U_1 - U_2||_{\mathcal{B}^-}.$$

It follows that τ is a contraction mapping and it has a unique fixed point U_0 such that $\tau(U_0) = U_0$. By restricting the domain of U_0 to $\mathcal{H}^+$, we uniquely obtain Z_0. Thus, Z_0 is a fixed point for T when $q = 0$, and it solves (45). □

If $v \leq v_0$, $Z(p) \leq Z_0(p)$, and

$$\begin{pmatrix} \tilde{v} \\ \tilde{Z}(p) \end{pmatrix} := T \begin{pmatrix} v \\ Z(p) \end{pmatrix},$$

then $\tilde{v} \leq c(0,0) + \alpha v \leq c_0 + \alpha v_0 = v_0$. Also,

$$\begin{aligned} \tilde{Z}(p) &\leq \langle c(.,0), p(.)\rangle + \alpha v \int_0^\infty \bar{F}(y)p(y)dy + \alpha Z(\rho(0,p)) \\ &\leq \langle c(.,0), p(.)\rangle + \alpha v_0 \int_0^\infty \bar{F}(y)p(y)dy + \alpha Z_0(\rho(0,p)) \\ &\overset{(45)}{=} Z_0(p). \end{aligned}$$

For $0 \leq v \leq v_0$ and $0 \leq Z(p) \leq Z_0(p)$, we have just shown that

$$T \begin{pmatrix} v \\ Z(p) \end{pmatrix} \leq \begin{pmatrix} v_0 \\ Z_0(p) \end{pmatrix}. \tag{47}$$

This inspires the next theorem, which proves the existence of a solution of the system (35)–(36) by using a value iteration scheme. The solution is denoted as $(\bar{v}, \bar{Z})$.

THEOREM 2. *Under assumption* (38)*, a solution of* (35)–(36) *exists.*

Proof. Let

$$\begin{pmatrix} v_{n+1} \\ Z_{n+1}(p) \end{pmatrix} := T \begin{pmatrix} v_n \\ Z_n(p) \end{pmatrix}.$$

Starting with v_0 and Z_0 defined by (45), we first claim that

$$\begin{pmatrix} v_{n+1} \\ Z_{n+1}(p) \end{pmatrix} \leq \begin{pmatrix} v_n \\ Z_n(p) \end{pmatrix}.$$

The claim can be established recursively. First consider $n = 0$. Then, by setting $q = 0$, we have

$$v_1 \leq c_0 + \alpha v_0 + \alpha Z_0(\rho(0, \delta)) = v_0,$$

$$Z_1(p) \leq \langle c(.,0), p(.) \rangle + \alpha v_0 \int_0^\infty \bar{F}(y)p(y)dy + \alpha Z_0(\rho(0,p)) = Z_0(p),$$

where the equalities are due to (45). Now we assume that the claim holds for $n = k$, and then establish it for $n = k + 1$. If $v \leq v'$ and $Z(p) \leq Z'(p)$, then it follows from the definition of T that

$$T\begin{pmatrix} v \\ Z(p) \end{pmatrix} \leq T\begin{pmatrix} v' \\ Z'(p) \end{pmatrix}.$$

By the recursion hypothesis $(v_{k+1}, Z_{k+1}) \leq (v_k, Z_k)$. Take $v = v_{k+1}$, $Z = Z_{k+1}$, $v' = v_k$, and $Z' = Z_k$ in the above inequality to finish the proof of the claim.

Since (v_n, Z_n) is a nonincreasing sequence with a lower bound of $(0, 0)$, it has a limit $(\bar{v}, \bar{Z})$: $(v_n, Z_n) \downarrow (\bar{v}, \bar{Z})$. In other words,

$$\begin{pmatrix} \bar{v} \\ \bar{Z}(p) \end{pmatrix} = \lim_{n\to\infty} T_n \begin{pmatrix} v_0 \\ Z_0(p) \end{pmatrix},$$

where T_n is the n times composition of T. Since T is not known to be continuous, the above equality does not yield $(\bar{v}, \bar{Z})$ as a fixed point. But, for a finite n,

$$\begin{pmatrix} v_{n+1} \\ Z_{n+1}(p) \end{pmatrix} \geq T\begin{pmatrix} \bar{v} \\ \bar{Z}(p) \end{pmatrix}.$$

Applying T infinitely many times on both sides, we arrive at

$$\begin{pmatrix} \bar{v} \\ \bar{Z}(p) \end{pmatrix} \geq T\begin{pmatrix} \bar{v} \\ \bar{Z}(p) \end{pmatrix}. \tag{48}$$

On the other hand, by arbitrarily picking a q in (44), we see that

$$v_{n+1} \leq c(0, q) + \alpha v_n \bar{F}(q) + \alpha Z_n(\rho(q, \delta)),$$

$$Z_{n+1}(p) \leq \langle c(., q), p(.) \rangle + \alpha v_n \int_0^\infty \bar{F}(y + q)p(y)dy + \alpha Z_n(\rho(q, p)).$$

Specializing these inequalities for $v_n = \bar{v}$ and $Z_n(p) = \bar{Z}(p)$, we obtain

$$\begin{pmatrix} \bar{v} \\ \bar{Z}(p) \end{pmatrix} \leq T\begin{pmatrix} \bar{v} \\ \bar{Z}(p) \end{pmatrix}. \tag{49}$$

Combining (48) and (49), we establish that $(\bar{v}, \bar{Z})$ is a fixed point. The fixed point is not necessarily unique, but it is the maximum solution in the following sense. Any solution (v, Z) that satisfies $(v, Z) \leq (v_0, Z_0)$ also satisfies $(v, Z) \leq (\bar{v}, \bar{Z})$. □

Backward interpretation of the monotone iterative process (v_n, Z_n). Recall the monotone iterative process on (v_n, Z_n), which starts with $v_0 = c_0/(1-\alpha)$ and Z_0 as given in (45), and continues with $(v_{n+1}, Z_{n+1}) = T(v_n, Z_n)$.

Now consider a new sequence $(v_{n,N+1}, Z_{n,N+1})$ constructed by starting with

$$Z_{N+1,N+1}(p) = Z_0(p), \quad v_{N+1,N+1} = v_0$$

and by moving backwards recursively:

$$\begin{pmatrix} v_{n+1,N+1} \\ Z_{n+1,N+1}(p) \end{pmatrix} \leq T \begin{pmatrix} v_{n,N+1} \\ Z_{n,N+1}(p) \end{pmatrix}.$$

Then $Z_{n,N+1}(p) = Z_{N+1-n}(p)$ and $v_{n,N+1} = v_{N+1-n}$.

We define the total discounted cost as

$$J_{n,N+1}(\zeta, \pi, \tilde{q}) := \mathrm{E}\left[\sum_{t=n}^{N} \alpha^{t-n} c(I_t, q_t) + \sum_{t=N+1}^{\infty} \alpha^{t-n} c(I_t, 0)\right],$$

with $z_n = \zeta$ and $\pi_n = \pi$. Then we have the validation of the backward monotone process:

$$\inf_{\tilde{q}} J_{n,N+1}(\zeta, \pi, \tilde{q}) = \zeta v_{n,N+1} + (1-\zeta) Z_{n,N+1}(\pi).$$

Note that the cost $c(.,.)$ is bounded. Then, $J_{1,N+1}(\zeta, \pi, \tilde{q})$ converges to $J(\zeta, \pi, \tilde{q})$ as N increases. Thus, one has the option of finding the optimal order quantities either by a forward or a backward recursion.

Bounding the optimal order quantity. Start by setting $q = 0$ in (35) and obtain

$$\begin{aligned}
Z(p) &\leq \langle c(.,0), p(.,0) \rangle \; + \alpha v \int_0^\infty \bar{F}(y) p(y) dy + \alpha Z(\rho(0,p)) \\
&\leq c_0 \int_0^\infty p(x) dx \\
&\quad + h \int_0^\infty x p(x) dx + \alpha \frac{c_0}{1-\alpha} \int_0^\infty p(x) dx + \alpha \left(\sup_{p \in \mathcal{H}^+} \frac{|Z(\rho(0,p))|}{||\rho(0,p)||} \right) ||\rho(0,p)|| \\
&\leq c_0 \int_0^\infty p(x) dx + h \int_0^\infty x p(x) dx + \alpha \frac{c_0}{1-\alpha} \int_0^\infty p(x) dx + \alpha \left(\frac{a_0}{1-\alpha} \right) ||p|| \\
&\leq c_0 \int_0^\infty p(x) dx + h \int_0^\infty x p(x) dx \\
&\quad + \alpha \frac{c_0}{1-\alpha} \int_0^\infty p(x) dx + \alpha \frac{a_0}{1-\alpha} \left(\int_0^\infty p(x) dx + \int_0^\infty x p(x) dx \right) \\
&\leq \left(\frac{c_0}{1-\alpha} + \frac{a_0 \alpha}{1-\alpha} \right) \int_0^\infty p(x) dx + \left(h + \frac{a_0 \alpha}{1-\alpha} \right) \int_0^\infty x p(x) dx \\
&\leq \frac{a_0}{1-\alpha} \left\{ \int_0^\infty p(x) dx + \int_0^\infty x p(x) dx \right\}.
\end{aligned}$$

With an arbitrary order quantity q, the cost has a lower bound of $cq \int_0^\infty p(x) dx$. If this bound exceeds $Z(p)$, then q cannot be optimal. Hence, the optimal order quantity

satisfies $cq\int_0^\infty p(x)dx \le Z(p)$, which along with the above inequality implies

$$q \le \frac{a_0}{c(1-\alpha)}\left\{1+\frac{\int_0^\infty xp(x)dx}{\int_0^\infty p(x)dx}\right\}. \tag{50}$$

Note that the bound depends on the unnormalized probability p and can be arbitrarily large as $p \to 0$. Because of this observation, we choose to assume a bound on the order quantity in the next subsection.

3.2. Bounded order quantities. In this section we assume that there is a finite bound on the order quantity q in addition to the cost bounds in the previous section. The finite bound can be due to the supplier's limited production or transportation capacity, or the storage capacity IM can use. Let the capacity be m and let the corresponding Z and v be denoted by Z^m and v^m. Then (35)–(36) is written as

$$\begin{aligned} &Z^m(p)\\ &= \inf_{q\le m}\left\{\langle c(.,q),p(.)\rangle + \alpha v^m\int_0^\infty \bar F(y+q)p(y)dy + \alpha Z^m(\rho(q,p))\right\} \quad \text{for all } p\in\mathcal{H}^+,\\ &v^m = \inf_{q\le m}\left\{c(0,q)+\alpha v^m\bar F(q)+\alpha Z^m(\rho(q,\delta))\right\}. \end{aligned} \tag{51}$$

We shall prove that the functional Z^m is Lipschitz continuous on $\mathcal{H}$, i.e., there exist constants A^m and B^m such that

$$|Z^m(p)-Z^m(p')| \le A^m\int_0^\infty |p(y)-p'(y)|dy + B^m\int_0^\infty y|p(y)-p'(y)|dy \tag{52}$$

for any $p, p' \in \mathcal{H}$. This additional smoothness property allows us to establish the uniqueness of a solution of the system in (35)–(36). The next lemma illustrates how the constants A^m and B^m can be chosen.

LEMMA 5. *For a fixed m, Z^m is Lipschitz continuous, where the constants A^m and B^m in* (52) *should be chosen as*

$$A^m = \frac{c_0}{(1-\alpha)^2} + \frac{m}{1-\alpha}\left(c_1+\frac{\alpha h}{1-\alpha}\right) \quad \textit{and} \quad B^m = \frac{h}{1-\alpha}.$$

Proof. To bound the differences of Z^m, we construct

$$Z^{m,n+1}(p) = \inf_{q\le m} G(q,p;v^m,Z^{m,n})$$

by starting with $Z^{m,0}$, which solves

$$\langle c(.,0),p(.)\rangle + \alpha v^m\int_0^\infty \bar F(y)p(y)dy + \alpha Z^{m,0}(\rho(0,p)) = Z^{m,0}(p). \tag{53}$$

Since $Z^{m,n}$ converges to Z^m as n goes to infinity, it suffices to prove that $Z^{m,n}$ is Lipschitz continuous with constants independent of n. This proof is by induction on n. In other words, Lipschitz continuity must be preserved as n grows and it must hold when we start at $n=0$. These two requirements correspond respectively to the following two claims, which are proved below.

Claim 1. $Z^{m,n+1}$ is Lipschitz continuous with constants A^m and B^m, if $Z^{m,n}$ is Lipschitz continuous with the same constants.

Claim 2. $Z^{m,0}$ is Lipschitz continuous with constants A^m and B^m.

Proof of Claim 1. We start by bounding $G(q,p;;v^m,Z^{m,n}) - G(q,p';v^m,Z^{m,n})$ as follows. In the first inequality below, we use $v^m \le c_0/(1-\alpha)$ and the fact that ρ is a linear operator:

$$\begin{aligned}
& |G(q,p;v^m,Z^{m,n}) - G(q,p';v^m,Z^{m,n})| \\
\le\ & (c_0 + c_1 q)\int_0^\infty |p(y)-p'(y)|dy + h\int_0^\infty y|p(y)-p'(y)|dy \\
& + \frac{\alpha c_0}{1-\alpha}\int_0^\infty |p(y)-p'(y)|dy \\
& + \alpha\left\{A^m\int_0^\infty |\rho(q,p-p')(y)|dy + B^m\int_0^\infty y|\rho(q,p-p')(y)|dy\right\} \\
\overset{(26,27)}{\le}\ & \left(\frac{c_0}{1-\alpha} + c_1 q\right)\int_0^\infty |p(y)-p'(y)|dy + h\int_0^\infty y|p(y)-p'(y)|dy \\
& + \alpha\left\{(A^m + B^m q)\int_0^\infty |p(y)-p'(y)|dy + B^m\int_0^\infty y|p(y)-p'(y)|dy\right\} \\
\le\ & \left(\frac{c_0}{1-\alpha} + c_1 m + \alpha(A^m + B^m m)\right)\int_0^\infty |p(y)-p'(y)|dy \\
& + (h + \alpha B^m)\int_0^\infty y|p(y)-p'(y)|dy.
\end{aligned}$$

Note that the right-hand side of the last inequality is independent of q, provided that $0 \le q \le m$. Hence,

$$\begin{aligned}
|Z^{m,n+1}(p) - Z^{m,n+1}(p')| &= |\inf_{q\le m} G(q,p;v^m,Z^{m,n}) - \inf_{q\le m} G(q,p';v^m,Z^{m,n})| \\
&\le \left(\frac{c_0}{1-\alpha} + c_1 m + \alpha(A^m + B^m m)\right)\int_0^\infty |p(y)-p'(y)|dy \\
&\quad + (h + \alpha B^m)\int_0^\infty y|p(y)-p'(y)|dy \\
&\le A^m\int_0^\infty |p(y)-p'(y)|dy + B^m\int_0^\infty y|p(y)-p'(y)|dy.
\end{aligned}$$

The last inequality holds only if

$$h + \alpha B^m \le B^m,$$

so we set $B^m = h/(1-\alpha)$. Then, we have the equality below and we require the inequality below:

$$\frac{c_0}{1-\alpha} + c_1 m + \alpha(A^m + B^m m) = \frac{c_0}{1-\alpha} + c_1 m + \alpha(A^m + mh/(1-\alpha)) \le A^m.$$

This inequality yields the condition on A^m and completes the proof of Claim 1. □

Proof of Claim 2. We prove the claim by using another induction. Consider the following iteration:

$$
\begin{aligned}
Z^{m,0,0}(p) &= \langle c(.,0), p(.)\rangle + \alpha v^m \int_0^\infty \bar{F}(y)p(y)dy,\\
Z^{m,0,n}(p) &= G(0,p;v^m, Z^{m,0,n-1})\\
&= \langle c(.,0), p(.)\rangle + \alpha v^m \int_0^\infty \bar{F}(y)p(y)dy + \alpha Z^{m,0,n-1}(\rho(0,p)).
\end{aligned}
$$

It is immediate that $Z^{m,0,n}(p) \geq Z^{m,0,n-1}(p)$, and we can also show that $Z^{m,0,n}(p)$ is bounded. Thus, $Z^{m,0,n}(p)$ converges to $Z^{m,0}(p)$, which is the unique solution of (53). We know from Claim 1 that G preserves Lipschitz continuity. Thus, it suffices to show that $Z^{m,0,0}$ is Lipschitz. This follows from

$$
\begin{aligned}
&|Z^{m,0,0}(p) - Z^{m,0,0}(p')|\\
&\quad\leq c_0 \int |p(y) - p'(y)|dy + h\int y|p(y) - p'(y)|dy + \frac{\alpha c_0}{1-\alpha}|p(y) - p'(y)|dy\\
&\quad\leq \underbrace{\frac{c_0}{1-\alpha}}_{\leq A^m}\int |p(y) - p'(y)|dy + \underbrace{h}_{\leq B^m}\int y|p(y) - p'(y)|dy.
\end{aligned}
$$

This completes the proof of Claim 2. Thus, the lemma is proved. □

Existence of a solution to (51) can be proved by following the steps in Theorem 2. The more interesting issue is whether the respective solutions of (35)–(36) and (51) coincide as the bound on the order quantity is removed. To make ideas concrete, define the map T^m similar to T in (44):

$$
T^m \begin{pmatrix} v \\ Z(p) \end{pmatrix} := \begin{pmatrix} \inf_{q\leq m} G(q,\delta;v,Z) \\ \inf_{q\leq m} G(q,p;v,Z) \end{pmatrix}. \tag{54}
$$

Let T_n^m be n compositions of T^m. Thus, (v_n^m, Z_n^m) can be obtained by applying T_n^m on (v_0, Z_0). By the arguments in Theorem 2, (v_n^m, Z_n^m) is a nonincreasing sequence, and it converges to, say, $(\bar{v}^m, \bar{Z}^m)$. We next establish that $(\bar{v}^m, \bar{Z}^m)$ converges to $(\bar{v}, \bar{Z})$, which is the maximal solution of the system in (35)–(36).

LEMMA 6. *$(\bar{v}^m, \bar{Z}^m) \downarrow (\bar{v}, \bar{Z})$ as m increases to ∞.*

Proof. Because of the constraint $q \leq m$,

$$
T^m \begin{pmatrix} v \\ Z(p) \end{pmatrix} \geq T^{m+1} \begin{pmatrix} v \\ Z(p) \end{pmatrix}
$$

for any (v, Z). Starting the value iteration with (v_0, Z_0),

$$
\begin{pmatrix} v_1^m \\ Z_1^m(p) \end{pmatrix} = T^m \begin{pmatrix} v_0 \\ Z_0(p) \end{pmatrix} \geq T^{m+1} \begin{pmatrix} v_0 \\ Z_0(p) \end{pmatrix} = \begin{pmatrix} v_1^{m+1} \\ Z_1^{m+1}(p) \end{pmatrix}.
$$

Applying T^m on the right and T^{m-1} on the left as many times as necessary, we set

$$
(\bar{v}^m, \bar{Z}^m) = \lim_{n\to\infty} T_n^m(v_0, Z_0) \geq \lim_{n\to\infty} T_n^{m+1}(v_0, Z_0) = (\bar{v}^{m+1}, \bar{Z}^{m+1}).
$$

In particular, $(\bar{v}^m, \bar{Z}^m) \geq (\bar{v}, \bar{Z})$. In addition, $(\bar{v}^m, \bar{Z}^m)$ is nonincreasing in m, so it has a limit, say $(\tilde{v}, \tilde{Z})$. Clearly, $(\tilde{v}, \tilde{Z}) \geq (\bar{v}, \bar{Z})$.

Recall that $(\bar{v}, \bar{Z})$ is the maximal fixed point of T. To finish the proof, it suffices to argue that $(\tilde{v}, \tilde{Z})$ is also a fixed point of T. Since $(\tilde{v}, \tilde{Z}) \leq (v_0, Z_0)$, we can repeat

the initial steps in the proof of Theorem 2 to obtain the following inequality, which is analogous to (48):

$$\begin{pmatrix} \tilde{v} \\ \tilde{Z}(p) \end{pmatrix} \geq T \begin{pmatrix} \tilde{v} \\ \tilde{Z}(p) \end{pmatrix}.$$

On the other hand, for any q and $q \leq m$, we have

$$\begin{aligned} Z^m(p) &\leq G(q, p; v^m, Z^m), \\ v^m &\leq G(q, \delta; v^m, Z^m). \end{aligned}$$

Hence,

$$\begin{pmatrix} \tilde{v} \\ \tilde{Z}(p) \end{pmatrix} \leq T \begin{pmatrix} \tilde{v} \\ \tilde{Z}(p) \end{pmatrix}.$$

Thus, $(\tilde{v}, \tilde{Z})$ is a fixed point of T. □

So far, we have studied the existence and the convergence of (v^m, Z^m). The next theorem validates the monotone iterative process, that is, (v^m, Z^m) minimizes the total discounted cost. As a side product of the theorem, v^m and Z^m turn out to be unique because they are equal to the minimum costs, which are unique by definition.

THEOREM 3. *The solution* (v^m, Z^m) *of* (51) *gives the minimum total discounted cost as follows:*

$$\begin{aligned} Z^m(\pi) &= \inf_{\tilde{q}: q_t \leq m} J(0, \pi, \tilde{q}), \\ v^m &= \inf_{\tilde{q}: q_t \leq m} J(1, \delta, \tilde{q}). \end{aligned}$$

Proof. The proof has two parts. We first show that $Z^m(\pi)$ and v^m are, respectively, smaller than $J(0, \pi, \tilde{q})$ and $J(1, \delta, \tilde{q})$ for any $\tilde{q}$ such that $q_t \leq m$. For an arbitrary $q \leq m$,

$$Z^m(p) \leq \langle c(., q), p(.) \rangle + \alpha v^m \int_0^\infty \bar{F}(y + q) p(y) dy + \alpha Z^m(\rho(q, p)).$$

Take $p = \pi_t$ and $q = q_t$ so that

$$Z^m(\pi_t) \leq \langle c(., q_t), \pi_t(.) \rangle + \alpha v^m \int_0^\infty \bar{F}(y + q_t) \pi_t(y) dy + \alpha Z^m(\rho(q_t, \pi_t)),$$

$$v^m \leq c(0, q_t) + \alpha v^m \bar{F}(q_t) + \alpha Z^m(\rho(q_t, \delta)).$$

Note that by (51) and (37), $Z^m(\lambda p) = \lambda Z^m(p)$ for any scalar $\lambda > 0$. Taking $\lambda = \int F(y + q_t) \pi_t(y) dy$, we obtain

$$\begin{aligned} Z^m(\pi_t) \leq \langle c(., q_t), \pi_t(.) \rangle &+ \alpha v^m \int_0^\infty \bar{F}(y + q_t) \pi_t(y) dy \\ &+ \alpha Z^m(\theta(q_t, \pi_t)) \int_0^\infty F(y + q_t) \pi_t(y) dy, \end{aligned}$$

$$v^m \leq c(0, q_t) + \alpha v^m \bar{F}(q_t) + \alpha Z^m(\theta(q_t, \delta)) F(q_t).$$

Now combine these two inequalities by using the weights $1 - z_t$ and z_t, respectively, and obtain

$$\begin{aligned}(1 - z_t)Z^m(\pi_t) + z_t v^m &\le z_t c(0, q_t) + (1 - z_t)\langle c(., q_t), \pi_t(.)\rangle \\ &\quad + \alpha \left\{ (1 - z_t) \left[v^m \int_0^\infty \bar{F}(y + q_t)\pi_t(y)dy \right.\right. \\ &\quad \left.\left. + Z^m(\theta(q_t, \pi_t)) \int_0^\infty F(y + q_t)\pi_t(y)dy \right] + z_t \left[v^m \bar{F}(q_t) + Z^m(\theta(q_t, \delta))F(q_t) \right] \right\}. \qquad (55)\end{aligned}$$

Consider the relation

$$\begin{aligned}&\mathrm{E}[(1 - z_{t+1})Z^m(\pi_{t+1}) + z_{t+1}v^m|\mathcal{Z}_t] \\ &\quad = \mathrm{P}(I_{t+1} > 0|\mathcal{Z}_t)Z^m(\pi_{t+1}) + \mathrm{P}(I_{t+1} = 0|\mathcal{Z}_t)v^m \\ &\quad = \left[z_t F(q_t) + (1 - z_t) \int_0^\infty F(q_t + y)\pi_t(y) \right] Z^m(\pi_{t+1}) \\ &\qquad + \left[z_t \bar{F}(q_t) + (1 - z_t) \int_0^\infty \bar{F}(q_t + y)\pi_t(y)dy \right] v^m \\ &\quad = \left[z_t F(q_t) + (1 - z_t) \int_0^\infty F(q_t + y)\pi_t(y) \right] Z^m(z_t\theta(q_t, \delta) + (1 - z_t)\theta(q_t, \pi_t)) \\ &\qquad + \left[z_t \bar{F}(q_t) + (1 - z_t) \int_0^\infty \bar{F}(q_t + y)\pi_t(y)dy \right] v^m \\ &\quad = z_t \left[Z^m(\theta(q_t, \delta))F(q_t) + v^m \bar{F}(q_t) \right] \\ &\qquad + (1 - z_t) \left[Z^m(\theta(q_t, \pi_t)) \int_0^\infty F(q_t + y)\pi_t(y)dy + v^m \int_0^\infty \bar{F}(q_t + y)\pi_t(y)dy \right].\end{aligned}$$

Now insert this equality into the curly brackets in (55) to obtain

$$\begin{aligned}(1 - z_t)Z^m(\pi_t) + z_t v^m &\le z_t c(0, q_t) + (1 - z_t)\langle c(., q_t), \pi_t(.)\rangle \\ &\quad + \alpha \left\{ \mathrm{E}[(1 - z_{t+1})Z^m(\pi_{t+1}) + z_{t+1}v^m|\mathcal{Z}_t] \right\}.\end{aligned}$$

By multiplying both sides by α^t and taking expected values, we get

$$\begin{aligned}\alpha^t \mathrm{E}[(1 - z_t)Z^m(\pi_t) + z_t v^m] &\le \alpha^t \mathrm{E}[z_t c(0, q_t) + (1 - z_t)\langle c(., q_t), \pi_t(.)\rangle] \\ &\quad + \alpha^{t+1} \mathrm{E}[(1 - z_{t+1})Z^m(\pi_{t+1}) + z_{t+1}v^m] \quad \text{for } t \ge 1.\end{aligned}$$

Now sum up both sides for $t \ge 1$ to obtain

$$\begin{aligned}\sum_{t=1}^\infty \alpha^t \mathrm{E}[(1 - z_t)Z^m(\pi_t) + z_t v^m] &\le \sum_{t=1}^\infty \alpha^t \mathrm{E}[z_t c(0, q_t) + (1 - z_t)\langle c(., q_t), \pi_t(.)\rangle] \\ &\quad + \sum_{t=1}^\infty \alpha^{t+1} \mathrm{E}[(1 - z_{t+1})Z^m(\pi_{t+1}) + z_{t+1}v^m]. \qquad (56)\end{aligned}$$

The second term on the right-hand side is next argued to be finite, so that it can be deducted from both sides of the inequality.

We now construct an upper bound for $\mathrm{E}Z^m(\pi_{t+1})$. First note that

$$\mathrm{E}Z^m(\pi_{t+1}) \le |\mathrm{E}Z^m(\pi_{t+1})| \le ||Z^m||_{\mathcal{B}} ||\mathrm{E}\pi_{t+1}|| = ||Z^m||_{\mathcal{B}} \mathrm{E}||\pi_{t+1}||,$$

where

$$\mathrm{E}||\pi_{t+1}|| = 1 + \mathrm{E}\int_0^\infty x\pi_{t+1}(x)dx.$$

But

$$\mathrm{E}\int_0^\infty x\pi_{t+1}(x)dx = \mathrm{E}(1-z_{t+1})\mathrm{E}[I_{t+1}|\mathcal{Z}_{t+1}] \le \mathrm{E}I_{t+1} \le I_0 + m(t+1),$$

where the last inequality follows from $q_t \le m$. Therefore, as $t \to \infty$,

$$\begin{aligned}\alpha^{t+1}\mathrm{E}[(1-z_{t+1})Z^m(\pi_{t+1}) + z_{t+1}v^m] &\le \alpha^{t+1}\mathrm{E}[Z^m(\pi_{t+1})] + \alpha^{t+1}v^m \\ &\le \alpha^{t+1}\frac{a_0}{1-\alpha}\mathrm{E}[\pi_{t+1}] + \alpha^{t+1}v^m \\ &\qquad\qquad \text{by Lemma 3} \\ &\le \alpha^{t+1}\frac{a_0}{1-\alpha}(I_0 + m(t+1)) + \alpha^{t+1}v^m.\end{aligned}$$

Since $\sum_t \alpha^{t+1}\frac{a_0}{1-\alpha}(I_0 + m(t+1)) + \alpha^{t+1}v^m < \infty$, the second term on the right-hand side of (56) is finite. Now deduct it from the left-hand side to obtain

$$\alpha\mathrm{E}[(1-z_1)Z^m(\pi_1) + z_1v^m] \le \sum_{t=1}^\infty \alpha^t\mathrm{E}[z_tc(0,q_t) + (1-z_t)\langle c(.,q_t),\pi_t(.)\rangle]. \tag{57}$$

From (57), it follows that

$$\begin{aligned}(1-\zeta)Z^m(\pi) + \zeta v^m &= \mathrm{E}[(1-z_1)Z^m(\pi) + z_1v^m] \\ &\le \sum_{t=1}^\infty \alpha^t[z_tc(0,q_t) + (1-z_t)\langle c(.,q_t),\pi_t(.)\rangle] = J(\zeta,\pi,\tilde{q}).\end{aligned}$$

This finishes the first part of the proof.

For the second part of the proof, we construct an optimal solution by considering an optimal feedback control $\hat{q}^m(p)$ and $\hat{q}^m$ such that

$$Z^m(p) = \langle c(.,\hat{q}^m(p)),p(.)\rangle + \alpha v^m\int_0^\infty \bar{F}(y+\hat{q}^m(p))p(y)dy + \alpha Z^m(\rho(\hat{q}^m(p),p)),$$

$$v^m = c(0,\hat{q}^m) + \alpha v^m\bar{F}(\hat{q}^m) + \alpha Z^m(\rho(\hat{q}^m,\delta)).$$

Existence of an optimal feedback control follows from the continuity of $c(\cdot,\cdot)$, the continuity of $\bar{F}(\cdot)$, and the Lipschitz continuity of Z^m in Lemma 5.

We associate these feedbacks with a stochastic process $\{\hat{\tilde{q}}_t^m\}$, which is adapted to $\mathcal{Z}_t$ and defined recursively as follows:

$$\hat{\tilde{q}}_1^m := \zeta\hat{q}^m + (1-\zeta)\hat{q}^m(\pi),$$

$$\hat{\tilde{q}}_{t+1}^m := z_t\hat{q}^m + (1-z_t)\hat{q}^m(\pi_t).$$

The definitions of $\hat{q}^m(p)$, $\hat{q}^m$, and $\hat{\tilde{q}}^m$ yield the next two equalities:

$$
\begin{aligned}
&(1-z_t)Z^m(\pi_t)+z_tv^m\\
&\quad=(1-z_t)\langle c(.,\hat{q}^m(\pi_t)),\pi_t(.)\rangle+\alpha(1-z_t)v^m\int_0^\infty \bar{F}(y+\hat{q}^m(\pi_t))\pi_t(y)dy\\
&\qquad+\alpha(1-z_t)Z^m(\rho(\hat{q}^m(\pi_t),\pi_t))+z_tc(0,\hat{q}^m)+\alpha z_tv^m\bar{F}(\hat{q}^m)+\alpha z_tZ^m(\rho(\hat{q}^m,\delta))\\
&\quad=(1-z_t)\langle c(.,\hat{\tilde{q}}^m_{t+1}),\pi_t(.)\rangle+\alpha(1-z_t)v^m\int_0^\infty \bar{F}(y+\hat{\tilde{q}}^m_{t+1})\pi_t(y)dy\\
&\qquad+\alpha(1-z_t)Z^m(\rho(\hat{\tilde{q}}^m_{t+1},\pi_t))+z_tc(0,\hat{\tilde{q}}^m_{t+1})+\alpha z_tv^m\bar{F}(\hat{\tilde{q}}^m_{t+1})+\alpha z_tZ^m(\rho(\hat{\tilde{q}}^m_{t+1},\delta))\\
&\quad=(1-z_t)\langle c(.,\hat{\tilde{q}}^m_{t+1}),\pi_t(.)\rangle+z_tc(0,\hat{\tilde{q}}^m_{t+1})\\
&\qquad+\alpha(1-z_t)\left[v^m\int_0^\infty \bar{F}(y+\hat{\tilde{q}}^m_{t+1})\pi_t(y)dy+Z^m(\rho(\hat{\tilde{q}}^m_{t+1},\pi_t))\right]\\
&\qquad+\alpha z_t[v^m\bar{F}(\hat{\tilde{q}}^m_{t+1})+Z^m(\rho(\hat{\tilde{q}}^m_{t+1},\delta))].
\end{aligned}
$$

Specializing for $t=1$ and taking the infimum over $\tilde{q}$ yields

$$(1-\zeta)Z^m(\pi)+\zeta v^m\geq J(\zeta,\pi,\hat{\tilde{q}}^m).$$

Combining the two parts of the proof we have

$$(1-\zeta)Z^m(\pi)+\zeta v^m=J(\zeta,\pi,\hat{\tilde{q}}^m).\qquad\square$$

Since $Z^m(\pi)$ and v^m are defined as a solution of (51) and they are given by the infima in Theorem 3, both $Z^m(\pi)$ and v^m are unique. As m increases, we have

$$\inf_{\tilde{q}:q_t\leq m}J(0,\pi,\tilde{q})\downarrow\inf_{\tilde{q}}J(0,\pi,\tilde{q}),$$

$$\inf_{\tilde{q}:q_t\leq m}J(1,\pi,\tilde{q})\downarrow\inf_{\tilde{q}}J(1,\pi,\tilde{q}).$$

These imply

$$Z(\pi)=\inf_{\tilde{q}}J(0,\pi,\tilde{q}),$$

$$v=\inf_{\tilde{q}}J(1,\pi,\tilde{q}).$$

Thus, $Z(\pi)$ and v are interpreted as the infima of the costs even when m disappears. However, a corresponding feedback solution that yields $Z(\pi)$ and v may not exist unless m is finite.

In this section, we have established that Z^m is continuous and that it converges to Z. However, these results do not guarantee the continuity of Z. Nevertheless, we have the weaker form of continuity as presented in the next lemma.

LEMMA 7. *Z is upper semicontinuous, i.e.,*

$$\limsup_{p_k\to p}Z(p_k)\leq Z(p).$$

Proof. Since $Z(p_k)\leq Z^m(p_k)$ for each k and Z^m is continuous,

$$\limsup_{p_k\to p}Z(p_k)\leq\limsup_{p_k\to p}Z^m(p_k)=Z^m(p).$$

Now take the limit as $m\to\infty$ to obtain the desired result. $\square$

4. A sufficiently small discount rate. We argue that T is a contraction map for a sufficiently small α. Namely, we let $M := 1 + a_0/(c(1-\alpha))$ and require that $\alpha(1+M) < 1$. In this case, the solution of the system (35)–(36), whose existence follows from Theorem 2, is unique.

Consider the difference

$$G(q,p;v,Z) - G(q,p;v',Z') = \alpha(v-v')\int_0^\infty \bar{F}(y+q)p(y)dy + \alpha[Z(\rho(q,p)) - Z'(\rho(q,p))].$$

Let us define a distance in the space $\Re \times \mathcal{B}$ by

$$\eta = d\left(\begin{pmatrix} v \\ Z(p) \end{pmatrix}, \begin{pmatrix} v' \\ Z'(p) \end{pmatrix}\right) := \max\left\{\sup_{p\in\mathcal{H}^+} \frac{|Z(p)-Z'(p)|}{||p||},\ |v-v'|\right\}.$$

Since $Z \in \mathcal{B}$ and $\eta < \infty$, we have

$$\begin{aligned} |G(q,p;v,Z) - G(q,p;v',Z')| &\le \alpha|v-v'|\int_0^\infty p(y)dy + \alpha||Z-Z'||_{\mathcal{B}}||\rho(q,p)|| \\ &\le \eta\alpha\int_0^\infty p(y)dy + \eta\alpha||\rho(q,p)||. \end{aligned} \tag{58}$$

On the other hand, by (50),

$$q \le \frac{a_0}{c(1-\alpha)}\frac{||p||}{\int_0^\infty p(y)dy},$$

or, equivalently,

$$q\int_0^\infty p(y)dy \le \frac{a_0}{c(1-\alpha)}||p||.$$

Using (26) and (27), we obtain

$$\begin{aligned} ||\rho(q,p)|| &\le \int_0^\infty |p(y)|dy + \int_0^\infty (y+q)|p(y)|F(y+q)dy \\ &\le ||p|| + q\int_0^\infty |p(y)|dy \\ &\le ||p||\left(1 + \frac{a_0}{c(1-\alpha)}\right) \\ &= M||p||. \end{aligned} \tag{59}$$

Now insert (59) into (58) to obtain

$$|G(q,p;v,Z) - G(q,p;v',Z')| \le \eta\alpha(1+M)||p||.$$

Hence, it follows that

$$|\inf_q G(q,p;v,Z) - \inf_q G(q,p;v',Z')| \le \eta\alpha(1+M)||p||.$$

Recalling $||\delta|| = 1$ and specializing to $p = \delta$, we have

$$|\inf_q G(q,\delta;v,Z) - \inf_q G(q,\delta;v',Z')| \le \eta\alpha(1+M).$$

In summary, we have proved

$$d\left(T\begin{pmatrix} v \\ Z(p) \end{pmatrix}, T\begin{pmatrix} v' \\ Z'(p) \end{pmatrix}\right) \leq \alpha(1+M)d\left(\begin{pmatrix} v \\ Z(p) \end{pmatrix}, \begin{pmatrix} v' \\ Z'(p) \end{pmatrix}\right).$$

In addition, if $\alpha(1+M) < 1$ as required at the beginning of this section, T is a contraction map on $\Re \times \mathcal{B}$. It is also a contraction on $\Re \times \mathcal{B}_c$, where $\mathcal{B}_c$ denotes the closed subset containing the continuous functions in $\mathcal{B}$. Therefore, when $\alpha(1+M) < 1$, (35)–(36) have a unique fixed point in $\Re \times \mathcal{B}_c$.

4.1. Relaxing the condition on the discount rate. We relax the condition $\alpha(1+M) < 1$ to $\alpha M < 1$ by measuring the distance d in $\Re \times \mathcal{B}$ for a fixed λ. Namely, we consider the projection of $\Re \times \mathcal{B}$ onto $\{\lambda\} \times \mathcal{B}$. For any λ, we define the projected value function Z^λ as the solution of

$$Z^\lambda(p) = \inf_q \left\{ \langle c(.,q), p(.)\rangle \ + \alpha\lambda \int_0^\infty \bar{F}(y+q)p(y)dy + \alpha Z^\lambda(\rho(q,p)) \right\}. \tag{60}$$

Note by (37) that $Z^\lambda(\mu p) = \mu Z^\lambda(p)$, and thus $Z^\lambda(0) = 0$.

LEMMA 8. *If $\lambda < v_0 = c_0/(1-\alpha)$, then $Z^\lambda(p) \leq Z_0(p)$ for all $p \in \mathcal{H}^+$.*

Proof. First note that

$$Z^\lambda(p) \leq \langle c(.,0), p(.)\rangle + \alpha v_0 \int_0^\infty \bar{F}(y)p(y)dy + \alpha Z^\lambda(\rho(0,p)),$$

$$Z_0(p) = \langle c(.,0), p(.)\rangle + \alpha v_0 \int_0^\infty \bar{F}(y)p(y)dy + \alpha Z_0(\rho(0,p)).$$

Now take the difference to obtain $Z^\lambda(p) - Z_0(p) \leq \alpha(Z^\lambda(\rho(0,p)) - Z_0(\rho(0,p)))$. Taking the positive parts, we obtain $[Z^\lambda(p) - Z_0(p)]^+ \leq \alpha[Z^\lambda(\rho(0,p)) - Z_0(\rho(0,p))]^+$. Divide both sides by $||p||$ and take the supremum over $p \in \mathcal{H}^+$ to obtain

$$\begin{aligned} \sup_{p\in\mathcal{H}^+} \frac{[Z^\lambda(p) - Z_0(p)]^+}{||p||} &\leq \alpha \sup_{p\in\mathcal{H}^+} \frac{[Z^\lambda(\rho(0,p)) - Z_0(\rho(0,p))]^+}{||p||} \\ &\leq \alpha||Z^\lambda - Z_0||_{\mathcal{B}} \sup_{p\in\mathcal{H}^+} \frac{||\rho(0,p)||}{||p||} \leq \alpha||Z^\lambda - Z_0||_{\mathcal{B}} \\ &\leq \alpha \left\{ \sup_{p\in\mathcal{H}^+} \frac{[Z^\lambda(p) - Z_0(p)]^+}{||p||} \right\}. \end{aligned} \tag{61}$$

By recursing (61) N times, we get

$$\sup_{p\in\mathcal{H}^+} \frac{[Z^\lambda(p) - Z_0(p)]^+}{||p||} \leq \alpha^N \left\{ \sup_{p\in\mathcal{H}^+} \frac{[Z^\lambda(p) - Z_0(p)]^+}{||p||} \right\}.$$

Thus, $[Z^\lambda(p) - Z_0(p)]^+ = 0$ or $Z^\lambda(p) \leq Z_0(p)$. ∎

Define, for $\lambda \leq v_0$ and $Z \leq Z_0$,

$$T^\lambda(Z)(p) := \inf_q \left\{ \langle c(.,q), p(.)\rangle \ + \alpha\lambda \int_0^\infty \bar{F}(y+q)p(y)dy + \alpha Z(\rho(q,p)) \right\}.$$

Use $v = v' = \lambda$ in the steps used in obtaining (58) to arrive at

$$|T^\lambda(Z)(p) - T^\lambda(Z')(p)| \leq \alpha||Z - Z'||_{\mathcal{B}}||\rho(q,p)|| \leq \alpha||Z - Z'||_{\mathcal{B}}M||p||.$$

The second inequality is due to $||\rho(q,p)|| \leq M||p||$, where $M = 1 + a_0/(c(1-\alpha))$. Then it is immediate that

$$||T^\lambda(Z_1) - T^\lambda(Z_2)||_{\mathcal{B}} \leq \alpha M||Z_1 - Z_2||_{\mathcal{B}}.$$

If $\alpha M < 1$, then $T^\lambda(Z)$ is a contraction mapping. Thus, $T^\lambda(Z)$ has a fixed point in $\Re \times \mathcal{B}_c$. Consequently, Z^λ is uniquely defined. It is shown to be nondecreasing with respect to λ in the next lemma.

LEMMA 9. *$Z^\lambda(p)$ is nondecreasing in λ for each $p \in \mathcal{H}^+$.*

Proof. For any λ, define

$$Z_n^\lambda(p) := T^\lambda Z_{n-1}^\lambda(p), \quad Z_0^\lambda(p) := Z_0(p).$$

Then $Z_n^\lambda(p) \downarrow Z^\lambda(p)$. Suppose $\lambda' > \lambda$. By induction, if

$$Z_{n-1}^{\lambda'}(p) \geq Z_{n-1}^\lambda(p),$$

then

$$Z_n^{\lambda'}(p) = T^{\lambda'}(Z_{n-1}^{\lambda'})(p) \geq T^\lambda(Z_{n-1}^{\lambda'})(p) \geq T^\lambda(Z_{n-1}^\lambda)(p) = Z_{n-1}^\lambda(p).$$

Going to the limit yields $Z^{\lambda'}(p) \geq Z^\lambda(p)$. □

Now consider the function $g(\lambda)$ for $\lambda \geq 0$ defined by

$$g(\lambda) := \inf_q \left\{ c(0,q) + \alpha\lambda\bar{F}(q) + \alpha Z^\lambda(\rho(q,\delta)) \right\}, \tag{62}$$

where Z^λ is given by (60). By Lemma 9, the map $g(\lambda)$ is nondecreasing. It is concave with a rate of increase of at most α, by the next theorem.

THEOREM 4. *The system in* (35)–(36) *has a unique solution.*

Proof. We first establish the concavity of $Z^\lambda(p)$ in λ for each fixed $p \in \mathcal{H}^+$:

$$Z^{\beta\lambda_1+(1-\beta)\lambda_2}(p) \geq \beta Z^{\lambda_1}(p) + (1-\beta)Z^{\lambda_2}(p).$$

This will be proved by induction on n in Z_n^λ, where $Z_n^\lambda = T^\lambda Z_{n-1}^\lambda$. For $n = 0$, $Z_0^\lambda(p) = Z_0(p)$, so it is a constant and hence concave. If the concavity holds for $Z_{n-1}^\lambda(p)$ then,

$$\begin{aligned}
Z_n^{\beta\lambda_1+(1-\beta)\lambda_2}(p) &= \inf_q \Bigg\{ \langle c(.,q), p(.)\rangle \; + \alpha(\beta\lambda_1 + (1-\beta)\lambda_2) \int_0^\infty \bar{F}(y+q)p(y)dy \\
&\qquad + \alpha Z_{n-1}^{\beta\lambda_1+(1-\beta)\lambda_2}(\rho(q,p)) \Bigg\} \\
&\geq \inf_q \Bigg\{ \beta\langle c(.,q), p(.)\rangle \; + \alpha\beta\lambda_1 \int_0^\infty \bar{F}(y+q)p(y)dy + \alpha\beta Z_{n-1}^{\lambda_1}(\rho(q,p)) \\
&\qquad + (1-\beta)\langle c(.,q), p(.)\rangle \; + \alpha(1-\beta)\lambda_2 \int_0^\infty \bar{F}(y+q)p(y)dy \\
&\qquad + \alpha(1-\beta)Z_{n-1}^{\lambda_2}(\rho(q,p)) \Bigg\} \\
&\geq \beta Z_n^{\lambda_1}(p) + (1-\beta)Z_n^{\lambda_2}(p).
\end{aligned}$$

Since T^λ preserves the concavity and Z_n^λ converge to Z^λ, we can conclude that $Z^\lambda(p)$ is concave in λ for every $p \in \mathcal{H}^+$.

We repeat the steps above with $g(\lambda)$ to see the following:

$$\begin{aligned}
&g(\beta\lambda_1 + (1-\beta)\lambda_2)\\
&\quad = \inf_q \Big\{ c(0,q) \ + \alpha(\beta\lambda_1 + (1-\beta)\lambda_2)\bar{F}(q) + \alpha Z^{\beta\lambda_1+(1-\beta)\lambda_2}(\rho(q,\delta)) \Big\}\\
&\quad \geq \inf_q \big\{ \beta c(0,q) \ + \alpha\beta\lambda_1\bar{F}(q) + \alpha\beta Z^{\lambda_1}(\rho(q,\delta))\\
&\qquad\qquad + (1-\beta)c(0,q) \ + \alpha(1-\beta)\lambda_2\bar{F}(q)p(y)dy + \alpha(1-\beta)Z^{\lambda_2}(\rho(q,\delta)) \big\}\\
&\quad \geq \beta g(\lambda_1) + (1-\beta)g(\lambda_2).
\end{aligned}$$

Hence, $g(\lambda)$ inherits concavity from $Z^\lambda(p)$ for every $p \in \mathcal{H}^+$.

By setting $q = 0$ in (62) and recalling that $Z^\lambda(\rho(0,\delta)) = Z^\lambda(0) = 0$, we obtain

$$g(\lambda) \leq c(0,0) + \alpha\lambda, \tag{63}$$

which shows that the function $g(\lambda)$ has a rate of growth of at most α. Furthermore, ignoring the last two terms in (62) and using (38), we get

$$g(0) \geq \inf_q c(0,q) \geq c_2 > 0. \tag{64}$$

We already know that there exists a solution to (60). Since $g(0) > 0$, and $g(\lambda)$ is concave and has a growth rate of at most α, the equation $g(\lambda) = \lambda$ has a unique solution v. Thus, setting $Z = Z^v$, we obtain a unique pair (v, Z) which solves (35)–(36). □

5. Abridged optimal control. Let $\hat{q}^k$ be the abridged feedback control defined as

$$\hat{q}^k = \{\hat{q}_1(\pi_1), \hat{q}_2(\pi_2), \dots, \hat{q}_k(\pi_k), \hat{q}^\infty, \dots, \hat{q}^\infty, \dots\}.$$

It is important to note that $\hat{q}^k$ is defined by k functions mapping $\mathcal{H}^+$ to $\Re$ and a scalar $\hat{q}^\infty$. After period k, the same scalar $\hat{q}^\infty$ is applied repeatedly without regard to the current state, so $\hat{q}_t$ is adapted to $\mathcal{Z}_t$ only for $t \leq k$.

Consider the *abridged* monotone iterative process starting with

$$\begin{aligned}
v_0 &= c(0,0) + \alpha v_0,\\
Z_0(p) &= \langle c(.,0), p(.)\rangle \ + \alpha v_0 \int_0^\infty \bar{F}(y)p(y)dy + \alpha Z_0(\rho(0,p))
\end{aligned}$$

by iterating with the feedback $\hat{q}_n(p)$ for $n \leq k$, which solves

$$\left.\begin{aligned}
v_{n+1} &= \inf_q \big\{ c(0,q) + \alpha v_n \bar{F}(q) + \alpha Z_n(\rho(q,\delta)) \big\}\\
Z_{n+1}(p) &= \inf_q \big\{ \langle c(.,q), p(.)\rangle \ + \alpha v_n \textstyle\int \bar{F}(y+q)p(y)dy + \alpha Z_n(\rho(q,p)) \big\}
\end{aligned}\right\} \quad \text{for } n \leq k. \tag{65}$$

Afterwards, the monotone iterative process applies the scalar $\hat{q}^\infty$, so we have

$$\left.\begin{aligned}
v_{n+1} &= \big\{ c(0,\hat{q}^\infty) + \alpha v_n \bar{F}(\hat{q}^\infty) + \alpha Z_n(\rho(\hat{q}^\infty,\delta)) \big\}\\
Z_{n+1}(p) &= \big\{ \langle c(.,\hat{q}^\infty), p(.)\rangle \ + \alpha v_n \textstyle\int \bar{F}(y+\hat{q}^\infty)p(y)dy + \alpha Z_n(\rho(\hat{q}^\infty,p)) \big\}
\end{aligned}\right\} \quad \text{for } n > k.$$

One can improve the value functions by choosing $\hat{q}^\infty$ as good as possible. For example, take

$$\hat{q}^\infty = \arg\inf_q \mathrm{E}\sum_{t=1}^{\infty} \alpha^t c(I_t, q) \text{ with } I_1 = I_{k+1}.$$

The scalar $\hat{q}^\infty$ exists because q can be bounded in the minimization of the sum of continuous functions $c(.,q)$ above.

The scalar $\hat{q}^\infty$ can be found by an alternative procedure. For a fixed q, define (v_q, Z_q) by

$$Z_q(p) = \langle c(.,q), p(.)\rangle + \alpha v_q \int_0^\infty \bar{F}(y+q)p(y)dy + \alpha Z_q(\rho(q,p)),$$
$$v_q = c(0,q) + \alpha v_q \bar{F}(q) + \alpha Z_q(\rho(q,\delta)).$$

Once the functional Z_q is obtained, we choose $\hat{q}^\infty$ as

$$\hat{q}^\infty = \arg\inf_q Z_q(\pi_{k+1}).$$

The abridged iterative process preserves the continuity of Z_n as established next.

LEMMA 10. *If Z_n is continuous over $\mathcal{H}^+ \setminus \{0\}$, then Z_{n+1} is continuous over the same set.*

Proof. Indeed if Z_n is continuous, then the function

$$G(q,p;v_n,Z_n) = \langle c(.,q), p(.)\rangle + \alpha v_n \int_0^\infty \bar{F}(y+q)p(y)dy + \alpha Z_n(\rho(q,p))$$

is continuous in $\Re^+ \times \mathcal{H}^+$. In view of (50), we can moreover assume that the order quantity q is bounded above by

$$q \le \frac{a_0}{c(1-\alpha)}\left\{1 + \frac{\int_0^\infty xp(x)dx}{\int_0^\infty p(x)dx}\right\}.$$

This bound depends on p and it is well defined for $p \in \mathcal{H}^+ \setminus \{0\}$. Then there exist $\hat{q}_n(p)$ and $\hat{q}_n$ achieving the infima such that

$$v_{n+1} = c(0,\hat{q}_n) + \alpha v_n \bar{F}(\hat{q}_n) + \alpha Z_n(\rho(\hat{q}_n,\delta)),$$
$$Z_{n+1}(p) = \langle c(.,\hat{q}_n(p)), p(.)\rangle + \alpha v_n \int_0^\infty \bar{F}(y+\hat{q}_n(p))p(y)dy + \alpha Z_n(\rho(\hat{q}_n(p),p)).$$

Hence, Z_{n+1} is continuous in $\hat{q}_n(p)$.

To complete the proof, we need to show continuity in p. Suppose that $p_k \to p$ in $\mathcal{H}^+$. For an arbitrary q,

$$Z_{n+1}(p_k) \le \langle c(.,q), p_k(.)\rangle + \alpha v_n \int_0^\infty \bar{F}(y+q)p_k(y)dy + \alpha Z_n(\rho(q,p_k)).$$

Take the lim sup of both sides and note that the right-hand side is continuous in p_k to obtain

$$\limsup_{k\to\infty} Z_{n+1}(p_k) \le \limsup_{k\to\infty}\left\{\langle c(.,q), p_k(.)\rangle + \alpha v_n \int_0^\infty \bar{F}(y+q)p_k(y)dy + \alpha Z_n(\rho(q,p_k))\right\}$$
$$\le \langle c(.,q), p(.)\rangle + \alpha v_n \int_0^\infty \bar{F}(y+q)p(y)dy + \alpha Z_n(\rho(q,p)).$$

Inserting $q = \hat{q}_n(p)$ into the right-hand side gives

$$\limsup_{k\to\infty} Z_{n+1}(p_k) \le Z_{n+1}(p). \tag{66}$$

Next

$$Z_{n+1}(p_k) = \langle c(.,\hat{q}(p_k)), p_k(.)\rangle + \alpha v_n \int_0^\infty \bar{F}(y+\hat{q}(p_k))p_k(y)dy + \alpha Z_n(\rho(\hat{q}(p_k),p)).$$

Since $p_k \to p$, we can extract a subsequence p_{k_l} indexed by k_l such that $p_{k_l} \to p$ and $\lim_{l\to\infty} Z_{n+1}(p_{k_l}) = \liminf_k Z_{n+1}(p_k)$. Using the standard arguments and the fact that $\hat{q}_n(p_{k_l})$ remains bounded, we can guarantee that the subsequence $\hat{q}_n(p_{k_l})$ converges to a number, say $\hat{q}'$. By continuity of Z_n,

$$\begin{aligned}\lim_{l\to\infty} Z_{n+1}(p_{k_l}) &= \langle c(.,\hat{q}'), p(.)\rangle + \alpha v_n \int_0^\infty \bar{F}(y+\hat{q}')p_k(y)dy + \alpha Z_n(\rho(\hat{q}',p)) \\ &\ge Z_{n+1}(p),\end{aligned}$$

where the inequality is due to the fact that $\hat{q}'$ may differ from $\hat{q}_n(p)$. Using the inequality along with the definition of the subsequence p_{k_l}, we have

$$\liminf_{k\to\infty} Z_{n+1}(p_k) = \lim_{l\to\infty} Z_{n+1}(p_{k_l}) \ge Z_{n+1}(p). \tag{67}$$

Combining (66) and (67), we obtain that $\limsup_k Z_{n+1}(p_k) = \liminf_k Z_{n+1}(p_k)$, so that Z_{n+1} is continuous. □

In this section, we have introduced an abridged optimal control for our inventory problem. The choice of k for the abridged control $\hat{q}^k$ is an important practical question. Note that v_n and $Z_{n+1}(p)$ decrease as k increases for any $p \in \mathcal{H}^+$. Moreover, these values cannot fall below the inventory cost of the corresponding fully observed system. The gap between these costs is an upper bound on the cost savings achievable by increasing k. A manager may use this bound to determine whether a given value of k is suitable. An alternative approach is to increase k until the reduction in the costs v_n and $Z_{n+1}(p)$ for any $p \in \mathcal{H}^+$ become less than a specified amount.

6. Conclusion and extensions. This paper has provided a rigorous treatment of inventory problems with partial observations. The observation process is a binary valued Markov chain, which arises from the zero-balance walk approach to inventory management. Since the inventory level is often not observed, its distribution is used to represent the state of the system. This approach immediately results in a dynamic program in a functional space. The dynamic programming equations are simplified by using unnormalized probabilities. By doing so, a Zakai-type system of equations are derived for our inventory problem. This simplification has allowed us to prove the existence of a value function under various assumptions. We also established the uniqueness of the function in some cases.

This paper is part of a greater effort [6] that aims to study inventory problems with partial observations. As such, it represents an important step in the study of inventory problems with partial observations. It treats a specific binary observation process. We plan to investigate the effect of other observation processes on inventory policies.

Search for a sufficient statistic. An interesting area of research is to classify those partial observation processes for which a sufficient statistic exists. While it

is not so in the general case, such statistics exist; see [7, 8], for example. When a sufficient statistic exists, the system becomes finite-dimensional, and the analysis can be carried much further.

Approximation algorithms. When a sufficient statistic does not exist, which is often the case, the analysis in this paper becomes very relevant. Furthermore, our analysis directly shows how successive approximations can lead to an optimal solution. In devising approximate optimal solutions and heuristic procedures, one can also benefit from the books [19, 20].

Finer interval observations. We have supposed in this paper that only events $[I_t = 0]$ and $[I_t > 0]$ are observed. This corresponds to observing if the inventory falls into the intervals $[0, 0]$ or $(0, \infty)$. When there is a finite storage capacity a, the inventory can take values in the interval $[0, a]$. This interval can be partitioned into $N + 2$ disjoint intervals, namely,

$$\mathcal{I}_0 := [a_0, a_0],\ \mathcal{I}_1 := (a_0, a_1],\ \mathcal{I}_2 := (a_1, a_2], \ldots,\ \mathcal{I}_N := (a_{N-1}, a_N),\ \mathcal{I}_{N+1} := [a_N, a_N]$$

for $0 = a_0 < a_1 < \cdots < a_{N-1} < a_N = a$. The signals associated with the observation of inventory in these intervals will be

$$z_t = n \quad \text{if } I_t \in \mathcal{I}_n \quad \text{for } 0 \le n \le N + 1.$$

Clearly this new signal process provides more information than the binary signal we have studied. However, we expect that the Bellman equations can still be linearized by using unnormalized probabilities.

In practice, the interval observations as defined above would happen when the inventory is stored in modules, e.g., bins, shelves, or different locations. The IM can see empty and full bins by simply walking in the storage area. In a typical case, the bins may be prioritized in such a way that the items in bin i are not used until the items in bin $i + 1$ are finished. Then a_1 would be the first bin's capacity, a_2 would be the first and second bins' cumulative capacity, etc. If three bins are full, the fourth is semifull, and the others are empty, the IM would conclude that $I_t \in \mathcal{I}_4 = (a_3, a_4]$ and observe the signal $z_t = 4$.

Stock-taking and inspections. Perpetual growth in the uncertainty of the actual inventories can lead to poor decisions and therefore higher costs. In practice, actual inventories are counted and/or inspected often periodically to reduce uncertainty [2]. Since stock-counting and inspections may require testing and disrupt other activities, they are expensive. Consequently, the inventory is not counted or inspected every period. Also, only a part of the inventory (partial inspection) may be examined to reduce the uncertainty to a reasonable level. The counting and inspection costs typically have both fixed and proportional components. The optimal timing and the amount of inspections, whose study would require the use of quasi-variational inequalities [4], is an interesting topic for future research.

Acknowledgments. We thank Guillermo Gallego for his constructive input during the development phase of our problem. The authors thank J. Adolfo Minjarés-Sosa, the anonymous referees, and the editor for their suggestions, which have improved the exposition of the paper.

REFERENCES

[1] K. J. Arrow, S. Karlin, and H. Scarf, *Studies in the Mathematical Theory of Inventory and Production*, Stanford University Press, Stanford, CA, 1958.

[2] A. Atali, H. Lee, and Ö. Özer, *If the Inventory Manager Knew: Value of RFID under Imperfect Inventory Information*, Working paper, Dept. of Management Science and Engineering, Stanford University, Stanford, CA, 2005.
[3] S. Axsäter, *Inventory Control*, Kluwer Academic Publishers, Norwell, MA, 2001.
[4] A. Bensoussan and J.-L. Lions, *Impulse Control and Quasi-Variational Inequalities*, Trans-Inter-Scientia, Kent, England, 1984.
[5] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.
[6] A. Bensoussan, M. Çakanyildirim, and S. P. Sethi, *On the optimal control of partially observed inventory systems*, C. R. Math. Acad. Sci. Paris, 341 (2005), pp. 419–426.
[7] A. Bensoussan, M. Çakanyildirim, and S. P. Sethi, *Optimal ordering policies for inventory problems with dynamic information delays*, Production and Operations Management, to appear.
[8] A. Bensoussan, M. Çakanyildirim, and S. P. Sethi, *Optimality of base stock and* (s, S) *policies for inventory problems with information delays*, J. Optim. Theory Appl., 130 (2006), pp. 153–172.
[9] A. Bensoussan, M. Çakanyildirim, and S. P. Sethi, *Economic evaluation of systems that expedite inventory information*, Production and Operations Management, to appear.
[10] D. Beyer, S. P. Sethi, and M. I. Taksar, *Inventory models with Markovian demands and cost functions of polynomial growth*, J. Optim. Theory Appl., 98 (1998), pp. 281–323.
[11] D. Beyer, F. Cheng, S. P. Sethi, and M. I. Taksar, *Markovian Demand Inventory Models*, Springer-Verlag, New York, 2007, in press.
[12] D. Bolger, *Managing inventory theft*, Furniture World Magazine, October (2002); available online at http://www.furninfo.com/absolutenm/templates/Article_Retailing.asp?articleid=2205&zoneid=7.
[13] S. Chopra and P. Meindl, *Supply Chain Management*, 2nd ed., Prentice–Hall, Upper Saddle River, NJ, 2004.
[14] X. Ding, M. L. Puterman, and A. Bisi, *The censored newsvendor and the optimal acquisition of information*, Oper. Res., 50 (2002), pp. 517–527.
[15] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*, Springer-Verlag, New York, 1995.
[16] M. L. Fisher, A. Raman, and A. S. McClelland, *Rocket science retailing is almost here: Are you ready?*, Harvard Business Review, 78 (2000), pp. 115–124.
[17] F. W. Harris, *Operations and Cost*, Factory Management Series, A.-W. Shaw, Chicago, 1913.
[18] H. J. Kushner, *Dynamic equations for nonlinear filtering*, J. Differential Equations, 3 (1967), pp. 179–190.
[19] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed., Springer-Verlag, New York, 2001.
[20] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Springer-Verlag, New York, 2003.
[21] M. A. Lariviere and E. L. Porteus, *Stalking information: Bayesian inventory management with unobserved lost sales*, Management Sci., 45 (1999), pp. 346–363.
[22] X. Lu, J. Song, and K. Zhu, *On "The censored newsvendor and the optimal acquisition of information,"* Oper. Res., 53 (2005), pp. 1024–1026.
[23] *Monitor workers to stop internal product theft*, Refrigerated Transporter, July (2001); available online at http://refrigeratedtrans.com/mag/transportation_monitor_workers_stop/index.html.
[24] S. Nahmias, *On ordering perishable inventory when both demand and lifetime are random*, Management Sci., 24 (1977), pp. 82–90.
[25] S. Nahmias, *Perishable inventory theory: A review*, Oper. Res., 30 (1982), pp. 680–708.
[26] A. Raman, N. DeHoratius, and Z. Ton, *Execution: The missing link in retail operations*, California Management Review, 43 (2001), pp. 136–152.
[27] J. T. Treharne and C. R. Sox, *Adaptive inventory control for nonstationary demand and partial observation*, Management Sci., 48 (2002), pp. 607–624.
[28] F. Wu, F. Kuo, and L. Liu, *The application of RFID on drug safety of inpatient nursing healthcare*, in Proceedings of the 7th International Conference on Electronic Commerce, ACM International Conference Proceeding Series 113, ACM, New York, 2005, pp. 85–92.
[29] C. A. Yano and H. L. Lee, *Lot sizing with random yields: A review*, Oper. Res., 43 (1995), pp. 311–333.
[30] M. Zakai, *On the optimal filtering of diffusion processes*, Z. Wahrsch. Verw. Gebiete, 11 (1969), pp. 230–243.
[31] P. H. Zipkin, *Foundations of Inventory Management*, McGraw–Hill, New York, 2000.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 210–230

TRACKING WITH PRESCRIBED TRANSIENT BEHAVIOR FOR NONLINEAR SYSTEMS OF KNOWN RELATIVE DEGREE*

ACHIM ILCHMANN†, EUGENE P. RYAN‡, AND PHILIP TOWNSEND‡

Abstract. Tracking of a reference signal (assumed bounded with essentially bounded derivative) is considered in the context of a class Σ_ρ of multi-input, multi-output dynamical systems, modelled by functional differential equations, affine in the control and satisfying the following structural assumptions: (i) arbitrary—but known—relative degree $\rho \geq 1$; (ii) the "high-frequency gain" is sign definite—but possibly of unknown sign. The class encompasses a wide variety of nonlinear and infinite-dimensional systems and contains (as a prototype subclass) all finite-dimensional, linear, m-input, m-output, minimum-phase systems of known strict relative degree. The first control objective is tracking, by the output y, with prescribed accuracy: given $\lambda > 0$ (arbitrarily small), determine a feedback strategy which ensures that, for every reference signal r and every system of class Σ_ρ, the tracking error $e = y - r$ is ultimately bounded by λ (that is, $\|e(t)\| < \lambda$ for all t sufficiently large). The second objective is guaranteed output transient performance: the tracking error is required to evolve within a prescribed performance funnel $\mathcal{F}_\varphi$ (determined by a function φ). Both objectives are achieved using a filter in conjunction with a feedback function of the tracking error, the filter states, and the funnel parameter φ.

Key words. output feedback, nonlinear systems, functional differential equations, transient behavior, tracking, high relative degree

AMS subject classifications. 93D15, 93C30, 34K20

DOI. 10.1137/050641946

1. Introduction.

In [5], a class of infinite-dimensional, m-input ($u(t) \in \mathbb{R}^m$), m-output ($y(t) \in \mathbb{R}^m$), nonlinear systems (with finite memory) given by a controlled functional differential equation of the form $\dot{y}(t) = g(p(t), (Ty)(t), u(t))$ is considered, where g is a continuous function, p represents a bounded disturbance, and T is a causal operator with a bounded-input bounded-output property: an output feedback control structure is developed which ensures approximate asymptotic tracking, with prescribed transient behavior, of any absolutely continuous bounded reference signal with essentially bounded derivative. Here we extend these investigations to incorporate higher-order systems, affine in the control, of the form

$$y^{(\rho)}(t) = R_1\, y(t) + R_2\, y^{(1)}(t) + \cdots + R_\rho\, y^{(\rho-1)}(t) + g(p(t), (Ty)(t)) + \Gamma u(t), \tag{1.1}$$

where $\rho \in \mathbb{N}$ is known, $y^{(i)}$ denotes the ith derivative of y, and the matrix Γ is assumed to be sign definite (equivalently, $\langle v, \Gamma v \rangle = 0 \;\Leftrightarrow\; v = 0$).

In an early contribution by Miller and Davison [12], the attainment of prescribed transient behavior is considered for a class of single-input, single-output, linear, minimum-phase systems with known high-frequency gain: a controller is introduced which guarantees the "error to be less than an (arbitrarily small) prespecified constant

*Received by the editors October 5, 2005; accepted for publication (in revised form) September 25, 2006; published electronically April 6, 2007. This research was based on work supported in part by the UK Engineering & Physical Sciences Research Council (GR/S94582/01).
http://www.siam.org/journals/sicon/46-1/64194.html

†Institute of Mathematics, Technical University Ilmenau, Weimarer Straße 25, 98693 Ilmenau, DE (achim.ilchmann@tu-ilmenau.de).

‡Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK (epr@maths.bath.ac.uk, p.townsend@bath.ac.uk).

after an (arbitrarily small) prespecified period of time, with an (arbitrarily small) prespecified upper bound on the amount of overshoot." However, the controller is adaptive with nondecreasing gain k, invokes a piecewise-constant switching strategy, and is less flexible in its scope for shaping transient behavior (in particular, an a priori bound on the initial data is required) when compared to the nonadaptive approach in [6].

The results of this paper generalize the main ideas in [6], where tracking with prescribed transient behavior is considered in a more restricted context of linear systems of known relative degree, subject to "mild" nonlinear perturbations: the generality of the operator T in (1.1) allows for a considerable diversity of nonlinear and infinite-dimensional effects, including delays and hysteresis phenomena. We implement a "backstepping" procedure in conjunction with a filter/precompensator in the construction of a nonadaptive controller. The backstepping procedure is akin to that of [17, 9, 12].

We briefly digress to review the literature on tracking and stabilization of high relative degree systems. Unless otherwise stated, all results relate to single-input, single-output systems. Bullinger and Allgöwer [1] introduce a high-gain observer in conjunction with an adaptive controller to achieve tracking with prescribed asymptotic accuracy $\lambda > 0$ (λ-tracking). This is achieved for a class of systems which are affine in the control, of known relative degree, and with affine linearly bounded drift term. Paper [17] considers linear minimum-phase systems with nonlinear perturbation; the control objective is (continuous) adaptive λ-tracking with nondecreasing gain. The class of allowable nonlinearities is considerably smaller than that of the present paper. Stabilization for systems of maximum relative degree in the so-called parametric strict feedback form is achieved in [18] via a piecewise constant adaptive switching strategy. Both these contributions use a backstepping procedure. Nonadaptive contributions are found in the work by Byrnes and Isidori [2] with extensions in [3]. They cover stabilization and tracking for a class of relative-degree-one nonlinear systems, with an exosystem, the positive orbits of which lie in a compact set: systems of higher relative degree are also considered (see in particular [2, eq. (33)]), and the authors state (without proof) that "these systems can [be] reduced to systems of (relative degree 1) by means of the semiglobal back-stepping Lemma." The main result in [2, Prop. 7.1] pertains to practical tracking and applies high-gain principles in conjunction with an internal model: the multilayered nature of the assumptions determining the system class makes it difficult to assess the overlap with the class considered in the present paper. Related investigations, based on high-gain properties and/or an internal model principle, can be found in [10, 13, 9]: we will have occasion to comment further on the latter in section 3.1.3 below.

The paper is organized as follows. Sections 2 and 3 introduce the control objectives and the system class: section 3.1 highlights several particular subclasses. In section 4, the control and feedback laws are constructed: an existence theorem for the resulting closed-loop system is provided in section 4.3. Our main results on transient and asymptotic behavior of the closed-loop are given in section 5 and illustrated in an example in section 6. All proofs are relegated to the appendix.

We close this introduction with remarks on notation. Throughout, $\mathbb{R}_+ := [0, \infty)$ and $\mathbb{C}_-$ denotes the open left half complex plane $\{\lambda \in \mathbb{C} |\ \operatorname{Re}\lambda < 0\}$. The Euclidean inner product and induced norm on $\mathbb{R}^n$ are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively. The open ball of radius $\delta > 0$ centered at $x \in \mathbb{R}^n$ is denoted by $\mathbb{B}_\delta(x)$. For an interval $I \subset \mathbb{R}$, $C(I, \mathbb{R}^n)$ is the space of continuous functions $I \to \mathbb{R}^n$, $L^\infty(I, \mathbb{R}^n)$

is the space of essentially bounded measurable functions $x\colon I \to \mathbb{R}^n$ with norm $\|x\|_\infty := \text{ess-sup}_{t\in I}\|x(t)\|$, $L^1(I,\mathbb{R}^n)$ is the space of integrable functions $x\colon I \to \mathbb{R}^n$ with norm $\|x\|_1 := \int_I \|x(t)\|\mathrm{d}t < \infty$, $L^\infty_{\text{loc}}(I,\mathbb{R}^n)$ (respectively, $L^1_{\text{loc}}(I,\mathbb{R}^n)$) is the space of measurable, locally essentially bounded (respectively, locally integrable) functions $I \to \mathbb{R}^n$, and $W^{1,\infty}(I,\mathbb{R}^n)$ is the space of absolutely continuous functions $x\colon I \to \mathbb{R}^n$ with $x, \dot{x} \in L^\infty(I,\mathbb{R}^n)$. The spectrum of $A \in \mathbb{R}^{n\times n}$ is denoted by $\text{spec}(A)$.

2. Control objectives and the performance funnel. There are two control objectives: (i) approximate tracking, by the output, of reference signals $r \in \mathcal{R} := W^{1,\infty}(\mathbb{R}_+,\mathbb{R}^m)$ (in particular, for arbitrary $\lambda > 0$, we seek an output feedback strategy which ensures that, for every $r \in \mathcal{R}$, the closed-loop system has bounded solution and the tracking error $e(t) = y(t) - r(t)$ is ultimately bounded by λ (that is, $\|e(t)\| \le \lambda$ for all t sufficiently large)) and (ii) prescribed transient behavior of the tracking error.

Both objectives are captured in the concept of a performance funnel

$$\mathcal{F}_\varphi := \big\{(t,e) \in \mathbb{R}_+ \times \mathbb{R}^m \big|\ \varphi(t)\|e\| < 1\big\}$$

associated with a function φ of the following class:

$$\Phi := \big\{\varphi \in W^{1,\infty}(\mathbb{R}_+,\mathbb{R}) \ \big|\ \ \varphi(0) = 0, \quad \varphi(s) > 0 \quad \forall\, s > 0 \text{ and } \liminf_{s\to\infty} \varphi(s) > 0\big\}.$$

The aim is an output feedback strategy ensuring that, for every reference signal $r \in \mathcal{R}$, the tracking error $e = y - r$ evolves within the funnel $\mathcal{F}_\varphi$; see Figure 1. For example, if $\liminf_{t\to\infty} \varphi(t) \ge 1/\lambda$, then evolution within the funnel ensures that the first control objective is achieved. If φ is chosen as the function $t \mapsto \min\{t/\tau, 1\}/\lambda$, then evolution within the funnel ensures that the prescribed tracking accuracy $\lambda > 0$ is achieved within the prescribed time $\tau > 0$. The feedback structure incorporates a filter and essentially exploits an intrinsic high-gain property of the system/filter interconnection to ensure that, if $(t, e(t))$ approaches the funnel boundary, then an appropriately generated gain attains values sufficiently large to preclude boundary contact.

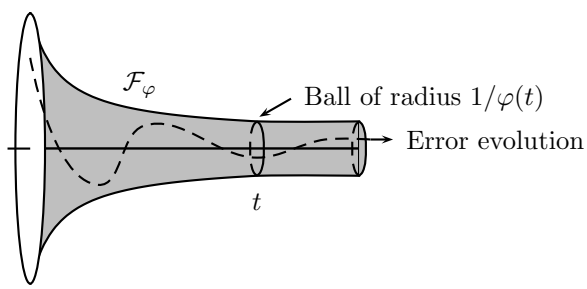


FIG. 1. *Prescribed performance funnel* $\mathcal{F}_\varphi$.

3. Class of systems. We subsume (1.1) in the following:

$$\left\{\begin{array}{l} \dot{x}(t) = Ax(t) + f(p(t),(Ty)(t),x(t)) + Bu(t), \\ y(t) = Cx(t), \\ x|_{[-h,0]} = x^0 \in C([-h,0],\mathbb{R}^{\rho m}), \end{array}\right. \tag{3.1}$$

$$A = \begin{bmatrix} 0 & I & 0 & \cdots & 0 \\ 0 & 0 & I & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & I \\ R_1 & R_2 & \cdots & R_{\rho-1} & R_\rho \end{bmatrix} \in \mathbb{R}^{\rho m \times \rho m}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \Gamma \end{bmatrix} \in \mathbb{R}^{\rho m \times m}, \tag{3.2}$$

$$C = \begin{bmatrix} I & 0 & \cdots & 0 & 0 \end{bmatrix} \in \mathbb{R}^{m \times \rho m}, \quad f\colon \mathbb{R}^m \times \mathbb{R}^q \times \mathbb{R}^{\rho m} \to \mathbb{R}^{\rho m} \text{ continuous.} \tag{3.3}$$

Observe that $\Gamma = CA^{\rho-1}B$. In the special case wherein f is given by

$$f(p, w, x) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ g(p, w) \end{bmatrix}, \tag{3.4}$$

it is clear that (1.1) and (3.1) are equivalent. Next, we define the class of operators T allowable in (3.1).

DEFINITION 3.1 (operator class $\mathcal{T}_h$). *Let $h \geq 0$. An operator T is said to be of class $\mathcal{T}_h$ if, and only if, for some $l, q \in \mathbb{N}$, the following hold:*

(i) $T\colon C([-h, \infty), \mathbb{R}^l) \to L^\infty_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^q)$.

(ii) *For every $\delta > 0$, there exists $\Delta > 0$ such that, for all $\zeta \in C([-h, \infty), \mathbb{R}^l)$,*

$$\sup_{t \in [-h, \infty)} \|\zeta(t)\| \leq \delta \quad \Longrightarrow \quad \|(T\zeta)(t)\| \leq \Delta \quad \text{for almost all } t \geq 0\,.$$

(iii) *For all $t \in \mathbb{R}_+$, the following hold:*

(a) *for all $\zeta, \psi \in C([-h, \infty), \mathbb{R}^l)$,*

$$\zeta(\cdot) \equiv \psi(\cdot) \text{ on } [-h, t] \quad \Longrightarrow \quad (T\zeta)(s) = (T\psi)(s) \quad \text{for almost all } s \in [0, t];$$

(b) *for all continuous functions $\beta\colon [-h, t] \to \mathbb{R}^l$, there exist $\tau, \delta, c > 0$ such that, for all $\zeta, \psi \in C([-h, \infty), \mathbb{R}^l)$ with $\zeta|_{[-h,t]} = \beta = \psi|_{[-h,t]}$ and $\zeta(s), \psi(s) \in \mathbb{B}_\delta(\beta(t))$ for all $s \in [t, t + \tau]$,*

$$\operatorname{ess-sup}_{s \in [t, t+\tau]} \|(T\zeta)(s) - (T\psi)(s)\| \;\leq\; c \sup_{s \in [t, t+\tau]} \|\zeta(s) - \psi(s)\|\,.$$

Remark 3.2. Property (ii) is a bounded-input, bounded-output assumption on the operator T. Property (iii)(a) is a natural assumption of causality. Property (iii)(b) is a technical assumption of local Lipschitz type which is used in establishing well-posedness of the closed-loop system (defined later in section 4.3).

We are now in a position to make precise the system class.

DEFINITION 3.3 (system class Σ_ρ). *For $\rho \in \mathbb{N}$, Σ_ρ is the class of m-input, m-output systems (A, B, C, f, p, T, h) of the form (3.1), where $h \geq 0$ quantifies the memory of the system, and A, B, and C are structured as in (3.2)–(3.3) and satisfy*

(A1) *sign-definite high-frequency gain: $\Gamma = CA^{\rho-1}B$ is either positive definite or negative definite (equivalently, $\langle v, \Gamma v \rangle = 0 \;\Leftrightarrow\; v = 0$).*

The functions f, p and operator T are such that

(A2) $p \in L^\infty(\mathbb{R}_+, \mathbb{R}^m)$,

(A3) *for some $q \in \mathbb{N}$, $T\colon C([-h, \infty), \mathbb{R}^m) \to L^\infty_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^q)$ is of class $\mathcal{T}_h$,*

(A4) *$f\colon \mathbb{R}^m \times \mathbb{R}^q \times \mathbb{R}^{\rho m} \to \mathbb{R}^{\rho m}$ is continuous and, for all nonempty compact sets $P \subset \mathbb{R}^m$, $W \subset \mathbb{R}^q$, and $Y \subset \mathbb{R}^m$, there exists a constant $c_0 = c_0(P, W, Y) > 0$ such that $\|f(p, w, x)\| \leq c_0$ for all $(p, w, x) \in P \times W \times \{v \in \mathbb{R}^{\rho m} \mid Cv \in Y\}$.*

Remark 3.4. (i) Due to the presence of the nonlinear function f, the (vector) relative degree of (3.1) at some point $x^0 \in \mathbb{R}^{\rho m}$ may not be defined; see [7, pp. 137 and 220]. However, if $f \equiv 0$, then it follows from assumption (A1) that the vector relative degree of the linear system (3.1) is $(\rho, \dots, \rho) \in \mathbb{R}^m$ at each point $x^0 \in \mathbb{R}^{\rho m}$ and, in particular,

$$CA^iB = 0 \quad \text{for } i = 1, \dots, \rho - 2 \text{ and } \Gamma = CA^{\rho-1}B \text{ is invertible.} \tag{3.5}$$

The linear system (A, B, C) is said to have *strict relative degree* ρ if, and only if, (3.5) holds. Note that assumption (A1) requires the strengthened assumption that $CA^{\rho-1}B$ is either positive definite or negative definite. In the multi-input, multi-output case, (A1) is rather restrictive. By contrast, in the single-input, single-output case, the assumption of sign definiteness is redundant and (A1) is simply equivalent to positing that the relative degree of the linear triple (A, B, C) is known.

(ii) Recall that a linear system (A, B, C) is said to be *minimum phase* if, and only if,

$$\det \begin{bmatrix} sI - A & B \\ C & 0 \end{bmatrix} \neq 0 \quad \forall\, s \in \mathbb{C} \text{ with } \mathrm{Re}(s) \geq 0. \tag{3.6}$$

Due to the structure of the matrices A, B, and C in (3.2)–(3.3) and assumption (A1), (A, B, C) is minimum phase.

(iii) Assumption (A4) constrains the nature of the dependence of f on its third argument: in particular, for compact sets P, W, and Y, it posits boundedness of f on $P \times W \times \{v \in \mathbb{R}^{\rho m} \mid Cv \in Y\}$. For example, (A4) holds if there exists a continuous function $\pi : \mathbb{R}^m \times \mathbb{R}^q \times \mathbb{R}^m \to \mathbb{R}_+$ such that $\|f(p, w, x)\| \leq \pi(p, w, Cx)$ for all (p, w, x). Assumption (A4) plays a crucial role in the later analysis: in its absence (i.e., if f is merely assumed to be continuous), it is not difficult to construct examples for which the performance objectives cannot be achieved (indeed, finite escape times can occur).

(iv) With reference to Figure 2, the system (3.1) can be thought of as the interconnection of two blocks. The dynamical system represented by block Λ_1, which can be influenced directly by the system control u, is also driven by the output w from the dynamic block Λ_2, as shown in Figure 2. The block Λ_2 can be considered as a causal operator mapping the system output y to w (an internal quantity, unavailable for feedback purposes); it allows for infinite-dimensional (e.g., delays, diffusions) and hysteresis (e.g., backlash) effects, some examples of which are given in section 3.1.

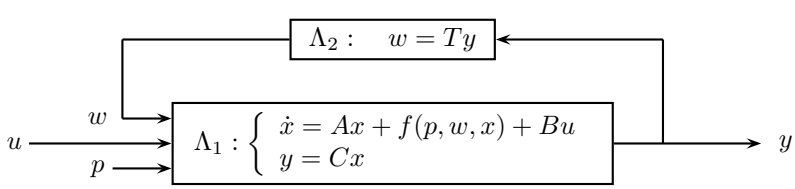


FIG. 2. *System of class* Σ_ρ.

3.1. Subclasses of Σ_ρ.

3.1.1. Finite-dimensional linear prototype. For motivational purposes, we first examine a prototype linear system and show that all finite-dimensional linear

systems of this form are incorporated into the class Σ_ρ. Consider an m-input, m-output linear system of the form

$$\dot{w}(t) = \tilde{A}\, w(t) + \tilde{B}\, u(t), \quad w(0) = w^0 \in \mathbb{R}^n, \quad y(t) = \tilde{C}\, w(t), \tag{3.7}$$

with strict relative degree $\rho \geq 1$, $\tilde{A} \in \mathbb{R}^{n\times n}$, $\tilde{B} \in \mathbb{R}^{n\times m}$, $\tilde{C} \in \mathbb{R}^{m\times n}$, $n \geq \rho m$, and positive-definite or negative-definite high-frequency gain $\tilde{C}\tilde{A}^{\rho-1}\tilde{B}$. To show that the system (3.7) belongs to the class Σ_ρ, we present the following lemma, a proof of which can be found in the appendix.

LEMMA 3.5. *Consider a linear system of the form* (3.7) *with strict relative degree* $\rho \in \mathbb{N}$. *Define*

$$\mathcal{C} := \begin{bmatrix} \tilde{C} \\ \tilde{C}\tilde{A} \\ \vdots \\ \tilde{C}\tilde{A}^{\rho-1} \end{bmatrix} \in \mathbb{R}^{\rho m\times n}, \quad \mathcal{B} := [\tilde{B} \,\vdots\, \tilde{A}\tilde{B} \,\vdots\, \cdots \,\vdots\, \tilde{A}^{\rho-1}\tilde{B}] \in \mathbb{R}^{n\times \rho m}$$

and let $\mathcal{V} \in \mathbb{R}^{n\times(n-\rho m)}$ *be such that* $\operatorname{im} \mathcal{V} = \ker \mathcal{C}$. *Then*

(i) $\mathbb{R}^n = \ker \mathcal{C} \oplus \operatorname{im} \mathcal{B}$;

(ii) *the matrix*

$$\mathcal{U} = \begin{bmatrix} \mathcal{C} \\ \mathcal{N} \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad \textit{where} \quad \mathcal{N} = (\mathcal{V}^T\mathcal{V})^{-1}\mathcal{V}^T[I - \mathcal{B}(\mathcal{C}\mathcal{B})^{-1}\mathcal{C}] \in \mathbb{R}^{(n-\rho m)\times n},$$

is invertible, with inverse $\mathcal{U}^{-1} = [\mathcal{B}(\mathcal{C}\mathcal{B})^{-1} \,\vdots\, \mathcal{V}]$, *and the triple*

$$(\hat{A}\,,\, \hat{B}\,,\, \hat{C}) := \big(U\tilde{A}U^{-1},\, U\tilde{B}\,,\, \tilde{C}U^{-1}\big) \tag{3.8}$$

has the following structure (wherein I *and* 0 *denote the* $m \times m$ *identity matrix and zero matrix, respectively):*

(3.9)

$$\hat{A} = \begin{bmatrix} 0 & I & 0 & \cdots & 0 & 0 \\ 0 & 0 & I & & & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & I & 0 \\ R_1 & R_2 & \cdots & R_{\rho-1} & R_\rho & S \\ P & 0 & \cdots & 0 & 0 & Q \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \Gamma \\ 0 \end{bmatrix}, \quad \hat{C} = [I \,\vdots\, 0 \,\vdots\, \cdots \,\vdots\, 0 \,\vdots\, 0 \,\vdots\, 0\,],$$

with $[R_1 \,\vdots\, \cdots \,\vdots\, R_\rho \,\vdots\, S] = \tilde{C}\tilde{A}^\rho\mathcal{U}^{-1}$, $\Gamma = \tilde{C}\tilde{A}^{\rho-1}\tilde{B}$, $P = \mathcal{N}\tilde{A}^\rho\tilde{B}\Gamma^{-1}$, *and* $Q = \mathcal{N}\tilde{A}\mathcal{V}$;

(iii) *if the system* (3.7) *is minimum phase, then* $\operatorname{spec}(Q) \subset \mathbb{C}_-$.

We remark that, in the case $\rho = 1$, (3.9) is to be interpreted as

$$\hat{A} = \begin{bmatrix} R_1 & S \\ P & Q \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} \Gamma \\ 0 \end{bmatrix}, \quad \hat{C} = [I \,\vdots\, 0\,]. \tag{3.10}$$

Invoking the similarity transformation (3.8)–(3.9) and writing $x^0 := \mathcal{C}w^0$, $z^0 := \mathcal{N}w^0$, $x(t) := \mathcal{C}w(t)$, it is readily verified that system (3.7) is equivalent to

$$\dot{x}(t) = Ax(t) + f(p(t), (Ty)(t), x(t)) + Bu(t), \quad x(0) = x^0, \quad y(t) = Cx(t), \tag{3.11}$$

where A, B, and C are as in (3.2)–(3.3), $p\colon t \mapsto S(\exp Qt)z^0$, T is the linear operator given by

$$(Ty)(t) = S\left(\int_0^t \exp(Q(t-s))Py(s)\mathrm{d}s\right), \quad t \geq 0,$$

and the function f takes the special form (3.4) with $g : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ given by $g(p,w) := p + w$.

If (3.7) has sign-definite high-frequency gain, then $\tilde{C}\tilde{A}^{\rho-1}\tilde{B} = \Gamma = CA^{\rho-1}B$ is either positive definite or negative definite, and hence assumption (A1) is satisfied. If we assume that (3.7) has the minimum-phase property, then by Lemma 3.5 (iii), Q has spectrum in $\mathbb{C}_-$: it follows that $p \in L^\infty(\mathbb{R}_+, \mathbb{R}^m)$ and T belongs to the class of operators $\mathcal{T}_0$, and so assumptions (A2) and (A3) are satisfied. Assumption (A4) is trivially satisfied. Therefore, the system class Σ_ρ contains all m-input, m-output, finite-dimensional, linear, minimum-phase systems of strict relative degree ρ with sign-definite high-frequency gain.

3.1.2. Infinite-dimensional linear systems. The finite-dimensional class of systems of the form (3.8) can be extended to infinite dimensions by reinterpreting the operators Q, P, and S as the generating operators of a regular linear system (regular in the sense of [16]). In the infinite-dimensional setting, Q is assumed to be the generator of a strongly continuous semigroup $\mathbf{S} = (\mathbf{S}_t)_{t\in\mathbb{R}_+}$ of bounded linear operators and a Hilbert space X with norm $\|\cdot\|_X$. Let X_1 denote the space $\mathrm{dom}(Q)$ endowed with the graph norm and let X_{-1} denote the completion of X with respect to the norm $\|z\|_{-1} = \|(s_0 I - Q)^{-1}z\|_X$, where s_0 is any fixed element of the resolvent set of Q. Then P is assumed to be a bounded linear operator from $\mathbb{R}^m$ to X_{-1} and S is assumed to be a bounded linear operator from X_1 to $\mathbb{R}^m$. Assuming that the semigroup $\mathbf{S}$ is exponentially stable and that S extends to a bounded linear operator (again denoted by S) from X to $\mathbb{R}^m$, then the operator T given by

$$(Ty)(t) := S\left(\int_0^t \mathbf{S}_{t-s}Py(s)\,\mathrm{d}s\right)$$

is of class $\mathcal{T}_0$ (see [14] for details) and, writing $p(t) := S\,\mathbf{S}_t z^0$, we again arrive at the structure of (3.11).

3.1.3. Nonlinear systems. In [9, eq. (1)] the following class of systems is studied:

$$\left\{\begin{aligned} \dot{x}_1(t) &= x_2(t) + f_1(w(t), y(t)) \\ &\vdots \\ \dot{x}_{\rho-1}(t) &= x_\rho(t) + f_{\rho-1}(w(t), y(t)) \\ \dot{x}_\rho(t) &= \gamma\, u(t) + f_\rho(w(t), y(t)) \\ \dot{w}(t) &= q(w(t), y(t)) \\ y(t) &= x_1(t) \\ (x_1(0), \ldots, x_\rho(0), w(0)) &= (x_1^0, \ldots, x_\rho^0, w^0) \end{aligned}\right. \tag{3.12}$$

where $\gamma \in \mathbb{R}\setminus\{0\}$, $q\colon \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^p$, and $f_i\colon \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, \rho$, are locally Lipschitz functions. Denote, by T, the mapping $y \mapsto w$ induced by the subsystem $\dot{w} = q(w, y)$ with initial condition $w(0) = w^0$. Then (3.12) is equivalent to (3.1)

(with $h = 0$ and $m = 1$). Moreover, if we assume that the subsystem $\dot{w} = q(w, y)$ is input-to-state stable (ISS), then, as shown in [4, sect. 2.3], the operator T is of class $\mathcal{T}_0$, in which case system (3.12), interpreted in its equivalent form (3.1), is of class Σ_ρ.

We remark that, in [9, eq. (1)], an assumption of *integral* input-to-state stability (iISS) (strictly weaker than our assumption of ISS) is imposed on the subsystem $\dot{w} = q(w, y)$. In this respect, the full generality of the system class in [9] is not captured by the class considered in the present paper.

3.1.4. Nonlinear delay systems. Let functions $\mathcal{G}_i \colon \mathbb{R} \times \mathbb{R}^l \to \mathbb{R}^q : (t, \zeta) \mapsto \mathcal{G}_i(t, \zeta)$, $i = 0, \dots, n$, be measurable in t and locally Lipschitz in ζ uniformly with respect to t: precisely, (i) for each fixed ζ, $\mathcal{G}_i(\cdot, \zeta)$ is measurable and (ii) for every compact $\mathcal{K} \subset \mathbb{R}^l$ there exists a constant c such that

$$\|\mathcal{G}_i(t, \zeta) - \mathcal{G}_i(t, \psi)\| \le c \, \|\zeta - \psi\| \quad \text{for almost all } t \text{ and} \quad \forall \, \zeta, \psi \in \mathcal{K} \, .$$

For $i = 0, \dots, n$, let $h_i \in \mathbb{R}_+$ and define $h := \max_i h_i$. For $\zeta \in C([-h, \infty), \mathbb{R}^l)$, let

$$(T\zeta)(t) := \int_{-h_0}^{0} \mathcal{G}_0(s, \zeta(t + s)) \, \mathrm{d}s + \sum_{i=1}^{n} \mathcal{G}_i(t, \zeta(t - h_i)) \quad \forall \, t \ge 0 \, .$$

The operator T, so defined, is of class $\mathcal{T}_h$: for details see [14].

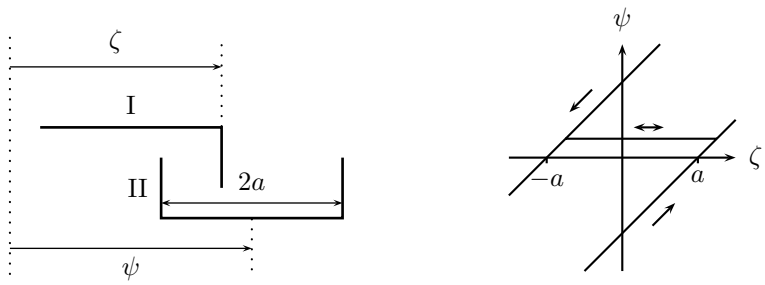
3.1.5. Systems with hysteresis. A general class of hysteresis operators, which includes many physically motivated hysteretic effects, is discussed in [11]. Examples of such operators include backlash hysteresis, elastic-plastic hysteresis, and Preisach operators. In [5], it is pointed out that these operators are of class $\mathcal{T}_0$. For illustration, we describe a particular example of a hysteresis operator.

Backlash hysteresis. Consider a one-dimensional mechanical link consisting of two components, denoted I and II (of width $2a$) and illustrated in Figure 3(a). The displacements of each part (with respect to some fixed datum) at time $t \ge 0$ are given by $\zeta(t)$ and $\psi(t)$ with $|\zeta(t) - \psi(t)| \le a$ for all t, and $\psi(0) := \zeta(0) + b$ for some prespecified $b \in [-a, a]$. Within the link there is mechanical play: that is to say, the position $\psi(t)$ of II remains constant as long as the position $\zeta(t)$ of I remains within the interior of II. Thus, assuming continuity of ζ, we have $\dot{\psi}(t) = 0$ whenever $|\zeta(t) - \psi(t)| < a$. Given a continuous input $\zeta \in C(\mathbb{R}_+, \mathbb{R})$, describing the evolution of the position of I, denote the corresponding position of II by $\psi = T\zeta$. The operator T (in effect we define a family $T_{a,b}$ of operators parameterized by $a > 0$ and $b \in [-a, a]$), so defined, is known as *backlash* or *play* and is of class $\mathcal{T}_0$.

4. The control. Let $\varphi \in \Phi$ determine a performance funnel $\mathcal{F}_\varphi$. We proceed to construct a feedback structure which ensures that, for every reference $r \in \mathcal{R}$ and when applied to any system of class Σ_ρ, the tracking error $e = y - r$ evolves within $\mathcal{F}_\varphi$. We initially assume $\rho \ge 2$; the case of systems with strict relative degree $\rho = 1$ will be treated separately in due course.

4.1. Filter. Fix $\mu > 0$ (arbitrarily) and introduce the filter

$$\begin{aligned} \dot{\xi}_i(t) &= -\mu \, \xi_i(t) + \xi_{i+1}, & \xi_i(0) &= \xi_i^0 \in \mathbb{R}^m \, , & i &= 1, \dots, \rho - 2, \\ \dot{\xi}_{\rho-1}(t) &= -\mu \, \xi_{\rho-1}(t) + u(t), & \xi_{\rho-1}(0) &= \xi_{\rho-1}^0 \in \mathbb{R}^m \, , \end{aligned}$$



(a) (b)

FIG. 3. *Backlash hysteresis.*

which, on writing (wherein I and 0 denote the $m \times m$ identity and zero matrices)

$$\xi(t) = \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \\ \xi_3(t) \\ \vdots \\ \xi_{\rho-2}(t) \\ \xi_{\rho-1}(t) \end{bmatrix}, \quad F = \begin{bmatrix} -\mu I & I & 0 & \cdots & 0 & 0 \\ 0 & -\mu I & I & \cdots & 0 & 0 \\ 0 & 0 & -\mu I & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\mu I & I \\ 0 & 0 & 0 & \cdots & 0 & -\mu I \end{bmatrix}, \quad G = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ I \end{bmatrix},$$

may be expressed as

$$\begin{cases} \dot{\xi}(t) = F\xi(t) + Gu(t), & \xi(0) = \xi^0 \in \mathbb{R}^{(\rho-1)m}, \\ \xi_1(t) = H\xi(t), & H := \big[I \vdots 0 \vdots 0 \vdots \cdots \vdots 0 \vdots 0\big]. \end{cases} \tag{4.1}$$

4.2. Feedback. Define

$$s(\Gamma) := \begin{cases} +1\,, & \Gamma \text{ positive definite,} \\ -1\,, & \Gamma \text{ negative definite.} \end{cases}$$

Let $\nu\colon \mathbb{R} \to \mathbb{R}$ be any C^∞ function with the following property:

$$\begin{cases} \text{there exists a strictly increasing unbounded sequence } (k_j) \text{ such that} \\ \text{the sequence } \big(s(\Gamma)\nu(k_j)\big) \text{ is strictly decreasing and unbounded.} \end{cases} \tag{4.2}$$

Introduce the projections

$$\pi_i\colon \mathbb{R}^{(\rho-1)m} \to \mathbb{R}^{im}, \quad \xi = (\xi_1, \ldots, \xi_{\rho-1}) \mapsto (\xi_1, \ldots, \xi_i), \qquad i = 1, \ldots, \rho-1,$$

and define the C^∞ function

$$\gamma_1\colon \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m, \quad (k, e) \mapsto \gamma_1(k, e) := -\nu(k)e\,, \tag{4.3}$$

with derivative (Jacobian matrix function) $D\gamma_1$. Next, for $i = 2, \ldots, \rho-1$, define the C^∞ function $\gamma_i\colon \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^{(i-1)m} \to \mathbb{R}^m$ by the recursion

$$\begin{aligned} \gamma_i(k, e, \pi_{i-1}\xi) := \gamma_{i-1}(k, e, \pi_{i-2}\xi) + \|D\gamma_{i-1}(k, e, \pi_{i-2}\xi)\|^2\, k^4\, (1 + \|\pi_{i-1}\xi\|^2) \\ \times \Big(\mu^{2-i}\xi_{i-1} + \gamma_{i-1}(k, e, \pi_{i-2}\xi)\Big)\,, \end{aligned} \tag{4.4}$$

wherein we adopt the notational convention $\gamma_1(k,e,\pi_0\xi) := \gamma_1(k,e)$. Define the C^∞ function $\gamma_\rho\colon \mathbb{R}\times\mathbb{R}^m\times\mathbb{R}^{(\rho-1)m}\to\mathbb{R}^m$ as follows:

$$\begin{aligned}\gamma_\rho(k,e,\xi) :=& \mu^{\rho-1}\gamma_{\rho-1}(k,e,\pi_{\rho-2}\xi) + \mu^{\rho-1}\|D\gamma_{\rho-1}(k,e,\pi_{\rho-2}\xi)\|^2 k^4\,(1+\|\xi\|^2)\\ &\times\Big(\mu^{2-\rho}\xi_{\rho-1}+\gamma_{\rho-1}(k,e,\pi_{\rho-2}\xi)\Big).\end{aligned}\tag{4.5}$$

Finally, we introduce the bijection

$$\alpha : [0,1)\to[1,\infty),\quad s\mapsto 1/(1-s).\tag{4.6}$$

For arbitrary $r\in\mathcal{R}$, the *control strategy* is given by

$$u(t) = -\gamma_\rho\big(k(t), Cx(t)-r(t), \xi(t)\big),\quad k(t)=\alpha\big(\varphi^2(t)\|Cx(t)-r(t)\|^2\big).\tag{4.7}$$

Remark 4.1. (i) If $s(\Gamma)$ is known a priori, then the function $\nu\colon k\mapsto -s(\Gamma)k$ is sufficient to ensure property (4.2); if $s(\Gamma)$ is unknown, then the function $\nu\colon k\mapsto k\cos k$ suffices. In the latter case, the role of the function ν is similar to that of a "Nussbaum" function in adaptive control. Note, however, that the requisite property (4.2) is less restrictive than (a) the "Nussbaum properties" as required in [17], for example, or (b) the stronger "scaling invariant Nussbaum properties", as required in [9], for example.

(ii) The function α in (4.6) may be generalized to any C^∞ bijection $\alpha:[0,1)\to[1,\infty)$ with the property that $\alpha' = d(\alpha)$ for some function d: the particular choice $d(\cdot)=(\cdot)^2$ yields the specific function adopted in (4.6) for simplicity of presentation. In the case of general α, the term k^4 in (4.4) and (4.5) should be replaced by $d^2(k)$.

(iii) In the specific case of a system of relative degree $\rho=2$, writing $e(t)=Cx(t)-r(t)$ and omitting the argument t for simplicity, the control strategy takes the explicit form

$$\begin{cases} u = \mu\,\nu(k)e - \mu\big[(\nu'(k)\|e\|)^2+(\nu(k))^2\big]\,k^4\,[1+\|\xi\|^2]\theta,\\ k=\alpha\big(\varphi^2\|e\|^2\big), \qquad \theta=\xi-\nu(k)e\,,\\ \dot\xi = -\mu\,\xi+u, \qquad \xi(0)=\xi^0.\end{cases}\tag{4.8}$$

We will adopt this controller in the example in section 6.

4.3. Well-posedness of the closed-loop system. The conjunction of the filter (4.1) and the feedback (4.7) applied to (3.1) yields the initial-value problem

$$\begin{cases}\dot x(t) = A\,x(t)+f(p(t),(TCx)(t),x(t)) - B\gamma_\rho(k(t),Cx(t)-r(t),\xi(t)),\\ \dot\xi(t)=F\xi(t)-G\gamma_\rho(k(t),Cx(t)-r(t),\xi(t)),\\ k(t)=\alpha\big(\varphi^2(t)\|Cx(t)-r(t)\|^2\big)\,,\\ x|_{[-h,0]}=x^0\in C([-h,0],\mathbb{R}^{\rho m}),\quad \xi(s)=\xi^0\in\mathbb{R}^{(\rho-1)m}\quad\forall\ s\in[-h,0]\,.\end{cases}\tag{4.9}$$

By a solution of (4.9) on $[-h,\omega)$ we mean a function $(x,\xi)\in C([-h,\omega),\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m})$, with $0<\omega\le\infty$, $x|_{[-h,0]}=x^0$, and $\xi(s)=\xi^0$ for all $s\in[-h,0]$, such that $(x,\xi)|_{[0,\omega)}$ is absolutely continuous, satisfies the differential equations in (4.9) for almost all $t\in[0,\omega)$, and avoids the singularity in α in the sense that $\varphi(t)\|Cx(t)-r(t)\|<1$ for all $t\in[0,\omega)$. To answer affirmatively the question of well-posedness of the closed loop, we provide an existence theorem for a class of initial-value problems of sufficient generality to incorporate (4.9). For $h\ge 0$, consider the initial-value problem

$$\begin{cases}\dot\zeta(t)=Z(t,(\widehat T\zeta)(t),\zeta(t)), & \zeta(t)\in\mathcal{D},\\ \zeta|_{[-h,0]}=\zeta^0\in C([-h,0],\mathbb{R}^N), & \zeta^0(0)\in\mathcal{D},\end{cases}\tag{4.10}$$

where $\mathcal{D} \subset \mathbb{R}^N$ is a nonempty, open set, $Z\colon [-h,\infty)\times\mathbb{R}^q\times\mathcal{D}\to\mathbb{R}^N$ is a Carathéodory function, and $\widehat{T}$ is a causal operator of class $\mathcal{T}_h$. By a solution of (4.10) on $[-h,\omega)$ we mean a function $\zeta\in C([-h,\omega),\mathbb{R}^N)$, with $0<\omega\le\infty$, and $\zeta|_{[-h,0]}=\zeta^0$ such that $\zeta|_{[0,\omega)}$ is absolutely continuous and satisfies the differential equations in (4.10) for almost all $t\in[0,\omega)$ and $\zeta(t)\in\mathcal{D}$ for all $t\in[0,\omega)$. A solution of (4.9) or of (4.10) is *maximal* if, and only if, it has no proper right extension that is also a solution.

THEOREM 4.2. *Let $\mathcal{D}\subset\mathbb{R}^N$ be nonempty and open, let $\widehat{T}$ be an operator of class $\mathcal{T}$, and let $Z\colon[-h,\infty)\times\mathbb{R}^q\times\mathcal{D}\to\mathbb{R}^N$ be a Carathéodory function. Then, for each $\zeta^0\in C([-h,0],\mathbb{R}^N)$ with $\zeta(0)\in\mathcal{D}$, there exists a solution $\zeta\colon[-h,\omega)\to\mathbb{R}^N$, $\zeta([0,\omega))\subset\mathcal{D}$, of the initial-value problem* (4.10) *and every solution can be extended to a maximal solution. Moreover, if Z is locally essentially bounded and $\zeta\colon[-h,\omega)\to\mathbb{R}^N$, $\zeta([0,\omega))\subset\mathcal{D}$, is a maximal solution with $\omega<\infty$, then, for every compact set $\mathcal{K}\subset\mathcal{D}$, there exists $\hat{t}\in[0,\omega)$ such that $\zeta(\hat{t})\notin\mathcal{K}$.*

Proof. The proof is a straightforward modification of that of [5, Thm. 5]. ∎

We apply this result to our closed-loop system (4.9).

COROLLARY 4.3. *Let $(A,B,C,f,p,T,h)\in\Sigma_\rho$ with $\rho\ge1$ and let $\varphi\in\Phi$. For every $r\in\mathcal{R}$ and $(x^0,\xi^0)\in C([-h,0],\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m})$, application of the feedback* (4.7) *in conjunction with the filter* (4.1) *to the system* (3.1) *yields the initial-value problem* (4.9)*, which has a solution, and every solution can be extended to a maximal solution. If a maximal solution of* (4.9) *on $[-h,\omega)$ is bounded and such that the associated gain function k is also bounded, then $\omega=\infty$.*

The proof is in the appendix.

5. Main results.

5.1. Preliminary lemmas. Let $(A,B,C,f,p,T,h)\in\Sigma_\rho$ with $\rho\ge2$. Rewriting the conjunction of the nonlinear system (3.1) and the filter (4.1) as

$$\left\{\begin{array}{l}\begin{bmatrix}\dot{x}(t)\\ \dot{\xi}(t)\end{bmatrix}=\begin{bmatrix}A&0\\0&F\end{bmatrix}\begin{bmatrix}x(t)\\ \xi(t)\end{bmatrix}+\begin{bmatrix}I\\0\end{bmatrix}f(p(t),(Ty)(t),x(t))+\begin{bmatrix}B\\G\end{bmatrix}u(t),\\ \qquad y(t)=[\,C\,\vdots\,0\,]\begin{bmatrix}x(t)\\ \xi(t)\end{bmatrix},\end{array}\right. \tag{5.1}$$

we have the following technicality, a proof of which can be found in the appendix.

LEMMA 5.1. *For system* (5.1)*, there exist $K\in\mathbb{R}^{\rho m\times(\rho-1)m}$ and $N\in\mathbb{R}^{(\rho-1)m\times\rho m}$ such that*

$$L:=\begin{bmatrix}C&0\\N&-NK\\0&I\end{bmatrix}\in\mathbb{R}^{(2\rho-1)m\times(2\rho-1)m}$$

is invertible and

$$L\begin{bmatrix}A&0\\0&F\end{bmatrix}L^{-1}=\begin{bmatrix}A_1&A_2&\tilde{\Gamma}\\A_3&A_4&0\\0&0&F\end{bmatrix},\quad L\begin{bmatrix}B\\G\end{bmatrix}=\begin{bmatrix}0\\G\end{bmatrix},\quad [\,C\,\vdots\,0\,]\,L^{-1}=[\,I\,\vdots\,0\,\vdots\,0\,],$$

where $\tilde{\Gamma}:=\begin{bmatrix}\Gamma\,\vdots\,0\end{bmatrix}\in\mathbb{R}^{m\times(\rho-1)m}$, $\Gamma:=CA^{\rho-1}B$, and $A_4\in\mathbb{R}^{(\rho-1)m\times(\rho-1)m}$ is such that $\operatorname{spec}(A_4)\subset\mathbb{C}_-$.

In view of Lemma 5.1, there exist K and N such that, under the coordinate change

$$\begin{bmatrix} y(t) \\ z(t) \\ \xi(t) \end{bmatrix} = L \begin{bmatrix} x(t) \\ \xi(t) \end{bmatrix}, \quad \begin{bmatrix} y^0 \\ z^0 \\ \xi^0 \end{bmatrix} = L \begin{bmatrix} x^0 \\ \xi^0 \end{bmatrix}, \quad L := \begin{bmatrix} C & 0 \\ N & -NK \\ 0 & I \end{bmatrix}, \tag{5.2}$$

the conjunction (5.1) of system (3.1) and filter (4.1) can be represented by

$$\begin{cases} \dot{y}(t) = A_1 y(t) + A_2 z(t) + Cf(p(t), (Ty)(t), x(t)) + \Gamma\xi_1(t), \\ \dot{z}(t) = A_3 y(t) + A_4 z(t) + Nf(p(t), (Ty)(t), x(t)), \\ \dot{\xi}(t) = F\xi(t) + Gu(t), \\ (y, z, \xi)|_{[-h,0]} = (y^0, z^0, \xi^0) \in C([-h, 0], \mathbb{R}^m \times \mathbb{R}^{(\rho-1)m} \times \mathbb{R}^{(\rho-1)m}), \end{cases} \tag{5.3}$$

where $A_4 \in \mathbb{R}^{(\rho-1)m\times(\rho-1)m}$ has spectrum in $\mathbb{C}_-$. If $(x, \xi)\colon [0, \omega) \to \mathbb{R}^{\rho m} \times \mathbb{R}^{(\rho-1)m}$ is a maximal solution of the nonlinearly perturbed closed-loop system (4.9), then, in view of (5.3) and writing

$$y(t) = Cx(t), \qquad e(t) = y(t) - r(t), \quad e|_{[-h,0]} = e^0(\cdot) = y^0(\cdot) - r(0)\,, \tag{5.4}$$

we arrive at the following equivalent to (4.9):

$$\begin{cases} \dot{e}(t) = A_1 e(t) + A_2 z(t) + f_1(t) + \Gamma\xi_1(t), \\ \dot{z}(t) = A_3 e(t) + A_4 z(t) + f_2(t), \\ \dot{\xi}(t) = F\xi(t) - G\gamma_\rho(k(t), e(t), \xi(t)), \\ k(t) = \alpha\big(\varphi^2(t)\|e(t)\|^2\big), \\ (e, z, \xi)|_{[-h,0]} = (e^0, z^0, \xi^0) \in C([-h, 0], \mathbb{R}^m \times \mathbb{R}^{(\rho-1)m} \times \mathbb{R}^{(\rho-1)m}), \end{cases} \tag{5.5}$$

where the functions f_1 and f_2 are given by

$$\begin{cases} f_1(t) := A_1 r(t) + Cf(p(t), (Ty)(t), x(t)) - \dot{r}(t), \\ f_2(t) := A_3 r(t) + Nf(p(t), (Ty)(t), x(t)). \end{cases} \tag{5.6}$$

Since $(\varphi(t)\|e(t)\|)^2 < 1$ for all $t \in [0, \omega)$, the properties of $\varphi \in \Phi$ yield boundedness of the function e which, together with boundedness of r, implies boundedness of y. Since T is of class $\mathcal{T}_h$ and y is bounded, Ty is essentially bounded. By boundedness of r, essential boundedness of $\dot{r}$ and p, and assumption (A4), we may now conclude (essential) boundedness of the functions f_1 and f_2. Observing that A_4 is Hurwitz and f_2 bounded, the second of the differential equations in (5.5) yields boundedness of z. These observations are recorded in the following lemma.

LEMMA 5.2. *Let $(A, B, C, f, p, T, h) \in \Sigma_\rho$ with $\rho \geq 2$. Let $\varphi \in \Phi$, $r \in \mathcal{R}$, and $(x^0, \xi^0) \in C([-h, 0], \mathbb{R}^{\rho m} \times \mathbb{R}^{(\rho-1)m})$. If $(x, \xi)\colon [-h, \omega) \to \mathbb{R}^{\rho m} \times \mathbb{R}^{(\rho-1)m}$ is a maximal solution of* (4.9)*, then the functions y, z, and e, given by* (5.2) *and* (5.4)*, are bounded. Furthermore, the functions f_1 and f_2, given by* (5.6)*, are essentially bounded and bounded, respectively.*

The proofs of our main results (Theorems 5.4 and 5.5 below) rely crucially on a further technicality: the signals $\theta_i = \mu^{1-i}\xi_i + \gamma_i(k, e, \pi_{i-1}\xi)$, $i = 1, \ldots, \rho - 1$, are bounded. More precisely, we have the following (with proof in the appendix).

LEMMA 5.3. *Let $(A,B,C,f,p,T,h) \in \Sigma_\rho$ with $\rho \geq 2$. Let $\varphi \in \Phi$, $r \in \mathcal{R}$, and $(x^0,\xi^0) \in C([-h,0],\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m})$. If $(x,\xi)\colon [-h,\omega) \to \mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m}$ is a maximal solution of (4.9), then the function $\theta = (\theta_1,\dots,\theta_{\rho-1})\colon [0,\omega) \to \mathbb{R}^{(\rho-1)m}$ is bounded, where*

$$\theta_i(t) := \mu^{1-i}\,\xi_i(t) + \gamma_i(k(t),e(t),\pi_{i-1}\xi(t))\,, \qquad i = 1,\dots,\rho-1, \tag{5.7}$$

with the notational convention $\gamma_1(k,e,\pi_0\xi) := \gamma_1(k,e)$.

5.2. Relative degree 1 case. We are now in a position to state our main result for the case when the system has strict relative degree 1; in this case, a filter is not necessary and the controller (4.7) simplifies to

$$u(t) = \nu(k(t))(Cx(t)-r(t)), \quad k(t) = \alpha\big(\varphi^2(t)\|Cx(t)-r(t)\|^2\big)\,. \tag{5.8}$$

The closed-loop initial-value problem then becomes

$$\begin{cases} \dot{x}(t) = Ax(t) + B\nu(k(t))(Cx(t)-r(t)) + f(p(t),T(Cx)(t),x(t)),\\ k(t) = \alpha\big(\varphi^2(t)\|Cx(t)-r(t)\|^2\big)\,,\\ x|_{[-h,0]} = x^0 \in C([-h,0],\mathbb{R}^m). \end{cases} \tag{5.9}$$

THEOREM 5.4. *Let $(A,B,C,f,p,T,h) \in \Sigma_1$ and $\varphi \in \Phi$ with associated performance funnel $\mathcal{F}_\varphi$. For each reference signal $r \in \mathcal{R}$, and initial data $(x^0,\xi^0) \in C([-h,0],\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m})$, application of the feedback (5.8) to (3.1) yields the initial-value problem (5.9), which has a solution, and every solution can be maximally extended. Every maximal solution $x\colon [-h,\omega) \to \mathbb{R}^m$ has the following properties:*

(i) *$\omega = \infty$;*

(ii) *x, k, and u are bounded;*

(iii) *the tracking error evolves within the funnel $\mathcal{F}_\varphi$ and is bounded away from the funnel boundary; i.e., there exists $\varepsilon > 0$ such that, for all $t \geq 0$, $\varphi(t)\|Cx(t)-r(t)\| \leq 1-\varepsilon$.*

The proof of Theorem 5.4 follows easily by modifying (all vestiges of the filter equations are excised) the proof of Theorem 5.5. The latter proof is in the appendix.

5.3. Relative degree $\rho \geq 2$ case. We now arrive at the main result of the paper (with proof in the appendix).

THEOREM 5.5. *Let $(A,B,C,f,p,T,h) \in \Sigma_\rho$ with $\rho \geq 2$ and let $\varphi \in \Phi$ with associated performance funnel $\mathcal{F}_\varphi$. For each reference signal $r \in \mathcal{R}$ and initial data $(x^0,\xi^0) \in C([-h,0],\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m})$, application of the feedback (4.7), in conjunction with the filter (4.1), to (3.1) yields the initial-value problem (4.9), which has a solution, and every solution can be maximally extended. Every maximal solution $(x,\xi)\colon [-h,\omega) \to \mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m}$ has the following properties:*

(i) *$\omega = \infty$;*

(ii) *x, ξ, k, and u are bounded;*

(iii) *the tracking error evolves within the funnel $\mathcal{F}_\varphi$ and is bounded away from the funnel boundary; i.e., there exists $\varepsilon > 0$ such that, for all $t \geq 0$, $\varphi(t)\|Cx(t)-r(t)\| \leq 1-\varepsilon$.*

6. Example. We illustrate the controller strategy (4.7) applied to the following single-input, single-output system of relative degree $\rho = 2$:

$$\ddot{y}(t) + b_0 \sin y(t) + b_1 y(t)|y(t)| + (T_{a,b}\,y)(t) = b_2\,u(t)\,, \tag{6.1}$$

where b_0, b_1, and $b_2 \neq 0$ are unknown real parameters and $T_{a,b}$ represents the backlash operator, as defined in section 3.1.5, with parameters $a > 0$ and $b \in [-a, a]$. Equation (6.1) is equivalent to (3.1) with

$$x(t) = \begin{bmatrix} y(t) \\ \dot{y}(t) \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ b_2 \end{bmatrix}, \quad C = [1 \,\vdots\, 0], \quad f(p, w, x) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} w,$$

and the operator T given by $(Ty)(t) = b_0 \sin y(t) + b_1 y(t)|y(t)| + (T_{a,b}y)(t)$, $t \in \mathbb{R}_+$. Setting $h = 0$ and $p = 0$, the resulting system $(A, B, C, f, 0, T, 0)$ is of class Σ_2.

Fix $\tau > 0$ arbitrarily and define $\varphi \in \Phi$ by

$$t \mapsto \varphi(t) = \begin{cases} 20(1 - (\frac{t}{\tau} - 1)^2), & 0 \leq t < \tau, \\ 20, & t \geq \tau. \end{cases} \tag{6.2}$$

Evolution within the associated performance funnel $\mathcal{F}_\varphi$ ensures a tracking accuracy $|e(t)| < 0.05$ for all $t \geq \tau$. Choosing $\nu\colon k \mapsto k \cos k$, $\xi^0 = 0$, writing $e(t) = y(t) - r(t)$, and suppressing the argument t for simplicity, the control strategy (4.8) is

$$\begin{cases} u = \mu(k \cos k)e - \mu\big[(\cos k - k \sin k)^2 e^2 + k^2 \cos^2 k\big] k^4 [1 + \xi^2]\theta\,, \\ k = \big[1 - \varphi^2 e^2\big]^{-1}\,, \qquad \theta = \xi - (k \cos k)e\,, \\ \dot{\xi} = -\mu\,\xi + u, \qquad\qquad \xi(0) = 0. \end{cases} \tag{6.3}$$

For purposes of illustration, as reference signal $r \in \mathcal{R}$, we take the first component ζ_1 of the solution (chaotic and bounded; see [15, appx. C]) of the following Lorenz system of equations:

$$\begin{cases} \dot{\zeta}_1(t) = \frac{1}{2}\zeta_2(t) - \zeta_1(t), & \zeta_1(0) = \frac{1}{2}, \\ \dot{\zeta}_2(t) = \frac{28}{5}\zeta_1(t) - \frac{1}{10}\zeta_2(t) - 2\zeta_1(t)\zeta_3(t), & \zeta_2(0) = 0, \\ \dot{\zeta}_3(t) = 2\zeta_1(t)\zeta_2(t) - \frac{8}{30}\zeta_3(t), & \zeta_3(0) = 3. \end{cases} \tag{6.4}$$

Setting $b_0 = \frac{1}{2}$, $b_1 = 1 = b_2$, $\mu = 10$, $\tau = 50$ and adopting backlash hysteresis with parameters $a = 1/2$, $b = 0$ and initial data $(y(0), \dot{y}(0)) = (0, 0)$, the behavior of the closed-loop system (6.1)–(6.3) is depicted in Figure 4.

7. Appendix.

7.1. Proof of Lemma 3.5. Parts of the following proof are implicit in the proofs of [7, Lem. 4.1.1] and [8, Prop. 11.5.1 and 11.5.2] (in a general context of nonlinear systems); here, we provide a simple, self-contained proof in the restricted context of linear systems.

Step (i). First, note that

$$\mathcal{CB} = \begin{bmatrix} 0 & & \Gamma \\ & \iddots & \\ \Gamma & & * \end{bmatrix},$$

and, since Γ is invertible, we see that $\mathcal{CB} \in \mathrm{GL}_{\rho m}(\mathbb{R})$. Furthermore, $\mathcal{NB} = 0$. Assertion (i) then follows from the observation that, for any $x \in \mathbb{R}^n$, we have $v := (I - \mathcal{B}(\mathcal{CB})^{-1}\mathcal{C})x \in \ker \mathcal{C}$ and $w := \mathcal{B}(\mathcal{CB})^{-1}\mathcal{C}x \in \operatorname{im} \mathcal{B}$, and so $x = v + w$.

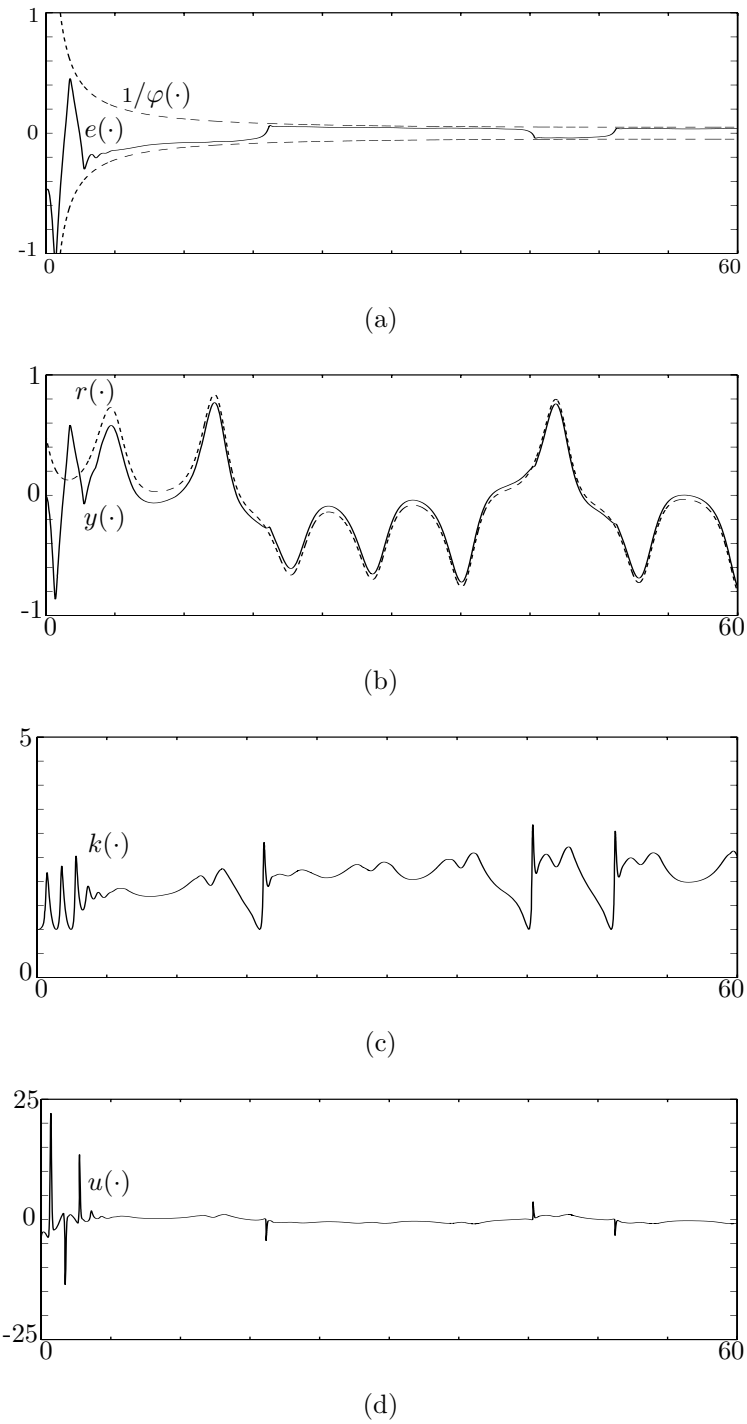


FIG. 4. *Tracking of a Lorenz component reference signal; system* (6.1) *with unknown sign* $b_2 \neq 0$ *and control strategy* (6.3). (a) *The funnel and tracking error* e. (b) *The reference* r *and output* y. (c) *The function* k. (d) *The control* u.

Step (ii). We now prove assertion (ii). It is clear that $\mathcal{U}^{-1} = [\mathcal{B}(\mathcal{CB})^{-1} \vdots \mathcal{V}]$. It is also immediate that $\hat{B} := \mathcal{U}\tilde{B}$ and $\hat{C} := \tilde{C}\mathcal{U}^{-1}$ have the structure given in (3.9). Furthermore,

$$\mathcal{U}\,\tilde{A} = \hat{A}\mathcal{U} \tag{7.1}$$

for some $\hat{A}$ of the form

$$\hat{A} = \begin{bmatrix} 0 & I & 0 & \dots & 0 & 0 \\ 0 & 0 & I & & & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & 0 & \dots & 0 & I & 0 \\ R_1 & R_2 & \dots & R_{\rho-1} & R_\rho & S \\ P_1 & P_2 & \dots & P_{\rho-1} & P_\rho & Q \end{bmatrix},$$

with $R_i \in \mathbb{R}^{m\times m}$, $P_i \in \mathbb{R}^{(n-\rho m)\times m}$, $i = 1,\dots,\rho$, $S \in \mathbb{R}^{m\times(n-\rho m)}$, $Q = \mathcal{N}\tilde{A}\mathcal{V} \in \mathbb{R}^{(n-\rho m)\times(n-\rho m)}$, and $[R_1 \vdots \cdots \vdots R_\rho \vdots S] = \tilde{C}\tilde{A}^\rho\mathcal{U}^{-1}$. If $\rho = 1$, then $\hat{A}$ takes the form shown in (3.10).

Recalling that $\mathcal{NB} = 0$, we see that

$$[P_1 \vdots \cdots \vdots P_\rho] = \mathcal{N}\tilde{A}\mathcal{B}(\mathcal{CB})^{-1} = [0 \vdots \cdots \vdots 0 \vdots \mathcal{N}\tilde{A}^\rho\tilde{B}] \begin{bmatrix} * & & \Gamma^{-1} \\ & \iddots & \\ \Gamma^{-1} & & 0 \end{bmatrix},$$

and hence $P_i = 0$ for $i = 2,\dots,\rho$. Writing $P = P_1$, it follows that $\hat{A}$ takes the form in (3.9) and $P = \mathcal{N}\tilde{A}^\rho\tilde{B}\Gamma^{-1}$.

Step (iii). Finally, we prove part (iii) of the lemma. Writing

$$M_1(s) = \begin{bmatrix} sI - \tilde{A} & \tilde{B} \\ \tilde{C} & 0 \end{bmatrix}, \quad M_2(s) = \begin{bmatrix} \mathcal{U} & 0 \\ 0 & I \end{bmatrix} M_1(s) \begin{bmatrix} \mathcal{U}^{-1} & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} sI - \hat{A} & \hat{B} \\ \hat{C} & 0 \end{bmatrix},$$

and

$$M_3(s) = \begin{bmatrix} \hat{C} & 0 \\ \hat{A} - sI & -\hat{B} \end{bmatrix} = \begin{bmatrix} I & 0 & 0 & \dots & 0 & 0 & 0 \\ -sI & I & 0 & \dots & 0 & 0 & 0 \\ 0 & -sI & I & & 0 & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots & \vdots \\ 0 & 0 & \dots & -sI & I & 0 & 0 \\ R_1 & R_2 & \dots & R_{\rho-1} & R_\rho - sI & S & -\Gamma \\ P & 0 & \dots & 0 & 0 & Q - sI & 0 \end{bmatrix},$$

we see that $|\det M_1(s)| = |\det M_2(s)| = |\det M_3(s)| = |\det\Gamma\,\det(sI - Q)|$.

By the minimum-phase property of $(\tilde{A}, \tilde{B}, \tilde{C})$, we have $\det(M_1(s)) \neq 0$ for all $s \in \mathbb{C} \setminus \mathbb{C}_-$, and so $\det(sI - Q) \neq 0$ for all $s \in \mathbb{C} \setminus \mathbb{C}_-$. It follows that $\operatorname{spec}(Q) \subset \mathbb{C}_-$, and hence assertion (iii) holds. ∎

7.2. Proof of Corollary 4.3. Introducing the open set

$$\mathcal{D} := \Big\{(x,\xi,\eta) \in \mathbb{R}^{\rho m} \times \mathbb{R}^{(\rho-1)m} \times \mathbb{R} \,\Big|\, \varphi(|\eta|)\,\|Cx - r(|\eta|)\| < 1\Big\}$$

and defining, on $\mathcal{D}$,

$$\gamma_\rho^*\colon (x,\xi,\eta)\mapsto \gamma_\rho\big(\alpha(\varphi^2(|\eta|)\,\|Cx-r(|\eta|)\|^2)\,,\,Cx-r(|\eta|)\,,\,\xi\big),$$

the initial-value problem (4.9) may be recast on $\mathcal{D}$ as

$$\text{(7.2)}\qquad \begin{cases} \dot{x}(t)=Ax(t)+f(p(t),T(Cx)(t),x(t))-B\gamma_\rho^*(x(t),\xi(t),\eta(t)),\\ \dot{\xi}(t)=F\xi(t)-G\gamma_\rho^*(x(t),\xi(t),\eta(t)),\\ \dot{\eta}(t)=1,\\ (x,\xi,\eta)|_{[-h,0]}=(x^0,\xi^0,0)\in C([-h,0],\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m}\times\mathbb{R}). \end{cases}$$

Setting $\zeta=(x,\xi,\eta)$ and defining the Carathéodory function

$$\begin{aligned} &Z\colon [-h,\infty)\times\mathbb{R}^q\times\mathbb{R}^{2(\rho-1)m+1}\to\mathbb{R}^{(2\rho-1)m+1},\\ &(t,w,\zeta)\mapsto Z(t,w,\zeta):=\begin{bmatrix}A&0&0\\0&F&0\\0&0&0\end{bmatrix}\zeta+\begin{bmatrix}I\\0\\0\end{bmatrix}f(p(t),w,x)-\begin{bmatrix}B\\G\\0\end{bmatrix}\gamma_\rho^*(\zeta)+\begin{bmatrix}0\\0\\1\end{bmatrix} \end{aligned}$$

we can rewrite (7.2) as follows:

$$\text{(7.3)}\qquad \dot{\zeta}(t)=Z(t,(\widehat{T}\zeta)(t),\zeta(t)),\quad \zeta|_{[-h,0]}=\zeta^0\in C([-h,0],\mathbb{R}^{(2\rho-1)m+1}),$$

where the operator $\widehat{T}$, given by $(\widehat{T}\zeta)(t)=(TCx)(t)$, is of class $\mathcal{T}_h$. We then apply the existence result, Theorem 4.2, to establish: (i) the existence of a solution $t\mapsto\zeta(t)\in\mathcal{D}$ to (7.2) and that (ii) every solution can be extended to a maximal solution $\zeta\colon[-h,\omega)\to\mathcal{D}$. Furthermore, if there exists a compact set $\mathcal{C}\subset\mathcal{D}$ such that $(x(t),\xi(t),\eta(t))\in\mathcal{C}$ for all $t\in[0,\omega)$, then $\omega=\infty$.

Clearly, if $\zeta=(x,\xi,\eta)\colon[-h,\omega)\to\mathcal{D}$ is a solution of (7.3), then $(x,\xi)\colon[-h,\omega)\to\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m}$ is a solution of (4.9); conversely, if $(x,\xi)\colon[-h,\omega)\to\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m}$ is a solution of (4.9), then $\zeta=(x,\xi,\eta)\colon[-h,\omega)\to\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m}\times\mathbb{R}$, with component η given by $\eta(t)=t$, is a solution of (7.3). We may now conclude that, for each $(x^0,\xi^0)\in C([-h,0],\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m})$, (4.9) has a solution and every solution can be maximally extended.

Let $(x,\xi)\colon[-h,\omega)\to\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m}$ be a maximal solution of (4.9) (and so $t\mapsto\zeta(t)=(x(t),\xi(t),t)$ is a maximal solution of (4.10)). Assume that (x,ξ) is bounded and that the gain function $t\mapsto k(t)=\alpha\big(\varphi^2(t)\|Cx(t)-r(t)\|^2\big)$ is also bounded. Then there exist $c>0$ and $\varepsilon>0$ such that $\|(x(t),\xi(t))\|\le c$ and $\varphi(t)\|Cx(t)-r(t)\|\le 1-\varepsilon$ for all $t\in[0,\omega)$. Seeking a contradiction, suppose that $\omega<\infty$. It then follows that $\mathcal{K}:=\big\{(\hat{x},\hat{\xi},\hat{\eta})\in\mathcal{D}\big|\ \varphi(|\hat{\eta}|)\|C\hat{x}-r(|\hat{\eta}|)\|\le 1-\varepsilon,\ \|(\hat{x},\hat{\xi})\|\le c,\ \hat{\eta}\in[-h,\omega]\big\}$ is a compact subset of $\mathcal{D}$ which contains the trajectory $\zeta([-h,\omega))$ of the maximal solution ζ of (4.10). This contradicts the last assertion of Theorem 4.2, and so $\omega=\infty$. ∎

7.3. Proof of Lemma 5.1. Define

$$K:=\big[[\mu\,I+A]^{\rho-2}B\,\vdots\,[\mu\,I+A]^{\rho-3}B\,\vdots\,\cdots\,\vdots\,[\mu\,I+A]B\,\vdots\,B\big]\in\mathbb{R}^{\rho m\times(\rho-1)m}$$

and note that

$$AK-KF=\big[[\mu\,I+A]^{\rho-1}B\,\vdots\,0\,\vdots\,\cdots\,\vdots\,0\big],\quad KG=B,\quad\text{and}\quad CK=0.$$

Writing $\tilde{B} := (\mu I + A)^{\rho-1}B$, we have $C\tilde{B} = CA^{\rho-1}B = \Gamma$, and so the triple $(A, \tilde{B}, C)$ defines a linear system of relative degree one. Let $V \in \mathbb{R}^{\rho m \times (\rho-1)m}$ be such that $\operatorname{im} V = \ker C$. By Lemma 3.5 applied in the context of the system $(A, \tilde{B}, C)$, the matrix $\left[\begin{smallmatrix} C \\ N \end{smallmatrix}\right]$, with $N := (V^TV)^{-1}V^T\,[I - \tilde{B}\Gamma^{-1}C]$, is invertible, with inverse $\left[\tilde{B}\Gamma^{-1} \vdots V\right]$. Writing

$$L = \begin{bmatrix} C & 0 \\ N & -NK \\ 0 & I \end{bmatrix} \quad \text{with} \quad L^{-1} = \begin{bmatrix} \tilde{B}\Gamma^{-1} & V & K \\ 0 & 0 & I \end{bmatrix}$$

and recalling that $KG = B$, $CB = 0$, and $CK = 0$, we have

$$L \begin{bmatrix} B \\ G \end{bmatrix} = \begin{bmatrix} 0 \\ G \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} C \vdots 0 \end{bmatrix} L^{-1} = \begin{bmatrix} I \vdots 0 \end{bmatrix}.$$

Moreover, noting that $CAK = \begin{bmatrix} \Gamma \vdots 0 \vdots \cdots \vdots 0 \end{bmatrix} =: \tilde{\Gamma}$ and $N[AK - KF] = 0$, we have

$$L \begin{bmatrix} A & 0 \\ 0 & F \end{bmatrix} L^{-1} = \begin{bmatrix} CA\tilde{B}\Gamma^{-1} & CAV & CAK \\ NA\tilde{B}\Gamma^{-1} & NAV & N[AK - KF] \\ 0 & 0 & F \end{bmatrix} = \begin{bmatrix} A_1 & A_2 & \tilde{\Gamma} \\ A_3 & A_4 & 0 \\ 0 & 0 & F \end{bmatrix},$$

where $\tilde{\Gamma} = [\Gamma \vdots 0 \vdots \cdots \vdots 0]$. It remains to show that A_4 has spectrum in $\mathbb{C}_-$. Writing

$$M_4(s) = \begin{bmatrix} sI - A & B \\ C & 0 \end{bmatrix} \quad \text{and} \quad M_5(s) = \begin{bmatrix} sI - A & 0 & B \\ 0 & sI - F & -G \\ C & 0 & 0 \end{bmatrix},$$

we have

$$M_6(s) := \begin{bmatrix} I & K & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} M_5(s) \begin{bmatrix} I & K & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} sI - A & AK - KF & 0 \\ 0 & sI - F & -G \\ C & 0 & 0 \end{bmatrix}.$$

In view of the particular structure of F, G, and $AK - KF$, it is readily verified that $|\det M_6(s)| = |\det M_7(s)|$, where

$$M_7(s) = \begin{bmatrix} sI - A & [\mu I + A]^{\rho-1}B \\ C & 0 \end{bmatrix}.$$

Define

$$M_8(s) := \begin{bmatrix} C & 0 \\ N & 0 \\ 0 & I \end{bmatrix} M_7(s) \begin{bmatrix} \tilde{B}\Gamma^{-1} & V & 0 \\ 0 & 0 & I \end{bmatrix} = \begin{bmatrix} sI - A_1 & -A_2 & \Gamma \\ -A_3 & sI - A_4 & 0 \\ I & 0 & 0 \end{bmatrix}.$$

By the minimum-phase property of the triple (A, B, C) (recall Remark 3.4(ii)), for all $s \in \mathbb{C} \setminus \mathbb{C}_-$, we have $\det M_4(s) \neq 0$. We may now conclude that, for all $s \in \mathbb{C} \setminus \mathbb{C}_-$,

$$\begin{aligned} |\det \Gamma \det(sI - A_4)| &= |\det M_8(s)| = |\det M_7(s)| \\ &= |\det M_6(s)| = |\det M_5(s)| = |\det(sI - F) \det M_4(s)| \neq 0, \end{aligned}$$

and so $\operatorname{spec}(A_4) \subset \mathbb{C}_-$. This completes the proof. ∎

7.4. Proof of Lemma 5.3. Assume that $(x,\xi)\colon [-h,\omega)\to\mathbb{R}^{\rho m}\times\mathbb{R}^{(\rho-1)m}$ is a maximal solution of (4.9). Write $y(t)=Cx(t)$ and $e(t)=y(t)-r(t)$ for all $t\in[-h,\omega)$. By Lemma 5.1, there exists an invertible linear transformation L under which the closed-loop system (4.9) may be expressed in the form (5.5), wherein, by Lemma 5.2, e and z are bounded and the functions f_1 and f_2 given by (5.6) are essentially bounded and bounded, respectively. By boundedness of z, essential boundedness of f_1, and the first of equations (5.5), we may infer the existence of $c_1>0$ such that

$$\|\dot e(t)\|\le c_1\big(1+\|\xi_1(t)\|\big)\quad\text{for a.a. } t\in[0,\omega).$$

By boundedness of φ, e, essential boundedness of $\dot\varphi$, and recalling that $\alpha'(s)=\alpha^2(s)\ge 1$ for all $s\in[0,1)$, there exists a constant $c_2>0$ such that

$$\begin{aligned}|\dot k(t)| &= 2\alpha'(\varphi^2(t)\|e(t)\|^2)\big|\varphi^2(t)\langle e(t),\dot e(t)\rangle+\varphi(t)\dot\varphi(t)\|e(t)\|^2\big|\\ &\le c_2k^2(t)\,(1+\|\xi_1(t)\|)\quad\text{for a.a. } t\in[0,\omega).\end{aligned}$$

Since $k(t)\ge 1$ for all $t\in[0,\omega)$, we may now conclude the existence of a constant $c_3>0$ such that

$$\|(\dot k(t),\dot e(t))\|^2\le c_3\,\Delta(t)\quad\text{for a.a. } t\in[0,\omega),\quad\text{where } \Delta(t):=k^4(t)\big(1+\|\xi_1(t)\|^2\big).$$

Then, invoking (4.4), (5.7), and writing $c_{4,1}:=c_3/\mu>0$, we have

$$\begin{aligned}\langle\theta_1(t),\dot\theta_1(t)\rangle &\le \langle\theta_1(t)\,,\,-\mu\xi_1(t)+\xi_2(t)\rangle+\|\theta_1(t)\|\|D\gamma_1(k(t),e(t))\|\|(\dot k(t),\dot e(t))\|\\ &\le \langle\theta_1(t)\,,\,-\mu\theta_1(t)+\mu\gamma_1(k(t),e(t))\rangle+\langle\theta_1(t),\xi_2(t)\rangle\\ &\qquad+\sqrt{\mu}\,\|\theta_1(t)\|\,\|D\gamma_1(k(t),e(t))\|\,\sqrt{(c_3/\mu)\,\Delta(t)}\\ &\le c_{4,1}-\mu\|\theta_1(t)\|^2+\langle\theta_1(t),\xi_2(t)\rangle+\mu\langle\theta_1(t),\gamma_1(k(t),e(t))\rangle\\ &\qquad+\mu\|\theta_1(t)\|^2\,\|D\gamma_1(k(t),e(t))\|^2\,\Delta(t)\\ &= c_{4,1}-\mu\|\theta_1(t)\|^2+\big\langle\theta_1(t),\xi_2(t)+\mu\gamma_2(k(t),e(t),\xi_1(t))\big\rangle\\ &= c_{4,1}-\mu\|\theta_1(t)\|^2+\mu\langle\theta_1(t),\theta_2(t)\rangle\quad\text{for a.a. } t\in[0,\omega).\end{aligned}$$

Analogous calculations yield the existence of constants $c_{4,2},\dots,c_{4,\rho-1}>0$, such that

$$\langle\theta_i(t),\dot\theta_i(t)\rangle\le c_{4,i}-\mu\|\theta_i(t)\|^2+\mu\langle\theta_i(t),\theta_{i+1}(t)\rangle\quad\text{for a.a. } t\in[0,\omega),\ \ i=2,\dots,\rho-2,$$

and, using (4.5), $\langle\theta_{\rho-1}(t),\dot\theta_{\rho-1}(t)\rangle\le c_{4,\rho-1}-\mu\|\theta_{\rho-1}(t)\|^2$ for almost all $t\in[0,\omega)$. Writing $c_4=c_{4,1}+\cdots+c_{4,\rho-1}$, we have

$$\begin{aligned}\tfrac12\tfrac{\mathrm{d}}{\mathrm{d}t}\|\theta(t)\|^2 &\le c_4-\mu\|\theta(t)\|^2+\mu\langle\theta_1(t),\theta_2(t)\rangle+\cdots+\mu\langle\theta_{\rho-2}(t),\theta_{\rho-1}(t)\rangle\\ &\le c_4-\mu\langle\theta(t),P\theta(t)\rangle\quad\text{for a.a. } t\in[0,\omega),\end{aligned}$$

where P is a positive-definite, symmetric, tridiagonal matrix with all diagonal entries equal to 1 and all sub- and superdiagonal entries equal to $-1/2$. By positivity of P, it follows that θ is bounded. This completes the proof of the lemma. □

7.5. Proof of Theorem 5.5. Let (x^0,ξ^0) be arbitrary. By Corollary 4.3, (4.9) has a solution and every solution can be maximally extended. Let (x,ξ) be a maximal solution of (4.9) with interval of existence $[-h,\omega)$. Writing $y(t)=Cx(t)$, $e(t)=y(t)-r(t)$ for all $t\in[0,\omega)$ and invoking Lemma 5.1, there exists an invertible linear transformation L which takes (4.9) into the equivalent form (5.5)–(5.6). Introducing

$\theta_1\colon [0,\omega) \to \mathbb{R}^m$ given by (5.7), namely, $\theta_1(t) = \xi_1(t) - \nu(k(t))e(t)$, then the first of equations (5.5) yields

$$\dot{e}(t) = f_3(t) + \nu(k(t))\,\Gamma\, e(t) \quad \text{for a.a. } t \in [0,\omega), \tag{7.4}$$

with $f_3(t) := A_1 e(t) + A_2 z(t) + \Gamma\theta_1(t) + f_1(t)$. By Lemmas 5.2 and 5.3, the functions y, z, e and $\theta = (\theta_1, \dots, \theta_{\rho-1})$, given by (5.7), are bounded, which, together with essential boundedness of f_1, implies essential boundedness of f_3. Therefore, there exists $c_5 > 0$ such that

$$\langle e(t), \dot{e}(t)\rangle \le c_5 + \nu(k(t))\,\langle e(t), \Gamma e(t)\rangle \quad \text{for a.a. } t \in [0,\omega). \tag{7.5}$$

We are now in a position to prove boundedness of k. Recalling that Γ is either positive definite or negative definite, there exist constants $\beta_0, \beta_1 > 0$ such that

$$\beta_0\|e\|^2 \le |\langle e, \Gamma e\rangle| \le \beta_1\|e\|^2 \quad \forall\, e \in \mathbb{R}^m.$$

Define the continuous function $\tilde{\nu}\colon \mathbb{R} \to \mathbb{R}$ as follows:

$$\tilde{\nu}(k) := \begin{cases} -\beta_1\nu(k), & s(\Gamma)\nu(k) \ge 0\,, \\ -\beta_0\nu(k), & s(\Gamma)\nu(k) < 0\,. \end{cases}$$

Observe that

$$\nu(k)\langle e, \Gamma e\rangle \le -s(\Gamma)\tilde{\nu}(k)\|e\|^2 \quad \forall\, e \in \mathbb{R}^m\,,\ \forall\, k \ge 0\,,$$

which, together with boundedness of e, φ, essential boundedness of $\dot{\varphi}$, and (7.5), implies the existence of $c_6 > 0$ such that

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\,(\varphi(t)\|e(t)\|)^2 \le c_6 - 2s(\Gamma)\,\tilde{\nu}(k(t))\,\big(\varphi(t)\|e(t)\|\big)^2 \quad \text{for a.a. } t \in [0,\omega).$$

In view of property (4.2) of ν, there exists a strictly increasing unbounded sequence (k_j) in $(1,\infty)$ such that the sequence $\big(s(\Gamma)\tilde{\nu}(k_j)\big)$ is also strictly increasing, unbounded, and such that $s(\Gamma)\tilde{\nu}(k_j) > 0$ for all $j \in \mathbb{N}$. Seeking a contradiction, suppose k is unbounded on $[0,\omega)$. For each $j \in \mathbb{N}$, define $\tau_j := \inf\{t \in [0,\omega)|\ k(t) = k_{j+1}\}$ and $\sigma_j := \sup\{t \in [0,\tau_j]|\ \tilde{\nu}(k(t)) = \tilde{\nu}(k_j)\}$. It is readily verified that $\sigma_j < \tau_j$ and $k(\sigma_j) < k(\tau_j)$; moreover, for all $j \in \mathbb{N}$ and all $t \in [\sigma_j, \tau_j]$, $k(t) \ge k_j$ and $s(\Gamma)\tilde{\nu}(k(t)) \ge s(\Gamma)\tilde{\nu}(k_j)$. Therefore,

$$(\varphi(t)\|e(t)\|)^2 \ge \alpha^{-1}(k_j) \ge \alpha^{-1}(k_1) = 1 - \frac{1}{k_1} =: c_7 > 0 \quad \forall\, t \in [\sigma_j, \tau_j],\ \forall\, j \in \mathbb{N},$$

where $\alpha^{-1}\colon [1,\infty) \to [0,1)$ is the inverse of the bijection α. Thus,

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\,(\varphi(t)\|e(t)\|)^2 \le c_6 - 2c_7 s(\Gamma)\tilde{\nu}(k(t)) \quad \forall\, t \in [\sigma_j, \tau_j],\ \ \forall\, j \in \mathbb{N}.$$

Let $j^* \in \mathbb{N}$ be sufficiently large so that $c_6 - 2c_7 s(\Gamma)\tilde{\nu}(k_{j^*}) < 0$. Then

$$\big(\varphi(\tau_{j^*})\|e(\tau_{j^*})\|\big)^2 < \big(\varphi(\sigma_{j^*})\|e(\sigma_{j^*})\|\big)^2,$$

whence the contradiction

$$0 > \alpha\big(\varphi^2(\tau_{j^*})\|e(\tau_{j^*})\|^2\big) - \alpha\big(\varphi^2(\sigma_{j^*})\|e(\sigma_{j^*})\|^2\big) = k(\tau_{j^*}) - k(\sigma_{j^*}) > 0.$$

This proves boundedness of k. Therefore, there exists $\varepsilon > 0$ such that $\varphi(t)\|e(t)\| \leq 1 - \varepsilon$ for all $t \in [0, \omega)$. By boundedness of θ, e, and k, together with continuity of the functions γ_i, it follows from the recursive construction in (5.7) that, for $i = 1, \ldots, \rho-1$, ξ_i is bounded. We may now deduce that x and ξ are bounded, and, by (4.3), (4.4), (4.5), and (4.7), we may also infer boundedness of u. Finally, by boundedness of x, ξ, and k, together with Corollary 4.3, we conclude that $\omega = \infty$. ∎

REFERENCES

[1] E. Bullinger and F. Allgöwer, *Adaptive λ-tracking for nonlinear higher relative degree systems*, Automatica, 41 (2005), pp. 1191–1200.

[2] C.I. Byrnes and A. Isidori, *Limit sets, zero dynamics and internal models in the problem of nonlinear output regulation*, IEEE Trans. Automat. Control, 48 (2003), pp. 1712–1723.

[3] C.I. Byrnes and A. Isidori, *Nonlinear internal models for output regulation*, IEEE Trans. Automat. Control, 49 (2004), pp. 2244–2247.

[4] A. Ilchmann, E.P. Ryan, and C.J. Sangwin, *Systems of controlled functional differential equations and adaptive tracking*, SIAM J. Control Optim., 40 (2002), pp. 1746–1764.

[5] A. Ilchmann, E.P. Ryan, and C.J. Sangwin, *Tracking with prescribed transient behaviour*, ESAIM Control Optim. Calc. Var., 7 (2002), pp. 471–493.

[6] A. Ilchmann, E.P. Ryan, and P.N. Townsend, *Tracking control with prescribed transient behavior for systems of known relative degree*, Systems Control Lett., 55 (2006), pp. 396–406.

[7] A. Isidori, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, London, 1995.

[8] A. Isidori, *Nonlinear Control Systems* II, 1st ed., Springer-Verlag, London, 1999.

[9] Z.-P. Jiang, I. Mareels, D.J. Hill, and J. Huang, *A unifying framework for global regulation via nonlinear output feedback: From ISS to iISS*, IEEE Trans. Automat. Control, 49 (2004), pp. 549–562.

[10] P. Krishnamurthy and F. Khorrami, *Dynamic high-gain scaling: State and output feedback with application to systems with ISS appended dynamics driven by all states*, IEEE Trans. Automat. Control, 49 (2004), pp. 2219–2239.

[11] H. Logemann and A.D. Mawby, *Low-gain integral control of infinite dimensional regular linear systems subject to input hysteresis*, in Advances in Mathematical Systems Theory, F. Colonius, U. Helmke, D. Prätzel-Wolters, and F. Wirth, eds., Birkhäuser Boston, Boston, MA, 2000, pp. 255–293.

[12] D.E. Miller and E.J. Davison, *An adaptive controller which provides an arbitrarily good transient and steady-state response*, IEEE Trans. Automat. Control, 36 (1991), pp. 68–81.

[13] L. Praly and Z.P. Jiang, *Linear output feedback with dynamic high gain for nonlinear systems*, Systems Control Lett., 53 (2004), pp. 107–116.

[14] E.P. Ryan and C.J. Sangwin, *Controlled functional differential equations and adaptive stabilization*, Internat. J. Control, 74 (2001), pp. 77–90.

[15] C. Sparrow, *The Lorenz Equations: Bifurcations, Chaos and Strange Attractors*, Springer-Verlag, New York, 1982.

[16] G. Weiss, *Transfer functions of regular linear systems, part* 1: *Characterization of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[17] X. Ye, *Universal λ-tracking for nonlinearly-perturbed systems*, Automatica, 35 (1999), pp. 109–119.

[18] X. Ye, *Switching adaptive output-feedback control of nonlinearly parameterized systems*, Automatica, 41 (2005), pp. 983–989.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 231–252

INFINITE HORIZON RISK SENSITIVE CONTROL OF DISCRETE TIME MARKOV PROCESSES UNDER MINORIZATION PROPERTY*

GIOVANNI B. DI MASI† AND ŁUKASZ STETTNER‡

Abstract. Risk sensitive control of Markov processes satisfying the minorization property is studied using splitting techniques. Existence of solutions to the multiplicative Poisson equation is shown. Approximation by uniformly ergodic controlled Markov processes is introduced, which allows us to show the existence of solutions to the infinite horizon risk sensitive Bellman equation.

Key words. risk sensitive control, discrete time Markov processes, splitting, Poisson equation, Bellman equation

AMS subject classifications. 93E20, 60J05, 93C55

DOI. 10.1137/040618631

1. Introduction.

On a probability space $(\Omega, \mathcal{F}, P)$ consider a controlled Markov process $X = (x_n)$ taking values on a complete separable metric state space E endowed with the Borel σ-algebra $\mathcal{E}$. Assume that x_n has a controlled transition operator $P^{a_n}(x_n, \cdot)$, where a_n is the control at time n taking values on a compact metric space U and adapted to the σ-algebra $\sigma\{x_0, x_1, \ldots, x_n\}$.

Let $c : E \times U \to R$ be a continuous and bounded function. Our aim is to minimize the exponential ergodic performance criterion,

$$J_x^{\gamma}\big((a_n)\big) = \lim_{t\to\infty} \sup \frac{1}{t} \ln E_x^{(a_n)} \left\{ \exp\left\{ \sum_{i=0}^{t-1} \gamma c(x_i, a_i) \right\} \right\}, \tag{1}$$

with risk factor $\gamma > 0$. In what follows we shall distinguish the following special classes of admissible controls (a_n): *Markov controls* $\mathcal{U}_M = \{(a_n) : a_n = u_n(x_n)\}$, where $u_n : E \mapsto U$ is a sequence of Borel measurable functions, and *stationary controls* $\mathcal{U}_s = \{(a_n) : a_n = u(x_n)\}$, where $u : E \mapsto U$ is a Borel measurable function.

Consider the following assumptions:

(A1) $\exists_{\beta>0}$ $\exists_{C compact \in \mathcal{E}}$ $\exists_{\nu\in\mathcal{P}(E)}$ with $\nu(C) = 1$ such that $\forall_{A\in\mathcal{E}}$

$$\inf_{x\in C} \inf_{a\in U} P^a(x, A) \geq \beta\nu(A).$$

(A2) C given in (A1) is ergodic, i.e., $\forall_{(a_n)\in\mathcal{U}_M}$ $\forall_{x\in E}$ $E_x^{(a_n)}, \{\tau_C\} < \infty$, where $\tau_C = \inf\{i > 0 : x_i \in C\}$, and furthermore $\sup_{x\in C} E_x^{(a_n)}\{\tau_C\} < \infty$.

In this paper the risk sensitive control problem with cost functional (1) and general state space is studied. The paper generalizes [12] and [13] where uniform ergodicity assumption was required. Instead of uniform ergodicity we require minorization property (A1), which allows us to use splitting technique arguments described in [21] and

*Received by the editors November 10, 2004; accepted for publication (in revised form) October 4, 2006; published electronically April 6, 2007.
http://www.siam.org/journals/sicon/46-1/61863.html

†Dipartimento di Matematica Pura ed Applicata, Università di Padova, Via Belzoni 7, 35131 Padova, Italy, and CNR-LADSEB (dimasi@math.upipd.it).

‡Institute of Mathematics, Polish Academy of Sciences, Sniadeckich 8, 00-956 Warsaw, Poland (stettner@impan.gov.pl). The research of this author was supported by PBZ KBN grant 016/P03/99 and MNiSzW grant 1P03A01328.

[22]. Risk sensitive discrete time control problems have been studied in a number of papers [1], [5], [6], [7], [8], [9], [10], [11], [14], [15], [16] for finite or countable state spaces. The general state space model in discrete time was considered in [12] and [13] only. Financial applications of risk sensitive control problems were introduced in [3] and continued in various papers; see, e.g., [23] and [25] and the references therein. Splitting techniques for controlled problems were used in a number of papers; see, e.g., [20] for the case of average cost per unit time and [4] for partially observed ergodic control problems. The first part of this paper studies the so-called Poisson equation. This equation was considered in particular in [17], [18], [5], [2], and [22]. Here we present a rather simple probabilistic characterization (based on splitting arguments) of its solution, which in some sense generalizes Theorem 5.1 of [22] and Proposition 2.8 of [18], where an operator form of the solution was considered. The main result of the paper concerns the existence of solutions to the Bellman equation corresponding to the risk sensitive control problem with a general state space (under minorization property) and is obtained using a uniformly ergodic approximating process. The idea of the approximation of general Markov processes by Markov processes with nice ergodic properties has been exploited in various papers, particularly in [18], where multiplicative Poisson equations are studied, while in [16] ideas based on the spectral theory are used. In this paper we show that the approximation by uniformly ergodic controlled Markov processes is also a useful tool to study risk sensitive Bellman equations.

2. Splitting of Markov processes. Let $\hat{E} = \{C \times \{0\} \cup C \times \{1\} \cup E \setminus C \times \{0\}\}$ and $\hat{x}_n = (x_n^1, x_n^2) \in \hat{E}$. Given a Markov control $a_n = u_n(x_n^1)$, where $u_n : E \mapsto U$ is a sequence of Borel measurable functions, consider the following Markov process defined on $\hat{E}$ (compare to [22] and [17]):

(i) When $(x_n^1, x_n^2) \in C \times \{0\}$, x_n^1 moves to y according to $(1-\beta)^{-1}(P^{a_n}(x_n^1, dy) - \beta\nu(dy))$, and whenever $y \in C$, x_n^2 is changed into $x_{n+1}^2 = \beta_{n+1}$, where β_n is independently and identically distributed (i.i.d.) and $P\{\beta_n = 0\} = 1 - \beta$, $P\{\beta_n = 1\} = \beta$.

(ii) When $(x_n^1, x_n^2) \in C \times \{1\}$, x_n^1 moves to y according to ν and $x_{n+1}^2 = \beta_{n+1}$.

(iii) When $(x_n^1, x_n^2) \in E \setminus C \times \{0\}$, x_n^1 moves to y according to $P^{a_n}(x_n^1, dy)$, and whenever $y \in C$, x_n^2 is changed into $x_{n+1}^2 = \beta_{n+1}$.

In what follows we shall write that the control (a_n) of $(\hat{x}_n)$ is in the class $\mathcal{U}_M$ whenever there is a sequence of Borel measurable functions $u_n : E \mapsto U$ such that $a_n = u_n(x_n^1)$.

Let $C_0 = C \times \{0\}$, $C_1 = C \times \{1\}$. By direct calculation we obtain the following.

LEMMA 1. *For $n = 1, 2 \ldots$ we have almost surely*

$$P\{\hat{x}_n \in C_0 | \hat{x}_n \in C_0 \cup C_1, \hat{x}_{n-1}, \ldots, \hat{x}_0\} = 1 - \beta,$$

$$P\{\hat{x}_n \in C_1 | \hat{x}_n \in C_0 \cup C_1, \hat{x}_{n-1}, \ldots, \hat{x}_0\} = \beta.$$

Furthermore we have the following.

LEMMA 2. *Under Markov control $(a_n) \in \mathcal{U}_M$ the process $(\hat{x}_n = (x_n^1, x_n^2))$ is Markov with transition operator $\hat{P}^{a_n}(\hat{x}_n, dy)$ defined by* (i)–(iii) *and if $(a_n) \in \mathcal{U}_s$, it has a unique invariant measure $\Psi^{(a_n)}$ given by*

$$\Psi^{(a_n)}(A) = \frac{\hat{E}_z^{(a_n)}\left\{\sum_{i=1}^{\tau_{C_1}} \chi_A(\hat{x}_i)\right\}}{\hat{E}_z^{(a_n)}\{\tau_{C_1}\}}, \tag{2}$$

with $z \in C_1$, for any Borel subset A of $\hat{E}$, where $\hat{E}_z^{(a_n)}$ stands for conditional law of Markov process $\hat{x}_n$ with initial state z. Furthermore the first coordinate (x_n^1) is also a Markov process with transition operator $P^{a_n}(x_n^1, dy)$.

Proof. The Markov property of $(\hat{x}_n)$ follows from the construction (i)–(iii) above of the split Markov process. One can easily verify that $\Psi^{(a_n)}$ is in fact an invariant measure if we know that $\hat{E}_z^{(a_n)}\{\tau_{C_1}\}$ is finite. To show this, notice first that for $x \notin C$ and positive integer m

$$\hat{E}_{(x,0)}^{(a_n)}\{\tau_C \chi_{\tau_C \le m}\} = E_x^{(a_n)}\{\tau_C \chi_{\tau_C \le m}\},$$

where τ_C in the left-hand side is the first time in which (x_n^1) hits C, while in the right-hand side it is the analogous hitting time for (x_n). Therefore by (A2) letting $n \to \infty$ we obtain that for $x \notin C$

$$\hat{E}_{(x,0)}^{(a_n)}\{\tau_C\} = E_x^{(a_n)}\{\tau_C\}. \tag{3}$$

For $\hat{x} = (x^1, 0) \in C_0$ by (i), (3), and (A2) there is an $M > 0$ such that

$$\hat{E}_{\hat{x}}\{\tau_C\} = \frac{P(x,C)-\beta}{1-\beta} + \hat{E}_{\hat{x}}\left\{\chi_{E\setminus C}(x_1^1)(1+\hat{E}_{\hat{x}_1}\{\tau_C\})\right\}$$

$$= \frac{P(x,C)-\beta}{1-\beta} + \hat{E}_{\hat{x}}\left\{\chi_{E\setminus C}(x_1^1)(1+E_{x_1^1}\{\tau_C\})\right\}$$

$$\le \frac{P(x,C)-\beta}{1-\beta} + \frac{1}{1-\beta} E_{x^1}\{\tau_C\} \le M.$$

Let $\tau_1 = \tau$ and $\tau_{n+1} = \tau_n + \theta_{\tau_n} \circ \tau$ for $n = 1, 2 \ldots$. Then for $\hat{x} \in C_0$ we have

$$\hat{E}_{\hat{x}}\{\tau_{C_1}\} = \hat{E}_x\left\{\sum_{i=1}^{\infty} \chi_{C_1^c}(\hat{x}_{\tau_1}) \cdots \chi_{C_1^c}(x_{\tau_{i-1}}) \chi_{C_1}(\hat{x}_{\tau_i}) \tau_i\right\}$$

$$\le \sum_{i=1}^{\infty} \beta(1-\beta)^{i-1} i M = \frac{M}{\beta},$$

and since for $\hat{x} \in C_1$

$$\hat{E}_x\{\tau_{C_1}\} = \beta + \hat{E}_x\left\{\chi_{C_0}(x_1^1)(1+\hat{E}_{(x_1^1,0)}\{\tau_{C_1}\})\right\},$$

we have that $\hat{E}_z^{(a_n)}\{\tau_{C_1}\}$ is in fact finite.

To prove the second statement of Lemma 2 notice that for $A \in \mathcal{E}$

$$\begin{aligned} &P\{x_{n+1}^1 \in A | x_n^1, x_{n-1}^1, \ldots, x_0^1\} \\ &= P\{x_{n+1}^1 \in A | x_n^1, x_n^2 = 0, x_{n-1}^1, \ldots, x_0^1\} P\{x_n^2 = 0 | x_n^1, x_{n-1}^1, \ldots, x_0^1\} \\ &\quad + P\{x_{n+1}^1 \in A | x_n^1, x_n^2 = 1, x_{n-1}^1, \ldots, x_0^1\} P\{x_n^2 = 1 | x_n^1, x_{n-1}^1, \ldots, x_0^1\}. \end{aligned} \tag{4}$$

In the case when $x_n^1 \in C$, (4) is equal to

$$\frac{P^{a_n}(x_n^1, A) - \beta\nu(A)}{1-\beta}(1-\beta) + \beta\nu(A) = P^{a_n}(x_n^1, A).$$

For $x_n^1 \notin C$, (4) is equal to $P^{a_n}(x_n^1, A)$, which completes the proof of the Markov property of (x_n^1). □

The following corollary explains the meaning of the splitting.

COROLLARY 1. *For any bounded Borel measurable function* $f : E^m \mapsto R$, $m = 1, 2, \ldots,$ *and control* $(a_n) \in \mathcal{U}_M$ *we have*

$$E_x^{(a_n)} \{f(x_1, x_2, \ldots, x_m)\} = \hat{E}_{\delta_x^*}^{(a_n)} \left\{f(x_1^1, x_2^1, \ldots, x_m^1)\right\}, \tag{5}$$

where $\delta_x^* = \delta_{(x,0)}$ *for* $x \in E \setminus C$ *and* $\delta_x^* = (1-\beta)\delta_{(x,0)} + \beta\delta_{(x,1)}$ *for* $x \in C$ *and* $\hat{E}_\mu$ *stands for the conditional law of Markov process* $(\hat{x}_n)$ *with initial law* $\mu \in \mathcal{P}(\hat{E})$.

Proof. It follows from (4) that for a bounded Borel measurable $g : E \mapsto R$

$$\begin{aligned}
&\hat{E}\left\{g(x_{i+1}^1)|x_i^1, x_{i-1}^1, \ldots, x_0^1\right\} \\
&= \hat{E}\left\{\hat{E}\left\{g(x_{i+1}^1)|\hat{x}_i, \hat{x}_{i-1}, \ldots, \hat{x}_0\right\}|x_i^1, x_{i-1}^1, \ldots, x_0^1\right\} \\
&= \hat{E}\left\{\hat{E}_{\hat{x}_i}\left\{g(x_1^1)\right\}|x_i^1, x_{i-1}^1, \ldots, x_0^1\right\} = \hat{E}_{\delta^*_{x_i^1}}\left\{g(x_1^1)\right\}. \qquad (6)
\end{aligned}$$

On the other hand by the Markov property of (x_n^1) we have

$$\hat{E}\left\{g(x_{i+1}^1)|x_i^1, x_{i-1}^1, \ldots, x_0^1\right\} = \int_E g(y)P^{a_i}(x_i^1, dy). \tag{7}$$

Consequently applying (6) and (7) to function $f : E^m \mapsto R$ we obtain (5). □

3. Multiplicative Poisson equation. In this section we shall assume that

(A3) $\forall_{(a_n)\in\mathcal{U}_s}$ $\exists_d$ such that $\forall_{x\in\hat{E}}$

$$\hat{E}_x^{(a_n)}\left\{\exp\left\{\sum_{i=1}^{\tau_{C_1}}\left(\gamma c(x_i^1, a_i) - d\right)\right\}\right\} < \infty$$

and for $x \in C_1$

$$\hat{E}_x^{(a_n)}\left\{\exp\left\{\sum_{i=1}^{\tau_{C_1}}\left(\gamma c(x_i^1, a_i) - d\right)\right\}\right\} \geq 1.$$

LEMMA 3. *Under* (A3) *for* $(a_n) \in \mathcal{U}_s$ *there is a unique* $\lambda^\gamma((a_n))$ *such that*

$$\hat{E}_x^{(a_n)}\left\{\exp\left\{\sum_{i=1}^{\tau_{C_1}}\left(\gamma c(x_i^1, a_i) - \lambda^\gamma((a_n))\right)\right\}\right\} = 1 \tag{8}$$

for $x \in C_1$.

Proof. Notice that for $x \in C_1$ the mapping

$$D : \lambda \mapsto \hat{E}_x^{(a_n)}\left\{\exp\left\{\sum_{i=1}^{\tau_{C_1}}\gamma c(x_i^1, a_i) - \lambda\right\}\right\}$$

is strictly decreasing whenever it is finite. Moreover by (A3) we have that $\infty > D(d) \geq 1$. Since $\lim_{b\to\infty} D(b) = 0$ by continuity of D (which follows by the monotone convergence theorem) there is a unique $\lambda^\gamma((a_n))$ for which (8) holds. □

Remark 1. Notice that by letting $d = \inf_{x\in E, a\in U}\gamma c(x, a)$ we have a sufficient condition for (A3) in the form

(D1) $\hat{E}_x^{(a_n)}\{\exp\{\gamma\|c\|_{sp}\tau_{C_1}\}\} < \infty$ for $x \in \hat{E}$,

where $\|c\|_{sp} := \sup_{x\in E, a\in U} c(x,a) - \inf_{x\in E, a\in U} c(x,a)$. In section 6 we shall formulate a sufficient condition for (D1) in terms of the expected value of the functional with respect to the original Markov process (x_n).

For Borel measurable $u : E \mapsto U$ let

$$e^{\hat{w}^u(x)} = \hat{E}_x^u\left\{\exp\left\{\sum_{i=0}^{\tau_{C_1}}\left(\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u)\right)\right\}\right\}, \tag{9}$$

where by $\lambda^\gamma(u)$ we denote the value of $\lambda^\gamma((u(x_n^1))$ and the expected value $\hat{E}_x^u$ stands for $\hat{E}_x^{(u(x_n^1))}$.

By (A3) clearly $\lambda^\gamma(u) \geq d$ and therefore $\hat{w}^u$ is well defined. For a Borel measurable function $\tilde{w} : \hat{E} \mapsto R$ define the operator $\Phi(\tilde{w})$ by the formula

$$e^{\Phi(\tilde{w})} = (1-\beta)\int_C e^{\tilde{w}(x,0)}\nu(dx) + \beta\int_C e^{\tilde{w}(x,1)}\nu(dx) \tag{10}$$

whenever it is well defined. Notice that

$$e^{\Phi(\tilde{w})} = \hat{E}_x\{\exp\{\tilde{w}(\hat{x}_1)\}\} \tag{11}$$

for $x \in C_1$. We have the following analogue of Theorem 5.1 of [22] and Proposition 2.8 of [18].

LEMMA 4. *Function $\hat{w}^u$ defined in* (9) *is a solution to the multiplicative Poisson equation (MPE) for the split Markov process $(\hat{x}_n)$:*

$$e^{\hat{w}^u(x)} = e^{\gamma c(x^1, u(x^1)) - \lambda^\gamma(u)}\int_{\hat{E}} e^{\hat{w}^u(y)}\hat{P}^{u(x^1)}(x, dy). \tag{12}$$

Moreover $\Phi(\hat{w}^u) = 0$, and for any other solution $\tilde{w}^u$ to (12) *we have*

$$\tilde{w}^u(x) - \Phi(\tilde{w}^u) \geq \hat{w}^u(x) \tag{13}$$

with equality for Ψ^u almost all $x \in \hat{E}$. Furthermore, if $\tilde{w}$ and λ satisfy the equation

$$e^{\tilde{w}(x)} = e^{\gamma c(x^1, u(x^1)) - \lambda}\int_{\hat{E}} e^{\tilde{w}(y)}\hat{P}^{u(x^1)}(x, dy), \tag{14}$$

then $\lambda \geq \lambda^\gamma(u)$.

Proof. In fact, using (6) we have

$$\hat{E}_x^u\{\exp\{w(\hat{x}_1)\}\} = \hat{E}_x^u\left\{\chi_{\hat{x}_1\in C_1}\hat{E}_{x_1}^u\left\{\exp\left\{\sum_{i=0}^{\tau_{C_1}}\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u)\right\}\right\}\right\}$$

$$+\ \hat{E}_x^u\left\{\chi_{\hat{x}_1\notin C_1}\hat{E}_{x_1}^u\left\{\exp\left\{\sum_{i=0}^{\tau_{C_1}}\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u)\right\}\right\}\right\} = \hat{E}_x^u\{\chi_{\hat{x}_1\in C_1},$$

$$\exp\left\{\gamma c(x_1^1, u(x_1^1)) - \lambda^\gamma(u)\right\}\} + \hat{E}_x^u\left\{\chi_{\hat{x}_1\notin C_1}\exp\left\{\sum_{i=1}^{\tau_{C_1}}\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u)\right\}\right\}$$

$$= \hat{E}_x^u\left\{\exp\left\{\sum_{i=0}^{\tau_{C_1}}\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u)\right\}\right\}\exp\left\{-(\gamma c(x^1, u(x^1)) - \lambda^\gamma(u))\right\},$$

from which (12) follows. Furthermore we clearly see that $\Phi(\hat{w}) = 0$. If $\tilde{w}^u$ is a solution to (12), then by (11) $\Phi(\tilde{w}^u)$ is well defined and by iteration, for a positive integer $m = 1, 2, \ldots,$

$$e^{\tilde{w}^u(x)} = \hat{E}^u_x \left\{ \exp\left\{ \sum_{i=0}^{\tau_{C_1}\wedge m} \left(\gamma c(x^1_i, u(x^1_i)) - \lambda^\gamma(u)\right)\right\} \hat{E}^u_{\hat{x}_{\tau_{C_1}\wedge m}} \left\{\exp\left\{\tilde{w}^u(\hat{x}(1))\right\}\right\}\right\}.$$

By Fatou's lemma

$$e^{\tilde{w}^u(x)} \geq \hat{E}^u_x \left\{ \exp\left\{ \sum_{i=0}^{\tau_{C_1}} \left(\gamma c(x^1_i, u(x^1_i)) - \lambda^\gamma(u)\right)\right\} \hat{E}^u_{\hat{x}_{\tau_{C_1}}} \left\{\exp\left\{\tilde{w}^u(\hat{x}(1))\right\}\right\}\right\}$$

$$= \hat{E}^u_x \left\{ \exp\left\{ \sum_{i=0}^{\tau_{C_1}} \left(\gamma c(x^1_i, u(x^1_i)) - \lambda^\gamma(u)\right)\right\}\right\} e^{\Phi(\tilde{w}^u)},$$

from which we immediately obtain (13). For $\varepsilon > 0$ let

$$Z_\varepsilon = \left\{ x \in \hat{E} : \tilde{w}^u(x) - \Phi(\tilde{w}^u) \geq \hat{w}^u(x) + \varepsilon \right\}.$$

If $\Psi^u(Z_\varepsilon) > 0$, then there is a positive integer n such that $\hat{P}^u_z\left\{\hat{x}_{\tau_{C_1}\wedge n}\right\} > 0$ for $z \in C_1$. Consequently

$$e^{\tilde{w}^u(z) - \Phi(\tilde{w}^u)} = \hat{E}^u_z \left\{ \exp\left\{ \sum_{i=0}^{\tau_{C_1}\wedge n-1} \left(\gamma c(x^1_i, u(x^1_i)) - \lambda^\gamma(u)\right)\right\} \exp\left\{\tilde{w}^u(\hat{x}_{\tau_{C_1}\wedge n})\right\}\right\}$$

$$> \hat{E}^u_z \left\{ \exp\left\{ \sum_{i=0}^{\tau_{C_1}\wedge n-1} \left(\gamma c(x^1_i, u(x^1_i)) - \lambda^\gamma(u)\right)\right\} \exp\left\{\hat{w}^u(\hat{x}_{\tau_{C_1}\wedge n})\right\}\right\} = e^{\hat{w}^u(z)}.$$

Since by (11) we have that $e^{\tilde{w}^u(z) - \Phi(\tilde{w}^u)} = e^{\hat{w}^u(z)} = e^{\gamma c(z^1, u(z^1)) - \lambda^\gamma(u)}$ we obtained a contradiction. Consequently $\Psi^u(Z_\varepsilon) = 0$ for each $\varepsilon > 0$.

Assume now that $\tilde{w}$ and λ satisfy (14). Let $\tau_1 = \tau_{C_1}$, $\tau_{n+1} = \tau_n + \theta_{\tau_n} \circ \tau_{C_1}$ for $n = 1, 2, \ldots$. Then for a positive integer m

$$e^{\tilde{w}(x)} = \hat{E}^u_x \left\{ \exp\left\{ \sum_{i=0}^{\tau_n\wedge m} \left(\gamma c(x^1_i, u(x^1_i)) - \lambda\right)\right\} \hat{E}^u_{\hat{x}_{\tau_n\wedge m}} \left\{\exp\left\{\tilde{w}(\hat{x}(1))\right\}\right\}\right\},$$

and letting $m \to \infty$, by Fatou's lemma

$$e^{\tilde{w}(x)} \geq \hat{E}^u_x \left\{ \exp\left\{ \sum_{i=0}^{\tau_n} \left(\gamma c(x^1_i, u(x^1_i)) - \lambda\right)\right\} \hat{E}^u_{\hat{x}_{\tau_n}} \left\{\exp\left\{\tilde{w}(\hat{x}(1))\right\}\right\}\right\}$$

$$= \hat{E}^u_x \left\{ \exp\left\{ \sum_{i=0}^{\tau_{C_1}} \left(\gamma c(x^1_i, u(x^1_i)) - \lambda\right)\right\}\right\} D^{n-1}(\lambda) e^{\Phi(\tilde{w})},$$

with function D defined in the proof of Lemma 3. Since the right-hand side of the above inequality remains bounded as $n \to \infty$ we should have $D(\lambda) \leq 1$, and this means that $\lambda \geq \lambda^\gamma(u)$. □

There is a $1:1$ correspondence between the solutions to the MPE for the split Markov process and the solution to the MPE corresponding to the original Markov process.

COROLLARY 2. *Given a solution $\tilde{w}^u : \hat{E} \mapsto R$ and λ to the MPE* (14) *we have that w^u defined for $x \in E$ by*

$$e^{w^u(x)} := e^{\tilde{w}^u(x,0)} + 1_C(x)\beta(e^{\tilde{w}^u(x,1)} - e^{\tilde{w}^u(x,0)}) \tag{15}$$

is a solution to the MPE for the original Markov process (x_n)

$$e^{w^u(x)} = e^{\gamma c(x,u(x))-\lambda} \int_E e^{w^u(y)} P^{u(x)}(x, dy). \tag{16}$$

Furthermore if w^u is a solution to (16)*, then $\tilde{w}^u$ defined by*

$$e^{\tilde{w}^u(x^1,x^2)} = e^{\gamma c(x^1,u(x^1))-\lambda} \hat{E}_{x^1,x^2}\left\{e^{w^u(x_1^1)}\right\} \tag{17}$$

is a solution to (14).

Proof. By Lemma 1 we have

$$\begin{aligned}\hat{E}_x^u\left\{e^{\tilde{w}^u(\hat{x}_1)}\right\} &= \hat{E}_x^u\left\{\hat{E}_x^u\left\{e^{\tilde{w}^u(\hat{x}_1)}|x_1^1\right\}\right\}\\ &= \hat{E}_x^u\left\{\chi_C(x_1^1)((1-\beta)e^{\tilde{w}^u(x_1^1,0)} + \beta e^{\tilde{w}^u(x_1^1,1)})\right.\\ &\left. + \chi_{E\setminus C}(x_1^1)e^{\tilde{w}^u(x_1^1,0)}\right\} = \hat{E}_x^u\left\{e^{w^u(x_1^1)}\right\}. \end{aligned} \tag{18}$$

Therefore by (14) taking into account (5) we obtain that w^u defined in (15) is a solution to (16). Assume now that w^u is a solution to (16). Then by (5)

$$\hat{E}_{\delta_x^*}^u\left\{e^{w^u(x_1^1)}\right\} = E_x^u\left\{e^{w^u(x_1^1)}\right\},$$

and for $\tilde{w}^u$ given in (17) we obtain (15). From (15) we obtain (18), which in turn by (17) shows that $\tilde{w}^u$ is a solution to (14). □

Furthermore we have the following.

COROLLARY 3. *If $\lambda(u)$ and w^u is a solution to the MPE*

$$e^{w^u(x)} = e^{\gamma c(x,u(x))-\lambda(u)} \int_E e^{w^u(y)} P^{u(x)}(x, dy),$$

w^u is a bounded Borel measurable function, and the family

$$\left\{\exp\left\{\sum_{i=1}^{\tau_{C_1}\wedge N}\left(\gamma c(x_i^1, u(x_i^1)) - \lambda(u)\right)\right\}; N = 1, 2, \ldots\right\}$$

is uniformly integrable with respect to $\hat{P}_z^u$ with $z \in C_1$, then $\lambda(u) = \lambda^\gamma(u)$.

Proof. By Corollary 2 the function $\tilde{w}^u$ defined in (17) and $\lambda(u)$ is a solution to the MPE (14). Clearly $\tilde{w}$ is also a bounded function. Moreover,

$$e^{\tilde{w}^u(x)} = \hat{E}_x^u\left\{\exp\left\{\sum_{i=0}^{\tau_{C_1}\wedge N}\left(\gamma c(x_i^1, u(x_i^1)) - \lambda(u)\right)\right\}\hat{E}_{\hat{x}_{\tau_{C_1}\wedge N}}^u\left\{\exp\left\{\tilde{w}^u(\hat{x}(1))\right\}\right\}\right\}.$$

By the dominated convergence theorem, letting $N \to \infty$ and using uniform integrability of $\{\exp\{\sum_{i=1}^{\tau_{C_1}\wedge N}(\gamma c(x_i^1, u(x_i^1)) - \lambda(u))\}; N = 1, 2, \ldots\}$ and the fact that $\tilde{w}$ is bounded we obtain

$$e^{\tilde{w}^u(z)} = \hat{E}_z^u \left\{ \exp\left\{ \sum_{i=0}^{\tau_{C_1}} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda(u)\right) \right\} \right\} e^{\Phi(\tilde{w}^u)}$$

for $z \in C_1$. Consequently, since by the MPE we have that $\tilde{w}^u(z) - \Phi(\tilde{w}^u) = c(z^1, u(z^1)) - \lambda(u)$, for $z = (z^1, 1) \in C_1$ we obtain

$$\hat{E}_z^u \left\{ \exp\left\{ \sum_{i=1}^{\tau_{C_1}} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda(u)\right) \right\} \right\} = 1,$$

from which, by Lemma 3, we immediately have that $\lambda(u) = \lambda^\gamma(u)$. □

Remark 2. Notice that $\lambda^\gamma(u)$ is the minimal solution to the MPE (14). However, solutions may be well defined $\forall\, \lambda \geq \lambda^\gamma(u)$, as shown in the example below.

Example. Consider the following model: $E = \{0, 1, 2, \ldots\}$, $\gamma = 1$, there is no control, $c(x) \equiv c$ for $x \in E$, and the transition probabilities are $P(0, 1) = 1 - \alpha_0$, $P(0, 0) = \alpha_0$, $P(x, 0) = \alpha$, $P(x, x + 1) = 1 - \alpha$ for $x = 1, 2, \ldots$, with $0 < \alpha_0, \alpha < 1$. The MPE Bellman equation is of the form

$$e^{w(x)} = e^{c-\lambda} E_x \left\{ e^{w(x_1)} \right\}. \tag{19}$$

Clearly the pair $\lambda = c$ and $w \equiv 0$ is a solution to (19). We shall identify other solutions. Taking into account the form of transition probabilities, from (19) we obtain

$$e^{w(0)} = e^{c-\lambda} \left(\alpha_0 e^{w(0)} + (1 - \alpha_0) e^{w(1)} \right),$$

and for $n = 1, 2, \ldots$

$$e^{w(n)} = e^{c-\lambda} \left(\alpha e^{w(0)} + (1 - \alpha) e^{w(n+1)} \right).$$

The first equation has a solution whenever

$$\lambda > c + \ln \alpha_0. \tag{20}$$

Iterating the second equation we obtain that

$$e^{w(n)} = \frac{1}{1 - \alpha} \left(e^{\lambda - c + \kappa} q^{n-2} - \alpha(q^{n-2} + \cdots + 1) \right) e^{w(0)}, \tag{21}$$

with $\kappa = \ln(\frac{e^{\lambda-c} - \alpha_0}{1-\alpha_0})$ and $q = \frac{e^{\lambda-c}}{1-\alpha}$. Whenever $q \leq 1$ the right-hand side of (21) is negative for a sufficiently large n, so that there are no solutions. Whenever $q > 1$ we have

$$e^{w(n)} = \frac{w^{w(0)}}{(1 - \alpha)(q - 1)} \left((e^{\lambda - c + \kappa}(q - 1) - \alpha q) q^{n-2} + \alpha \right), \tag{22}$$

and the above equation has a solution only when $e^{\lambda-c+\kappa}(q-1) - \alpha q \geq 0$. Taking into account the form of κ, this leads to the quadratic inequality

$$e^{2(\lambda-c)} - (\alpha_0 + 1 - \alpha) e^{\lambda - c} + \alpha_0 - \alpha \geq 0,$$

from which we in turn obtain that $e^{\lambda-c} \leq \alpha_0 - \alpha$ or $e^{\lambda-c} \geq 1$. In view of (20) the first inequality is not satisfied. From the second inequality we obtain $\lambda \geq c$, and then we clearly see that (20) also holds. Finally, $\forall\, \lambda \geq c$ we have solutions to (14) which are unbounded whenever $\lambda > c$.

The value $\lambda^\gamma(u)$ defined in Lemma 3 has (under some additional assumptions) the following important interpretation.

PROPOSITION 1. *If for Borel measurable $u : E \mapsto U$*

(B1) *the family*

$$\left\{ \exp\left\{ \sum_{i=1}^{\tau_{C_1}\wedge N} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u)\right) \right\}; N = 1, 2, \dots \right\} \tag{23}$$

is uniformly integrable with respect to $\hat{P}_z^u$ measure with $z \in C_1$,

(B2) *for $x \in \hat{E}$*

$$\inf_N \hat{E}_x^u \left\{ \exp\left\{ \sum_{i=1}^{\sigma_{C_1}\wedge N-1} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u)\right) \right\} \right\} > 0 \tag{24}$$

with $\sigma_{C_1} = \inf\{i \geq 0 : \hat{x}(i) \in C_1\}$, and

(B3) *for $x \in \hat{E}$*

$$\sup_N \hat{E}_x^u \left\{ \exp\left\{ \sum_{i=1}^{\sigma_{C_1}\wedge N-1} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u)\right) \right\} \right\} < \infty, \tag{25}$$

then for $x \in E$

$$\lambda^\gamma(u) = \lim_{n\to\infty} \frac{1}{n} \ln E_x^u \left\{ \exp\left\{ \sum_{i=0}^{n-1} \gamma c(x_i, u(x_i)) \right\} \right\}. \tag{26}$$

Proof. Let $\lambda > \lambda^\gamma(u)$. For $z \in C_1$ we have

$$\hat{E}_z^u \left\{ \exp\left\{ \sum_{i=1}^{\tau_{C_1}} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda\right) \right\} \right\} < 1$$

and consequently by the dominated convergence theorem and (23) for $N \geq N_0$, with N_0 sufficiently large,

$$\hat{E}_z^u \left\{ \exp\left\{ \sum_{i=1}^{\tau_{C_1}\wedge N} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda\right) \right\} \right\} \leq 1. \tag{27}$$

Let

$$e^{w_N^u(x)} = \hat{E}_x^u \left\{ \exp\left\{ \sum_{i=0}^{\sigma_{C_1}\wedge N-1} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda\right) \right\} \right\}. \tag{28}$$

By (B3) w_N^u is well defined. For $x \notin C_1$

$$e^{w^u_{N+1}(x)} = \hat{E}^u_x \left\{ e^{\gamma c(x_0^1, u(x_0^1)) - \lambda} \hat{E}^u_{\hat{x}_1} \left\{ \exp\left\{ \sum_{i=0}^{\sigma_{C_1} \wedge N-1} \big(\gamma c(x_i^1, u(x_i^1)) - \lambda\big) \right\} \right\} \right\}$$

$$= \hat{E}^u_x \left\{ e^{\gamma c(x_0^1, u(x_0^1)) - \lambda} e^{w^u_N(\hat{x}_1)} \right\}, \tag{29}$$

and for $x \in C_1$ by (27) we have

$$e^{w^u_{N+1}(x)} = e^{\gamma c(x_0^1, u(x_0^1)) - \lambda}$$

$$\geq e^{\gamma c(x_0^1, u(x_0^1)) - \lambda} \hat{E}^u_x \left\{ \exp\left\{ \sum_{i=1}^{\sigma_{C_1} \wedge N} \big(\gamma c(x_i^1, u(x_i^1)) - \lambda\big) \right\} \right\}$$

$$= \hat{E}^u_x \left\{ e^{\gamma c(x_0^1, u(x_0^1)) - \lambda} e^{w^u_N(\hat{x}_1)} \right\}. \tag{30}$$

Consequently

$$e^{w^u_{N+1}(x)} \geq \hat{E}^u_x \left\{ e^{\gamma c(x_0^1, u(x_0^1)) - \lambda} e^{w^u_N(\hat{x}_1)} \right\},$$

and by iteration for $N \geq N_0$

$$e^{w^u_{N+k}(x)} \geq \hat{E}^u_x \left\{ e^{\sum_{i=0}^{k-1} (\gamma c(x_i^1, u(x_i^1)) - \lambda)} e^{w^u_N(\hat{x}_k)} \right\}$$

$$\geq \hat{E}^u_x \left\{ e^{\sum_{i=0}^{k-1} (\gamma c(x_i^1, u(x_i^1)) - \lambda)} e^{-\gamma \|c\| N} \right\}.$$

Therefore

$$\frac{1}{k} \ln \hat{E}^u_x \left\{ e^{\gamma \sum_{i=0}^{k-1} c(x_i^1, u(x_i^1))} \right\} \leq \frac{1}{k} \gamma \|c\| N$$

$$+ \frac{1}{k} \sup_N \ln \hat{E}^u_x \left\{ \exp\left\{ \sum_{i=0}^{\sigma_{C_1} \wedge N-1} \big(\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u)\big) \right\} \right\} + \lambda,$$

and by (B3)

$$\limsup_{k \to \infty} \frac{1}{k} \ln \hat{E}^u_x \left\{ e^{\gamma \sum_{i=0}^{k-1} c(x_i^1, u(x_i^1))} \right\} \leq \lambda.$$

Consequently, letting λ decrease to $\lambda^\gamma(u)$, we obtain

$$\limsup_{k \to \infty} \frac{1}{k} \ln \hat{E}^u_x \left\{ e^{\gamma \sum_{i=0}^{k-1} c(x_i^1, u(x_i^1))} \right\} \leq \lambda^\gamma(u). \tag{31}$$

Assume now that $\lambda < \lambda^\gamma(u)$. By Fatou's lemma for $z \in C_1$ and $N \geq N_0$, with N_0 sufficiently large, we have

$$\hat{E}^u_z \left\{ \exp\left\{ \sum_{i=1}^{\tau_{C_1} \wedge N} \big(\gamma c(x_i^1, u(x_i^1)) - \lambda\big) \right\} \right\} \geq 1. \tag{32}$$

Therefore for w_N^u given by (28) (with λ as above) similarly as in (29) and (30) we have

$$e^{w_{N+1}^u(x)} \le \hat{E}_x^u \left\{ e^{\gamma c(x_0^1, u(x_0^1)) - \lambda} e^{w_N^u(\hat{x}_1)} \right\}, \tag{33}$$

and by iteration for $N \ge N_0$

$$\begin{aligned} e^{w_{N+k}^u(x)} &\le \hat{E}_x^u \left\{ e^{\sum_{i=0}^{k-1} (\gamma c(x_i^1, u(x_i^1)) - \lambda)} e^{w_N^u(\hat{x}_k)} \right\} \\ &\le \hat{E}_x^u \left\{ e^{\sum_{i=0}^{k-1} (\gamma c(x_i^1, u(x_i^1)) - \lambda)} e^{\gamma \|c\| N} \right\}. \end{aligned}$$

Therefore

$$\begin{aligned} &\frac{1}{k} \ln \hat{E}_x^u \left\{ e^{\gamma \sum_{i=0}^{k-1} c(x_i^1, u(x_i^1))} \right\} \ge -\frac{1}{k} \gamma \|c\| N \\ &+ \frac{1}{k} \inf_N \ln \hat{E}_x^u \left\{ \exp \left\{ \sum_{i=0}^{\sigma_{C_1} \wedge N - 1} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u) \right) \right\} \right\} + \lambda, \end{aligned}$$

and by (B2)

$$\liminf_{k \to \infty} \frac{1}{k} \ln \hat{E}_x^u \left\{ e^{\gamma \sum_{i=0}^{k-1} c(x_i^1, u(x_i^1))} \right\} \ge \lambda,$$

and finally

$$\liminf_{k \to \infty} \frac{1}{k} \ln \hat{E}_x^u \left\{ e^{\gamma \sum_{i=0}^{k-1} c(x_i^1, u(x_i^1))} \right\} \ge \lambda^\gamma(u), \tag{34}$$

which together with (31) using (5) completes the proof. □

Remark 3. Notice that under (D1) assumptions (B1) and (B3) are satisfied. By the Hölder inequality, using the fact that $\lambda^\gamma(u) \le \sup_{x \in E, a \in U} \gamma c(x, a)$ we have that

$$\begin{aligned} 1 \le &\left(\hat{E}_x^u \left\{ \exp \left\{ \sum_{i=1}^{\sigma_{C_1} \wedge N - 1} \left(-\gamma c(x_i^1, u(x_i^1)) + \lambda^\gamma(u) \right) \right\} \right\} \right)^{\frac{1}{2}} \\ &\times \left(\hat{E}_x^u \left\{ \exp \left\{ \sum_{i=1}^{\sigma_{C_1} \wedge N - 1} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u) \right) \right\} \right\} \right)^{\frac{1}{2}} \le \left(\hat{E}_x^u \left\{ e^{\gamma \|c\|_{sp} \tau_{C_1}} \right\} \right)^{\frac{1}{2}} \\ &\times \left(\hat{E}_x^u \left\{ \exp \left\{ \sum_{i=1}^{\sigma_{C_1} \wedge N - 1} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u) \right) \right\} \right\} \right)^{\frac{1}{2}}, \end{aligned}$$

from which, under (D1), inequality (24) holds, and in conclusion (26) holds.

Furthermore notice that

$$M_N := e^{-\hat{w}^u(z)} \exp \left\{ \sum_{i=0}^{\tau_{C_1} \wedge N} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u) \right) \right\} e^{\hat{w}^u(x_{\tau_{C_1} \wedge N + 1})}$$

is a positive martingale that converges almost surely to M_∞ as $N \to \infty$. Since by the choice of $\lambda^\gamma(u)$ we have that $\hat{E}_z^u M_N = 1 = \hat{E}_z^u M_\infty$, from Theorem II 21 of [19] it follows that the family $\{M_N, N = 1, 2, \ldots\}$ is uniformly integrable. When $\hat{w}^u$ is bounded we then immediately obtain (B1).

Finally, by the Jensen inequality we can easily formulate a sufficient condition for (B2) as follows: for $x \in \hat{E}$

$$\inf_N \hat{E}_x^u \left\{ \sum_{i=1}^{\sigma_{C_1} \wedge N-1} \left(\gamma c(x_i^1, u(x_i^1)) - \lambda^\gamma(u) \right) \right\} > -\infty.$$

4. Uniformly ergodic approximation of controlled Markov processes. We shall now assume that

(A4) for $x \in E$, $A \in \mathcal{E}$

$$P^a(x, A) = \int_A p(x, a, y)\nu(dy), \tag{35}$$

where p is a positive continuous function of its coordinates.

Denote by $|x|$ the value of $\rho(x, \theta)$, where ρ is a metric on E compatible with the topology of E and $\theta \in E$ is a fixed point.

Let

$$\tilde{p}_N(x, a, y) = \begin{cases} \frac{p(x,a,y)}{\Delta_N^a(x)} & \text{for} \quad |y| \le N, \\ \frac{p(\theta,\bar{a},y)}{\Delta_N^a(x)} & \text{for} \quad |y| \ge N+1, \\ \frac{p(x,a,y)(N+1-|y|)+p(\theta,\bar{a},y)(|y|-N)}{\Delta_N^a(x)} & \text{elsewhere,} \end{cases}$$

with $\Delta_N^a(x) = P^a(x, B_N) + P^{\bar{a}}(\theta, B_{N+1}^c) + \int_{B_{N+1}\setminus B_N} [p(x, a, y)(N+1-|y|) + p(\theta, \bar{a}, y)(|y| - N)]\nu(dy)$, where $B_N = \{x \in E : |x| \le N\}$ and $\bar{a}$ is a fixed element of U.

Let

$$p_N(x, a, y) = \tilde{p}_N(x, a, y) \text{ if } |x| \le N,$$
$$p_N(x, a, y) = \tilde{p}_N\left(\frac{x}{|x|}N, a, y\right) \text{ for } |x| > N$$

and define

$$P_N^a(x, dy) = p_N(x, a, y)\nu(dy). \tag{36}$$

We clearly have the following.

LEMMA 5.

$$\sup_{a \in U} \|P_N^a(x, \cdot) - P^a(x, \cdot)\|_{var} \to 0 \tag{37}$$

as $N \to \infty$, uniformly in x from compact sets. Furthermore for each N

$$\sup_{a,a' \in U} \sup_{x,x' \in E} \sup_{y \in E} \frac{p_N(x, a, y)}{p_N(x', a', y)} < \infty. \tag{38}$$

For $(a_n) \in \mathcal{U}_s$ let

$$F_{Nx}^{(a_n)}(\lambda) = \hat{E}_x^{(a_n),N} \left\{ \exp\left\{ \sum_{i=1}^{\tau_{C_1}} (\gamma c(x_i^1, a_i) - \lambda) \right\} \right\} \tag{39}$$

and

$$F_x^{(a_n)}(\lambda) = \hat{E}_x^{(a_n)}\left\{\exp\left\{\sum_{i=1}^{\tau_{C_1}}(\gamma c(x_i^1, a_i) - \lambda)\right\}\right\}, \tag{40}$$

where $\hat{P}_x^{(a_n),N}$ is the conditional probability of the split Markov process $(\hat{x}_n)$ corresponding to Markov process (x_n) with transition probability $P_N^{a_n}$ at time n.

Notice that whenever $x \in C_1$ the functions in (39) and (40) do not depend on x and will be denoted by $F_N^{(a_n)}$ and $F^{(a_n)}$. We have the following.

PROPOSITION 2. *Assume that there exist N_0 and $d_1 < d_2$ such that for $N \geq N_0$, $(a_n) \in \mathcal{U}_s$, and $x \in C_1$ we have*

$$F_N^{(a_n)}(d_2) = F_{Nx}^{(a_n)}(d_2) \leq 1 \leq F_{Nx}^{(a_n)}(d_1) < \infty, \tag{41}$$

$F_{Nx}^{(a_n)}(\lambda) \to F_x^{(a_n)}(\lambda)$ for $x \in C_1$ uniformly in $(a_n) \in \mathcal{U}_s$ and $\lambda \in [d_1, d_2]$, and furthermore

$$\sup_{(a_n)} |F^{(a_n)\prime}(d_1)| < \infty, \tag{42}$$

where $'$ stands for the derivative with respect to λ. Then

$$\lambda_N^\gamma((a_n)) := \left(F_N^{(a_n)}\right)^{-1}(1) \to \left(F^{(a_n)}\right)^{-1}(1) = \lambda^\gamma((a_n)) \tag{43}$$

uniformly in $(a_n) \in \mathcal{U}_s$ as $N \to \infty$.

Proof. Assume that there exist $\varepsilon > 0$, a sequence (a_n^k) of strategies from $\mathcal{U}_s$, and a sequence $N_k \to \infty$ such that

$$|\lambda_{N_k}^\gamma((a_n^k)) - \lambda^\gamma((a_n^k))| > \varepsilon. \tag{44}$$

By assumption we have that

$$|F_{N_k x}^{(a_n^k)}(\lambda_{N_k}((a_n^k))) - F_x^{(a_n^k)}(\lambda_{N_k}((a_n^k)))| \to 0$$

and therefore

$$F_x^{(a_n^k)}(\lambda_{N_k}^\gamma((a_n^k))) \to 1 = F_{N_k x}^{(a_n^k)}(\lambda_{N_k}^\gamma((a_n^k))) \tag{45}$$

as $k \to \infty$. Since $F_x^{(a_n^k)}(\lambda^\gamma((a_n^k))) = 1$ and $\sup_{(a_n^k)} |F^{(a_n^k)\prime}(\lambda)|$ is bounded for $\lambda \in [d_1, d_2]$ (by (42)), we should have $|\lambda_{N_k}^\gamma((a_n^k)) - \lambda^\gamma((a_n^k))| \to 0$ as $k \to \infty$, which contradicts (44). □

Remark 4. Notice that the choice of d_1 and d_2 in (41) is uniform with respect to $(a_n) \in \mathcal{U}_s$. Two natural candidates are $d_1 = \inf_{x \in E, a \in U} \gamma c(x, a)$ and $d_2 = \sup_{x \in E, a \in U} \gamma c(x, a)$.

To have convergence $F_{Nx}^{(a_n)}(\lambda) \to F_x^{(a_n)}(\lambda)$ for $x \in C_1$ uniform in $(a_n) \in \mathcal{U}_s$ and $\lambda \in [d_1, d_2]$ we have to assume for $x \in C_1$ that

$$\sup_{(a_n) \in \mathcal{U}_s} \hat{E}_x^{(a_n)}\left\{\exp\left\{\gamma \|c\|_{sp} \tau_{C_1}\right\}\right\} < \infty,$$

and

(D2) $\sup_N \sup_{(a_n)\in\mathcal{U}_s} \hat{E}_x^{(a_n),N}\{\exp\{\gamma\|c\|_{sp}\tau_{C_1}\}\} < \infty$.

Since

$$|F_x^{(a_n)\prime}(\lambda)| = \hat{E}_x^{(a_n)}\left\{\tau_{C_1}\exp\left\{\sum_{i=1}^{\tau_{C_1}}(\gamma c(x_i^1,h_i)-\lambda)\right\}\right\}$$

$$\leq \hat{E}_x^{(a_n)}\{\tau_{C_1}\exp\{\gamma\|c\|_{sp}\tau_{C_1}\}\}$$

$$\leq K\hat{E}_x^{(a_n)}\{\exp\{(1+\varepsilon)\gamma\|c\|_{sp}\tau_{C_1}\}\}$$

for $\varepsilon > 0$ and $K > 0$, to have (42) it is sufficient to assume for $x \in C_1$ that

(D3) $\sup_{(a_n)\in\mathcal{U}_s} \hat{E}_x^{(a_n)}\{\exp\{(1+\varepsilon)\gamma\|c\|_{sp}\tau_{C_1}\}\} < \infty$ for a sufficiently small $\varepsilon > 0$.

We have the following.

PROPOSITION 3. *For each N there are λ_N^γ and $w_N \in C(E)$ such that*

$$e^{w_N(x)} = \inf_{a\in U}\left[e^{\gamma c(x,a)-\lambda_N^\gamma}\int_E e^{w_N(y)}P_N^a(x,dy)\right], \tag{46}$$

and consequently

$$\lambda_N^\gamma = \inf_{a_n} J_x^{\gamma,N}((a_n)), \tag{47}$$

where

$$J_x^{\gamma,N}((a_n)) := \limsup_{t\to\infty}\frac{1}{t}\ln E_x^{(a_n),N}\left\{\exp\left\{\sum_{i=0}^{t-1}\gamma c(x_i,a_i)\right\}\right\}$$

and the infimum is taken over all admissible controls (a_n).

Moreover the strategy $\hat{a}_n^N = u_N(x_n)$, where $u_N : E \mapsto U$ is a Borel measurable function for which the infimum in equation (46) *is attained, is optimal and if $\hat{E}_x^{(\hat{a}_n^N),N}\{\exp\{\gamma\|c\|_{sp}\tau_{C_1}\}\} < \infty$, we have additionally that $\lambda_N^\gamma = \lambda_N^\gamma(u_N)$ with $\lambda_N^\gamma(u_N)$ defined in Lemma* 3 *for a Markov process with transition operator $P_N^{u_N}$.*

Proof. In view of (38) and Theorem 1 of [12] it remains only to show that $\lambda_N^\gamma = \lambda_N^\gamma(u_N)$, which in turn follows directly from Corollary 3. □

COROLLARY 4. *Under* (D2) *and* (D3) *we have that*

$$\lambda^\gamma := \inf_{(a_n)\in\mathcal{U}_s} J_x^\gamma((a_n)) = \lim_{N\to\infty}\lambda_N^\gamma \tag{48}$$

for $x \in E$.

Proof. By Remark 4 and Proposition 2

$$\sup_{(a_n)\in\mathcal{U}_s}|\lambda_N^\gamma((a_n)) - \lambda^\gamma((a_n))| \to 0.$$

Since by Remark 3 and Propositions 1 and 2 $\lambda_N^\gamma((a_n)) = \inf_{(a_n)\in\mathcal{U}_s}\lambda_N^\gamma((a_n)) = \lambda_N^\gamma$, we obtain (48). □

5. Risk sensitive Bellman equation. Let u_N be an optimal control function corresponding to $P_N^a(x,dy)$. Furthermore assume that

(A5) $\exists_{\epsilon>0}$ such that $\forall_{K\ compact\subset \hat{E}}$

$$\sup_{a\in U}\sup_{x\in K}\sup_{N}\hat{E}_x^{a,N}\left\{\exp\left\{\sum_{i=1}^{\tau_{C_1}}(\gamma c(x_i^1,u_N(x_i^1))-\lambda_N^\gamma(u_N))(1+\epsilon)\right\}\right\}$$

$$= M(K)<\infty, \tag{49}$$

where above we control the split Markov process $(\hat{x}_n)$ using at time 0 control $a_0=a$ and then $a_n=u_N(x_n^1)$ for $n\geq 1$, with u_N as in Proposition 3.

THEOREM 1. *Under* (A1)–(A5) *there exist* λ^γ *and a continuous function* $w: E\mapsto R$ *such that*

$$e^{w(x)}=\inf_{a\in U}\left[e^{\gamma c(x,a)-\lambda^\gamma}\int_E e^{w(y)}P^a(x,dy)\right]. \tag{50}$$

Moreover, under (D2)–(D3) *we have that*

$$\lambda^\gamma=\inf_{(a_n)\in\mathcal{U}_s}J_x^\gamma((a_n))=\lim_{N\to\infty}\lambda_N^\gamma(u_N). \tag{51}$$

Assuming additionally that (*D*1) *is satisfied for* $\hat{a}_n=\hat{u}(x_n)$, *where* $\hat{u}$ *is a Borel measurable function for which the infimum on the right-hand side of* (50) *is attained, we have that* $\lambda^\gamma=\lambda^\gamma(\hat{u})$. *Furthermore, if for an admissible control* (a_n) *we have*

$$\limsup_{t\to\infty}E_x^{(a_n)}\left\{\left(E_{x_t}^{a_t}\left\{e^{w(x_1)}\right\}\right)^\alpha\right\}<\infty$$

for every $\alpha>1$, *then* $\lambda^\gamma\leq J_x^\gamma((a_n))$.

Proof. The proof consists of several steps.

Step 1. We prove first that $\sup_N\hat{E}_x^{a,N}\{\exp\{\hat{w}_N^{u_N}(\hat{x}_1\}\}$ is bounded uniformly on compact subsets of $(E_0\cup C_1)\times U$, where $\hat{w}_N^{u_N}$ is a solution to the multiplicative Poisson equation corresponding to the transition operator $P_N^a(x,dy)$ with control function u_N with $\Phi(\hat{w}_N^{u_N})=0$.

In fact,

$$\hat{E}_x^{a,N}\{\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\}=\hat{E}_x^{a,N}\left\{\chi_{C_1}(\hat{x}_1)e^{\gamma c(x_1^1,u_N(x_1^1))-\lambda_N^\gamma(u_N)}\right\}$$

$$+\ \hat{E}_x^{a,N}\left\{\chi_{C_1^c}(\hat{x}_1)\exp\left\{\sum_{i=1}^{\tau_{C_1}}(\gamma c(x_i^1,u_N(x_i^1))-\lambda_N^\gamma(u_N))\right\}\right\}, \tag{52}$$

and by (A5) the required boundedness follows.

Step 2. We show now that for $N=1,2\ldots$, functions $\hat{E}_x^{a,N}\{\chi_{C_1}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\}$, $\hat{E}_x^{a,N}\{\chi_{C_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\}$, and $\hat{E}_x^{a,N}\left\{\chi_{(E\setminus C)_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\right\}$ are equicontinuous in x and a from compact subsets of $E_0\cup C_1$ and U, respectively.

Notice first that by (37) for each compact set $K\subset E_0\cup C_1$, $\varepsilon'>0$ there is a compact set $K_1\supset C_0\cup C_1$ such that

$$\sup_{a\in U}\sup_{x\in K}\sup_{N}\hat{P}_x^{aN}\{\hat{x}_1\in K_1^c\}<\varepsilon'. \tag{53}$$

Furthermore by the Hölder inequality

$$\sup_{a\in U}\sup_{x\in K}\sup_{N}\hat{E}_x^{a,N}\left\{\chi_{K_1^c}(\hat{x}_1)\exp\left\{\sum_{i=1}^{\tau_{C_1}}(\gamma c(x_i^1,u_N(x_i^1))-\lambda_N^\gamma(u_N))\right\}\right\} \tag{54}$$

$$\leq \sup_{a\in U}\sup_{x\in K}\sup_{N}\left(\hat{P}_x^{a,N}\{\hat{x}_1\in K_1^c\}\right)^{\frac{\varepsilon}{1+\varepsilon}},$$

$$\sup_{a\in U}\sup_{x\in K}\sup_{N}\left(\hat{E}_x^{a,N}\left\{\exp\left\{\sum_{i=1}^{\tau_{C_1}}(\gamma c(x_i^1,u_N(x_i^1))-\lambda_N^\gamma(u_N))(1+\epsilon)\right\}\right\}\right)^{\frac{1}{1+\varepsilon}}$$

$$\leq {\varepsilon'}^{\frac{\varepsilon}{1+\varepsilon}}(M(K))^{\frac{1}{1+\varepsilon}}.$$

Consequently

$$|\hat{E}_x^{a,N}\{\chi_{C_1}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\}-\hat{E}_{x'}^{a',N}\{\chi_{C_1}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\}|$$

$$\leq e^{|\gamma|\|c\|}\|\hat{P}^{aN}(x,C_1\cap\cdot)-\hat{P}^{a'N}(x',C_1\cap\cdot)\|_{var}$$

and by (53)–(54)

$$|\hat{E}_x^{a,N}\{\chi_{C_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\}-\hat{E}_{x'}^{a',N}\{\chi_{C_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\}|$$

$$\leq 2{\varepsilon'}^{\frac{\varepsilon}{1+\varepsilon}}(M(K))^{\frac{1}{1+\varepsilon}}+|\hat{E}_x^{a,N}\{\chi_{K_1}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\}$$

$$-\hat{E}_{x'}^{a',N}\{\chi_{K_1}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1')\}\}|$$

and

$$|\hat{E}_x^{a,N}\left\{\chi_{(E\setminus C)_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\right\}-\hat{E}_{x'}^{a',N}\left\{\chi_{(E\setminus C)_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\right\}|$$

$$\leq 2{\varepsilon'}^{\frac{\varepsilon}{1+\varepsilon}}(M(K))^{\frac{1}{1+\varepsilon}}+|\hat{E}_x^{a,N}\left\{\chi_{K_1\cap(E\setminus C)_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\right\}$$

$$-\hat{E}_{x'}^{a',N}\left\{\chi_{K_1\cap(E\setminus C)_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1')\}\right\}|.$$

For $\delta>0$ choose K_1 in (52) such that ${\varepsilon'}^{\frac{\varepsilon}{1+\varepsilon}}(M(K))^{\frac{1}{1+\varepsilon}}<\frac{\delta}{3}$. Since

$$\max\Big\{|\hat{E}_x^{a,N}\{\chi_{K_1}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\}-\hat{E}_{x'}^{a',N}\{\chi_{K_1}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1')\}\}|,$$

$$|\hat{E}_x^{a,N}\left\{\chi_{K_1\cap(E\setminus C)_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\right\}-\hat{E}_{x'}^{a',N}\left\{\chi_{K_1\cap(E\setminus C)_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1')\}\right\}|\Big\}$$

$$\leq \sup_{x\in K_1}\exp\{\hat{w}_N^{u_N}(x)\}\,\|\hat{P}^{aN}(x,K_1\cap\cdot)-\hat{P}^{a'N}(x',K_1\cap\cdot)\|_{var}$$

for x, $x'\in E_0\cup C_1$, and a, $a'\in U$ such that

$$\|\hat{P}^{aN}(x,C_1\cap\cdot)-\hat{P}^{a'N}(x',C_1\cap\cdot)\|_{var}\leq\frac{\delta}{3e^{\gamma\|c\|}} \tag{55}$$

and

$$\|\hat{P}^{aN}(x,K_1\cap\cdot)-\hat{P}^{a'N}(x',K_1\cap\cdot)\|_{var}\leq\frac{\delta}{3\sup_{z\in K_1}e^{\hat{w}_N^{u_N}(z)}} \tag{56}$$

we obtain that

$$|\hat{E}_x^{a,N}\{\chi_{C_1}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\} - \hat{E}_{x'}^{a',N}\{\chi_{C_1}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\}| \le \delta,$$

$$|\hat{E}_x^{a,N}\{\chi_{C_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\} - \hat{E}_{x'}^{a',N}\{\chi_{C_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\}| \le \delta,$$

and

$$|\hat{E}_x^{a,N}\left\{\chi_{(E\setminus C)_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\right\} - \hat{E}_{x'}^{a',N}\left\{\chi_{(E\setminus C)_0}(\hat{x}_1)\exp\{\hat{w}_N^{u_N}(\hat{x}_1)\}\right\}| \le \delta.$$

Now by (A5) $\sup_{z\in K_1} e^{\hat{w}_N^{u_N}(z)}$ is bounded in N, and therefore by (37) we can choose x, x' and a, a' in (55) and (56) uniformly in N, which completes the proof of the equicontinuity.

Step 3. By Steps 1 and 2 and by (5) and (15) we immediately see that

$$E_x^{a,N}\{\exp\{w_N^{u_N}(x_1)\}\}$$

is uniformly (in N) bounded and equicontinuous on compact subsets of $E\times U$. Since u_N is optimal for $P_N^a(x,dy)$ we have that $w_N^{u_N} = w_N$. Therefore by Ascoli's theorem (Theorem 33 of [24]) there is a subsequence N_k such that

$$E_x^{a,N_k}\{\exp\{w_{N_k}(x_1)\}\}$$

converges uniformly in $a\in U$ and x from compact subsets of E and $\lambda_{N_k}^{\gamma}(u_{N_k})\to\lambda$ (since $\lambda_N^{\gamma}(u_n)\in[\inf_{x\in E,a\in U}\gamma c(x,a), \sup_{x\in E,a\in U}\gamma c(x,a)]$). Consequently there is a continuous function w such that

$$e^{w(x)} = \inf_{a\in U}\left[e^{\gamma c(x,a)-\lambda}\lim_{k\to\infty}\int_E e^{w_{N_k}(y)}P_{N_k}^a(x,dy)\right]. \tag{57}$$

Moreover, using the fact that $w_N^{u_N} = w_N$ is a solution to the Bellman equation (46) we obtain

$$\begin{aligned} e^{w(x)} &= \lim_{k\to\infty}\inf_{a\in U}\left[e^{\gamma c(x,a)-\lambda}\int_E e^{w_{N_k}(y)}P_{N_k}^a(x,dy)\right] \\ &= \lim_{k\to\infty} e^{\lambda-\lambda_{N_k}^{\gamma}(u_{N_k})}e^{w_{N_k}(x)} = \lim_{k\to\infty} e^{w_{N_k}(x)} \end{aligned} \tag{58}$$

with the convergence uniform on compact sets.

Step 4. To prove that function w defined in (57) is a solution to the Bellman equation (50) it remains to show that

$$\lim_{k\to\infty} E_x^{a,N_k}\{\exp\{w_{N_k}(x_1)\}\} = E_x^a\left\{e^{w(x_1)}\right\}. \tag{59}$$

In fact, by Fatou's lemma

$$E_x^a\left\{e^{w(x_1)}\right\} \le \lim_{k\to\infty} E_x^{a,N_k}\{\exp\{w_{N_k}(x_1)\}\} < \infty. \tag{60}$$

By Steps 1 and 2 one can find a compact set $K_1\supset C$ such that

$$\sup_N\sup_{a\in U} E_x^{a,N}\left\{\chi_{K_1^c}(x_1)\exp\{w_N(x_1)\}\right\} \le \frac{\varepsilon}{3} \tag{61}$$

and

$$\sup_{a\in U} E_x^a\left\{\chi_{K_1^c}(x_1)\exp\{w(x_1)\}\right\} \le \frac{\varepsilon}{3}. \tag{62}$$

Therefore

$$\begin{aligned}
&|E_x^a\{\exp\{w(x_1)\}\} - E_x^{a,N_k}\{\exp\{w_{N_k}(x_1)\}\}|\\
&\le |E_x^a\{\chi_{K_1}(x_1)\exp\{w(x_1)\}\} - E_x^{a,N_k}\{\chi_{K_1}(x_1)\exp\{w(x_1)\}\}|\\
&\quad + |E_x^{a,N_k}\{\chi_{K_1}(x_1)(\exp\{w(x_1)\} - \exp\{w_{N_k}(x_1)\})\}|\\
&\quad + E_x^{a,N_k}\left\{\chi_{K_1^c}(x_1)\exp\{w_{N_k}(x_1)\}\right\} + E_x^a\left\{\chi_{K_1^c}(x_1)\exp\{w(x_1)\}\right\}\\
&\le \sup_{x\in K_1} e^{w(x)}\|P^a(x,K_1\cap\cdot) - P^{aN}(x,K_1\cap\cdot)\|_{var} + \sup_{x\in K_1}|e^{w(x)} - e^{w_{N_k}(x)}| + \frac{2\varepsilon}{3}.
\end{aligned}$$

Consequently letting $k\to\infty$ and taking into account that ε may be arbitrarily small we obtain the convergence (59). By the continuity in x and a of the right-hand side of (50) we have the existence of a Borel measurable function $\hat{u}$ for which the infimum is attained. Identity (51) follows immediately from Corollary 4. From Remark 3 and Proposition 1 (under (D1)) we obtain that $\lambda^\gamma = \lambda^\gamma(\hat{u})$.

Step 5. If for an admissible control (a_n) we have $\limsup_{t\to\infty} E_x^{(a_n)}\{(E_{x_t}^{a_t}\{e^{w(x_1)}\})^\alpha\} < \infty$ for every $\alpha > 1$, then by the Hölder inequality we have from (50)

$$\begin{aligned}
w(x) &\le \ln E_x^{(a_n)}\left\{\exp\left\{\sum_{i=0}^{t-1}(\gamma c(x_i,a_i) - \lambda^\gamma)\right\} E_{x_t}^{a_t}\left\{e^{w(x_1)}\right\}\right\}\\
&\le -t\lambda^\gamma + \ln\left(E_x^{(a_n)}\left\{\exp\left\{\sum_{i=0}^{t-1}\gamma c(x_i,a_i)(1+\epsilon)\right\}\right\}\right)^{\frac{1}{1+\epsilon}}\\
&\quad + \ln\left(E_x^{(a_n)}\left\{\left(E_{x_t}^{a_t}\left\{e^{w(x_1)}\right\}\right)^{1+\frac{1}{\epsilon}}\right\}\right)^{\frac{\epsilon}{1+\epsilon}}.
\end{aligned}$$

Dividing both sides of the last inequality by t and letting t go to infinity, we obtain that $\frac{1}{1+\epsilon}J_x^{\gamma(1+\epsilon)}((a_n)) \ge \lambda^\gamma$ for any $\epsilon > 0$. It remains to show that the mapping $\gamma \mapsto J_x^\gamma((a_n))$ is a continuous function for $\gamma > 0$ since then letting $\epsilon\to 0$ we obtain $J_x^\gamma((a_n)) \ge \lambda^\gamma$. To prove continuity notice that for $\gamma_1,\gamma_2 > 0$, $\gamma_1 \le \gamma_2$

$$\begin{aligned}
|J_x^{\gamma_1}((a_n)) - J_x^{\gamma_2}((a_n))| &\le \limsup_{t\to\infty}\frac{1}{t}\left|\ln E_x^{(a_n)}\left\{\exp\left\{\sum_{i=0}^{t-1}\gamma_1 c(x_i,a_i)\right\}\right\}\right.\\
&\left. - \ln E_x^{(a_n)}\left\{\exp\left\{\sum_{i=0}^{t-1}\gamma_2 c(x_i,a_i)\right\}\right\}\right| \le \limsup_{t\to\infty}\frac{1}{t}|\gamma_1-\gamma_2|\sup_{\gamma\in[\gamma_1,\gamma_2]}|g_t'(\gamma)|\\
&\le \|c\||\gamma_1-\gamma_2|,
\end{aligned}$$

since the derivative of the function

$$g_t(\gamma) := \ln E_x^{(a_n)}\left\{\exp\left\{\sum_{i=0}^{t-1}\gamma c(x_i,a_i)\right\}\right\}$$

is bounded by $t\|c\|$. □

Remark 5. A sufficient condition for (A5) can be formulated as follows:

(D4) There is $\varepsilon > 0$ such that for each compact set $K \subset \hat{E}$

$$\sup_{a\in U}\sup_{x\in K}\sup_{N} \hat{E}_x^{a,N}\left\{\exp\left\{(1+\varepsilon)|\gamma|\|c\|_{sp}\tau_{C_1}\right\}\right\} < \infty,$$

where the split Markov process $(\hat{x}_n)$ after control a at time 0 is controlled using the control function u_N.

6. Remarks on assumptions and an example. We shall formulate first a sufficient condition for (A3). It is worth noting that the assumption (63) below corresponds to geometrical regularity of the set C with a constant $\gamma\|c\|_{sp}$, as considered in [18].

PROPOSITION 4. *If for $x \in E$ and $(a_n) \in \mathcal{U}_s$*

$$E_x^{(a_n)}\left\{\exp\left\{\gamma\|c\|_{sp}\tau_C\right\}\right\} < \infty \tag{63}$$

and

$$\sup_{x\in C} E_x^{(a_n)}\left\{\exp\left\{\gamma\|c\|_{sp}\tau_C\right\}\right\} - \beta\exp\left\{\gamma\|c\|_{sp}\right\} < 1, \tag{64}$$

then (A3) *holds.*

Proof. Notice first that by Corollary 1 for $z \notin C$ and positive integer m we have

$$\hat{E}_{(z,0)}^{(a_n)}\left\{\exp\left\{\gamma\|c\|_{sp}\tau_C\right\}\chi_{\tau_C\le m}\right\} = E_z^{(a_n)}\left\{\exp\left\{\gamma\|c\|_{sp}\tau_C\right\}\chi_{\tau_C\le m}\right\}. \tag{65}$$

Letting $m \to \infty$ by (63) we obtain

$$\hat{E}_{(z,0)}^{(a_n)}\left\{\exp\left\{\gamma\|c\|_{sp}\tau_C\right\}\right\} = E_z^{(a_n)}\left\{\exp\left\{\gamma\|c\|_{sp}\tau_C\right\}\right\}. \tag{66}$$

Now for $(a_n) \in \mathcal{U}_M$ and $x \in C$, using the definition of split Markov process and (64) we have

$$\begin{aligned}
&\hat{E}_{(x,0)}^{(a_n)}\left\{\exp\left\{\gamma\|c\|_{sp}\tau_C\right\}\right\} \\
&= \hat{E}_{(x,0)}^{(a_n)}\left\{\chi_C(x_1^1)e^{\gamma\|c\|_{sp}} + \chi_{C^c}(x_1^1)\hat{E}_{\hat{x}_1}^{(a_n)}\left\{\exp\left\{\gamma\|c\|_{sp}\tau_C\right\}\right\}\right\} \\
&= e^{\gamma\|c\|_{sp}}\frac{P^{a_0}(x,C)-\beta}{1-\beta} + e^{\gamma\|c\|_{sp}}\int_{C^c} E_z^{(a_n)}\left\{\exp\left\{\gamma\|c\|_{sp}\tau_C\right\}\right\}\frac{P^{a_0}(x,dz)}{1-\beta} \\
&= \frac{1}{1-\beta}\left[e^{\gamma\|c\|_{sp}}(P^{a_0}(x,C)-\beta) - E_x^{(a_n)}\left\{\chi_C(x_1)e^{\gamma\|c\|_{sp}}\right\} + E_x\left\{e^{\gamma\|c\|_{sp}\tau_C}\right\}\right] \\
&< \frac{1}{1-\beta}.
\end{aligned} \tag{67}$$

Let $\tau_1 = \tau_C := \{i \ge 0 : x_i^1 \in C\}$, $\tau_{n+1} = \tau_n + \tau_1 \circ \theta_{\tau_n}$.

For $x \in \hat{E}$ and $L = \sup_{z\in C}\hat{E}_{(z,0)}^{(a_n)}\left\{e^{\gamma\|c\|_{sp}\tau_C}\right\}$, using Lemma 1 we have

$$\begin{aligned}
\hat{E}_x^{(a_n)}\left\{e^{\gamma\|c\|_{sp}\tau_{C_1}}\right\} &= \hat{E}_x\left\{\sum_{i=1}^{\infty}\chi_{C_1^c}(\hat{x}_{\tau_1})\dots\chi_{C_1^c}(x_{\tau_{i-1}})\chi_{C_1}(\hat{x}_{\tau_i})e^{\gamma\|c\|_{sp}\tau_i}\right\} \\
&\le \sum_{i=1}^{\infty}\hat{E}_x^{(a_n)}\left\{\chi_{C_1^c}(\hat{x}_{\tau_1})\dots\chi_{C_1^c}(\hat{x}_{\tau_{i-1}})e^{\gamma\|c\|_{sp}\tau_{i-1}}\right\}L\beta \\
&\le \hat{E}_x^{(a_n)}\left\{e^{\gamma\|c\|_{sp}\tau_C}\right\}\sum_{i=1}^{\infty}(1-\beta)^{i-1}\beta L^{i-1} = \hat{E}_x^{(a_n)}\left\{e^{\gamma\|c\|_{sp}\tau_C}\right\}\frac{\beta}{1-(1-\beta)L}.
\end{aligned}$$

Taking into account (66), (63), (64) we obtain (D1), which completes the proof. □

Taking Remark 4 into account, by the proof of Proposition 4 we easily obtain a sufficient condition for (A5).

COROLLARY 5. *If there is an $\varepsilon > 0$ such that for any compact set $K \subset E$, we have*

$$\sup_{a\in U}\sup_{x\in K}\sup_{N} E_x^{a,N}\left\{\exp\left\{(1+\varepsilon)\gamma\|c\|_{sp}\tau_C\right\}\right\} < \infty \tag{68}$$

and

$$\sup_{a\in U}\sup_{x\in C}\sup_{N} E_x^{a,N}\left\{\exp\left\{(1+\varepsilon)\gamma\|c\|_{sp}\tau_C\right\}\right\} - \beta\exp\left\{\varepsilon\gamma\|c\|_{sp}\right\} < 1, \tag{69}$$

where the Markov process (x_n) is controlled using constant a at time 0 and $a_n = u_N(x_n)$ afterwards, then (A5) *holds.*

Consequently we see that the assumptions imposed in the paper are satisfied for a class of processes for which $f(\gamma) := E_x\{e^{\gamma\tau_C}\}$ is finite, provided we choose γ sufficiently small (to guarantee (64) and (69)). As an example one can consider a discretized ergodic diffusion (x_n) in R^d given by the following equation:

$$x_{n+1} = x_n + Ax_n + b(x_n, a_n) + D(x_n, a_n)w_n, \tag{70}$$

where (w_n) is a sequence of i.i.d. standard normal random vectors in R^d, A is a stable matrix, $b(x,a)$ is a continuous bounded vector function of $x \in R^d$ and $a \in U$, and $D(x,a)$ is a continuous bounded matrix-valued function which is uniformly elliptic, i.e., $\inf_{x\in R^d}\inf_{a\in U} trD(x,a)D(x,a)^T > 0$. Notice that by the nondegeneracy of D the minorization property (A1) is satisfied for any closed ball C in R^d. The stability of the matrix A and boundedness of b imply that if C is sufficiently large, the controlled process, no matter what control is used, is pushed to C. For completeness we add the following Lyapunov-type criterion (more detailed analysis of geometric regularity assumption can be found in Chapter 15 of [21]).

LEMMA 6. *If for $(a_n) \in \mathcal{U}_s$*

$$\sup_{x\notin C} E_x^{(a_n)}\left\{\|x_{\tau_C}\|^{-1}\right\} < \infty \tag{71}$$

and for $\gamma > 0$

$$\sup_{x\notin C}\sup_{a\in U} E_x^a\{\|x_1\|\} \le e^{-4\gamma}\|x\|, \tag{72}$$

then

$$E_x^{(a_n)}\{e^{\gamma\tau_C}\} < \infty. \tag{73}$$

Proof. Define a Lyapunov function $V(s,x) := e^{2\gamma(s+1)}\|x\|$. For $x \notin C$ by (72) we have

$$E_x^{(a_n)}\{V(s+1,x_1)\} - V(s,x) \le e^{2\gamma(s+2)}E_x^{(a_n)}\{\|x_1\|\} - e^{2\gamma(s+1)}\|x\|$$
$$\le -(e^{2\gamma(s+1)} - e^{2\gamma s})\|x\|.$$

Consequently

$$E_{x_m}^{(a_n)}\{V(m+1,x_1)\} - V(m,x_m) \le -(e^{2\gamma(m+1)} - e^{2\gamma m})\|x_m\|,$$

$$E^{(a_n)}_{x_{m-1}}\{V(m,x_1)\} - V(m-1,x_{m-1}) \le -(e^{2\gamma m} - e^{2\gamma(m-1)})\|x_{m-1}\|,$$

and summing up the above inequalities till the process (x_n) enters the set C and taking the expected value, we obtain

$$E^{(a_n)}_{x_{\tau_C}}\{V(\tau,x_\tau)\} - V(0,x) \le -E^{(a_n)}_x\left\{\|x_{\tau_C}\|(e^{2\gamma\tau_C}-1)\right\}, \tag{74}$$

from which by nonnegativity of V we have that

$$E^{(a_n)}_x\left\{e^{2\gamma\tau_C}\|x_{\tau_C}\|\right\} \le V(0,x). \tag{75}$$

By the Hölder inequality

$$E^{(a_n)}_x\{e^{\gamma\tau_C}\} \le \left(E^{(a_n)}_x\left\{e^{2\gamma\tau_C}\|x_{\tau_C}\|\right\}\right)^{\frac{1}{2}}\left(E^{(a_n)}_x\left\{\|x_{\tau_C}\|^{-1}\right\}\right)^{\frac{1}{2}},$$

and from (71) we obtain (73). □

In section 4 we introduced a uniformly ergodic approximation. One can consider more general, i.e., v-separable, approximations, as studied in [18]. Following Theorem 2.4 of [18] we then have that the Lyapunov drift criterion (DV3) holds and consequently the local multiplicative mean ergodic Theorem 3.4 of [18] is satisfied. The transition kernel of the original Markov process (x_n), although quite regular under assumption (A4), may not be itself v-separable. The latter property would require uniform approximation of the transition kernel, while we can show only an approximation which is uniform on compact subsets. A useful sufficient condition for v-separability is formulated in Lemma B.3 of [18].

Acknowledgments. The authors would like to thank both reviewers and the Associate Editor for valuable comments and references and in particular for pointing out a gap in the proofs of Lemma 4 and Theorem 1.

REFERENCES

[1] G. Avila-Goday and E. Fernández-Gaucherand, *Controlled Markov chains with exponential risk sensitive criteria: Modularity, structured policies and applications*, in Proceedings of the 37th IEEE Conference on Decision and Control (CDC), Tampa, FL, 1998, pp. 778–783.

[2] S. Balaji and S. P. Meyn, *Multiplicative ergodicity and large deviations for an irreducible Markov chain*, Stochastic Process. Appl., 90 (2000), pp. 123–144.

[3] T. R. Bielecki and S. Pliska, *Risk sensitive dynamic asset management*, Appl. Math. Optim., 39 (1999), pp. 337–360.

[4] V. S. Borkar and A. Budhiraja, *A further remark on dynamic programming for partially observed Markov processes*, Stochastic Process. Appl., 112 (2004), pp. 79–93.

[5] V. S. Borkar and S. P. Meyn, *Risk-sensitive optimal control for Markov decision processes with monotone cost*, Math. Oper. Res., 27 (2002), pp. 192–209.

[6] A. Brau-Rojas, R. Cavazos-Cadena, and E. Fernández-Gaucherand, *Controlled Markov chains with risk sensitive criteria: Some (counter) examples*, in Proceedings of the 37th IEEE Conference on Decision and Control (CDC), Tampa, FL, 1998, pp. 1853–1858.

[7] R. Cavazos-Cadena, *Solution to the risk-sensitive average cost optimality in a class of Markov decision processes with finite state space*, Math. Methods Oper. Res., 57 (2003), pp. 263–285.

[8] R. Cavazos-Cadena and E. Fernandez-Gaucherand, *Risk-sensitive control in communicating average Markov decision chains,* in Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications, M. Dror, P. L'Ecuyer, and F. Szidarovzky, eds., Kluwer Academic, Dordrecht, The Netherlands, 2002, pp. 515–553.

[9] R. Cavazos-Cadena and E. Fernandez-Gaucherand, *Controlled Markov chains with risk sensitive criteria: Average cost, optimality equations, and optimal solutions*, Math. Methods Oper. Res., 49 (1999), pp. 299–324.

[10] R. Cavazos-Cadena and E. Fernandez-Gaucherand, *The vanishing discount approach in a stable Markov reward process with a risk-sensitive average cost criterion*, IEEE Trans. Automat. Control, 45 (2000), pp. 1800–1816.
[11] R. Cavazos-Cadena and D. Hernandez-Hernandez, *Solution to the risk-sensitive average optimality equation in communicating Markov decision chains with finite state space: An alternative approach*, Math. Methods Oper. Res., 56 (2002), pp. 473–479.
[12] G. B. Di Masi and L. Stettner, *Risk-sensitive control of discrete-time Markov processes with infinite horizon*, SIAM J. Control Optim., 38 (1999), pp. 61–78.
[13] G. B. Di Masi and L. Stettner, *Infinite horizon risk sensitive control of discrete time Markov processes with small risk*, Systems Control Lett., 40 (2000), pp. 305–321.
[14] W. H. Fleming and D. Hernández-Hernández, *Risk-sensitive control of finite state machines on an infinite horizon* I, SIAM J. Control Optim., 35 (1997), pp. 1790–1810.
[15] D. Hernandez-Hernandez and S. J. Marcus, *Risk sensitive control of Markov processes in countable state space*, Systems Control Lett., 29 (1996), pp. 147–155.
[16] W. Huisinga, S. P. Meyn, and Ch. Schutte, *Phase transitions and metastability in Markovian and molecular systems*, Ann. Appl. Probab., 14 (2004), pp. 419–458.
[17] I. Kontoyiannis and S. P. Meyn, *Spectral theory and limit theorems for geometrically ergodic Markov processes*, Ann. Appl. Probab., 13 (2003), pp. 304–362.
[18] I. Kontoyiannis and S. P. Meyn, *Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes*, Electron. J. Probab., 10 (2005), pp. 61–123.
[19] P. A. Meyer, *Probability and Potentials*, Blaisdell, Waltham, MA, 1966.
[20] S. P. Meyn, *Stability, performance evaluation, and optimization*, in Markov Decision Processes: Models, Methods, Directions, and Open Problems, E. Feinberg and A. Schwartz, eds., Kluwer Academic, Dordrecht, The Netherlands, 2001, pp. 43–82.
[21] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer, London, 1993.
[22] E. Nummelin, *General Irreducible Markov Chains and Nonnegative Operators*, Cambridge University Press, Cambridge, UK, 1984.
[23] H. Pham, *Large deviations approach to optimal long term investment*, Finance Stoch., 7 (2003), pp. 169–195.
[24] H. L. Royden, *Real Analysis*, Macmillan, New York, 1968.
[25] L. Stettner, *Duality and risk sensitive portfolio optimization*, in Mathematics of Finance, Proceedings of the AMS-IMS-SIAM Joint Summer Res. Conference, G. Yin and Q. Zhang, eds., AMS, Providence, RI, 2004, pp. 333–347.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 253–273

STABILITY AND UNIQUENESS FOR THE CRACK IDENTIFICATION PROBLEM*

ZAKARIA BELHACHMI† AND DORIN BUCUR†

Abstract. This paper deals with the identifiability of nonsmooth defects by boundary measurements, and the stability of their detection. We introduce and analyze a new pointwise regularity concept at the boundary of an open set which turns out to play a crucial role in the identifiability of defects by two boundary measurements. As a consequence, we prove the unique identifiability for a large class of closed sets, including sets with an infinite number of connected components of positive capacity and totally disconnected sets. In order to rigorously justify numerical approximation results of defects by optimal design methods, we prove a geometric stability result of the defect identification process, without any a priori smoothness assumptions.

Key words. defect identification, conductivity, uniqueness, stability

AMS subject classifications. 35R30, 49Q10

DOI. 10.1137/S0363012904441179

1. Introduction.

This paper deals with the defect identification problem by boundary measurements. Roughly speaking, the problem can be formulated as follows: Given a smooth bounded domain $\Omega \subseteq \mathbb{R}^2$, find a closed set $K \subseteq \Omega$ knowing the traces on the boundary $w_{i|\partial\Omega}$ of the solutions of

$$\left\{ \begin{array}{ll} -\Delta w_i & = 0 \text{ in } \Omega \setminus K, \\ \frac{\partial w_i}{\partial n} & = 0 \text{ on } \partial K, \\ \frac{\partial w_i}{\partial n} & = \psi_i \text{ on } \partial\Omega \end{array} \right. \tag{1}$$

for *several inputs* ψ_i. We refer the reader to the paper [4] for a complete review of the most important and up-to-date results concerning this problem.

There are three main challenges when dealing with such a problem:

- The uniqueness of the defect for a given number of measures: May different defects give the same measures?
- Stability with respect to the measurements: Do close measures give "close" cracks? There is a subsequent question: What is the right sense of closeness for defects: close in geometry or in behavior (like γ-convergence)?
- (Numerical) reconstruction of the defects and rigorous convergence results.

A way to tackle this geometric inverse problem is to use optimal design methods. From this viewpoint, one needs to understand the three items above in the context of minimal regularity assumptions for defects. Dropping a priori regularity assumptions for the stability purpose allows us, for example, to give a formal justification to the convergence of the approximation process by a shape optimization approach.

The purpose of this paper is twofold. First, we introduce and analyze a new pointwise regularity concept at the boundary of an open set, called *conductivity*, which plays a crucial role in the identifiability of defects by two boundary measurements.

*Received by the editors February 16, 2004; accepted for publication (in revised form) October 9, 2006; published electronically April 6, 2007.
http://www.siam.org/journals/sicon/46-1/44117.html

†Département de Mathématiques, UMR-CNRS 7122, Université de Metz, Ile du Saulcy, 57045 Metz Cedex 01, France (belhach@math.univ-metz.fr, bucur@math.univ-metz.fr).

It is known that one measure cannot uniquely determine even a smooth curve K, and, following Alessandrini and Diaz Valenzuela [1], two suitably chosen inputs can uniquely determine closed sets K which can be decomposed in a finite union of disjoint continua (see also [17]). Roughly speaking, we prove that unique identification holds for the family of defects which are conductive at quasi-every point of their boundaries (see Theorem 3.9). As a consequence, we prove unique identifiability by two boundary measurements for a large class of closed sets, including sets with an infinite number of connected components of positive capacity and totally disconnected sets. Our proof uses the scheme of Alessandrini and Diaz Valenzuela based on nonexistence of critical points for suitable holomorphic functions. The construction of critical points relies on the conductivity regularity concept. With respect to the proof of Alessandrini and Diaz Valenzuela several new technical difficulties appear, which are related to the fact that harmonic conjugates of solutions are not Hölder continuous up to the boundary and information given by the unique continuation principle cannot be propagated "across" the defects. The conductivity regularity concept has several features in common with the Dirichlet regularity related to the Wiener criterion, but we are not able to prove or disprove their equivalence. Nevertheless, our result associated with the Kellogg property also shows that the equivalence of the two regularity concepts (conductivity and Dirichlet regularity) would straightforwardly imply the conjecture that *all* closed sets are uniquely identifiable, up to a set of zero capacity, by two boundary measurements.

The second purpose of the paper is to investigate the stability of the detection from the shape optimization point of view. Precisely, we prove that asymptotic geometric stability holds in the class of defects having a uniform bound on the number of connected components (see Theorem 4.3). Roughly speaking, convergence of the measures in the space of traces implies geometric convergence of the defects (this is the framework of the so-called Tikhonov principle [20]). All previous stability results (see [2, 4, 11, 18] and references therein) require us to know a priori a quantitative estimate of the smoothness of the defects, and provide quantitative estimates for the stability. By dropping the a priori smoothness hypotheses we lose any quantitative estimate but—and here is the main interest of such a result—we can rigorously justify that suitable numerical approximations of the defects are convergent (see Theorems 5.1 and 5.3). This result is to be compared to the one obtained in [8] for shape optimization problems associated with the Dirichlet–Laplacian and relies deeply on the shape stability result of [6] and on the elimination of the smoothness hypotheses (see also [7, 9, 15]). Stability results based on a priori smoothness cannot be used to achieve shape convergence for numerical approximations in the optimal design framework.

2. Setting the problem. Throughout the paper, Ω denotes a bounded simply connected open set in $\mathbb{R}^2$ with smooth boundary. By $|E|$ we denote the Lebesgue measure of the set E and by cap (E) its capacity, i.e.,

$$\text{cap}\,(E) = \inf\left\{\int_{\mathbb{R}^2} |\nabla u|^2 + |u|^2 dx, \quad u \in \mathcal{U}_E\right\},$$

where $\mathcal{U}_E$ is the class of all functions $u \in C_c^\infty(\mathbb{R}^2)$ such that $u \geq 1$ a.e. in a neighborhood of E. It is said that a property $p(x)$ holds quasi everywhere on E (q.e. on E) if the set of all points $x \in E$ for which $p(x)$ does not hold has capacity zero. We refer to [14] for details concerning capacity.

A function u is said to be *quasi-continuous* if for every $\epsilon > 0$ there exists an open set A_ϵ such that cap $(A_\epsilon) < \epsilon$ and $u|_{\Omega \setminus A_\epsilon}$ is continuous in $\Omega \setminus A_\epsilon$. Throughout the

paper, every time we refer to pointwise properties of Sobolev functions, we implicitly consider quasi-continuous representatives.

The usual Sobolev space is denoted by $H^1(\Omega)$. Recall that every function $u \in H^1(\Omega)$ has a quasi-continuous representative, unique up to a set of zero capacity. Considering quasi-continuous representatives, one can define the trace (as restriction) of a function $u \in H^1(\Omega)$ on every continuum of positive diameter. We recall the following result (see [6]).

LEMMA 2.1. *Let $u \in H^1(\Omega)$ and let K_1, K_2 be two compact connected sets in Ω with positive diameter. If there exist two different constants $c_1, c_2 \in \mathbb{R}$ such that $u(x) = c_1$ q.e. on K_1 and $u(x) = c_2$ q.e. on K_2, then $K_1 \cap K_2 = \emptyset$.*

We also recall the definition of the following functional space. Let $U \subseteq \mathbb{R}^2$ be an open set; the Dirichlet space $\mathcal{L}^{1,2}(U)$ is defined as [14].

$$\mathcal{L}^{1,2}(U) = \{u \in L^2_{loc}(U) : \nabla u \in [L^2(U)]^2\}, \tag{2}$$

where the gradient of u is taken in the sense of distributions. Introducing the equivalence relation

$$u\mathcal{R}v \text{ if } \int_U |\nabla(u-v)|^2 dx = 0,$$

we see that the quotient space $\mathcal{L}^{1,2}(U)_{/\mathcal{R}} := L^{1,2}(U)$ is a Hilbert space for the scalar product

$$(u,v)_{L^{1,2}(U)} = \int_U \nabla u \nabla v dx.$$

Let C be a connected component of U and let $u, v \in \mathcal{L}^{1,2}(U)$ such that $u\mathcal{R}v$. Then $u - v$ is constant a.e. on C.

Following [12, Corollary 2.2], if U is smooth enough (e.g., with Lipschitz continuous boundary), then $\mathcal{L}^{1,2}(U) = H^1(U)$. If U is not smooth, then $H^1(U)$ might be strictly contained in $\mathcal{L}^{1,2}(U)$. Observe also that if U is not smooth enough, several "well-known" properties of H^1-spaces fail to be true, e.g., the Poincaré–Wirtinger inequality.

For an arbitrary set $F \subseteq \mathbb{R}^2$ and for $\varepsilon > 0$ let us denote the dilation of F by ε, $F^\varepsilon = \cup_{x \in F} B(x, \varepsilon)$ being the union of all open balls centered in points of F with radius ε, and denote by $\overline{F}^\varepsilon$ its closure. Clearly the following holds for $\varepsilon < \nu$: $F^\varepsilon = (\overline{F})^\varepsilon \subseteq \overline{(F^\varepsilon)} \subseteq F^\nu$.

DEFINITION 2.2. *The Hausdorff distance between two compact sets $K_1, K_2 \subseteq \mathbb{R}^2$ is defined by*

$$d_H(K_1, K_2) = \inf\{\varepsilon > 0 : K_1 \subseteq K_2^\varepsilon \, , K_2 \subseteq K_1^\varepsilon\}.$$

Note that the family of closed subsets of a given compact of $\mathbb{R}^2$ is compact for the Hausdorff metric. We refer to [6] and [19] for more details on the Hausdorff metric and on the following.

LEMMA 2.3. *Let $\{u_n\}_{n\in\mathbb{N}} \subseteq H^1(\Omega)$, $\{K_n\}_{n\in\mathbb{N}}$ be a sequence of compact connected sets in Ω, and $\{c_n\}_{n\in\mathbb{N}}$ be a sequence of constants such that $u_n(x) = c_n$ q.e. on K_n. If $K_n \xrightarrow{H} K$, then K is connected. Suppose that $u_n \overset{H^1(\Omega)}{\rightharpoonup} u$. Then there exists a constant $c \in \mathbb{R}$ such that $c_n \longrightarrow c$ and $u(x) = c$ q.e. on $K \cap \Omega$.*

Let $K \subset \Omega$ be compact and $\psi \in L^2(\partial\Omega)$ be such that $\int_{\partial\Omega} \psi = 0$. We consider in what follows *the perfectly insulating problem*

$$\begin{cases} -\Delta w &= 0 \text{ in } \Omega \setminus K, \\ \frac{\partial w}{\partial n} &= 0 \text{ on } \partial K, \\ \frac{\partial w}{\partial n} &= \psi \text{ on } \partial\Omega. \end{cases} \tag{3}$$

Since K is the unknown of the problem and may vary, if ambiguity occurs on the choice of K, this solution will be denoted $w_{\psi,K}$. It is clear that using the usual tools (e.g., the Lax–Milgram theorem [3]; see also [6]) one has the following.

PROPOSITION 2.4. *There exists a unique solution $u \in L^{1,2}(\Omega \setminus K)$ obtained by solving the following minimization problem:*

$$\min_{\phi \in L^{1,2}(\Omega\setminus K)} \frac{1}{2} \int_{\Omega\setminus K} |\nabla\phi|^2 dx - \int_{\partial\Omega} \phi\psi d\sigma.$$

Let us denote by 1_U the characteristic function of the set U. The following result has a simple proof (see section 4 and [6, 10]).

PROPOSITION 2.5. *The following holds for $\varepsilon \to 0$:*

$$\nabla w_{\psi,\overline{K}^\varepsilon} 1_{\Omega\setminus\overline{K}^\varepsilon} \overset{L^2(\Omega,\mathbb{R}^2)}{\longrightarrow} \nabla w_{\psi,K} 1_{\Omega\setminus K}.$$

Note also that for $\varepsilon > 0$ the function $w_{\psi,\overline{K}^\varepsilon}$ has a harmonic conjugate in $\Omega\setminus\overline{K}^\varepsilon$ (see [1], for example), i.e., is the real part of a holomorphic function. Following Proposition 2.5 and the usual properties of holomorphic functions, the function $w_{\psi,K}$ also has a harmonic conjugate. The problem which is solved by the conjugate functions will be clarified by studying the following.

The perfectly conducting problem. A dual problem, called "the perfectly conducting case" has been introduced in [11] for one connected crack and extended in [1] for a finite number of connected cracks, say, $K = \cup_{i=1}^k K_i$. In this case, the problem is formulated as follows (see, for instance, [1]):

$$\begin{cases} -\Delta u = 0 \text{ in } \ \Omega \setminus K, \\ u = c_i \text{ q.e. on } K_i, \\ \frac{\partial u}{\partial n} = \psi \text{ on } \partial\Omega. \end{cases} \tag{4}$$

The constants c_i are uniquely determined by the no-flux condition that the solution u has to satisfy: for every smooth Jordan curve $\gamma \subseteq \Omega \setminus K$ $\int_\gamma \frac{\partial u}{\partial n} d\sigma = 0$.

Moreover, the solution of this problem is given by the minimization of the following energy functional:

$$\min \left\{ \frac{1}{2} \int_{\Omega\setminus K} |\nabla u|^2 dx - \int_{\partial\Omega} u\psi d\sigma : u \in H^1(\Omega), u \text{ q.e. constant on } \ K_i \right\}. \tag{5}$$

Details concerning the equivalence of the formulations (4) and (5) can be found in [1] (see also [6] for more details concerning the formulation via quasi-continuous functions).

Here is the main key to understanding the uniqueness of the inverse problem for arbitrary compact sets. We manage the perfectly conductive case for arbitrary K by introducing the following Sobolev-like space. Let Ω be a bounded open set, and $F \subseteq \overline{\Omega}$ an arbitrary set. We define

$$H^1_{cond,F}(\Omega) = cl_{H^1(\Omega)}\{u \in H^1(\Omega) : \exists \varepsilon > 0, \nabla u = 0 \text{ a.e. on } F^\varepsilon \cap \Omega\}. \tag{6}$$

Let us denote by $u(F)$ the image of the set F by u. For sets F which have a certain regularity (e.g., a finite number of Lipschitz connected components), the previous definition coincides with

$$cl_{H^1(\Omega)}\{u \in H^1(\Omega) : u(F) \text{ is finite}\} \tag{7}$$

and

$$cl_{H^1(\Omega)}\{u \in C^\infty(\Omega) \cap H^1(\Omega) : u(F) \text{ finite}\},$$

but it is not clear whether this holds for arbitrary F. Of course, in (7) we consider quasi-continuous representatives. Observe also that if $u \in H^1_{cond,F}(\Omega)$, then $|u| \in H^1_{cond,F}(\Omega)$.

Note that the following inequality holds for every function $u \in H^1(\Omega)$ such that $\int_{\partial\Omega} u d\sigma = 0$:

$$\int_{\partial\Omega} u^2 dx \le C \int_\Omega |\nabla u|^2 dx, \tag{8}$$

where C is a constant depending only on Ω. This is a consequence of the trace theorem and the Poincaré inequality in $H^1(\Omega)$. Note also that $u \mapsto \int_\Omega |\nabla u|^2 dx$ is a norm on $\{u \in H^1(\Omega), \int_{\partial\Omega} u d\sigma = 0\}$ and that for every $u \in H^1_{cond,K}(\Omega)$ we have $\nabla u = 0$ a.e. on K.

We see problem (4) for an arbitrary K only by its variational formulation

$$\min\left\{\frac{1}{2}\int_{\Omega\setminus K} |\nabla u|^2 dx - \int_{\partial\Omega} u\psi d\sigma : u \in H^1_{cond,K}(\Omega)\right\}. \tag{9}$$

Proposition 2.6. *Problem* (9) *has a unique solution such that* $\int_{\partial\Omega} u d\sigma = 0$.

Note that the gradient of the solution is unique. We can fix a representative such that $\int_{\partial\Omega} u d\sigma = 0$.

Proof. To prove the existence of a solution for problem (9), the Lax–Milgram theorem can be directly used. Nevertheless, in order to familiarize the reader with the space $H^1_{cond,K}(\Omega)$, we show this by using the direct methods of the calculus of variations.

Let $u_n \in H^1_{cond,K}(\Omega)$ be a minimizing sequence. We can assume that $\int_{\partial\Omega} u_n d\sigma = 0$; if not, we simply add suitable constants. Since $0 \in H^1_{cond,K}(\Omega)$, we can also assume

$$\frac{1}{2}\int_{\Omega\setminus K} |\nabla u_n|^2 dx - \int_{\partial\Omega} u_n\psi d\sigma \le 0.$$

Using the Cauchy inequality together with (8) there exists a constant M depending only on Ω such that

$$\int_{\Omega\setminus K} |\nabla u_n|^2 dx \le M.$$

There exists $u \in H^1_{cond,K}(\Omega)$ such that $\nabla u_n \overset{L^2(\Omega,\mathbb{R}^2)}{\rightharpoonup} \nabla u$ and $u_{n|\partial\Omega} \overset{L^2(\partial\Omega)}{\longrightarrow} u_{|\partial\Omega}$. Consequently,

$$\frac{1}{2}\int_{\Omega\setminus K} |\nabla u|^2 dx - \int_{\partial\Omega} u\psi d\sigma \le \liminf_{n\to\infty} \frac{1}{2}\int_{\Omega\setminus K} |\nabla u_n|^2 dx - \int_{\partial\Omega} u_n\psi d\sigma,$$

and hence u is a solution of (9).

The uniqueness of the solution comes from the convexity of the energy functional. □

Note the following facts: The solution given by Proposition 2.6 satisfies $-\Delta u = 0$ on $\Omega \setminus K$ in the sense of distributions and $\frac{\partial u}{\partial n} = \psi$ on $\partial\Omega$ in the weak sense of traces; for every $x \in K$ such that $x \in U_x \subseteq K$, where U_x is a continuum, the solution of (9) is constant q.e. on U_x.

We give the following proposition which has a simple direct proof, and refer to section 4 for a more detailed discussion of the stability question.

PROPOSITION 2.7. *The following holds for $\varepsilon \to 0$:*

$$\nabla u_{\psi,\overline{K}^\varepsilon} \overset{L^2(\Omega,\mathbb{R}^2)}{\longrightarrow} \nabla u_{\psi,K}.$$

Proof. For simplicity, let us denote $u_\varepsilon = u_{\psi,\overline{K}^\varepsilon}$. As in Proposition 2.6, there exists a constant M independent on ε such that

$$\int_\Omega |\nabla u_\varepsilon|^2 dx \le M.$$

There exists $u \in H^1(\Omega)$ such that $\nabla u_\varepsilon \overset{L^2(\Omega,\mathbb{R}^2)}{\rightharpoonup} \nabla u$ and $u_{\varepsilon|\partial\Omega} \overset{L^2(\partial\Omega)}{\longrightarrow} u_{|\partial\Omega}$. We get

$$\frac{1}{2}\int_\Omega |\nabla u|^2 dx - \int_{\partial\Omega} u\psi d\sigma \le \liminf_{\varepsilon\to 0} \frac{1}{2}\int_\Omega |\nabla u_\varepsilon|^2 dx - \int_{\partial\Omega} u_\varepsilon \psi d\sigma. \tag{10}$$

Note that $u \in H^1_{cond,K}(\Omega)$ since $u_\varepsilon \in H^1_{cond,\overline{K}^\varepsilon}(\Omega) \subseteq H^1_{cond,K}(\Omega)$. Let u^* be the solution of (9) in $H^1_{cond,K}(\Omega)$. Then

$$\frac{1}{2}\int_\Omega |\nabla u^*|^2 dx - \int_{\partial\Omega} u^*\psi d\sigma \le \frac{1}{2}\int_\Omega |\nabla u|^2 dx - \int_{\partial\Omega} u\psi d\sigma.$$

From the definition of $H^1_{cond,K}(\Omega)$, there exists a sequence $\phi_n \in H^1(\Omega)$, such that $\nabla\phi_n = 0$ a.e. on $K^{1/n}$, such that $\int_{\partial\Omega}\phi_n d\sigma = 0$ and $\phi_n \longrightarrow u^*$ in $H^1(\Omega)$-strong. Choosing suitable couples (ε, n) such that $\varepsilon < 1/n$, we get $\phi_n \in H^1_{cond,\overline{K}^\varepsilon}(\Omega)$. Consequently,

$$\begin{aligned}\limsup_{\varepsilon\to 0} \frac{1}{2}\int_\Omega |\nabla u_\varepsilon|^2 dx - \int_{\partial\Omega} u_\varepsilon\psi d\sigma &\le \lim_{n\to\infty} \frac{1}{2}\int_\Omega |\nabla\phi_n|^2 dx - \int_{\partial\Omega}\phi_n\psi d\sigma \\ &= \frac{1}{2}\int_\Omega |\nabla u^*|^2 dx - \int_{\partial\Omega} u^*\psi d\sigma.\end{aligned} \tag{11}$$

From (10) and (11) we get $u = u^* = u_{\psi,K}$, and from the convergence of the L^2-norms of the gradients we get that

$$\nabla u_\varepsilon \longrightarrow \nabla u_{\psi,K}\text{-strong } L^2. \qquad \square$$

The result of Proposition 2.7 is still true if on $\partial\Omega$ one considers Dirichlet boundary conditions.

PROPOSITION 2.8 (existence of stream functions). *Let w and u be the solutions of* (3) *and* (4)*, respectively. There exists holomorphic functions W and U in $\Omega \setminus K$ such that $w = \operatorname{Re} W$ and $u = \operatorname{Re} U$. Moreover,* $\operatorname{Im} W$ *and* $\operatorname{Im} U$ *solve problems* (12)

and (13) *below, respectively. Let* Ψ *a primitive of* ψ *on* $\partial\Omega$. *The problem solved by* $\operatorname{Im} W$ *is*

$$\min\left\{\int_\Omega |\nabla\phi|^2 dx : \phi \in H^1_{cond,K}(\Omega), \phi = \Psi \text{ on } \partial\Omega\right\}, \tag{12}$$

and the problem solved by $\operatorname{Im} U$ *is*

$$\begin{cases} -\Delta\phi &= 0 \text{ in } \Omega\setminus K,\\ \frac{\partial\phi}{\partial n} &= 0 \text{ on } \partial K,\\ \phi &= \Psi \text{ on } \partial\Omega. \end{cases} \tag{13}$$

Proof. For $\varepsilon > 0$, the result is true in $\Omega\setminus\overline{K}^{\varepsilon}$ by [1, 6]. Making $\varepsilon\to 0$, the result is true in $\Omega\setminus K$ as a consequence of Propositions 2.5 and 2.7. □

3. Unique identifiability by two boundary measurements. In a first step we introduce a regularity notion, called *conductivity*, for a point of the boundary of an open set U; this kind of regularity should be related to the notion of Wiener regular point, rather than to the usual smoothness of the boundary.

In a second step, we prove that all sets which q.e. satisfy this regularity assumption on their boundaries are uniquely (up to a set of zero capacity) identifiable by two boundary measurements. The proof follows the same steps as in [1], and essentially is obtained by approximating these sets with sets that have a finite number of connected components.

Definition 3.1. *Let* K *be a compact subset of* Ω. *A point* $x\in K$ *is a capacity point for* K *if* $\forall r>0$, $\operatorname{cap}(B_{x,r}\cap K)>0$.

Proposition 3.2. *The set* K^* *of capacity points of a compact set* K *is compact and* $\operatorname{cap}(K\setminus K^*)=0$.

Proof. The compactness comes directly and the relation $\operatorname{cap}(K\setminus K^*)=0$ follows from the Lindelöf property and the subadditivity of the capacity. □

Remark 3.3. From now on, every time we consider a compact set K, we replace it implicitly with K^*. Since $\operatorname{cap}(K\setminus K^*)=0$, problems (3) and (4) have the same solutions on $\Omega\setminus K$ and $\Omega\setminus K^*$, respectively. From a practical point of view, every time an open set $U\subseteq\Omega$ is considered, it is replaced with $\Omega\setminus(\overline{\Omega}\setminus U)^*$.

Definition 3.4. *Let* U *be an open subset of* Ω *and* $x\in\partial U$. *We say that* x *is conductive for* U *if for every* $r>0$ *and for every* $\varphi\in C(\overline{U})\cap H^1_{cond,\partial U\cap B_{x,r}}(\Omega)$,

$$\liminf_{\substack{y\to x\\ y\in\partial U}} \frac{|\varphi(y)-\varphi(x)|}{|y-x|} = 0. \tag{14}$$

Roughly speaking, x is a conductivity point if there exists a "path" of conductivity on ∂U passing through x and having locally positive capacity. Note that every conductivity point is a capacity point for U^c.

Proposition 3.5. *Let* K *be a compact subset of* Ω *such that* $\Omega\setminus K$ *is connected and* F *a continuum of positive diameter such that* $x\in F\subseteq\partial(\Omega\setminus K)$. *Then* x *is a conductivity point for* $\Omega\setminus K$.

Proof. Let $\varphi\in C(\overline{\Omega\setminus K})\cap H^1_{cond,\partial(\Omega\setminus K)\cap B_{x,r}}(\Omega)$. Then, φ is q.e. constant on the continuum of positive diameter $\tilde{F}$ passing through x and contained in $F\cap B_{x,r}$. Indeed, if ϕ is a quasi-continuous representative of φ on Ω, then ϕ is finely continuous q.e. (see [16]) and coincides q.e. with φ on $\Omega\setminus K$. Since every point of $\partial(\Omega\setminus K)$ is

thick with respect to $\Omega \setminus K$ (which is connected), we conclude that $\phi(x) = \varphi(x)$ q.e. on $\partial(\Omega \setminus K)$. Therefore φ is q.e. constant on $\tilde{F}$ and relation (14) holds. □

In what follows we give two examples of the form $U = \Omega \setminus K$ with points $x \in \partial K$ which are disconnected from the rest of the set K; in one of them we show that such a point may be conductive despite the fact it is not contained in any continuum of positive diameter (subset of K).

Example 3.6. Let $\Omega = [-2, 2] \times [-2, 2]$ and

$$K = \{(0,0)\} \bigcup_{n \in \mathbb{N}^*} \frac{1}{n} \times \left[0, \frac{1}{n}\right].$$

Then $(0,0)$ is not a conductivity point for $\Omega \setminus K$; consider, for example, $u(x,y) = x$. Obviously $u \in C(\overline{\Omega})$ and it is easy to see (cf., e.g., [5, 6]) that $u \in H^1_{cond,K}(\Omega)$ by approaching u strongly in $H^1(\Omega)$ by the sequence u_n of solutions of the following equations. Let $K_n = \bigcup_{k=1}^{n} [\frac{1}{k} - \frac{1}{n^2}, \frac{1}{k} + \frac{1}{n^2}] \times [0, \frac{1}{k}]$ and let u_n solve $-\Delta u_n = 0$ in $\Omega \setminus K_n$, $u_n = u$ on $\partial\Omega$ and $u_n = \frac{1}{k}$ on $[\frac{1}{k} - \frac{1}{n^2}, \frac{1}{k} + \frac{1}{n^2}] \times [0, \frac{1}{k}]$.

On the other hand, (14) does not hold, since $\frac{|u(x,y)-u(0)|}{|(x,y)|} \geq \sqrt{2}^{-1}$, for every $(x,y) \in K$.

Example 3.7. Now let

$$K = \{(0,0)\} \bigcup_{n \in \mathbb{N}^*} \{b_n\} \times \left[0, r_n\right],$$

where

$$b_n = \sum_{k=n}^{\infty} \frac{1}{k^2(k+1)^3}$$

and

$$r_n = \frac{1}{n}.$$

Then $(0,0)$ is a conductivity point for $\Omega \setminus K$ which is not contained in a continuum of positive capacity of K.

The proof needs some computation. We give it in the appendix at the end of the paper.

We also prove in the appendix the following proposition, which is an extension of Proposition 3.5.

PROPOSITION 3.8. *Let K be a compact subset of Ω such that $\Omega \setminus K$ is connected. Every $x \in \partial(\Omega \setminus K)$ for which there exists a continuum of positive diameter U_x, such that $x \in U_x \subseteq K$, is a conductivity point for $\Omega \setminus K$.*

In particular, if K is a continuum of positive diameter, then $\Omega \setminus K$ is conductive at every point of $\partial(\Omega \setminus K)$. As well, if K is a compact set having a finite number of connected components, then $\Omega \setminus K$ is conductive at quasi-every point of its boundary (except the isolated points).

Note that if K is a compact subset of Ω, the only detectable part of ∂K is the one contained in the boundary of the connected component of $\Omega \setminus K$ which touches $\partial\Omega$. For this reason, we shall assume (only) in the following theorem that $\Omega \setminus K$ is connected.

The fluxes we consider are defined as in [1]. Consider a division of $\partial\Omega$ into three connected disjoint parts $\Gamma_0, \Gamma_1, \Gamma_2$. For $i = 0, 1, 2$ we consider on $\partial\Omega$ a nonnegative

function η_i such that supp $\eta_i \subseteq \Gamma_i$, $\eta_i \in L^2(\partial\Omega)$, $\int_{\partial\Omega} \eta_i = 0$. We take for $k = 1, 2$, $\psi_k = \eta_0 - \eta_k$.

THEOREM 3.9. *Let $K, \tilde{K}$ be two compact subsets of Ω such that $\Omega \setminus K$, $\Omega \setminus \tilde{K}$ are connected. Let ψ_1, ψ_2 be two fluxes on $\partial\Omega$ chosen as above. Suppose that for $k = 1, 2$ either $w_{\psi_k,K|\partial\Omega} = w_{\psi_k,\tilde{K}|\partial\Omega}$ or $u_{\psi_k,K|\partial\Omega} = u_{\psi_k,\tilde{K}|\partial\Omega}$. If $\Omega \setminus K$ and $\Omega \setminus \tilde{K}$ are conductive at quasi-every point of their boundaries, then $K = \tilde{K}$ q.e.*

Proof. The proof relies on the nonexistence of geometrical critical points for particular holomorphic functions. Let us first prove that if $K \neq \tilde{K}$ q.e., then we may find a geometrical critical point for the solution of (3) (or (4)) with a boundary data of the form $\alpha\psi_1 + \beta\psi_2$, for a certain couple α, β which satisfies $\alpha^2 + \beta^2 = 1$. The proof follows along the same lines as in [1], in the new hypotheses on the conductivity of the sets K and $\tilde{K}$. A new kind of difficulty appears, since the unique continuation property does not give information over all $\Omega \setminus (K \cup \tilde{K})$.

We shall consider only problem (3) (the case (4) follows the same ideas). For $k = 1, 2$, let w_k^*, $\tilde{w}_k^*$ be the conjugate functions of $w_k = w_{\psi_k,K}$, $\tilde{w}_k = w_{\psi_k,\tilde{K}}$, such that $w_k + iw_k^*$ are holomorphic in $\Omega \setminus K$ and $\tilde{w}_k + i\tilde{w}_k^*$ are holomorphic in $\Omega \setminus \tilde{K}$, respectively. Note that for the boundary condition $\alpha\psi_1 + \beta\psi_2$ the solution of (3) on $\Omega \setminus K$ is $\alpha w_1 + \beta w_2$ and that the harmonic conjugate of this function in $\Omega \setminus K$ is $\alpha w_1^* + \beta w_2^*$.

From the unique continuation property, we get as in [1] that $w_k = \tilde{w}_k$ on G, where G is the connected component of $\Omega \setminus (K \cup \tilde{K})$ satisfying $\partial\Omega \subseteq \partial G$ (the functions w_k and $\tilde{w}_k$ have the same Cauchy data on $\partial\Omega$). The main difficulty is that the information is not obtained over all $\Omega \setminus (K \cup \tilde{K})$. In [1], using the particular structure of K and $\tilde{K}$, the information could be extended in $\Omega \setminus (K \cup \tilde{K})$.

Let us suppose that $\Omega \setminus K \neq \Omega \setminus \tilde{K}$, say, $\Omega \setminus K \not\subseteq \Omega \setminus \tilde{K}$. There exists $x \in \Omega \setminus K$ such that $x \notin \Omega \setminus \tilde{K}$ (i.e., $x \in \tilde{K}$). Since $\Omega \setminus K$ is connected and $\partial\Omega$ is smooth, there exists a smooth curve $\gamma : [0, 1] \to \overline{\Omega} \setminus K$ such that $\gamma(0) = x$, $\gamma((0, 1)) \subseteq \Omega \setminus K$, $\gamma(1) \in \partial\Omega$. Let $x_0 = \gamma(t_0)$, where

$$t_0 = \sup\{t \in [0, 1] : \gamma(t) \in \tilde{K}\}.$$

Obviously, $x_0 \in \partial\tilde{K}$ and also $x_0 \in \partial G$. Since $d(x_0, K) > 0$, there exists a ball $B_{x_0,r}$ such that $B_{x_0,r} \cap K = \emptyset$.

We prove the following.

LEMMA 3.10. *For every $\delta > 0$, G has a conductivity point on $\partial G \cap B_{x_0,\delta}$.*

Proof. For every $\varepsilon > 0$ we consider the open set $\tilde{K}^\varepsilon$. There exists an open polygonal set V_ε such that

$$\tilde{K}^{\varepsilon/2} \subseteq V_\varepsilon \subseteq \tilde{K}^\varepsilon.$$

Let U_ε be the connected component of V_ε which contains x_0. Choosing a sequence (ε_n) such that $\varepsilon_n \to 0$, $\varepsilon_{n+1} < \varepsilon_n/2$ we get

$$U_{\varepsilon_{n+1}} \subseteq U_{\varepsilon_n}.$$

There are two possibilities:

1. diam $(U_{\varepsilon_n}) \to 0$;
2. diam $(U_{\varepsilon_n}) \to \eta > 0$.

The first case. Suppose that diam $(U_{\varepsilon_n}) \to 0$. For n large enough we have $U_{\varepsilon_n} \subseteq B_{x_0,r/2}$. Let A_n be the connected component of $\Omega \setminus \overline{U}_{\varepsilon_n}$ such that $\partial\Omega \subseteq \partial A_n$.

Let $P_n = \partial A_n \setminus \partial\Omega$. Then P_n is a closed polygonal Jordan curve, which separates Ω in two regions. We observe that $P_n \subseteq \Omega \setminus (K \cup \tilde{K})$ because $P_n \cap K = \emptyset$ (since $P_n \subseteq \overline{B}_{x_0,r/2}$) and $P_n \cap \tilde{K} = \emptyset$ since ($P_n \subseteq \partial U_{\varepsilon_n}$ and $d(\partial U_{\varepsilon_n}, \tilde{K}) = \varepsilon_n/2$).

Since P_n intersects γ, and γ lies in G, the connectedness of G implies that P_n is entirely in G. Therefore, for ξ small enough, we have that

$$G \cap B_{x_0,\xi} = (\Omega \setminus \tilde{K}) \cap B_{x_0,\xi}, \tag{15}$$

$$\partial G \cap B_{x_0,\xi} = \partial(\Omega \setminus \tilde{K}) \cap B_{x_0,\xi}. \tag{16}$$

In this case two possibilities may hold: either x_0 is a conductivity point or it is not. If it is not a conductivity point, we replace it with a close point of ∂G which is conductive. Such a point exists, since following [5, Lemma 4.5] x_0 is a capacity point also for ∂G, and the family of points of ∂G which are not conductive is, by hypothesis, of zero capacity (note that ∂G coincides locally with $\partial(\Omega \setminus \tilde{K})$).

The second case. Suppose that diam $(U_{\varepsilon_n}) \to \eta > 0$. We observe that $\bigcap_n U_{\varepsilon_n} = C$, where C is a continuum such that $x \in C \subseteq \tilde{K}$, diam $C = \eta$. Let $0 < \xi < \eta/2$. Then $C \cap \partial B_{x_0,\xi} \neq \emptyset$.

Denoting again by A_n the connected component of $\Omega \setminus \overline{U}_{\varepsilon_n}$ such that $\partial\Omega \subseteq \partial A_n$, let us set again $P_n = \partial A_n \setminus \partial\Omega$. Then P_n is a polygonal Jordan curve satisfying $P_n \cap \tilde{K} = \emptyset$. Let us denote $z_n = \gamma(t_n)$, where

$$t_n = \min\{t \in [0,1], \gamma(t) \in P_n\}. \tag{17}$$

We observe that z_n is well defined and $z_n \to x_0$ for $n \to \infty$.

Since P_n contains in its "interior" region the continuum C and $P_n \cap \tilde{K} = \emptyset$ and $(P_n \cap B_{x_0,\xi}) \cap K = \emptyset$, there exists a connected component F_n of P_n passing through z_n which is contained in G and cuts $\partial B_{x_0,\xi}$ into at least two points. For $n \to \infty$, we have that F_n converges in the Hausdorff sense to a continuum F which contains x_0 and lies in the boundary of G. From Proposition 3.5 x_0 is a conductivity point for G. □

Proof of Theorem 3.9 (continuation). Let x_0 be the conductive point given by Lemma 3.10. Up to translation by constants, we can assume that for $k = 1, 2$ $w_k(x_0) = w_k^*(x_0) = 0$. Note that the function $|\tilde{w}_1^*| + |\tilde{w}_2^*|$ belongs to $H^1_{cond,\tilde{K}}(\Omega)$ and equals $|w_1^*| + |w_2^*|$ on G. This last function is continuous in a neighborhood of x_0, and hence we can apply the conductivity property to $|w_1^*| + |w_2^*|$ in x_0.

There exists a sequence of points x_n such that $x_n \in \partial G$, $x_n \to x_0$, and

$$\frac{|w_1^*(x_n)| + |w_2^*(x_n)|}{|x_n - x_0|} \to 0. \tag{18}$$

Hence, for $k = 1, 2$,

$$\frac{w_k^*(x_n)}{|x_n - x_0|} \to 0. \tag{19}$$

We suitably chose values α_n, β_n, such that $\alpha_n^2 + \beta_n^2 = 1$ and $\alpha_n w_1(x_n) + \beta_n w_2(x_n) = 0$. Choosing a subsequence of $(\alpha_n)_n, (\beta_n)_n$ such that $\alpha_n \to \alpha_0$, $\beta_n \to \beta_0$ and using relations (19), we have that the sequence of holomorphic functions

$$f_n = (\alpha_n w_1 + \beta_n w_2) + i(\alpha_n w_1^* + \beta_n w_2^*)$$

satisfies

$$\frac{f_n(x_n) - f_n(x_0)}{|x_n - x_0|} \to 0.$$

Consequently, x_0 is a geometrical critical point for $f_0 = (\alpha_0 w_1 + \beta_0 w_2) + i(\alpha_0 w_1^* + \beta_0 w_2^*)$. Indeed, we have, for $n \to \infty$,

$$\begin{aligned}&\frac{f_0(x_n) - f_0(x_0)}{|x_n - x_0|}\\ &= \frac{f_n(x_n) - f_n(x_0)}{|x_n - x_0|} - (\alpha_n - \alpha_0)\frac{w_1(x_n) + iw_1^*(x_n)}{|x_n - x_0|} - (\beta_n - \beta_0)\frac{w_2(x_n) + iw_2^*(x_n)}{|x_n - x_0|} \to 0.\end{aligned}$$

The last two terms converge to zero thanks to the holomorphy in x_0 of the functions $w_k + iw_k^*$.

To get the contradiction we observe that f_0 cannot have geometrical critical points in $\Omega \setminus K$. Indeed, this is a consequence of the result of [1] applied on $\Omega \setminus \overline{K^\varepsilon}$ by passing to the limit $K^\varepsilon \to K$ and using the continuity property of critical points. □

The main difficulty in the proof of this theorem is the fact that the unique continuation property gives information only in the connected component of $\Omega \setminus (K \cup \tilde{K})$ touching $\partial\Omega$. In [1], this information is extended over all $\Omega \setminus (K \cup \tilde{K})$ by using the connectedness of the cracks. Here we are not able to do that, and for this reason we can use information only "on one side" of the crack $\tilde{K}$, namely, on G. Since the conductivity hypothesis is formulated in $\Omega \setminus \tilde{K}$ and not in G, we are brought to discussing the two cases of Lemma 3.10. □

Remark 3.11. The conductivity property is somehow related to the thickness property relying on the Wiener criterion. It would be of interest to characterize all sets which are q.e. conductive at the boundary, and in this way to characterize all detectable sets by two boundary measurements. Even an example of a compact set which is not conductive q.e. would be of interest.

Remark 3.12. Notice from relations (15)–(16) that the density of the conductive points of ∂G into ∂G is sufficient for carrying out the proof. In the appendix, we give an example of a Cantor set which is conductive in a dense set of its boundary points, and consequently the unique identifiability for this totally disconnected set holds true (see Example 6.2).

4. Sequential stability of the inverse problems. Let ψ_1, ψ_2 be two fluxes on $\partial\Omega$ which uniquely identify q.e. conductive sets (Theorem 3.9) for problem (3) as well as for (4). For a compact set $K \subseteq \Omega$ and a sequence of compacts $(K_n)_n$ such that

$$w_{\psi_i, K_n|\partial\Omega} \overset{L^2(\partial\Omega)}{\longrightarrow} w_{\psi_i, K|\partial\Omega}, \quad i = 1, 2,$$

or

$$u_{\psi_i, K_n|\partial\Omega} \overset{L^2(\partial\Omega)}{\longrightarrow} u_{\psi_i, K|\partial\Omega}, \quad i = 1, 2,$$

we wonder if $K_n \overset{H}{\to} K$.

This assertion is in general false. First, the convergence in the Hausdorff metric does not have much in common with the behavior of the PDE on "moving" cracks.

For this reason, it is not senseless to think of stability in terms of *behavior*, i.e., two close measurements should give cracks such that *all* measurements are close. This approach is to be compared to the γ-convergence of sets (see [5]) which has a certain relation to the geometric convergence but is not at all equivalent. This kind of approach seems necessary as soon as one deals with "wild" cracks without any a priori structure. Nevertheless, we restrict ourselves to the Hausdorff metric because it seems quite difficult to describe the general behavior of sets. Note that for homogeneous Neumann boundary conditions the general behavior of the direct problem for moving domains is not, to our knowledge, known.

Second, uniqueness holds for sets $\Omega \setminus K$ which are q.e. conductive with the convention that $\Omega \setminus K$ is connected. If $\Omega \setminus K$ is disconnected, the only identifiable part is the connected component "touching" $\partial\Omega$. From a purely geometric point of view, this means that different geometries for K may give similar measures. Here we explain what can happen, from the information we have, namely, the coincidence of the identifiable connected components. Under mild assumptions on K, our result becomes a standard stability result.

In order to understand the sequential stability for the crack identification problem, the usual tool relies on the stability of the direct problem associated with compactness and uniqueness of the identification. Compactness is a geometric property of the Hausdorff convergence, and for uniqueness we rely on Theorem 3.9. The geometric stability of the direct problems (3) and (4) relies on the Mosco convergence of the Sobolev spaces $H^1(\Omega \setminus K_n)$ for problem (3) and $H^1_{cond,K_n}(\Omega)$ for problem (4). Ultimately, because of the existence of harmonic conjugates for solutions of problem (3), the only important case to be studied is the Mosco convergence of $H^1_{cond,K_n}(\Omega)$ (see [6]).

Let X be a Hilbert space and $\{G_n\}_{n\in\mathbb{N}}$ a sequence of subsets of X. The weak upper and strong lower limits in the sense of Kuratowski are defined as follows:

$$w - \limsup_{n\to\infty} G_n = \{u \in X : \exists\{n_k\}_k, \exists u_{n_k} \in G_{n_k} \text{ such that } u_{n_k} \overset{w-X}{\rightharpoonup} u\},$$

$$s - \liminf_{n\to\infty} G_n = \{u \in X : \exists u_n \in G_n \text{ such that } u_n \overset{s-X}{\longrightarrow} u\}.$$

If $\{G_n\}_{n\in\mathbb{N}}$ are closed subspaces in X, it is said that G_n converges in the sense of Mosco to G if

(M_1) $G \subseteq s - \liminf_{n\to\infty} G_n$,

(M_2) $w - \limsup_{n\to\infty} G_n \subseteq G$.

Note that in general $s - \liminf_{n\to\infty} G_n \subseteq w - \limsup_{n\to\infty} G_n$. Therefore, if G_n converges in the sense of Mosco to G, then

$$s - \liminf_{n\to\infty} G_n = G = w - \limsup_{n\to\infty} G_n.$$

For our purposes, we consider the compact sets $K_n, K \subseteq \Omega$ and wonder if $G_n := H^1_{cond,K_n}(\Omega)$ converges in the sense of Mosco to $H^1_{cond,K}(\Omega)$ into the space $X := H^1(\Omega)$.

Assume that $K_n \overset{H}{\to} K$. Then condition M_1 is immediately satisfied. Indeed, using density, it is enough to consider $u \in H^1_{cond,K}(\Omega)$ such that $\nabla u = 0$ a.e. on K^ε, for some $\varepsilon > 0$. Following the Hausdorff convergence, for n large enough we have $K_n \subseteq K^{\frac{\varepsilon}{2}}$; hence $K_n^{\frac{\varepsilon}{2}} \subseteq K^\varepsilon$ and so $\nabla u = 0$ a.e. on $K_n^{\frac{\varepsilon}{2}}$, and therefore $u \in H^1_{cond,K_n}(\Omega)$.

In general, condition M_2 is not true. Take for example $\Omega = (-2,2)\times(-2,2)$, $K_n = \cup_{k=0}^{n}\{\frac{k}{n}\}\times[0,1]$, and $u_n(x,y) = x$. A second example which typically restricts M_2 from holding is when K_n consists of many small disconnected sets, e.g., $K_n = \cup_{k,p=0}^{n}\overline{B}((\frac{k}{n},\frac{p}{n}),\varepsilon_n)$, $\varepsilon_n > 0$, $\varepsilon_n \to 0$. Then

$$K_n \xrightarrow{H} [0,1]\times[0,1],$$

but for a suitable choice of ε_n every function of $H^1(\Omega)$ can be written as a limit of a sequence of $H^1_{cond,K_n}(\Omega)$ (choose ε_n such that cap $(K_n)\to 0$).

THEOREM 4.1. *Let $K_n, K \subseteq \Omega$, $K_n \xrightarrow{H} K$. If M_2 occurs, then for every $\psi \in L^2(\partial\Omega)$ such that $\int_{\partial\Omega}\frac{\partial\psi}{\partial n}d\sigma = 0$, we have*

1. $u_{K_n,\psi|\partial\Omega} \xrightarrow{L^2(\partial\Omega)} u_{K,\psi|\partial\Omega}$,
2. $w_{K_n,\psi|\partial\Omega} \xrightarrow{L^2(\partial\Omega)} w_{K,\psi|\partial\Omega}$.

Proof. Let us prove assertion 1. For simplicity, we set $u_n = u_{K_n,\psi}$ and $u = u_{K,\psi}$. As in Propositions 2.6, and 2.7 there exists a uniform bound M such that

$$\int_\Omega |\nabla u_n|^2 dx \le M.$$

For a subsequence (still denoted using the same index), we can write $u_n \rightharpoonup \tilde{u}$ weakly $H^1(\Omega)$. From the second Mosco condition, which is assumed by hypothesis, we get $\tilde{u}\in H^1_{cond,K}(\Omega)$. In order to prove that $\tilde{u} = u$, we observe that

$$\begin{aligned}(20)\qquad &\frac{1}{2}\int_\Omega |\nabla u|^2 dx - \int_{\partial\Omega} u\psi d\sigma \\ &\quad\le \frac{1}{2}\int_\Omega |\nabla\tilde{u}|^2 dx - \int_{\partial\Omega}\tilde{u}\psi d\sigma \le \liminf_{n\to\infty}\frac{1}{2}\int_\Omega |\nabla u_n|^2 dx - \int_{\partial\Omega} u_n\psi d\sigma.\end{aligned}$$

We also note that the first Mosco condition is a direct consequence of the geometric convergence $K_n \xrightarrow{H} K$. Indeed, for proving M_1 it is enough to consider $\phi\in H^1_{cond,K}(\Omega)$ such that $\nabla\phi = 0$ a.e. on K^ε, for some $\varepsilon > 0$ (this set is dense in $H^1_{cond,K}(\Omega)$). Indeed, for n large enough such that $K_n \subseteq K^{\varepsilon/2}$ we get that $\nabla\phi = 0$ a.e. on $K_n^{\varepsilon/2}$; hence $\phi\in H^1_{cond,K_n}(\Omega)$.

Let $\phi_n \in H^1_{cond,K_n}(\Omega)$ such that $\phi_n \longrightarrow u$ in $H^1(\Omega)$-strong. We get

$$\begin{aligned}(21)\qquad \limsup_{n\to\infty}\frac{1}{2}\int_\Omega |\nabla u_n|^2 dx - \int_{\partial\Omega} u_n\psi d\sigma &\le \lim_{n\to\infty}\frac{1}{2}\int_\Omega |\nabla\phi_n|^2 dx - \int_{\partial\Omega}\phi_n\psi d\sigma \\ &= \frac{1}{2}\int_\Omega |\nabla u|^2 dx - \int_{\partial\Omega} u\psi d\sigma.\end{aligned}$$

From (20) and (21) we get $\tilde{u} = u$ and the strong H^1-convergence $u_n \longrightarrow u$. The convergence $u_{K_n,\psi|\partial\Omega} \xrightarrow{L^2(\partial\Omega)} u_{K,\psi|\partial\Omega}$ follows from the trace theorem.

To prove assertion 2 of the theorem, namely, $w_{K_n,\psi|\partial\Omega} \xrightarrow{L^2(\partial\Omega)} w_{K,\psi|\partial\Omega}$, one has to use the following duality argument, which was already applied in [6].

Following Proposition 2.8, let v_n be the conjugate function of $w_{K_n,\psi|\partial\Omega}$ in $\Omega\setminus K_n$. Then v_n solves the following problem (set as an energy minimization):

$$\min\left\{\int_\Omega |\nabla v|^2 dx : v\in H^1_{cond,K_n}(\Omega), v = \Psi \text{ on } \partial\Omega\right\},$$

where Ψ is a primitive of ψ on $\partial\Omega$.

Note that v_n solves a problem very similar to (9), with the only difference that on the fixed boundary $\partial\Omega$, the Neumann condition is replaced with a Dirichlet one. The same proof as for the first point of this theorem can be repeated. We get

$$\nabla v_n \overset{L^2(\Omega,\mathbb{R}^2)}{\longrightarrow} \nabla v$$

and

$$\nabla v_n 1_{\Omega\setminus K_n} \overset{L^2(\Omega,\mathbb{R}^2)}{\longrightarrow} \nabla v 1_{\Omega\setminus K},$$

since $\nabla v_n = 0$ a.e. on K_n. In terms of conjugate functions, this gives

$$\nabla w_{K_n,\psi|\partial\Omega} 1_{\Omega\setminus K_n} \overset{L^2(\Omega,\mathbb{R}^2)}{\longrightarrow} w_{K,\psi|\partial\Omega} 1_{\Omega\setminus K}.$$

Applying the trace theorem for $w_{K_n,\psi|\partial\Omega}$ into a smooth neighborhood U of $\partial\Omega$, we get $w_{K_n,\psi|\partial\Omega} \overset{L^2(\partial\Omega)}{\longrightarrow} w_{K,\psi|\partial\Omega}$. □

In the next proposition we prove that condition M_2 is satisfied, provided that the number of connected components of K_n and K are uniformly bounded (we denote by $\sharp K$ the number of connected compoenents of K).

PROPOSITION 4.2. *Let $K \subseteq \Omega, K_n \overset{H}{\longrightarrow} K$, $\sharp K_n \leq M$. Then the second Mosco condition holds for $H^1_{cond,K_n}(\Omega)$ and $H^1_{cond,K}(\Omega)$.*

Proof. Let $\phi_n \in H^1_{cond,K_n}(\Omega)$ such that $\phi_n \rightharpoonup \phi$ in $H^1(\Omega)$-weak. Let K_α be a connected component of K. We observe first that ϕ is q.e. constant on K_α. Indeed, from the Hausdorff convergence, K_α can be written $K_\alpha = \cup_{i=1}^M K^i_\alpha$, where (up to subsequences) $K^i_\alpha = H - \lim_{n\to\infty} K^i_n$, where K^i_n are connected components of K_n. In our notation, some of these components can be chosen empty sets. Following Lemma 2.3 (see also [19] and [6]) we get ϕ q.e. constant on K^i_α, and following Lemma 2.1 we get ϕ q.e. constant on K_α.

In order to prove that $\phi \in H^1_{cond,K}(\Omega)$ we use Hedberg's result [13], which asserts that ϕ can be approached strongly in $H^1(\Omega)$ by a sequence of functions which for every α are constant q.e. (hence a.e.) on a neighborhood of K_α. □

In what follows we give several situations when stability occurs. To simplify the notation, for every compact $K \subseteq \Omega$ we denote by G_K the connected component of $\Omega \setminus K$ which "touches" $\partial\Omega$.

THEOREM 4.3. *Let ψ_1, ψ_2 be as in Theorem 3.9. Suppose F is a compact subset of Ω and that (K_n) is a family of compact subsets of F such that*

$$\exists M > 0 \quad \forall n \in \mathbb{N} \quad \sharp K_n \leq M.$$

If either

$$u_{K_n,\psi_i|\partial\Omega} \overset{L^2(\partial\Omega)}{\longrightarrow} u_i, \quad i = 1, 2,$$

or

$$w_{K_n,\psi|\partial\Omega} \overset{L^2(\partial\Omega)}{\longrightarrow} w_i, \quad i = 1, 2,$$

holds, then there exists a compact set $K \subseteq \Omega$ such that we have $\sharp K \leq M$ and a subsequence $K_{n_k} \overset{H}{\to} K$, and that for $i = 1, 2$, we have $u_i = u_{\psi_i,K|\partial\Omega}$ (respectively, $w_i = w_{\psi_i,K|\partial\Omega}$).

If for another subsequence, we have $K_{n'_k} \overset{H}{\to} \tilde{K}$, *then* $G_K = G_{\tilde{K}}$ *q.e.*

Proof. By the compactness of the Hausdorff convergence we can write $K_{n_k} \xrightarrow{H} K$, with $K \subseteq F$, and $\sharp K \leq M$. Using Theorem 4.1 and Proposition 4.2 we get $u_{K_{n_k},\psi_i|\partial\Omega} \overset{L^2(\partial\Omega)}{\longrightarrow} u_{K,\psi_i|\partial\Omega}$ (and the same for w). Hence $u_i = u_{K,\psi_i|\partial\Omega}$ (and the same for w).

If for another subsequence we have $K_{n'_k} \overset{H}{\to} \tilde{K}$, we use the uniqueness theorem, Theorem 3.9, and get the conclusion. □

Remark 4.4. Theorem 4.3 is not a standard stability result, as one might expect. It is rather a description of possible situations regarding stability. Nevertheless, under mild assumptions on K, this becomes a usual sequential stability result.

In the following, K, K_n, and $\tilde{K}$ are as in Theorem 4.3.

COROLLARY 4.5. *Let* K *be such that* $\sharp K = M$, $\Omega \setminus K$ *is connected, and* $\overset{\circ}{K} = \emptyset$. *Then* $\tilde{K} = K$ *and the hole sequence* K_n *converges into the Hausdorff metric to* K.

Proof. From Theorem 4.3,

$$G_{\tilde{K}} = G_K = \Omega \setminus K \text{ q.e.} \tag{22}$$

Moreover, thanks to the hypothesis $\sharp K = M$, K has M connected components and each one has a positive diameter. Consequently, equality (22) holds everywhere. From the definition of $G_{\tilde{K}}$, relation (22) implies $\tilde{K} \subseteq K$. The converse is also true since $K = \partial K = \partial G_{\tilde{K}} \setminus \partial\Omega \subseteq \partial\tilde{K} \subseteq \tilde{K}$. □

Example 4.6. In order to give geometric intuition on the sense in which Theorem 4.3 should be understood, we give in Figures 1 and 2 two examples of cracks and cavities which give close measurements.

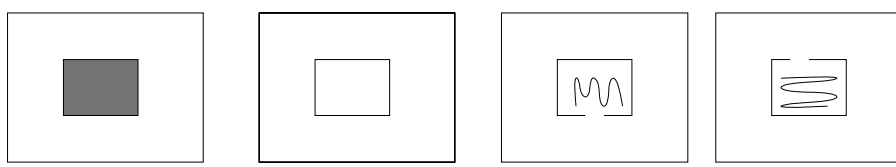


FIG. 1. *Four compacts giving "close" measures: Cavity* (a) *gives the same measure as crack* (b)*; asymptotically, cracks* (c) *and* (d) *give the same measures (as soon as the apertures of the rectangular cracks go to zero).*

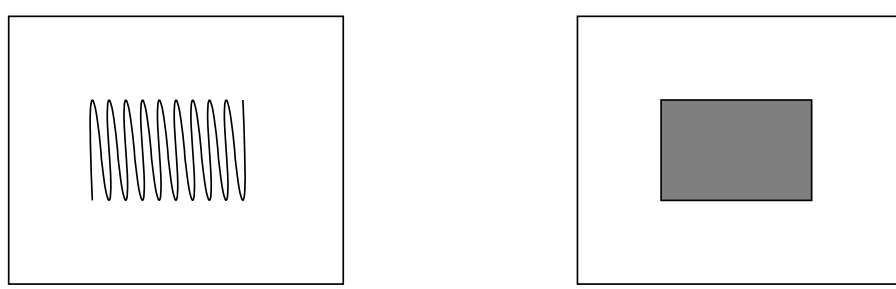


FIG. 2. *Two compacts giving "close" measures; on the left a "long and dense" curve and on the right a cavity.*

We give an example in which stability comes from the structure of K and requires that all K_n satisfy a uniform identifiability assumption. From a practical point of view this result might be helpful if all cracks do not have interior points, and locally their diameters are beyond a detectability level.

DEFINITION 4.7. *Let $\varepsilon > 0$. A compact set K is ε-detectable if for every $x \in K$, the diameter of the connected component of K containing x is greater than or equal to ε. A compact set $K \subseteq \Omega$ is called ε-stable if*

(23)
$$H^1_{cond,K}(\Omega) = \{u \in H^1(\Omega) \forall x \in K, \exists U_x \text{ continuum, diam } (U_x) \geq \varepsilon \text{ s.t. } u = c_x \text{ q.e. on } U_x\}.$$

To simplify, let us denote the space on the right-hand side as $H^1_\varepsilon(\Omega)$. Notice that if K is arbitrary, equality (23) does not occur; e.g., if K has an interior point. Take, for example, $\Omega = [-2,2] \times [-2,2]$, $K = [0,1] \times [0,1]$, and $u(x,y) = x$.

An example of ε-stable K is $K = \{\bigcup_{n=1}^{\infty} \{\frac{1}{n}\} \times [0,1]\} \bigcup \{0\} \times [0,1]$. Indeed, continua on lines are intervals. Hence on each vertical segment a function $u \in H^1_\varepsilon(\Omega)$ can take only a finite number of values. Using Lemma 2.1 it follows that u is q.e. constant on each vertical segment. Using the same argument as in Example 3.6 we get $u \in H^1_{cond,K}(\Omega)$.

PROPOSITION 4.8. *Suppose there exists $\varepsilon > 0$ such that K_n are ε-detectable. The conclusion of Theorem* 4.3 *holds, provided that K and $\tilde{K}$ are ε-stable.*

5. Application: Approximation by finite elements. We prove in this section that the unknown defects can be formally approached using finite elements, regardless of their regularity. Basically, this is one of the main applications of the stability result established in the previous section. All previous stability results, which give finer estimates for the stability, assume a priori the smoothness of the defects and suppose known their (uniform) Lipschitz character. In this regard, Theorem 4.3 does not give a quantitative estimate for the stability, but provides a rigorous justification of the approach by finite elements. For a similar argument related to shape optimization problems with homogeneous Dirichlet conditions on the free boundaries, we refer to [8].

We discuss both problems (3) and (4). It will be quite surprising to notice that, formally, problem (4) is easier to treat from a numerical point of view, since a unique mesh can be used at each step for both capturing the defect Γ and computing the finite element approximation of the solution. This is also the case for homogeneous Dirichlet problems [8]. For problem (3) the defect Γ is captured on a mesh, while the finite element approximation of the solution needs a refinement of the mesh. This is precisely what is done in practice.

Let F be "the design region," i.e., a subdomain of Ω containing all defects. Let $(\mathcal{T}_h)_h$ denote a family of triangulations of Ω made of elements which are triangles (the extension to quadrilaterals is standard). The maximal size of elements is the discretization parameter, denoted by h. In addition, we assume that each triangulation satisfies the usual admissibility assumptions, i.e., the intersection of two different elements is either empty, a vertex, or a whole edge, and $\mathcal{T}_h$ is assumed to be "regular," i.e., the ratio between the diameter of any element $K \in \mathcal{T}_h$ and the diameter of its largest inscribed ball is bounded by a constant σ independent of K and h.

Let $K^* \subseteq F$ be a defect such that $\sharp K^* \leq M$ which gives the measures w_1, w_2 corresponding to the input fluxes, ψ_1, ψ_2, respectively. We solve the finite dimensional

problem

$$\min_{\substack{K\subset\mathcal{T}_h\cap F\\ \#K\le M}} \int_{\partial\Omega} |w_{K,\psi_1}-w_1|^2\,d\sigma + \int_{\partial\Omega} |w_{K,\psi_2}-w_2|^2\,d\sigma, \tag{24}$$

which admits at least one solution, denoted K_h. The following convergence result holds.

THEOREM 5.1. *For $h\to 0$, there exists a subsequence such that*

$$K_h \overset{H}{\to} \tilde{K} \quad \textit{and} \quad G_{K^*} = G_{\tilde{K}}.$$

Proof. By compactness we can extract a subsequence $K_h \overset{H}{\to} \tilde{K}$. First, we notice that

$$\int_{\partial\Omega} |w_{K_h,\psi_1}-w_1|^2 + \int_{\partial\Omega} |w_{K_h,\psi_2}-w_2|^2 \longrightarrow 0 \text{ as } h\longrightarrow 0. \tag{25}$$

Indeed, we define

$$K_h^* = \bigcup_{\substack{T\in\mathcal{T}_h\cap F\\ \overline{T}\cap K^*\neq\emptyset}} \overline{T}. \tag{26}$$

Then, $d(K_h^*, K^*)\le h$, $\#K_h^*\le M$, and $K_h^*\subset F$. Moreover, $K_h^*\overset{H}{\to} K^*$, and following the stability result for the direct problem [6], we have

$$\int_{\partial\Omega} |w_{K_h,\psi_1}-w_1|^2 + \int_{\partial\Omega} |w_{K_h,\psi_2}-w_2|^2 \longrightarrow 0 \text{ as } h\longrightarrow 0.$$

By the choice of K_h in (24) we get (25). Second, since (25) holds, we use Theorem 4.3 and get $G_{K^*}=G_{\tilde{K}}$, which means that K_h is an approximation of K^*. ∎

Remark 5.2. Notice that in the least square approximation (problem (24)), the continuous solutions $w_{K,\psi_1}, w_{K,\psi_2}$ are chosen to be compared to the measures w_1, w_2. In practice, instead of w_{K_h,ψ_i}, we use a finite element approximation, say $w^j_{K_h,\psi_i}$, obtained on a finer mesh. This approximation can be chosen such that $\|w^j_{K_h,\psi_i} - w_{K_h,\psi_i}\|_{L^2(\partial\Omega)}\le j$, $j<h$. Consequently, as h goes to zero, the result of Theorem 5.1 still holds.

In what follows we consider the approximation problem for the perfectly conducting case. Let $K^*\subseteq F$ be a defect such that $\sharp K^*\le M$ which gives the measures u_1, u_2 corresponding to the input fluxes ψ_1, ψ_2, respectively. We solve the following finite dimensional problem:

$$\min_{\substack{K\subset\mathcal{T}_h\cap F\\ \#K\le M}} \int_{\partial\Omega} |u^h_{K,\psi_1}-u_1|^2\,d\sigma + \int_{\partial\Omega} |u^h_{K,\psi_2}-u_2|^2\,d\sigma. \tag{27}$$

THEOREM 5.3. *For $h\longrightarrow 0$, there exists a subsequence such that*

$$K_h \overset{H}{\to} \tilde{K} \quad \textit{and} \quad G_{K^*} = G_{\tilde{K}}.$$

Proof. By compactness, we can extract a subsequence $K_h \overset{H}{\to} \tilde{K}$. We prove that for the continuous solutions we have

$$\int_{\partial\Omega} |u_{K_h,\psi_1} - u_1|^2 \, d\sigma + \int_{\partial\Omega} |u_{K_h,\psi_2} - u_2|^2 \, d\sigma \longrightarrow 0 \text{ as } h \longrightarrow 0. \tag{28}$$

For $i = 1, 2$ we have that

$$\int_{\partial\Omega} |u_{K_h,\psi_i} - u_i|^2 \, d\sigma \leq 2 \left(\int_{\partial\Omega} |u^h_{K_h,\psi_i} - u_i|^2 \, d\sigma + \int_{\partial\Omega} |u_{K_h,\psi_i} - u^h_{K_h,\psi_i}|^2 \, d\sigma \right).$$

We construct K_h^* as in (26). Then, for $i = 1, 2$,

$$\begin{aligned} \int_{\partial\Omega} |u^h_{K_h,\psi_i} - u_i|^2 \, d\sigma &\leq \int_{\partial\Omega} |u^h_{K_h^*,\psi_i} - u_i|^2 \, d\sigma \\ &\leq 2 \left(\int_{\partial\Omega} |u_{K_h^*,\psi_i} - u_i|^2 \, d\sigma + \int_{\partial\Omega} |u^h_{K_h^*,\psi_i} - u_{K_h^*,\psi_i}|^2 \, d\sigma \right). \end{aligned}$$

From the stability of the direct problem (see [6]) we get

$$\int_{\partial\Omega} |u_{K_h^*,\psi_i} - u_i|^2 \, d\sigma \longrightarrow 0 \text{ as } h \longrightarrow 0.$$

In order to get (28) we have to prove that

$$\int_{\partial\Omega} |u_{K_h,\psi_i} - u^h_{K_h,\psi_i}|^2 \, d\sigma + \int_{\partial\Omega} |u_{K_h^*,\psi_i} - u^h_{K_h^*,\psi_i}|^2 \, d\sigma \longrightarrow 0 \text{ as } h \longrightarrow 0.$$

In fact, it is enough to prove that if

$$K_h \overset{H}{\to} \tilde{K}, \ \#K_h \leq M, \ K_h \subset F,$$

then $u^h_{K_h,\psi_i} \overset{L^2(\partial\Omega)}{\to} u_{\tilde{K},\psi_i}$. This is a consequence of the Mosco convergence of the spaces

$$V_h = \left\{ u \in C(\overline{\Omega}), \ u \in P_1(T) \, \forall T \in \mathcal{T}_h, u = \text{constant on each connected component of } K_h \right\}$$

to $H^1_{cond,\tilde{K}}(\Omega)$.

Indeed, let $v_h \in V_h$, $v_h \overset{H^1(\Omega)}{\rightharpoonup} u$. Following [6], $u \in H^1(\Omega)$, u is constant on each connected component of $\tilde{K}$, and hence $u \in H^1_{cond,\tilde{K}}(\Omega)$. Now let $u \in H^1_{cond,\tilde{K}}(\Omega)$. Applying Hedberg's result [13] locally in a neighborhood of each connected component of $\tilde{K}$, for every $\epsilon > 0$, there exists $\delta > 0$ and $u_\delta \in H^1_{cond,\overline{\tilde{K}^\delta}}(\Omega) \cap C^\infty(\overline{\Omega})$, such that $|u_\delta - u|_{H^1(\Omega)} < \epsilon$. Then, $u_\delta \in H^1_{cond,K_h}(\Omega)$ for h small enough. Thus, for the finite element approximation $u^h_\delta \in V_h$ we have the error estimate $|u_\delta - u^h_\delta| \leq h \|u_\delta\|_{H^2(\Omega)}$. By a diagonal procedure, we construct $u^h_{\delta_h} \in V_h$, and $u^h_{\delta_h} \overset{H^1(\Omega)}{\to} u$. ∎

Remark 5.4. In Theorem 5.3, we have the approximation of $u_{K^*,\psi}$ obtained on $\mathcal{T}_h$, which means that no refinement is necessary. This is mainly due to the Dirichlet-type boundary conditions, which are easier to handle than the Neumann ones.

6. Appendix.

Proof of Example 3.7. We start with the following simple result.

LEMMA 6.1. *Let $(c_n)_n$ and $(r_n)_n$ be two sequences of real numbers such that $\forall\, n$, $0 < r_n \le c_n$, $(r_n)_n$ is decreasing, and $(c_n)_n$ converges to zero. Then, there exists a subsequence $(c_{n_k})_{n_k}$ such that $|c_{n_k} - c_{n_k+1}| \ge \frac{r_{n_k} - r_{n_k+1}}{2}$.*

Proof. Assume for contradiction that there exists n_0 such that

$$\forall n \ge n_0, \quad |c_n - c_{n+1}| < \frac{r_n - r_{n+1}}{2}.$$

Then, $\forall\, k > n_0$, we have

$$|c_{n_0} - c_k| < \frac{r_{n_0} - r_k}{2},$$

which yields, when k goes to $+\infty$, $c_{n_0} \le \frac{r_{n_0}}{2}$, in contradiction with the hypothesis of the lemma. □

Now choose r_n as in Example 3.7. Then $\overline{\Omega \setminus K} = \overline{\Omega}$ and take $\phi \in C(\overline{\Omega}) \cap H^1_{cond,K\cap B_{0,r}}(\Omega)$ such that $\phi(0) = 0$. Suppose for contradiction that 0 is not conductive. Then there exists $C, \delta > 0$ such that $\phi(x) \ge C|x|$ for $x \in K \cap B_{0,\delta}$. Since $\phi \in H^1_{cond,K\cap B_{0,r}}(\Omega)$ we have that ϕ is constant on every vertical line. Let us denote by c_n this constant. Writing $\phi(b_n, r_n) \ge C\sqrt{b_n^2 + r_n^2}$ we get that $c_n \ge \sqrt{b_n^2 + r_n^2} \ge r_n$.

We prove that the gradient of ϕ has an infinite L^2-norm, and this will contradict the hypothesis $\phi \in H^1_{cond,K\cap B_{0,r}}(\Omega)$. We have the following:

$$\begin{aligned}
\int_\Omega |\nabla\phi|^2 dx &\ge \sum_{n_0}^{\infty} \int_{[b_{n+1}, b_n]\times r_{n+1}} |\nabla\phi|^2 dx \\
&\ge \sum_{n_0}^{\infty} \int_{[b_{n+1}, b_n]\times r_{n+1}} \Big(\frac{c_{n+1} - c_n}{b_n - b_{n+1}}\Big)^2 dx \\
&\ge \sum_{n_0}^{\infty} \frac{r_{n+1}}{b_n - b_{n+1}} (c_{n+1} - c_n)^2.
\end{aligned}$$

Using Lemma 6.1, there exists a subsequence such that

$$\frac{r_{n_k+1}}{b_{n_k} - b_{n_k+1}} (c_{n_k+1} - c_{n_k})^2 \ge \frac{r_{n_k+1}}{4(b_{n_k} - b_{n_k+1})} (r_{n_k+1} - r_{n_k})^2.$$

We observe from the definition of b_n and r_n that $b_n - b_{n+1} = r_{n+1}(r_{n+1} - r_n)^2$; therefore the series above cannot converge since its general term does not converge to zero.

Proof of Proposition 3.8. Let K be a compact subset of Ω such that $\Omega \setminus K$ is connected, and let $x \in \partial(\Omega \setminus K)$ such that $x \in U_x \subseteq K$, where U_x is a continuum of positive diameter. Following Proposition 3.5, in order to prove that x is conductive for $\Omega \setminus K$ it is enough to prove the existence of a continuum of positive diameter F such that $x \in F \subseteq \partial(\Omega \setminus K)$.

One can mimic the proof of the second case in the proof of Theorem 3.9. The only difference is that a curve γ joining x to $\partial\Omega$ and lying in $\Omega \setminus K$ may not exist. Nevertheless, there exists a sequence of points $y_n \in \Omega \setminus K$, $y_n \to x$ and smooth curves γ_n joining y_n to a point of the $\partial\Omega$ and lying in $\Omega \setminus K$. Choosing ε_n as in Theorem 3.9 and choosing y_n such that $|x - y_n| < \varepsilon_n/2$, we define

$$t_n = \min\{t \in [0,1], \gamma_n(t) \in P_n\}. \tag{29}$$

We observe that z_n is well defined, but we do not have necessarily $z_n \to x$. Nevertheless, $\gamma_n([0,t_n])$ is a continuum containing y_n and z_n, and lying in $(\Omega \setminus K) \cap \overline{K}^{\varepsilon_n}$. Two possibilities occur: either for a subsequence we have $z_n \to x$ and apply the same argument as in Theorem 3.9, or $|z_n - x| \geq \alpha > 0$ and any Hausdorff limit of $\gamma_n([0,t_n])$ is a continuum of diameter greater than or equal to α contained in $\partial(\Omega \setminus K)$ and passing through x.

Example 6.2 (example of a Cantor set which is uniquely identifiable by two boundary measurements). Let $\Omega = B(0,2)$ and define

$$F_1 = [0,1] \times \{0\},$$

$$F_2 = \left\{ \left[0, \frac{1}{2} - \varepsilon_1\right] \cup \left[\frac{1}{2} + \varepsilon_1, 1\right] \right\} \times \{0\},$$

$$F_3 = \left\{ \left[0, \frac{1}{2}\left(\frac{1}{2} - \varepsilon_1\right) - \varepsilon_2\right] \cup \left[\frac{1}{2}\left(\frac{1}{2} - \varepsilon_1\right) + \varepsilon_2, \frac{1}{2} - \varepsilon_1\right] \right.$$
$$\left. \cup \left[\frac{1}{2} + \varepsilon_1, \frac{1}{2}\left(\frac{1}{2} + \varepsilon_1 + 1\right) - \varepsilon_2\right] \cup \left[\frac{1}{2}\left(\frac{1}{2} + \varepsilon_1 + 1\right) + \varepsilon_2, 1\right] \right\} \times \{0\},$$

etc. Let $(c_n)_n$ be an increasing sequence of positive numbers converging to 1. The value of ε_1 is chosen such that $\forall \varphi \in H^1_{cond,F_1}(\Omega \setminus F_1) \cap C(\overline{\Omega})$, $\varphi(0,0) = 0, \int_\Omega |\nabla \varphi|^2 dx \leq 1$ we have for every $t \in [0,1]$ $\int_0^t \varphi^2(s,0) ds \leq c_1 t^4$. Such a constant ε_1 exists; if it does not, for a sequence φ_k corresponding to $\varepsilon_1^k \to 0$ we would have

$$\int_0^{t_k} \varphi_k^2(s,0) ds > c_1 t_k^4.$$

Assuming $t_k \to t$, we clearly have $t \geq 1/2$ and for the limit function we get (by the continuity of the trace operator) $\int_0^t \varphi^2(s,0) ds \geq c_1 t^4$. But from Lemma 2.1 we have $\varphi = 0$ on $[0,1] \times \{0\}$, and hence we get a contradiction.

Note that ε_1 can be chosen such that

$$\forall \varphi \in H^1_{cond,F_1}(\Omega \setminus F_1) \cap C(\overline{\Omega}),\ \varphi(1,0) = 0, \int_\Omega |\nabla \varphi|^2 dx \leq 1, \tag{30}$$

we have

$$\forall t \in [0,1], \int_t^1 \varphi^2(s,0) ds \leq c_1 t^4. \tag{31}$$

As well, we define $\varepsilon_2 > 0$ such that $\forall \varphi \in H^1_{cond,F_1}(\Omega \setminus F_1) \cap C(\Omega)$, $\varphi(0,0) = 0, \int_\Omega |\nabla \varphi|^2 dx \leq 1$, we have for every $t \in [0,1]$ $\int_0^t \varphi^2(s,0) ds \leq c_2 t^4$. Note that ε_2 can be chosen such that similar relations as in (30)–(31) hold for the points $(1/2 - \varepsilon_1, 0)$, $(1,0)$ in the "left" direction and for $(1/2 + \varepsilon_1, 0)$ in the "right" direction.

By induction, we define F_n and set $F = \cap_{n \in \mathbb{N}} F_n$, which is a totally disconnected Cantor set. Since $\cup_{n \in \mathbb{N}} \partial F_n$ is dense in F, it is enough to prove that F is conductive at every point of ∂F_n. So fix n and choose $x_0 \in \partial F_n$. There are two possibilities: either x_0 is a left end point of an interval of F_n or a right end point. Thanks to the construction of ε_k, both situations are treated in the same way. If it is a left end point, the proof is similar to the conductivity of 0 that we give in the following.

Let $u \in H^1_{cond,F}(\Omega \setminus F) \cap C(\overline{\Omega})$. There exists $\varphi_\varepsilon \in H^1(\Omega)$ such that $\nabla\varphi_\varepsilon = 0$ on F^ε, $\varphi_\varepsilon \to u$ in $H^1(\Omega)$. Moreover, the functions φ_ε can be chosen continuous and may be translated by a constant such that $\varphi_\varepsilon(0) = 0$. It is clear that, even after translations, $\nabla\varphi_\varepsilon \to \nabla u$ strongly in L^2. Since for every $\varepsilon > 0$ we have for n large enough $F_n \subseteq F^\varepsilon$ we have from the previous construction that for every $t \in [0,1]$ $\int_0^t \varphi_\varepsilon^2(s,0)ds \leq (M+1)t^4$, where $M = \limsup_{n\to\infty} \int_\Omega |\nabla\varphi_n|^2 dx$. This implies (for a subsequence) that φ_ε converges weakly in $H^1(\Omega)$ to $\tilde{u} = u + c$, where c is a constant. The continuity of the trace operator gives

$$\forall t \in [0,1],\ \int_0^t \tilde{u}^2(s,0)ds \leq (M+1)t^4, \tag{32}$$

and the continuity of $\tilde{u}$ implies $\tilde{u}(0,0) = 0$, hence $c = 0$, and thus u satisfies (32). Consequently $\liminf_{t\to 0} \frac{|u(t,0)|}{t} = 0$, and hence $(0,0)$ is conductive; otherwise $|u(t,0)| \geq c|t|$ in a neighborhood of 0, which contradicts (32).

REFERENCES

[1] G. Alessandrini and A. Diaz Valenzuela, *Unique determination of multiple cracks by two measurements*, SIAM J. Control Optim., 34 (1996), pp. 913–921.

[2] G. Alessandrini and L. Rondi, *Optimal stability for the inverse problem of multiple cavities*, J. Differential Equations, 176 (2001), pp. 356–386.

[3] H. Brezis, *Analyse Fonctionelle*, Masson, Paris, 1983.

[4] K. Bryan and M. Vogelius, *A review of selected works on crack identification*, in Geometric Methods in Inverse Problems and PDE Control, IMA Vol. Math. Appl. 137, Springer, New York, 2004, pp. 25–41.

[5] D. Bucur, *Characterization for the Kuratowski limits of a sequence of Sobolev spaces*, J. Differential Equations, 151 (1999), pp. 1–19.

[6] D. Bucur and N. Varchon, *A duality approach for the boundary variation of Neumann problems*, SIAM J. Math. Anal., 34 (2002), pp. 460–477.

[7] D. Bucur and J.-P. Zolésio, *N-Dimensional shape optimization under capacitary constraints*, J. Differential Equations, 123 (1995), pp. 504–522.

[8] D. Chenais and E. Zuazua, *Finite element approximation of 2D elliptic optimal design*, J. Math. Pures Appl., 85 (2006), pp. 225–249.

[9] G. Dal Maso, *Some necessary and sufficient conditions for the convergence of unilateral convex sets*, J. Funct. Anal., 62 (1985), pp. 119–159.

[10] G. Dal Maso, F. Ebobisse, and M. Ponsiglione, *A stability result for nonlinear Neumann problems under boundary variations*, J. Math. Pures Appl., 82 (2003), pp. 503–532.

[11] A. Friedman and M. Vogelius, *Determining cracks by boundary measurements*, Indiana Univ. Math. J., 38 (1989), pp. 527–556.

[12] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[13] L. I. Hedberg, *Spectral synthesis in Sobolev spaces and uniqueness of solutions of Dirichlet problems*, Acta Math., 147 (1981), pp. 237–264.

[14] J. Heinonen, T. Kilpelainen, and O. Martio, *Nonlinear Potential Theory of Degenerate Elliptic Equations*, Clarendon Press, Oxford, New York, Tokyo, 1993.

[15] M. V. Keldys, *On the Solvability and Stability of the Dirichlet Problem*, Amer. Math. Soc. Transl., 51 (1966), pp. 1–73.

[16] T. Kilpelainen and J. Maly, *Supersolutions to degenerate elliptic equation on quasi open sets*, Comm. Partial Differential Equations, 17 (1992), pp. 371–405.

[17] H. Kim and J. K. Seo, *Unique determination of a collection of a finite number of cracks from two boundary measurements*, SIAM J. Math. Anal., 27 (1996), pp. 1336–1340.

[18] L. Rondi, *Optimal stability of reconstruction of plane Lipschitz cracks*, SIAM J. Math. Anal., 36 (2005), pp. 1282–1292.

[19] V. Šveràk, *On optimal shape design*, J. Math. Pures Appl., 72 (1993), pp. 537–551.

[20] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, Scripta Series in Mathematics, V. H. Winston & Sons, Washington, DC; John Wiley & Sons, New York-Toronto, Ont.-London, 1977.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 274–287

OPTIMAL BILINEAR CONTROL OF AN ABSTRACT SCHRÖDINGER EQUATION*

KAZUFUMI ITO† AND KARL KUNISCH‡

Abstract. Well-posedness of abstract quantum mechanical systems is considered and the existence of optimal control of such systems is proved. First order optimality systems are derived. Convergence of the monotone scheme for the solution of the optimality system is proved.

Key words. Schrödinger equation, C_0-groups, optimal control, optimality systems, monotone scheme

AMS subject classifications. 35Q40, 49K20, 65M12

DOI. 10.1137/05064254X

1. Introduction.

We consider a quantum mechanical system with internal Hamiltonian H_0 prepared in the initial state $\Psi_0(x)$, where x denotes the relevant spatial coordinate. The state $\Psi(x,t)$ satisfies the time-dependent Schrödinger equation (we set $h=1$). In the presence of an external interaction taken as an electric field, modeled by a coupling operator with amplitude $\epsilon(t) \in \mathbb{R}$ and a time-independent dipole moment operator $\hat{\mu}$, the new Hamiltonian $H_0 - \mu(t)$ gives rise to the control system

$$i\,\frac{\partial}{\partial t}\Psi(x,t) = (H_0 - \mu(t))\Psi(x,t), \quad \Psi(x,0) = \Psi_0(x), \tag{1.1}$$

where $\mu(t) = \epsilon(t)\hat{\mu}$. Here $\mu(t)$ represents a controlled Hamiltonian which can be a distributed control. The optimal control approach (see, e.g., [MT], [PDR], [TKO], [ZR]) allows us to assess the fitness of the final state $\Psi(T)$ to a prescribed goal. This is achieved through the introduction of a performance index J which is maximized. One possible choice for a cost functional is given by

$$J(\mu) = \frac{1}{2}\langle \Psi(T)|O|\Psi(T)\rangle - \frac{\alpha}{2}\int_0^T |\mu(t)|^2\,dt, \tag{1.2}$$

where $\alpha > 0$ and O is the observable operator that encodes the goal. The larger the value $\langle \Psi(T)|O|\Psi(T)\rangle$ is, the better the control objective is met. Here we used the notation $\langle \Psi(T)|O|\Psi(T)\rangle = \int_\Omega \overline{\Psi(T,x)}\,O\Psi(T,x)\,dx$. The conditions that we utilize for H_0, $\mu(\cdot)$ and O will be given in the following section. Maximization of $\langle \Psi(T)|O|\Psi(T)\rangle$ is at the price of a large laser influence $\int_0^T |\mu(t)|^2\,dt$. The optimally controlled evolution must therefore balance between the expense for the laser influence and the desire that the observable has an acceptably large value.

*Received by the editors October 12, 2005; accepted for publication (in revised form) September 6, 2006; published electronically April 13, 2007.
http://www.siam.org/journals/sicon/46-1/64254.html
†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (kito@unity.ncsu.edu). This research was partially supported by the Army Research Office under DAAD19-02-1-0394.
‡Institute for Mathematics and Scientific Computing, Karl-Franzens-Universität Graz, A-8010 Graz, Austria (karl.kunisch@uni-graz.at). This research was supported in part by the Radon Institute, Linz, Austria.

An alternative cost is given by

$$J(\mu) = -\frac{1}{2}\left(|\Psi(T) - \bar{\Psi}(T)|^2 + \alpha \int_0^T |\mu(t)|^2 \, dt \right),$$

where $\bar{\Psi}$ is a target state. Since $|\Psi|_{L^2} = 1$, it is equivalent to

$$J(\mu) = \mathrm{Re}\,(\Psi, \bar{\Psi}) - \frac{\alpha}{2} \int_0^T |\mu(t)|^2 \, dt. \tag{1.3}$$

In section 2 we shall establish well-posedness results for (1.1) based on a semigroup framework in a form that will facilitate the optimal control treatment. Section 3 is devoted to the precise statement of the optimal control problem, including the class of admissible control operators μ which are considered, and a proof for the existence of optimal solutions. First order necessary optimality conditions are derived in section 4. In section 5 we describe the monotone scheme for the general class of optimal problems that is considered in this paper. Well-posedness and subsequential convergence of the scheme are proved.

To point at some of the relevant literature for the problem under investigation we mention the pioneering work of Rabitz and collaborators; see, e.g., [PDR], [ZR], and the references given there. For existence of optimal controls we refer to [BP]. Differently from our semigroup approach, the work in [BP] is based on partial differential equation techniques, and requires higher regularity in time. Many important aspects of the monotone scheme for the solution of the optimality system were investigated in, e.g., [MST], [MT], [S], [TKO]. However, except for [S], which treats the case of scalar-valued controls, convergence proofs of the optimal controls and states have received little attention so far. The technique of proof in [S] and in the present work are different. While the key ingredient for the convergence proof in [S] is a convergence result in [BMS] for the convolution of a Hilbert-space valued function with a sequence of weakly convergent scalar-valued functions, our results are based on compactness arguments. This allows for finite dimensional (in space) as well as infinite dimensional (distributed) control action.

2. Well-posedness. Setting $\Psi(t,x) = \Psi_1(t,x) + \Psi_2(t,x)$ and $\Psi = (\Psi_1, \Psi_2)$, system (1.1) can equivalently be written as

$$\begin{aligned} \frac{\partial}{\partial t}\Psi_1(t,x) &= (H_0 - \mu(t))\,\Psi_2(t,x), \\ \frac{\partial}{\partial t}\Psi_2(t,x) &= -(H_0 - \mu(t))\,\Psi_1(t,x) \quad \text{for } (t,x) \in (0,T] \times \Omega. \end{aligned} \tag{2.1}$$

Here $T > 0$ and $\Omega = \mathbb{R}^n$ or Ω is a bounded subset of $\mathbb{R}^n$. The behavior of Ψ at the boundary of Ω is defined through the domain of the operator H_0. We refer to section 3 for specific examples. Throughout it is assumed that H_0 is a densely defined, self-adjoint positive semidefinite operator in a real Hilbert space H, consisting of functions defined over the domain Ω. Typically H is $L^2(\Omega)$. If H_0 satisfies the above assumptions it is necessarily closed. We define the closed linear operator A_0 in $X = H \times H$ by

$$A_0 = \begin{pmatrix} 0 & H_0 \\ -H_0 & 0 \end{pmatrix},$$

with $dom\,(A_0) = dom\,(H_0) \times dom\,(H_0)$. Note that A_0 is skew-adjoint, i.e.,

$$(A_0\Psi, \hat{\Psi}) = -(A_0\hat{\Psi}, \Psi) \text{ for all } \Psi, \hat{\Psi} \in dom\,(A_0).$$

Consequently by Stone's theorem [HP] A_0 generates a C_0-group $S(t)$ on X satisfying $|S(t)\Psi|_X = |\Psi|_X$ for all $\Psi \in X$ and $t \geq 0$. Let

$$V = dom\,(H_0^{\frac{1}{2}}) \quad \text{and} \quad \mathcal{V} = V \times V.$$

Then $H_0 \in \mathcal{L}(V, V^*)$ and thus $A_0 \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$, where

$$V^* \quad \text{and} \quad \mathcal{V}^* = V^* \times V^*$$

denote the dual space of V and $\mathcal{V}$, respectively, with H and X as pivot spaces. V is equipped with

$$|\phi|_V^2 = \langle H_0\phi, \phi\rangle_{V^*\times V} + |\phi|_H^2$$

as norm. Then the restriction of $S(t)$ to $\mathcal{V}$ is again a C_0-group. The dual $S^*(-t)$ is the extension of $S(t)$ to $\mathcal{V}^*$ and forms a C_0-group on $\mathcal{V}^*$. Moreover, for the extension group on $\mathcal{V}^*$ the domain of the generator is given by $dom_{\mathcal{V}^*}(A_0) = \mathcal{V}^*$.

Suppose that $\mu(t) \in \mathcal{L}(H)$ is self-adjoint for almost every $t \in (0, T)$ and define

$$B(t) = \begin{pmatrix} 0 & \mu(t) \\ -\mu(t) & 0 \end{pmatrix}.$$

In the context of an external interaction with an electric field, as mentioned in the introduction, $\mu(t) = \epsilon(t)\hat{\mu}$, where ϵ denotes a scalar-valued amplitude and $\hat{\mu} = \hat{\mu}(x)$ is a multiplication operator representing the dipole moment [MT], [MST], [ZR].

By a fixed point argument it can be argued that for every $T > 0$, $\mu \in L^2(0, T; \mathcal{L}(H))$, and $\Psi_0 \in X$ there exists a unique mild solution $\Psi \in C(0, T; X)$ to (2.1) satisfying

$$\Psi(t) = S(t)\Psi_0 - \int_0^t S(t-s)\, B(s)\Psi(s)\, ds \quad \text{for} \quad t \in [0, T]. \tag{2.2}$$

Here $C(0, T; X)$ stands for $C([0, T]; X)$. Moreover, if $\hat{\Psi} \in C(0, T; X)$ denotes the mild solution to (2.1) corresponding to $(\hat{\Psi}_0, \hat{\mu}) \in X \times L^2(0, T; H)$, then by Gronwall's inequality

$$|\Psi - \hat{\Psi}|_{C(0,T;X)} \leq \tilde{M} \left(|\Psi_0 - \hat{\Psi}_0|_X + \int_0^T |\mu(t) - \hat{\mu}(t)|_{\mathcal{L}(H)}\, dt \right), \tag{2.3}$$

for a constant $\tilde{M}$ depending continuously on $|\mu|_{L^1(0,T;\mathcal{L}(H))}$ and $|\Psi_0|_X$.

THEOREM 2.1. *If $\Psi_0 \in \mathcal{V}$ and $\mu \in L^2(0, T; \mathcal{L}(V) \cap \mathcal{L}(H))$, then the mild solution $\Psi \in C(0, T; X)$ to (2.2) satisfies*

$$\Psi(t) \in H^1(0, T; \mathcal{V}^*) \cap C(0, T; \mathcal{V})$$

and

$$\frac{d}{dt}\Psi(t) = (A_0 - B(t))\Psi(t) \text{ a.e. in } (0, T).$$

Moreover $|\Psi(t)|_X = |\Psi_0|_X$ *for all* $t \in [0,T]$*;*

$$|\Psi(t)|_{\mathcal{V}} \le K_1 \exp\left(K_2 \int_0^t |\mu(s)|_{\mathcal{L}(V)}\, ds\right) |\Psi_0|_{\mathcal{V}}$$

for constants K_i *independent of* μ *and* Ψ_0*, and for some* M_1 *depending continuously on its arguments*

$$\left|\frac{d}{dt}\Psi(t)\right|_{L^2(0,T;\mathcal{V}^*)} \le M_1\left(|\mu|_{L^2(0,T;\mathcal{L}(\mathcal{V})\cap\mathcal{L}(H))}, |\Psi_0|_{\mathcal{V}}\right). \tag{2.4}$$

Proof. Consider

$$A_0\Psi(t) = S(t)A_0\Psi_0 - \int_0^t S(t-s)\, A_0 B(s)\Psi(s)\, ds \text{ in } \mathcal{V}^*. \tag{2.5}$$

Adding this equation to (2.2) we find the a priori estimate

$$|\Psi(t)|_{\mathcal{V}} \le K_1|\Psi_0|_{\mathcal{V}} + K_2 \int_0^t |B(s)|_{\mathcal{L}(\mathcal{V})}|\Psi(s)|_{\mathcal{V}}\, ds$$

for embedding constants K_1, K_2. By Gronwall's inequality we have

$$|\Psi(t)|_{\mathcal{V}} \le K_1|\Psi_0|_{\mathcal{V}} \exp\left(K_2 \int_0^t |\mu(s)|_{\mathcal{L}(V)}\, ds\right) \text{ for } t \in (0,T).$$

This estimate allows us to verify existence of a solution to (2.5) in $C(0,T;\mathcal{V})$, which coincides with the solution to (2.2). By construction we have that $\Psi \in C(0,T; dom_{\mathcal{V}^*}(A_0))$. It follows with standard arguments (see, e.g., [P, p. 107]) applied to (2.2) that Ψ is differentiable almost everywhere in $(0,T)$ and that

$$\frac{d}{dt}\Psi(t) = A_0\Psi(t) - B(t)\Psi(t) \text{ in } \mathcal{V}^* \text{ for a.e. in } (0,T).$$

Hence $\Psi \in H^1(0,T;X) \cap C(0,T;\mathcal{V})$. In fact we have

$$\left|\frac{d}{dt}\Psi\right|_{L^2(0,T;\mathcal{V}^*)} \le K(|\Psi|_{L^2(0,T;\mathcal{V})} + |\mu|_{L^2(0,T;\mathcal{L}(H))}|\Psi|_{C(0,T;H)}),$$

which implies (2.4). Since

$$\frac{1}{2}\frac{d}{dt}|\Psi(t)|_X^2 = \left\langle \frac{d}{dt}\Psi(t), \Psi(t)\right\rangle_{\mathcal{V}^*,\mathcal{V}} = \langle (A_0 - B(t))\,\Psi(t), \Psi(t)\rangle_{\mathcal{V}^*,\mathcal{V}} = 0$$

for a.e. $t \in (0,T)$, it follows that $|\Psi(t)|_X = |\Psi_0|_X$ for all $t \in [0,T]$. □

3. Existence of an optimal solution. In this section we provide sufficient conditions for the existence of a solution to

$$\begin{cases} \max\, J(\mu) \text{ over } \mu \in L^2(0,T;U) \\ \text{subject to (2.2),} \end{cases} \tag{3.1}$$

where $J(\mu) = \frac{1}{2}\langle\Psi(T)|O|\Psi(T)\rangle - \frac{\alpha}{2}\int_0^T |\mu(t)|^2\, dt$, with $O \in \mathcal{L}(X)\cap\mathcal{L}(\mathcal{V})$ a self-adjoint positive definite operator. Here $\langle\Psi(T)|O|\Psi(T)\rangle$ stands for $(\Psi(T), O\Psi(T))_X$, with $(\cdot,\cdot)_X$ denoting the inner product in X.

Here U is a closed Hilbert space continuously embedded in $\{\mu \in \mathcal{L}(H) \cap \mathcal{L}(V) : \mu \text{ is self-adjoint}\}$. We assume that there exists a closed subspace $H_1 \subset H$ such that for $X_1 = H_1 \times H_1$ we have

$$\mathcal{V} \cap X_1 \text{ is compactly embedded into } X \tag{3.2}$$

and

$$|\Psi|_{L^2(0,T;\mathcal{V}\cap X_1)} \leq M\,(|\Psi_0|_{\mathcal{V}\cap X_1},\, |\mu|_{L^2(0,T;U)}), \tag{3.3}$$

where M depends continuously on its arguments, and Ψ denotes the solution to (2.2). Since

$$J(\mu) \to -\infty \text{ as } |\mu|_{L^2(0,T;U)} \to \infty,$$

there exists a maximizing sequence $\{\mu_n\}$ to (3.1), i.e.,

$$\lim_{n\to\infty} J(\mu_n) = \sup_{\mu\in L^2(0,T;U)} J(\mu) \quad \text{and} \quad |\mu_n|_{L^2(0,T;U)} \leq K,$$

for some K independent of n. Hence there exists a subsequence of $\{\mu_n\}$ denoted by the same symbol and $\bar{\mu} \in L^2(0,T;U)$ such that

$$\mu_n \to \bar{\mu} \text{ weakly in } L^2(0,T;U). \tag{3.4}$$

By (2.4) and (3.3) the sequence $\{\Psi_n\}$ is bounded in $L^2(0,T;X_1\cap\mathcal{V})$ and the sequence $\{\frac{d}{dt}\Psi_n\}$ is bounded in $L^2(0,T;\mathcal{V}^*)$, where $\Psi_n = \Psi(\mu_n)$ denotes the solution to (2.2) with μ replaced by μ_n. By Aubin's lemma, e.g., [CF], there exists $\bar{\Psi} \in H^1(0,T;\mathcal{V}^*)\cap L^2(0,T;\mathcal{V}\cap X_1)$ such that for a further subsequence

$$\Psi_n \to \bar{\Psi} \text{ strongly in } L^2(0,T;X) \tag{3.5}$$

and weakly in $L^2(0,T;\mathcal{V})$. For φ and ψ in X the mapping $B \to (B\varphi,\psi)_X, B \in U$, defines a bounded linear functional in U. Hence by the Riesz representation theorem there exists $F = F(\varphi,\psi) \in U$ such that

$$(B\varphi,\psi)_X = (F(\varphi,\psi),\mu)_U \text{ for all } \mu \in U, \text{ where } B = \begin{pmatrix} 0 & \mu \\ -\mu & 0 \end{pmatrix}. \tag{3.6}$$

Note that $F : X \times X \to U$ is a continuous, bilinear mapping satisfying

$$F(\varphi,\psi) = -F(\psi,\varphi).$$

Moreover, if $\psi_n \to \psi$ strongly in $L^2(0,T;X)$ and $\varphi_n \to \varphi$ strongly in $C(0,T;X)$ we have

$$F(\varphi_n,\psi_n) \to F(\varphi,\psi) \text{ in } L^2(0,T;U). \tag{3.7}$$

Taking the inner product in $L^2(0,T;X)$ of

$$\Psi_n(t) = S(t)\Psi_0 - \int_0^t S(t-s)\,B_n(s)\,\Psi_n(s)\,ds$$

with an arbitrary $\Phi \in L^2(0,T;X)$ implies that

$$\int_0^T (\Psi_n(t),\Phi(t))_X\,dt = \int_0^T (S(t)\Psi_0,\Phi(t))_X\,dt + \int_0^T \left(F\left(\int_\cdot^T S^*(t-\cdot)\Phi(t)\,dt,\Psi_n\right),B_n\right)_U ds.$$

From (3.4), (3.5), and (3.7) we deduce that

$$\begin{aligned}\int_0^T (\Psi(t),\Phi(t))_X\,dt &= \int_0^T (S(t)\Psi_0,\Phi(t))_X\,dt \\ &\quad + \int_0^T \left(F\left(\int_\cdot^T S^*(t-\cdot)\Phi(t)\,dt,\Psi\right),B\right)_U ds \\ &= \int_0^T (S(t)\Psi_0,\Phi(t))_X\,dt \\ &\quad - \int_0^T \left(\int_0^t S(t-s)\,B(s)\,\Psi(s)\,ds,\Phi(t)\right)_X dt.\end{aligned}$$

Since Φ was arbitrary we find

$$\bar\Psi(t) = S(t)\Psi_0 - \int_0^t S(t-s)\,\bar B(s)\,\bar\Psi(s)\,ds,$$

and thus $\bar\Psi$ is the unique solution to (2.2) with μ replaced by $\bar\mu$.

We next verify that

$$\Psi_n(T) \to \bar\Psi(T) \text{ strongly in } X. \tag{3.8}$$

For this purpose set $\Phi_n = \Psi_n - \bar\Psi$, and choose K such that

$$\max(|\Phi_n|_{L^2(0,T;\mathcal{V}^*)}, |\Phi_n|_{C(0,T;\mathcal{V}\cap X_1)}) \le K.$$

Due to (3.2) there exists [CF, p. 96], for every $\epsilon > 0$, a constant c_ϵ such that

$$|\Phi_n(T)|_X \le \epsilon|\Phi_n(T)|_{\mathcal{V}\cap X_1} + c_\epsilon|\Phi_n(T)|_{\mathcal{V}^*} \le \epsilon K + c_\epsilon|\Phi_n(T)|_{\mathcal{V}^*}. \tag{3.9}$$

By Hölder's inequality we have

$$\begin{aligned}|\,\Phi_n(T)|_{\mathcal{V}^*} &= \left|\frac{1}{\epsilon}\int_{T-\epsilon}^T \Phi_n(s)\,ds + \frac{1}{\epsilon}\int_{T-\epsilon}^T (s-T+\epsilon)\frac{d}{ds}\Phi_n'(s)\,ds\right|_{\mathcal{V}^*} \\ &\le \frac{1}{\epsilon}\left|\int_{T-\epsilon}^T \Phi_n(s)\,ds\right|_{\mathcal{V}^*} + \frac{1}{\epsilon}\left(\int_{T-\epsilon}^T (s-T+\epsilon)^2\,ds\right)^{\frac12}\left(\int_{T-\epsilon}^T \left|\frac{d}{ds}\Phi'(s)\right|_{\mathcal{V}^*} ds\right)^{\frac12} \\ &\le \frac{1}{\epsilon}\left|\int_{T-\epsilon}^T \Phi_n(s)\,ds\right|_{\mathcal{V}^*} + \frac{\sqrt{\epsilon}\,K}{\sqrt{3}},\end{aligned}$$

and with (3.9)

$$|\Phi_n(T)|_X \le \epsilon K + \frac{\sqrt{\epsilon}\,K}{\sqrt{3}} + \frac{1}{\epsilon}\left|\int_{T-\epsilon}^{T} \Phi_n(s)\,ds\right|_{\mathcal{V}^*}.$$

Since $\Phi_n \to 0$ weakly in $L^2(0,T;\mathcal{V})$ we have $\frac{1}{\epsilon}|\int_{T-\epsilon}^{T} \Phi_n(s)\,ds|_{\mathcal{V}^*} \to 0$ as $n \to \infty$ for every fixed $\epsilon > 0$. We conclude that (3.8) holds.

Weak lower-semicontinuity of norms and (3.8) imply that

$$J(\bar{\mu}) \ge \sup_{\mu} J(\mu)$$

and hence $\bar{\mu}$ is an optimal solution to (3.1). We thus proved the following result.

THEOREM 3.1. *If $\Psi_0 \in X_1 \cap \mathcal{V}$ and* (3.2), (3.3) *hold, then* (3.1) *admits a solution $\bar{\mu} \in L^2(0,T;U)$.*

Example 3.1. The control space in (3.1) is $\mathcal{U} = L^2(0,T;U)$. Here we consider the special case of a scalar-valued control coupling a time-dependent control amplitude ϵ with a fixed time-independent self-adjoint moment operator $\tilde{\mu} \in \mathcal{L}(V) \cap \mathcal{L}(H)$, i.e., we consider the closed subspace of $\mathcal{U}$ given by

$$\widehat{\mathcal{U}} = \{\epsilon\tilde{\mu} : \epsilon \in L^2(0,T;\mathbb{R})\},$$

which is isomorphic to $L^2(0,T;\mathbb{R})$. In this case U is the one dimensional space $\{\epsilon\tilde{\mu} : \epsilon \in \mathbb{R}\}$, which is endowed with the inner product of $\mathbb{R}$. The resulting control cost is $\frac{\alpha}{2}\int_0^T |\epsilon(t)|^2 dt$ and the bilinear mapping $F : X \times X \to U = \mathbb{R}$ is given by

$$F(\phi,\psi) = (\tilde{B}\phi,\psi)_X = (\tilde{\mu}(\phi_2),\psi_1)_H - (\tilde{\mu}(\phi_1),\psi_2)_H,$$

with $\tilde{B} = \left(\begin{smallmatrix} 0 & \tilde{\mu} \\ -\tilde{\mu} & 0 \end{smallmatrix}\right)$. The resulting optimality condition has the form

$$\alpha\bar{\epsilon} + (\tilde{B}\bar{\Psi}(t),\bar{\chi})_X = 0.$$

Example 3.2. Let $H = L^2(\Omega)/\mathbb{R}$ with $\Omega = (0,1)$ and $H_0 = -\Delta$ with periodic boundary conditions. Then $V = H_P^1(\Omega)$, the space of $H^1(\Omega)$ functions with periodic boundary conditions $\phi(0) = \phi(1)$. The control space is taken as multiplication operators by elements $\mu \in H_P^1(\Omega)$ and we identify U with $H^1(\Omega)_p$. Note that $\phi \to \mu\phi$ defines a self-adjoint element in $\mathcal{L}(H) \cap \mathcal{L}(V)$, since $H^1(\Omega)$ is a Banach algebra in dimension one. For $\phi \in X = H \times H$ and $\psi \in X = H \times H$ the element $F(\phi,\psi) \in V$ is the solution to

$$(F(\phi,\psi),\mu)_{H^1} = (\mu\phi^2,\psi^1)_H - (\mu\phi^1,\psi^2)_H \text{ for all } \mu \in V.$$

Thus the optimality condition can be expressed as

$$\alpha\mu(t) + (-\Delta + I)^{-1}(\Psi^2(t)\chi^1(t) - \Psi^1(t)\chi^2(t)) = 0,$$

where $\Psi, \chi \in C(0,T;\mathcal{V})$. Note that this implies additional spatial regularity of the optimal solution.

Example 3.3. Let H_0, H, V, and Ω be as in the previous example. Define

$$\tilde{U} = \{\tilde{\mu} \in L^2(\Omega) : \tilde{\mu}(x) = \tilde{\mu}(-x) \text{ for } x \in \Omega\}$$

endowed with the canonical inner product. Each $\tilde{\mu}$ can be uniquely identified with a self-adjoint operator $\mu \in \mathcal{L}(H)$ given by

$$\mu(\phi)(x) = \int_\Omega \tilde{\mu}(x-y)\phi(y)\,dy,$$

where $\tilde{\mu}$ is extended periodically from Ω to $\mathbb{R}$. All such operators also satisfy $\mu \in \mathcal{L}(V)$. The set of all these operators constitutes the control space U. The resulting penalty term in the cost functional J has the form

$$\frac{\alpha}{2}\int_0^T |\tilde{\mu}(t,\cdot)|^2_{L^2(\Omega)}\,dt.$$

Using symmetry of $\tilde{\mu}$ it can be shown that for ϕ, ψ in $X = H \times H$ the element $F \in L^2(\Omega)$ satisfying

$$(F(\phi,\psi),\tilde{\mu})_{L^2(\Omega)} = (\mu\phi^2,\psi^1)_{L^2(\Omega)} - (\mu\phi^1,\psi^2)_{L^2(\Omega)} \text{ for all } \mu \in U$$

is given by

$$F(\phi,\psi)(x) = \frac{1}{2}\int_\Omega \big(\phi^2(y)\psi^1(x+y) + \phi^2(y)\psi^1(-x+y) - \phi^1(y)\psi^2(x+y) - \phi^1(y)\psi^2(-x+y)\big)\,dy.$$

The resulting optimality condition is

$$\alpha\tilde{\mu}(t,x) + \frac{1}{2}\int_\Omega \big(\Psi^2(t,y)\chi^1(t,x+y) + \Psi^2(t,y)\chi^1(t,-x+y) - \Psi^1(t,y)\chi^2(t,x+y) - \Psi^1(t,y)\chi^2(t,-x+y)\big)\,dy = 0.$$

Analogous results can be obtained with $\Omega = (0,1)$ replaced by bounded cubes in $\mathbb{R}^n$ with H_0 satisfying periodic boundary conditions, or with $\Omega = \mathbb{R}^n$.

Example 3.4. Let $H_0 = -\Delta$ in $H = L^2(\mathbb{R}^n)$. Then H_0 is densely defined with $dom(H_0) = H^2(\mathbb{R}^n)$ and self-adjoint (see, e.g., [K]), with spectrum consisting of continuous spectrum given by $[0,\infty)$. We set $\mathcal{H} = \{\varphi \in L^2(\mathbb{R}^n) : \int_{\mathbb{R}^n}(1+|x|^2)\varphi(x)^2 < \infty\}$. To verify (3.2) let $\{f_n\}_{n=1}^\infty$ be a bounded sequence in $H_1 = V \cap \mathcal{H} = dom(H_0^{\frac{1}{2}}) \cap \mathcal{H}$. For $r \in N$ set $\Omega_r = \{x \in \mathbb{R}^n : |x|_{\mathbb{R}^n} \le r\}$. Extract a subsequence of $\{f_n\}$ that converges weakly in H_1 to some $f \in H_1$. Using compactness of $\{\phi|_{\Omega_r} : \phi \in V\}$ in $L^2(\Omega_r)$ successively extract further subsequences whose restriction to Ω_r converges strongly in $L^2(\Omega_r)$ to f, for $r = 1, 2, \ldots$. Let $\{f_{n_k}\}$ denote the sequence which arises from diagonalization of the above procedure. The restriction to Ω_r of this sequence converges strongly in $L^2(\Omega_r)$ to the restriction of f to Ω_r for each $r \in N$. Strong convergence of $\{f_{n_k}\}$ to $\{f\}$ in $L^2(\mathbb{R}^n)$ follows from the following estimate:

$$\int_{\mathbb{R}^n} |f - f_{n_k}|^2\,dx = \int_{\Omega_r} |f - f_{n_k}|^2\,dx + \int_{\mathbb{R}^n\setminus\Omega_r} |f - f_{n_k}|^2\,|x|^2\,\frac{dx}{|x|^2}$$

$$\le \int_{\Omega_r} |f - f_{n_k}|^2\,dx + \frac{1}{r^2}\int_{\mathbb{R}^n\setminus\Omega_r} |f - f_{n_k}|^2|x|^2\,dx \le \int_{\Omega_r} |f - f_{n_k}|^2\,dx + \frac{4}{r^2}C,$$

where C is the common bound for $\{f_{n_k}\}$ and f in H_1. Hence $dom(H_0^{\frac{1}{2}}) \cap \mathcal{H}$ is compactly embedded in H and (3.2) follows.

Turning to (3.3), consider, for $X = L^2(\mathbb{R}^n) \times L^2(\mathbb{R}^n)$,

$$\langle \tfrac{d}{dt}\Psi(t), |x|^2\Psi(t)\rangle = \tfrac{1}{2}\tfrac{d}{dt}\langle \Psi(t), |x|^2\Psi(t)\rangle$$
$$= \langle A_0\Psi(t), |x|^2\Psi(t)\rangle - (B(t)\Psi(t), |x|^2\Psi(t))_X$$
$$= (-\Delta\Psi_2, |x|^2\Psi_1)_H + (\Delta\Psi_1, |x|^2\Psi_2)_H - ((\mu(t)\Psi_2, |x|^2\Psi_1)_H - (\mu(t)\Psi_1, |x|^2\Psi_2)_H)$$
$$= 2(\nabla\Psi_2(t), x\Psi_1(t))_H - 2(\nabla\Psi_1(t), x\Psi_2(t))_H$$
$$\le |\nabla\Psi_1(t)|_H^2 + |\nabla\Psi_2(t)|_H^2 + \big|\, |x|\Psi_1(t)\big|_H^2 + \big|\, |x|\Psi_2(t)\big|_H^2,$$

and hence

$$\frac{1}{2}\frac{d}{dt}\big|\, |x|\Psi(t)\big|_X^2 \le K|\Psi|_{\mathcal{V}}^2 + \big|\, |x|\Psi(t)\big|_X^2$$

for a constant K satisfying $|\nabla\phi| \le K|\phi|_{\mathcal{V}}$ for all $\phi \in \mathcal{V}$. Gronwall's inequality and Theorem 2.1 imply the existence of a constant $\tilde{M} = \tilde{M}(|\Psi_0|_{\mathcal{V}\cap X_1}, |B|_{L^2(0,T;\mathcal{L}(\mathcal{V})\cap\mathcal{L}(X))})$ such that

$$|\Psi|_{C(0,T;\mathcal{V}\cap X_1)} \le \tilde{M},$$

which, in particular, implies (3.3).

4. Necessary optimality condition. We now derive a first order necessary optimality system for (3.1).

THEOREM 4.1. *Let* $(\bar\mu, \bar\Psi) = (\bar\mu, \Psi(\bar\mu))$ *be an optimal pair for* (3.1) *and assume that* $\Psi_0 \in \mathcal{V}$ *and* $O\,\bar\Psi(T) \in \mathcal{V}$. *Then*

$$\tfrac{d}{dt}\,\bar\Psi(t) = (A_0 - \bar B(t))\bar\Psi(t), \qquad \bar\Psi(0) = \Psi_0 \qquad \textit{(primal equation)},$$
$$\tfrac{d}{dt}\,\bar\chi(t) = (A_0 - \bar B(t))\bar\chi(t), \qquad \chi(T) = O\bar\Psi(T) \qquad \textit{(adjoint equation)},$$
$$\alpha\bar\mu(t) + F(\bar\Psi(t),\, \bar\chi(t)) = 0 \qquad \textit{(optimality)},$$

where the adjoint state satisfies $\bar\chi \in H^1(0,T;\mathcal{V}^*) \cap C(0,T;\mathcal{V})$ *and* $\bar B = \left(\begin{smallmatrix} 0 & \bar\mu \\ -\bar\mu & 0 \end{smallmatrix}\right)$.

Proof. For any $\mu \in L^2(0,T;U)$ we have

$$J(\mu) - J(\bar\mu) = -\alpha\,(\bar\mu, \mu - \bar\mu)_{L^2(0,T;U)} - \frac{\alpha}{2}|\mu - \bar\mu|^2_{L^2(0,T;U)}$$
$$+(\Psi(T) - \bar\Psi(T), O\bar\Psi(T))_X + \frac{1}{2}(\Psi(T) - \bar\Psi(T), O(\Psi(T) - \bar\Psi(T))\,)_X.$$

Let $\bar\chi(t) \in H^1(0,T;\mathcal{V}^*) \cap C(0,T;dom\,\mathcal{V})$ be the solution to the adjoint equation

$$\frac{d}{dt}\bar\chi(t) = (A_0 - \bar B(t))\bar\chi(t), \quad \chi(T) = O\bar\Psi.$$

Then,

$$(\Psi(T) - \bar\Psi(T), O\bar\Psi(T)) = \int_0^T \left\langle \frac{d}{dt}(\Psi(t) - \bar\Psi(t)), \bar\chi(t) \right\rangle + \left\langle \Psi(t) - \bar\Psi(t), \frac{d}{dt}\bar\chi(t) \right\rangle dt$$
$$= \int_0^T [\langle (A_0 - B(t))\Psi(t) - (A_0 - \bar B(t))\bar\Psi(t), \bar\chi(t)\rangle + \langle \Psi(t) - \bar\Psi(t), (A_0 - \bar B(t))\bar\chi(t)\rangle]\, dt$$
$$= -\int_0^T ((B(t) - \bar B(t))\Psi(t), \bar\chi(t))_X\, dt$$
$$= -\int_0^T \big((B(t) - \bar B(t))(\Psi(t) - \bar\Psi(t)), \bar\chi(t)\big)_X\, dt - \int_0^T \big((B(t) - \bar B(t))\bar\Psi(t), \bar\chi(t)\big)_X\, dt,$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $\mathcal{V}$ and $\mathcal{V}^*$. Hence

$$J(\mu) - J(\bar\mu) = -\int_0^T \big(\alpha\bar\mu(t) + F(\bar\Psi(t),\bar\chi(t)), \mu(t)-\bar\mu(t)\big)_U$$
$$-\int_0^T \big((B(t)-\bar B(t))(\Psi(t)-\bar\Psi(t)),\bar\chi(t)\big)_X\, dt$$
$$-\frac{\alpha}{2}\,|\mu-\bar\mu|^2_{L(0,T;U)} + \frac{1}{2}\,(\Psi(T)-\bar\Psi(T), O(\Psi(T)-\bar\Psi(T))\,)_X.$$

Taking the limit $\mu\to\bar\mu$ and using (2.3) we obtain the claim. □

5. An algorithm and its convergence. The following algorithm for solving the optimality system in case of scalar-valued controls was proposed in [ZR] and further developed in [TKO], [MT].

ALGORITHM.

(i) Choose $\delta\in[0,2]$, $\eta\in[0,2]$, $\tilde\mu^0\in L^2(0,T;U)$, $\chi^0\in C(0,T;X)$.

For $k=1,2,\dots$ until convergence

(ii)

$$\tfrac{d}{dt}\Psi^k(t) = (A_0 - B^k(t))\Psi^k(t), \quad \Psi^k(0)=\Psi_0,$$

$$\mu^k = (1-\delta)\tilde\mu^{k-1} - \tfrac{\delta}{\alpha}F(\Psi^k,\chi^{k-1}),$$

(iii)

$$\tfrac{d}{dt}\chi^k(t) = (A_0 - \tilde B^k(t))\chi^k(t), \quad \chi^k(T) = O\Psi^k(T),$$

$$\tilde\mu^k = (1-\eta)\,\mu^k - \tfrac{\eta}{\alpha}F(\Psi^k,\chi^k).$$

First we prove the well-posedness of the algorithm.

PROPOSITION 5.1. *Let $\psi_0\in\mathcal{V}$, $\mu\in L^2(0,T;U)$, and $\chi\in C(0,T;X)$. Then there exists a unique solution $\Psi\in H^1(0,T;\mathcal{V}^*)\cap C(0,T;\mathcal{V})$ to*

$$\Psi(t) = S(t)\Psi_0 - \int_0^t S(t-s)B(\Psi)(s)\Psi(s)\,ds, \tag{5.1}$$

where $B=B(\mu)$, with $\mu(\Psi)(t) = (1-\delta)\mu(t) - \frac{\delta}{\alpha}F(\Psi(t),\chi(t))$. Analogously, if $\Psi\in C(0,T;X)$, then there exists a unique solution $\chi\in H^1(0,T;\mathcal{V}^*)\cap C(0,T;\mathcal{V})$ to

$$\chi(t) = S^*(T-t)O\Psi(T) + \int_t^T S^*(s-t)\tilde\mu(\chi)(s)\,ds,$$

where $\tilde\mu(\chi)(t) = (1-\eta)\mu(t) - \frac{\eta}{\alpha}F(\Psi(t),\chi(t))$.

Proof. We verify the first claim by a continuation argument. The second one can be proved analogously. For any Ψ and $\hat\Psi$ in $C(0,T;X)$ we have

$$|B(\Psi)(t) - B(\hat\Psi)(t)| \le M|\Psi(t)-\hat\Psi(t)|_X,$$

where $M = \tilde M\frac{\delta}{\alpha}|\chi|_{C(0,T;X)}$ and $\tilde M$ is an embedding constant. Consider the iteration

$$\Psi_n = S(t)\Psi_0 + \int_0^t S(t-s)B(\Psi_{n-1})(s)\Psi_n(s)\,ds,$$

which is initialized by the constant with value Ψ_0. It is well defined by Theorem 2.1, and $|\Psi_n(t)|_X = |\Psi_0|_X$ for all n and $t \geq 0$. For consecutive iterates we find

$$\begin{aligned}
&\tfrac{1}{2}\tfrac{d}{dt}|\Psi_{n+1}(t) - \Psi_n(t)|_X^2 \\
&= ((A_0 - B(\Psi_n))(t)(\Psi_{n+1}(t) - \Psi_n(t)) \\
&\qquad -(B(\Psi_n)(t) - B(\Psi_{n-1})(t))\Psi_n(t), \Psi_{n+1}(t) - \Psi_n(t)) \\
&\qquad \leq \tfrac{M^2}{2}|\Psi_{n+1}(t) - \Psi_n(t)|^2 + \tfrac{1}{2}\,|\Psi_n(t) - \Psi_{n-1}(t)|^2.
\end{aligned}$$

Hence for every $\tau \in (0, T]$ and $t \in (0, \tau]$

$$|\Psi_{n+1}(t) - \Psi_n(t)|_X^2 \leq \frac{1}{M^2}(e^{M^2\tau} - 1) \max_{t\in[0,\tau]} |\Psi_n(t) - \Psi_{n-1}(t)|_X.$$

Selecting $\tau > 0$ sufficiently small so that $\theta = \frac{1}{M^2}(e^{M^2\tau} - 1) < 1$ implies that

$$|\Psi_{n+1} - \Psi_n|_{C(0,\tau;X)} \leq \theta^n\,|\Psi_1 - \Psi_0|_{C(0,T;X)} \to 0$$

as $n \to \infty$. By standard arguments existence of a solution to (5.1) on $[0, \tau]$ follows. Since τ only depends on M, this solution can be extended to a solution $\Psi \in C(0, T; X)$. Uniqueness follows by Gronwall's inequality. Another application of Theorem 2.1 guarantees that $\Psi \in H^1(0, T; \mathcal{V}^*) \cap C(0, T; \mathcal{V})$. □

THEOREM 5.2. *Assume that $(\delta, \eta) \neq (0, 0)$, that $\Psi_0 \in \mathcal{V} \cap X_1$, and that (3.2), (3.3) hold. Then the sequence $\{\mu^k, \tilde{\mu}^k, \Psi^k, \chi^k\}$ contains a subsequence which converges strongly in $L^2(0, T; U) \times L^2(0, T; U) \times C(0, T; X) \times C(0, T; X)$ and every such subsequence converges to some (μ, μ, Ψ, χ), where (μ, Ψ, χ) is a solution of the optimality system.*

Proof. For $k \geq 2$ and $\delta, \eta \in [0, 2]$

$$\begin{aligned}
&J(\mu^k) - J(\mu^{k-1}) = \tfrac{1}{2}(\,\Psi^k(T) - \Psi^{k-1}(T), O\,(\Psi^k(T) - \Psi^{k-1}(T))\,)_X \\
&+\tfrac{\alpha}{2}\int_0^T (\tfrac{2}{\delta} - 1)\,|\mu^k - \tilde{\mu}^{k-1}|_U^2 + (\tfrac{2}{\eta} - 1)\,|\mu^{k-1} - \tilde{\mu}^{k-1}|_U^2\,dt \geq 0.
\end{aligned} \tag{5.2}$$

If $\delta = 0$ or $\eta = 0$, then $\mu^k = \tilde{\mu}^{k-1}$, respectively, $\mu^{k-1} = \tilde{\mu}^{k-1}$, and the corresponding terms in (5.2) are dropped. This inequality will be verified at the end of the proof, in an analogous way as in the scalar case which was treated in [ZR], [MT].

From (5.2) it follows that $J(\mu^k)$ is monotonically increasing. Since $J(\mu)$ is bounded from above this implies that $\lim_{k\to\infty} J(\mu^k)$ exists. Recall that $|\Psi^k(t)|_X = |\Psi_0|_X$ and $|\chi^k(t)|_X \leq \|O\||\Psi_0|_X$ for all k and $t \in [0, T]$. It thus follows that

$$J(\mu^0) \leq J(\mu^k) = \frac{1}{2}(\Psi^k(T), O\Psi^k(T))_X - \frac{\alpha}{2}|\mu^k|^2_{L^2(0,T;U)},$$

and hence

$$\frac{\alpha}{2}|\mu^k|^2_{L^2(0,T;U)} \leq \frac{1}{2}|\Psi_0|_X^2\|O\|_{\mathcal{L}(X)} - J(\mu^0).$$

Moreover,

$$|\tilde{\mu}^k|_{L^2(0,T;U)} \leq |1 - \eta|\,|\mu^k|_{L^2(0,T;U)} + \frac{\eta}{2}|\Psi_0|_X^2\|O\|_{\mathcal{L}(X)},$$

and hence

$$\{\mu^k\}_{k=1}^\infty \text{ and } \{\tilde{\mu}^k\}_{k=1}^\infty \text{ are bounded in } L^2(0, T; U). \tag{5.3}$$

From Theorem 2.1 and assumptions (3.2) and (3.3), therefore,

$$\{\Psi^k\}_{k=1}^{\infty} \text{ and } \{\chi^k\}_{k=1}^{\infty} \text{ are bounded in } H^1(0,T;\mathcal{V}^*) \cap L^2(0,T;\mathcal{V} \cap X_1).$$

By Aubin's lemma there exists a subsequence $\{k^n\}$ of $\{k\}$ and $\Psi \in C(0,T;X)$, $\chi \in C(0,T;X)$ such that

$$\Psi^{k_n} \to \Psi \text{ and } \chi^{k_n} \to \chi \text{ strongly in } L^2(0,T;X).$$

Using the boundedness of $\{\Psi^{k_n}\}$ and $\{\chi^{k_n}\}$ in $C(0,T;X)$ and the properties of F one argues that $F(\Psi^{k_n},\chi^{k_n}) \to F(\Psi,\chi)$ strongly in $L^2(0,T;U)$. From (iii) of the algorithm we have

$$\eta\mu^{k_n} = \mu^{k_n} - \tilde{\mu}^{k_n} - \frac{\eta}{\alpha}F(\Psi^{k_n},\chi^{k_n}).$$

Since $\mu^{k_n} - \tilde{\mu}^{k_n} \to 0$ in $L^2(0,T;U)$ by (5.2), it follows, for $\eta \neq 0$, that μ^{k_n} converges strongly in $L^2(0,T;U)$ to some μ, as $k_n \to \infty$. For each k_n we have the following by (ii) of the algorithm:

$$\Psi^{k_n}(t) = S(t)\Psi_0 - \int_0^t S(t-s)B^{k_n}(s)\Psi^{k_n}(s)\,ds. \tag{5.4}$$

Let $\Psi \in C(0,T;X)$ denote the solution to

$$\Psi(t) = S(t)\Psi_0 - \int_0^t S(t-s)B(s)\Psi(s)\,ds. \tag{5.5}$$

From Gronwall's inequality it follows that $\Psi^{k_n} \to \Psi$ in $C(0,T;X)$. By (5.2) the sequence $\{\tilde{\mu}^{k_n}\}$ converges strongly in $L^2(0,T;U)$ to μ. Step (iii) of the algorithm implies that

$$\chi^{k_n}(t) = S^*(T-t)O\Psi^{k_n}(T) + \int_t^T S^*(s-t)\tilde{B}^{k_n}(s)\chi^{k_n}(s)\,ds. \tag{5.6}$$

Let χ in $C(0,T;X)$ denote the solution to

$$\chi(t) = S^*(T-t)O\Psi(T) + \int_t^T S^*(s-t)B(s)\chi(s)\,ds. \tag{5.7}$$

Again by Gronwall's lemma we find that $\chi^{k_n} \to \chi$ in $C(0,T;X)$. Passing to the limit in the second equation of (iii) implies that

$$\alpha B + F(\Psi,\chi) = 0. \tag{5.8}$$

For $\eta = 0$ there exists a subsequence $\{k^n\}$ of $\{k\}$ and $\bar{\Psi} \in C(0,T;X)$, $\bar{\chi} \in C(0,T;X)$ such that

$$\Psi^{k_n} \to \bar{\Psi} \text{ and } \chi^{k_n} \to \bar{\chi} \text{ strongly in } L^2(0,T;X).$$

By (5.2) and since $\tilde{\mu}^k = \mu^k$ for $\eta = 0$ we have $\lim_{n\to\infty} \mu^{k_n-1} - \mu^{k_n} = 0$ in $L^2(0,T;U)$. From

$$\mu^{k_n} = (1-\delta)\mu^{k_n-1} - \frac{\delta}{\alpha}F(\Psi^{k_n},\chi^{k_n-1})$$

it therefore follows that μ^{k_n-1} converges strongly to some μ in $L^2(0,T;U)$. By (5.2) also $\lim_{n\to\infty}\mu^{k_n} = \mu$ in $L^2(0,T;U)$. As before, the solutions to (5.4) and (5.6) converge strongly in $C(0,T;X)$ to the solutions of (5.5) and (5.7), and (5.8) also holds for $\eta = 0$. From (5.5), (5.7), and (5.8) we conclude that (μ,Ψ,χ) is a solution to the optimality system.

We now provide the proof of (5.2) for the case $\eta \neq 0, \delta \neq 0$. The remaining cases follow easily. We have

$$J(\mu^{k+1}) - J(\mu^k) = \frac{1}{2}(\,\Psi^{k+1}(T) - \Psi^k(T), O\,(\Psi^{k+1}(T) - \Psi^k(T))\,)_X$$
$$+(\Psi^{k+1}(T) - \Psi^k(T), O\,\Psi^k(T))_X + \frac{\alpha}{2}\int_0^T |\mu^{k+1}|^2 - \frac{\alpha}{2}\int_0^T |\mu^k|^2.$$

Suppressing the dependence of Ψ^k and μ^k on t we find

$$(\Psi^{k+1}(T) - \Psi^k(T), O\,\Psi^k(T))_X = (\Psi^{k+1}(T) - \Psi^k(T), \chi^k(T))_X$$
$$= \int_0^T \left(\frac{\partial}{\partial t}(\Psi^{k+1} - \Psi^k), \chi^k\right)_X + \left(\Psi^{k+1} - \Psi^k, \frac{\partial}{\partial t}\chi^k\right)_X$$
$$= \int_0^T ((A_0 - B^{k+1})\Psi^{k+1} - (A_0 - B^k)\Psi^k, \chi^k)_X + (\Psi^{k+1} - \Psi^k, (A_0 - \tilde{B}^k)\chi^k)_X$$
$$= \int_0^T ((\tilde{B}^k - B^{k+1})\Psi^{k+1}, \chi^k)_X + ((B^k - \tilde{B}^k)\Psi^k, \chi^k)_X$$
$$= \int_0^T (F(\Psi^{k+1}, \chi^k), \tilde{\mu}^k - \mu^{k+1})_U + (F(\Psi^k, \chi^k), \mu^k - \tilde{\mu}^k)_U$$
$$= \alpha\int_0^T \frac{1}{\delta}(\tilde{\mu}^k - \mu^{k+1}, (1-\delta)\tilde{\mu}^k - \mu^{k+1})_U + \frac{1}{\eta}(\tilde{\mu}^k - \mu^k, (1-\eta)\mu^k - \tilde{\mu}^k)_U$$
$$= \alpha\int_0^T \frac{1}{\delta}|\tilde{\mu}^k - \mu^{k+1}|_U^2 + \frac{1}{\eta}|\tilde{\mu}^k - \mu^k|_U^2 - (\tilde{\mu}^k - \mu^{k+1}, \tilde{\mu}^k)_U - (\mu^k - \tilde{\mu}^k, \mu^k)_U.$$

Hence we find

$$J(\mu^{k+1}) - J(\mu^k) = \frac{1}{2}(\,\Psi^{k+1}(T) - \Psi^k(T), O\,(\Psi^{k+1}(T) - \Psi^k(T))\,)_X$$
$$+(\,\Psi^{k+1}(T) - \Psi^k(T), O\,\Psi^k(T)\,)_X - \frac{\alpha}{2}\int_0^T |\mu^{k+1}|_U^2 + \frac{\alpha}{2}\int_0^T |\mu^k|^2$$
$$= \frac{1}{2}(\,\Psi^{k+1}(T) - \Psi^k(T), O\,(\Psi^{k+1}(T) - \Psi^k(T))\,)_X$$
$$+\frac{\alpha}{2}\int_0^T \left(\left(\frac{2}{\delta} - 1\right)|\tilde{\mu}^k - \mu^{k+1}|_U^2 + \left(\frac{2}{\eta} - 1\right)|\mu^k - \tilde{\mu}^k|_U^2\right) dt \geq 0. \qquad \square$$

In [S] it is argued that the set of limit points is in fact compact. Moreover, if the penalty parameter α is sufficiently large, then the limit set consists of a singleton.

In the previous theorem subsequential convergence followed under the assumption of compactness of the orbits implied by (3.2), (3.3). Alternatively a compactness assumption for U as a subset of $\mathcal{L}(X)$ also implies convergence.

THEOREM 5.3. *Assume that $(\delta,\eta) \neq (0,0)$, that $\Psi_0 \in \mathcal{V}$, $\chi^0 \in H^1(0,T;X)$, $\tilde{\mu} \in H^1(0,T;U)$, and that U is a compact subset of $\mathcal{L}(X)$. Then the conclusion of Theorem* 5.2 *holds.*

Proof. As in the proof of Theorem 5.2 $\{\mu^k\}_{k=1}^\infty$ and $\{\tilde\mu^k\}_{k=1}^\infty$ are bounded in $L^2(0,T;U)$, and by Theorem 3.1

$$\{\Psi^k\}_{k=1}^\infty \text{ and } \{\chi^k\}_{k=1}^\infty \text{ are bounded in } H^1(0,T;\mathcal{V}^*)\cap C(0,T;\mathcal{V}).$$

This implies that $\{F(\Psi^k,\chi^{k-1})\}_{k=1}^\infty$ and $\{F(\Psi^k,\chi^k)\}_{k=1}^\infty$ are bounded in $H^1(0,T;U)$. If $\delta=\eta=1$, then $\{\mu^k\}$ and $\{\tilde\mu^k\}$ are bounded in $H^1(0,T;U)$. Otherwise

$$\tilde\mu^k=(1-\eta)(1-\delta)\tilde\mu^{k-1}+(1-\eta)\frac{\delta}{\alpha}F(\Psi^k,\chi^{k-1})-\frac{\eta}{\alpha}F(\Psi^k,\chi^k),$$

with $|(1-\eta)(1-\delta)|<1$. It follows that μ^k and $\tilde\mu^k$ are bounded in $H^1(0,T;U)$. Hence there exists a subsequence $\{k_n\}$ of $\{k\}$ and $\mu\in H^1(0,T;U), \tilde\mu\in H^1(0,T;U)$ such that

$$\mu^{k_n}\to\mu, \quad \text{and} \quad \tilde\mu^{k_n}\to\tilde\mu \text{ strongly in } L^2(0,T;\mathcal{L}(X)).$$

By (5.2) we have $|\mu_k-\tilde\mu_k|_{L^2(0,T;\mathcal{L}(X))}\to 0$ if $\eta\neq 0$, whereas $\mu_k=\tilde\mu_k$ if $\eta=0$. In either case it follows that $\mu=\tilde\mu$. The proof can now be completed as the one for Theorem 5.2. ∎

Acknowledgments. We would like to thank Prof. A. Borzi for interesting discussions on various aspects of this paper and Dr. J. Salomon for providing us with a preprint of [S].

REFERENCES

[BMS] J. M. Ball, J. M. Marsden, and M. Slemrod, *Controllability for distributed bilinear systems*, SIAM J. Control Optim., 20 (1982), pp. 575–597.

[BP] L. Baudouin, O. Kavian, and J.-P. Puel, *Regularity for a Schrödinger equation with singular potentials and application to bilinear optimal control*, J. Differential Equations, 216 (2005), pp. 188–222.

[CF] P. Constantin and C. Foias, *Navier-Stokes Equations*, University of Chicago Press, Chicago, 1988.

[HP] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups*, AMS, Providence, RI, 1957.

[K] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.

[MST] Y. Maday, J. Salomon, and G. Turinici, *Monotonic time-discretized schemes*, Numer. Math., 103 (2006), pp. 323–338.

[MT] Y. Maday and G. Turinici, *New formulations of monotonically convergent quantum control algorithms*, J. Chem. Phys., 118 (2003), pp. 8191–8196.

[P] A. Pazy, *Semi-Groups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Heidelberg, 1983.

[PDR] A. P. Peirce, M. A. Dahleh, and H. Rabitz, *Optimal control of quantum-mechanical systems: Existence, numerical approximation, and applications*, Phys. Rev. A, 37 (1988), pp. 4950–4967.

[S] J. Salomon, *Limit points of the monotonic schemes in quantum control*, in Proceedings of the 44th Annual IEEE Conference on Decision and Control, Seville, Spain, 2005, CD-ROM. Available online at http://www.ceremade.dauphine.fr/~salomon/ar/proceeding.pdf.

[TKO] D. Tannor, V. Kazakov, and V. Orlov, *Control of photochemical branching: Novel procedures for finding optimal pulses and global upper bounds*, in Time Dependent Quantum Molecular Dynamics, J. Broeckhove and L. Lathouwers, eds., Plenum Press, New York, 1992, pp. 347–360.

[ZR] W. Zhu and H. Rabitz, *A rapid monotonically convergent iteration algorithm for quantum optimal control over the expectation value of a positive definite operator*, J. Chem. Phys., 109 (1998), pp. 385–391.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 288–307

STATE AGREEMENT FOR CONTINUOUS-TIME COUPLED NONLINEAR SYSTEMS*

ZHIYUN LIN†, BRUCE FRANCIS†, AND MANFREDI MAGGIORE†

Abstract. Two related problems are treated in continuous time. First, the state agreement problem is studied for coupled nonlinear differential equations. The vector fields can switch within a finite family. Associated to each vector field is a directed graph based in a natural way on the interaction structure of the subsystems. Generalizing the work of Moreau, under the assumption that the vector fields satisfy a certain subtangentiality condition, it is proved that asymptotic state agreement is achieved if and only if the dynamic interaction digraph has the property of being sufficiently connected over time. The proof uses nonsmooth analysis. Second, the rendezvous problem for kinematic point-mass mobile robots is studied when the robots' fields of view have a fixed radius. The circumcenter control law of Ando et al. [*IEEE Trans. Robotics Automation*, 15 (1999), pp. 818–828] is shown to solve the problem. The rendezvous problem is a kind of state agreement problem, but the interaction structure is state dependent.

Key words. state agreement, rendezvous, interacting nonlinear systems, time-varying interaction, asymptotical stability

AMS subject classifications. 93C10, 93D20, 37N35, 05C20

DOI. 10.1137/050626405

1. Introduction.

This paper studies a dynamical system that is the interconnection of subsystems. Examples are abundant in biology, physics, engineering, ecology, and social science: e.g., a biochemical reaction network [14], coupled Kuramoto oscillators [17, 39], arrays of chaotic systems [44, 45], a swarm of organisms [12, 13], and a group of autonomous agents [16, 22, 23]. We model such systems by coupled nonlinear differential equations in state form. Pioneering work on such coupled dynamical systems from a structural point of view is that of Siljak, e.g., [35, 36].

State agreement means that the states of the subsystems are all equal. For example, [11] studies a group of individuals who must act together as a team; each individual has its own subjective probability distribution for the unknown value of some parameter. How the group might reach a consensus and form a common subjective probability distribution for the parameter is a state agreement problem. In other contexts, state agreement arises as *synchronization* in theoretical physics, e.g., [5, 30, 39, 42, 43], and *consensus* in computer science, particularly in distributed computing, e.g., [25].

Central to the state agreement problem is the graph describing the interaction structure in the interconnected system—that is, who is coupled to whom. And a central question is, What properties of the interaction graph lead to state agreement? Most existing work has dealt with static graphs with a particular topology, such as rings [6, 30], cyclic digraphs [32], and fully connected graphs [12, 13, 34], or with static graphs having an unspecified topology but a certain connectedness. Example frameworks are coupled cell systems [38], coupled oscillators [17, 45], multiagent systems [4, 31], and formations of unicycles [23]. Of course, a static graph simplifies the

*Received by the editors March 9, 2005; accepted for publication (in revised form) November 4, 2006; published electronically April 13, 2007.
http://www.siam.org/journals/sicon/46-1/62640.html

†Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto M5S 3G4, ON, Canada (linz@control.utoronto.ca, francis@control.utoronto.ca, maggiore@control.utoronto.ca).

state agreement problem and allows one to focus on the difficulties caused by the nonlinear dynamics of the nodes.

The more interesting situation is when the interaction graph is time varying. From the point of view of control theory, the most suitable mathematical model for these setups is a switched interconnected system. However, attempts to understand how the switching affects the collective system behavior had been hampered by the lack of suitable analysis tools. Recently, however, great strides have been made [16] by characterizing the convergence of infinite products of certain types of nonnegative matrices in a linear discrete-time setup with an undirected interaction graph. For the switched linear continuous-time system model and a directed graph, [22] uses the graph Laplacian and the properties of some special matrices to prove asymptotic state agreement under certain graphical conditions. In addition, [29] uses the common Lyapunov function technique for the switched linear continuous-time system and shows that balanced digraphs play a key role in addressing the average-consensus problem. Two other works on state agreement for linear continuous-time systems are [15], which deals with random networks, and [26], which addresses the deterministic time-varying case.

However, many real systems are nonlinear in addition to having time-varying interaction among subsystems. Examples are systems of coupled oscillators. For nonlinear interconnected systems with time-varying interaction, new tools are required. A novel approach is taken by Moreau in [27]: The framework is nonlinear and discrete-time, and the stability analysis is based upon a blend of graph-theoretic and system-theoretic tools, with the notion of convexity being key. The idea is, roughly speaking, that if every agent always moves toward the relative interior of the convex hull of the set of neighbor agents at each step, state agreement will be achieved. The result in [27] was recently generalized in [2] as follows: The setup is still a discrete-time system, but each agent moves towards the relative interior of a set which is a function, not necessarily the convex hull, of the present and past states of neighbor agents. In this way communication delays can be accommodated.

One concrete instance of the state agreement problem is the rendezvous problem for autonomous mobile robots. Suppose the robots' fields of view have a fixed radius. Then the robots may come into and go out of sensor range of each other, and the interaction graph is therefore state dependent instead of time dependent. For this problem, some distributed algorithms were proposed in [1, 41], with the objective of getting the robots to congregate at a common location (achieving *rendezvous*). These algorithms were extended to various synchronous and asynchronous stop-and-go strategies in [9, 19, 20].

This paper makes two main contributions. The first is the continuous-time counterpart to the result of Moreau [27]. We borrow heavily from Moreau's geometric concepts and proof structure; we suggest, however, that the continuous-time case presents some considerable challenges, as one will see from the details of our proof. Thus our contribution to this problem is primarily technical in nature. As an example application, we apply our result to make new conclusions about synchronization of coupled Kuramoto oscillators. The second contribution of this paper is a solution of the continuous-time rendezvous problem for kinematic point-mass robots; we use the circumcenter control law of Ando et al. [1] and give the first proof of convergence in continuous time.

2. Preliminaries. Here we assemble some known and some novel concepts related to convex sets and tangent cones, directed graphs, and Dini derivatives. In

addition, we provide some fundamental properties associated with them.

2.1. Convex sets and tangent cones. References for this subsection are [3,37].

The convex hull of $\mathcal{S} \subset \mathbb{R}^m$ is denoted $\mathrm{co}(\mathcal{S})$. The convex hull of a finite set of points $x_1, \ldots, x_n \in \mathbb{R}^m$ is a *polytope*, denoted $\mathrm{co}\{x_1, \ldots, x_n\}$.

Let $\mathcal{S} \subset \mathbb{R}^m$ be convex. If $\mathcal{S}$ contains the origin, the smallest subspace containing $\mathcal{S}$ is the *carrier subspace*, denoted $\mathrm{lin}(\mathcal{S})$. The *relative interior* of $\mathcal{S}$, denoted $\mathrm{ri}(\mathcal{S})$, is the interior of $\mathcal{S}$ when it is regarded as a subset of $\mathrm{lin}(\mathcal{S})$ and the relative topology is used, and likewise for the *relative boundary*, denoted $\mathrm{rb}(\mathcal{S})$. If $\mathcal{S}$ does not contain the origin, it must be translated by an arbitrary vector: Let v be any point in $\mathcal{S}$ and let $\mathrm{lin}(\mathcal{S})$ denote the smallest subspace containing $\mathcal{S} - v$. Then $\mathrm{ri}(\mathcal{S})$ is the interior of $\mathcal{S}$ when it is regarded as a subset of the affine subspace $v + \mathrm{lin}(\mathcal{S})$, and similarly for $\mathrm{rb}(\mathcal{S})$.

A nonempty set $\mathcal{K} \subset \mathbb{R}^m$ is a *cone* if $\lambda y \in \mathcal{K}$ when $y \in \mathcal{K}$ and $\lambda > 0$. Let $\mathcal{S} \subset \mathbb{R}^m$ be a closed convex set and $y \in \mathcal{S}$. The *tangent cone* (often referred to as *contingent cone*) to $\mathcal{S}$ at y is the set

$$\mathcal{T}(y, \mathcal{S}) = \left\{ z \in \mathbb{R}^m : \liminf_{\lambda \to 0} \frac{\|y + \lambda z\|_{\mathcal{S}}}{\lambda} = 0 \right\},$$

where $\|y + \lambda z\|_{\mathcal{S}}$ denotes the distance from $y + \lambda z$ to $\mathcal{S}$. The *normal cone* to $\mathcal{S}$ at y is

$$\mathcal{N}(y, \mathcal{S}) = \{ z^* : \langle z, z^* \rangle \leq 0 \ \forall z \in \mathcal{T}(y, \mathcal{S}) \}.$$

Note that if y is in the interior of $\mathcal{S}$, then $\mathcal{T}(y, \mathcal{S}) = \mathbb{R}^m$. Thus the set $\mathcal{T}(y, \mathcal{S})$ is nontrivial only on $\partial\mathcal{S}$, the boundary of $\mathcal{S}$. In particular, if $\mathcal{S}$ contains only one point, y, then $\mathcal{T}(y, \mathcal{S}) = \{0\}$. In geometric terms the tangent cone for $y \in \partial\mathcal{S}$ is a cone centered at the origin which contains all vectors whose directions point from y "inside" (or they are "tangent to") the set $\mathcal{S}$.

LEMMA 2.1 (see [3]). *Let $\mathcal{S}_i, i = 1, \ldots, n$ be convex sets in $\mathbb{R}^m$.*

(i) *If $y \in \mathcal{S}_1 \subset \mathcal{S}_2$, then*

$$\mathcal{T}(y, \mathcal{S}_1) \subset \mathcal{T}(y, \mathcal{S}_2) \quad \textit{and} \quad \mathcal{N}(y, \mathcal{S}_2) \subset \mathcal{N}(y, \mathcal{S}_1).$$

(ii) *If $x_i \in \mathcal{S}_i$ $(i = 1, \ldots, n)$, then*

$$\begin{aligned} \mathcal{T}\left((x_1, \ldots, x_n), \mathcal{S}_1 \times \cdots \times \mathcal{S}_n\right) &= \mathcal{T}(x_1, \mathcal{S}_1) \times \cdots \times \mathcal{T}(x_n, \mathcal{S}_n), \\ \mathcal{N}\left((x_1, \ldots, x_n), \mathcal{S}_1 \times \cdots \times \mathcal{S}_n\right) &= \mathcal{N}(x_1, \mathcal{S}_1) \times \cdots \times \mathcal{N}(x_n, \mathcal{S}_n). \end{aligned}$$

2.2. Directed graphs. For a directed graph (digraph for short) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, n\}$ is the set of nodes and $\mathcal{E}$ is the set of arcs, we write $i \to j$ if there is a path from node i to node j. By definition, $i \to i$ for every node i. A *center* is a node i such that $i \to j$ for every node j, and $\mathcal{G}$ is *quasi-strongly connected* (QSC) if it has a center [7]. Finally, $\mathcal{G}$ is *fully connected* if for every two nodes i and j there is an arc from i to j.

2.3. Dini derivatives. Consider the nonautonomous system

$$\dot{y} = f(t, y), \tag{2.1}$$

where $\mathcal{D} \subset \mathbb{R}^m$ is a domain and $f : \mathbb{R} \times \mathcal{D} \to \mathbb{R}^m$. Let $V(t, y) : \mathbb{R} \times \mathcal{D} \to \mathbb{R}$ be a continuous function satisfying a local Lipschitz condition for y, uniformly with respect to t. Then we define

$$D_f^+ V(t, y) = \limsup_{\tau \to 0^+} \frac{V(t + \tau, y + \tau f(t, y)) - V(t, y)}{\tau}.$$

The function D_f^+V is called the *upper Dini derivative* of V along the trajectory of (2.1). Suppose that for an initial condition $y(0) = y^0$, (2.1) has a solution $y(t)$ defined on an interval $[0, \epsilon)$ and let $D^+V(t, y(t))$ be the upper Dini derivative of $V(t, y(t))$ with respect to t, i.e.,

$$D^+V(t, y(t)) = \limsup_{\tau \to 0^+} \frac{V(t+\tau, y(t+\tau)) - V(t, y(t))}{\tau}.$$

Let $t^* \in [0, \epsilon)$ and put $y(t^*) = y^*$. Then one has that (see [33])

$$D^+V(t^*, y(t^*)) = D_f^+V(t^*, y^*).$$

LEMMA 2.2. *Let $\mathcal{I}_0 = \{1, 2, \ldots, n\}$ and suppose for each $i \in \mathcal{I}_0$, $V_i : \mathbb{R} \times \mathcal{D} \to \mathbb{R}$ is of class C^1; let $V(t, y) = \max_{i \in \mathcal{I}_0} V_i(t, y)$; and let $\mathcal{I}(t) = \{i \in \mathcal{I}_0 : V_i(t, y(t)) = V(t, y(t))\}$ be the set of indices where the maximum is reached at time t. Then $D^+V(t, y(t))$ satisfies*

$$D^+V(t, y(t)) = \max_{i \in \mathcal{I}(t)} \dot{V}_i(t, y(t)).$$

The proof can be obtained from Danskin's theorem [8, 10].

3. The state agreement problem: Main results. Our setup is a general interconnection of nonlinear subsystems, where the vector fields can switch within a finite family. We associate to each vector field a directed graph based in a natural way on the interaction structure of the subsystems; this is called an *interaction digraph* in the present paper. Assuming that the vector fields satisfy a certain subtangentiality condition, we show that asymptotic state agreement is achieved if and only if the dynamic interaction digraph has the property of being sufficiently connected over time, in a certain technical sense. Most of the proofs are deferred to section 5.

To formalize the notion of a switched interconnected system, first consider a family of systems

$$\begin{aligned} \dot{x}_1 &= f_p^1(x_1, \ldots, x_n) \\ &\vdots \\ \dot{x}_n &= f_p^n(x_1, \ldots, x_n), \end{aligned}$$

where $x_i \in \mathbb{R}^m$ is the state of subsystem i and where the index p belongs to a finite set $\mathcal{P}$. Notice that the subsystems share a common state space, $\mathbb{R}^m$. Introducing the *aggregate state* $x \in \mathbb{R}^{mn}$, we have the concise form

$$\dot{x} = f_p(x), \quad p \in \mathcal{P}, \tag{3.1}$$

where for each $p \in \mathcal{P}$, $f_p : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$.

We now associate to each vector field f_p an interaction digraph $\mathcal{G}_p$ capturing the interaction structure of the n subsystems (agents).

DEFINITION 3.1 (interaction digraph). *The interaction digraph $\mathcal{G}_p = (\mathcal{V}, \mathcal{E}_p)$ consists of*

- *a finite set $\mathcal{V}$ of n nodes, each node i modeling agent i;*
- *an arc set $\mathcal{E}_p$ representing the links between agents. An arc from node j to node i indicates that agent j is a neighbor of agent i in the sense that f_p^i depends on x_j; i.e., there exist $x_j^1, x_j^2 \in \mathbb{R}^m$ such that*

$$f_p^i(x_1, \ldots, x_j^1, \ldots, x_n) \neq f_p^i(x_1, \ldots, x_j^2, \ldots, x_n).$$

The set of neighbors of agent i is denoted $\mathcal{N}_i(p)$.

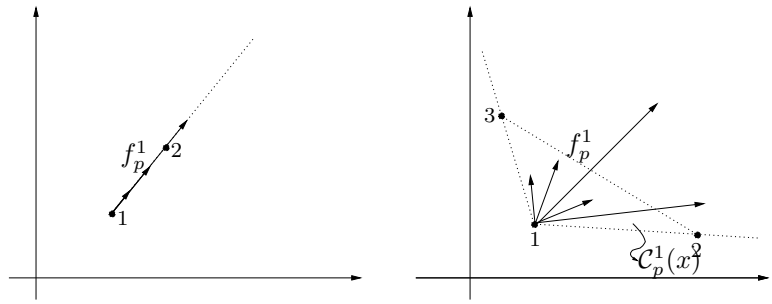


FIG. 3.1. *Some examples of vector fields f_p^i satisfying assumption* A2.

Let $\mathcal{C}_p^i(x) = \text{co}\{x_i, x_j : j \in \mathcal{N}_i(p)\}$ denote the polytope in $\mathbb{R}^m$ formed by the states of agent i and its neighbors. Also, it is convenient to introduce a subset $\mathcal{S} \subset \mathbb{R}^m$ of the common state space that plays the role of a region of focus. In our state agreement problem, initial states of the agents will be in $\mathcal{S}$ and agreement will occur in $\mathcal{S}$. Let $\mathcal{I}_0$ denote the index set $\{1, \ldots, n\}$ and assume that, for each $i \in \mathcal{I}_0$ and each $p \in \mathcal{P}$, the vector fields $f_p^i : \mathbb{R}^{mn} \to \mathbb{R}^m$ satisfy the following two assumptions:

A1. f_p^i is locally Lipschitz on $\mathcal{S}^n$.

A2. For all $x \in \mathcal{S}^n$, $f_p^i(x) \in \text{ri}\left(\mathcal{T}(x_i, \mathcal{C}_p^i(x))\right)$.

Assumption A2 is sometimes referred to as a *strict subtangentiality condition*. Figure 3.1 illustrates two example situations of A2. In the left-hand example, agent 1 has only one neighbor, agent 2; the convex hull $\mathcal{C}_p^1(x)$ is the line segment joining x_1 and x_2; the tangent cone $\mathcal{T}(x_1, \mathcal{C}_p^1(x))$ is the closed ray $\{\lambda(x_2 - x_1) : \lambda \geq 0\}$ (in the figure it is shown translated to x_1); the relative interior $\text{ri}\left(\mathcal{T}(x_1, \mathcal{C}_p^1(x))\right)$ is the open ray $\{\lambda(x_2 - x_1) : \lambda > 0\}$; and A2 means that f_p^1 is nonzero and points in the direction of $x_2 - x_1$. In the right-hand example, agent 1 has two neighbors, agents 2 and 3; the convex hull $\mathcal{C}_p^1(x)$ is the triangle with vertices x_1, x_2, x_3; the tangent cone $\mathcal{T}(x_1, \mathcal{C}_p^1(x))$ is

$$\{\lambda_1(x_2 - x_1) + \lambda_2(x_3 - x_1) : \lambda_1, \lambda_2 \geq 0\}$$

(again, it is shown translated to x_1); the relative interior $\text{ri}\left(\mathcal{T}(x_1, \mathcal{C}_p^1(x))\right)$ is

$$\{\lambda_1(x_2 - x_1) + \lambda_2(x_3 - x_1) : \lambda_1, \lambda_2 > 0\};$$

and A2 means that f_p^1 points into this open cone. In general, A2 requires that $f_p^i(x)$ have the form

$$\sum_{j \in \mathcal{N}_i(p)} \alpha_j(x)(x_j - x_i),$$

where $\alpha_j(x)$ are nonnegative scalar functions, and that $f_p^i(x)$, now viewed as a vector applied at the vertex x_i, not be tangent to the relative boundary of the convex set $\mathcal{C}_p^i(x)$.

When the index p in (3.1) is replaced by a piecewise constant function $\sigma : [0, \infty) \to \mathcal{P}$, we obtain a *switched interconnected system*

$$\dot{x}(t) = f_{\sigma(t)}\left(x(t)\right). \tag{3.2}$$

The function σ is called a *switching signal*. The case of infinitely fast switching (chattering), which would call for a concept of generalized solution, is not considered here. As a matter of fact, it can be shown that even piecewise constant switching signals $\sigma(t)$ do not have sufficient regularity for asymptotic agreement of the switched interconnected system (3.2) [21]. Let $\mathcal{S}_{dwell}$ denote the class of piecewise constant switching signals such that any consecutive discontinuities are separated by no less than some fixed positive constant τ_D, the *dwell time*. We make the following assumption:

A3. $\sigma(t) \in \mathcal{S}_{dwell}$.

Having replaced p by a switching signal $\sigma(t)$, we similarly replace the interaction digraph $\mathcal{G}_p$ by a dynamic interaction digraph $\mathcal{G}_{\sigma(t)}$.

DEFINITION 3.2 (dynamic interaction digraph and union digraph). *Given a switching signal $\sigma(t)$, the* dynamic interaction digraph $\mathcal{G}_{\sigma(t)}$ *is the pair* $\left(\mathcal{V}, \mathcal{E}_{\sigma(t)}\right)$. *Given two real numbers $t_1 \leq t_2$,* the union digraph $\mathcal{G}\left([t_1, t_2]\right)$ *is the digraph whose arcs are obtained from the union of the arcs in $\mathcal{G}_{\sigma(t)}$ over the time interval $[t_1, t_2]$.*

DEFINITION 3.3. *A dynamic interaction digraph $\mathcal{G}_{\sigma(t)}$ is* uniformly quasi-strongly connected (UQSC) *if there exists $T > 0$ such that for all $t \geq 0$, the union digraph $\mathcal{G}([t, t+T])$ is QSC.*

Our main result, Theorem 3.8, is that the switched interconnected system achieves asymptotic state agreement on $\mathcal{S}$ if and only if the dynamic interaction digraph $\mathcal{G}_{\sigma(t)}$ is UQSC.

But first, the precise meaning of state agreement is given in the following definition.

DEFINITION 3.4. *The switched interconnected system* (3.2) *has the property of*

(i) state agreement on $\mathcal{S}$ *if* $\forall \zeta \in \mathcal{S},\ \forall \varepsilon > 0,\ \exists \delta > 0$ *such that* $\forall t_0 \geq 0$,

$$(\forall i)\ \ (\|x_i(t_0) - \zeta\| \leq \delta) \wedge (x_i(t_0) \in \mathcal{S}) \Longrightarrow (\forall t \geq t_0)(\forall i)\ \|x_i(t) - \zeta\| \leq \varepsilon;$$

(ii) asymptotic state agreement on $\mathcal{S}$ *if it has the property of state agreement on $\mathcal{S}$ and in addition* $\forall \varepsilon > 0,\ \forall c > 0,\ \exists T > 0$ *such that* $\forall t_0 \geq 0$,

$$(\forall i)\ (\|x_i(t_0)\| \leq c) \wedge (x_i(t_0) \in \mathcal{S}) \Longrightarrow (\exists \zeta \in \mathcal{S})(\forall t \geq t_0 + T)(\forall i)\ \|x_i(t) - \zeta\| \leq \varepsilon;$$

(iii) global asymptotic state agreement *if it has the property of asymptotic state agreement on $\mathbb{R}^m$.*

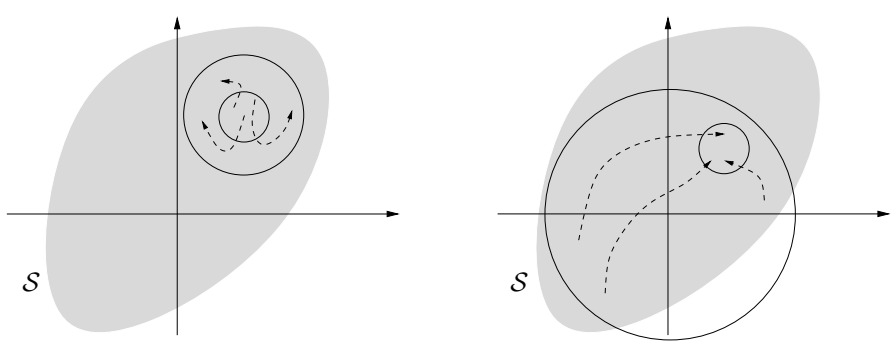


FIG. 3.2. *Asymptotic state agreement on $\mathcal{S}$.*

These definitions are illustrated in Figure 3.2 and can be stated, roughly speaking, as follows. State agreement (the left-hand figure) means that, for every point ζ in $\mathcal{S}$, the agents stay arbitrarily close to ζ if they start sufficiently close to ζ, uniformly with

respect to the starting time. Asymptotic state agreement (the two figures together) means, in addition, that the agents converge to a common location in $\mathcal{S}$.

These state agreement definitions are related to stability with respect to a set. Let Ω denote the set of aggregate states such that the subsystem states are all equal and in $\mathcal{S}$, i.e.,

$$\Omega = \{x \in \mathbb{R}^{nm} : x_1 = \cdots = x_n \in \mathcal{S}\}.$$

Then state agreement is equivalent to uniform stability with respect to Ω.

Finally, we give the following new definition of positive invariance specially for interconnected systems.

DEFINITION 3.5. *A set $\mathcal{A} \subset \mathbb{R}^m$ is said to be* positively invariant *for the switched interconnected system* (3.2) *if*

$$(\forall t_0 \geq 0)(\forall i)\ x_i(t_0) \in \mathcal{A} \implies (\forall t \geq t_0)(\forall i)\ x_i(t) \in \mathcal{A}.$$

Our first result establishes the positive invariance property of any compact convex set in $\mathcal{S}$ without needing any property of the interaction digraph. This result can perhaps be understood intuitively as follows. For $m = 2$, all agents move in the plane. Let $\mathcal{A}$ be a compact convex set in $\mathcal{S}$ and assume all agents start in $\mathcal{A}$. Let $\mathcal{C}(t)$ denote the convex hull of the agents' locations at time t. Because $\mathcal{A}$ is convex, clearly $\mathcal{C}(0) \subset \mathcal{A}$. Now invoke assumption A2. An agent that is initially in the interior of $\mathcal{C}(0)$ can head off in any direction at $t = 0$, but an agent that is initially on the boundary of $\mathcal{C}(0)$ is constrained to head into its interior. In this way, $\mathcal{C}(t)$ is nonincreasing (if $t_2 > t_1$, then $\mathcal{C}(t_2) \subset \mathcal{C}(t_1)$), and $\mathcal{A}$ is therefore positively invariant for the switched interconnected system (3.2).

THEOREM 3.6. *Let $\mathcal{A} \subset \mathcal{S}$ be a compact convex set. Then $\mathcal{A}$ is positively invariant for the switched interconnected system* (3.2).

The second result establishes state agreement of the system, again without needing any property of the interaction digraph.

THEOREM 3.7. *Suppose $\mathcal{S}$ is closed and convex. The switched interconnected system* (3.2) *has the property of state agreement on $\mathcal{S}$.*

Proof. Let $\zeta \in \mathcal{S}$ and $\varepsilon > 0$ be arbitrary, and let

$$\mathcal{A}_\varepsilon(\zeta) = \{y \in \mathcal{S} : \|y - \zeta\| \leq \varepsilon\}. \tag{3.3}$$

By Theorem 3.6, it follows that $\mathcal{A}_\varepsilon(\zeta)$ is positively invariant since it is a compact convex set in $\mathcal{S}$. We have thus proved that $\forall \zeta \in \mathcal{S},\ \forall \varepsilon > 0,\ \exists \delta = \varepsilon$ such that $\forall t_0 \geq 0$,

$$(\forall i)\ \ (\|x_i(t_0) - \zeta\| \leq \delta) \wedge (x_i(t_0) \in \mathcal{S}) \Longrightarrow (\forall t \geq t_0)(\forall i)\ \|x_i(t) - \zeta\| \leq \varepsilon.$$

The conclusion follows by Definition 3.4. ∎

Now comes our main result.

THEOREM 3.8. *Suppose $\mathcal{S}$ is closed and convex. The switched interconnected system* (3.2) *has the property of asymptotic state agreement on $\mathcal{S}$ if and only if the dynamic interaction digraph $\mathcal{G}_{\sigma(t)}$ is UQSC.*

This section concludes with a few remarks.

If $\mathcal{S} = \mathbb{R}^m$ in assumptions A1 and A2, then the switched interconnected system (3.2) has the global asymptotic state agreement property if and only if $\mathcal{G}_{\sigma(t)}$ is UQSC.

When the vector fields in the family (3.1) are nonautonomous, suppose assumptions A1 and A2 are replaced by the following (keeping assumption A3 the same):

A1′. $f_p^i(t,x)$ is locally Lipschitz with respect to x on $\mathcal{S}^n$ and piecewise continuous with respect to t.

A2′. For all $x \in \mathcal{S}^n$ and all $t \in \mathbb{R}$, $f_p^i(t,x) \in \text{ri}\left(\mathcal{T}(x_i, \mathcal{C}_p^i(x))\right)$.

It can be shown [21] that Theorem 3.8 no longer holds in general.

In the special case that the interaction graph is fixed ($\sigma(t)$ is a constant signal), then the property of UQSC is equivalent to QSC. Thus, we arrive at the following special result.

COROLLARY 3.9. *Suppose $\sigma(t) = p$ and $\mathcal{S} = \mathbb{R}^m$. Then, the interconnected system* (3.2) *has the globally asymptotic state agreement property if and only if $\mathcal{G}_p$ is QSC.*

For this special case we can actually relax the assumptions on the vector fields $f_p^i : \mathbb{R}^{mn} \to \mathbb{R}^m$ as follows:

A1″. f_p^i is continuous on $\mathbb{R}^{mn}$.

A2″. For all $x \in \mathbb{R}^{mn}$, $f_p^i(x) \in \mathcal{T}\left(x_i, \mathcal{C}_p^i(x)\right)$. Moreover, $f_p^i(x) \neq 0$ if $\mathcal{C}_p^i(x)$ is not a singleton and x_i is its vertex.

A sketch of the proof can be found in [24]. Unlike the proof of Theorem 3.8 here (see section 5), the proof in [24] relies on LaSalle's invariance principle. Finally, it is worth pointing out that assumption A1″ is too weak for sufficiency in Theorem 3.8 when the interaction digraph is dynamic [21].

Application: Synchronization of coupled oscillators. The Kuramoto model [17, 39] describes the dynamics of a set of n oscillators with angles θ_i with natural frequencies ω_i. The time evolution of the ith oscillator is given by

$$\dot{\theta}_i = \omega_i + k_i \sum_{j \in \mathcal{N}_i(t)} \sin(\theta_j - \theta_i),$$

where $k_i > 0$ is the coupling strength and $\mathcal{N}_i(t)$ is the set of neighbors of oscillator i at time t. The interaction structure can be general up to this point in the paper; that is, $\mathcal{N}_i(t)$ can be an arbitrary set of other nodes and can be dynamic.

The neighbor sets $\mathcal{N}_i(t)$ define $\mathcal{G}_{\sigma(t)}$ and the switched interconnected system

$$\dot{\theta}(t) = f_{\sigma(t)}\left(\theta(t)\right),$$

where $\theta = (\theta_1, \ldots, \theta_n)$ and $\sigma(t)$ is a suitable switching signal. For identical frequencies (i.e., $\omega_i = \omega \ \forall i$), the transformation $x_i = \theta_i - \omega t$ yields

$$\dot{x}_i = k_i \sum_{j \in \mathcal{N}_i(t)} \sin(x_j - x_i), \quad i = 1, \ldots, n. \tag{3.4}$$

Let a, b be any real numbers such that $0 \leq b - a < \pi$, and define $\mathcal{S} = [a, b]$. It can be checked that A1 and A2 are satisfied. Suppose $\sigma(t)$ here is regular enough to satisfy A3. Then from Theorem 3.8 it follows that, if and only if $\mathcal{G}_{\sigma(t)}$ is UQSC, the switched interconnected system (3.4) has the property of asymptotic state agreement on $\mathcal{S}$. This implies that there exists $\bar{x} \in \mathbb{R}$ such that the oscillators asymptotically synchronize:

$$\theta_i(t) \to \bar{x} + \omega t, \quad \dot{\theta}_i(t) \to \omega.$$

This extends Theorem 1 in [17], which assumes the interaction graph is undirected and static and the initial state $\theta_i(0) \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ for all i.

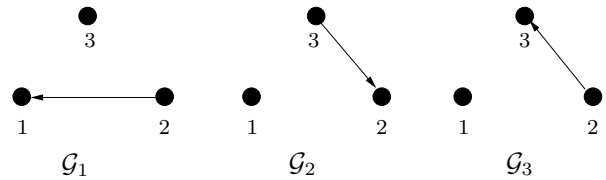


FIG. 3.3. *Three interaction digraphs* $\mathcal{G}_p$, $p = 1, 2, 3$.

As an example, three Kuramoto oscillators with time-varying interaction are simulated. The initial conditions are $\theta_1 = 0$, $\theta_2 = 1$, $\theta_3 = -1$. The natural frequency ω_i equals 1, and the coupling strength k_i is set to 1 for all i. The interaction structure switches among three possible interaction structures periodically, as shown in Figure 3.3. It can be checked that $\mathcal{G}_{\sigma(t)}$ is UQSC. Thus these three oscillators achieve asymptotical synchronization by the main theorem. Figure 3.4 shows the plots of $\sin(\theta_i)$, $i = 1, 2, 3$, and of the switching signal $\sigma(t)$. Synchronization is evident.

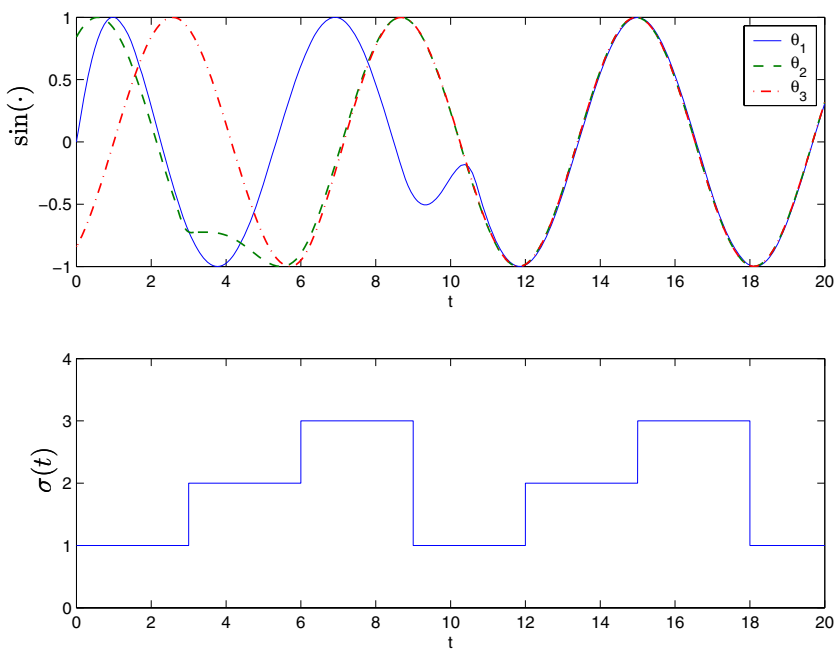


FIG. 3.4. *Synchronization of three oscillators with a dynamic interaction structure.*

4. The rendezvous problem. Now we turn to the second main topic: the rendezvous problem for autonomous mobile robots moving in continuous time. The problem here is different because connectivity is state dependent instead of time dependent a priori.

Suppose there are n robots, each having the simple kinematic model of velocity control: $\dot{x}_i = u_i$, where $x_i \in \mathbb{R}^m$ is the position of robot i. Assume that, due to the limited field of view of its sensor, each robot can sense only the relative positions of its neighbor robots within radius r. Letting $\mathcal{N}_i(x)$ denote the set of neighbors of robot i, where x is the aggregate state of n robots, we thus have that $\{y_{ij} = x_j - x_i : \ j \in$

$\mathcal{N}_i(x)\}$ is the information available to robot i. The *rendezvous problem* is to design local distributed control laws u_i, functions of $\{y_{ij} : j \in \mathcal{N}_i(x)\}$, such that all states $\{x_i : i = 1, \dots, n\}$ converge to a common value $\bar{x} \in \mathbb{R}^m$.

The interaction digraph is state dependent, $\mathcal{G}_{\sigma(x)}$, because of the proximity sensors, and the switched interconnected system takes the form

$$\dot{x} = f_{\sigma(x)}(x), \tag{4.1}$$

where $\sigma : \mathbb{R}^{mn} \to \mathcal{P}$. Let us *fix an initial state* $x^0 \in \mathbb{R}^{mn}$ and assume that (4.1) has a solution $x(t)$ defined for all $t \geq 0$. Then the state-dependent switching rule can be viewed as a time-dependent switching rule $\sigma(x(t))$, and the interaction graph becomes time dependent too, $\mathcal{G}_{\sigma(x(t))}$.

If some robots are initialized so far away from the rest that they never acquire information from them, then the rendezvous problem obviously cannot be solved. This corresponds to the situation where $\mathcal{G}_{\sigma(x(0))}$ is not QSC. Thus it is natural to assume that $\mathcal{G}_{\sigma(x(0))}$ *is* QSC. Moreover, we wish the control laws u_i to be devised such that $\mathcal{G}_{\sigma(x(t))}$ does not lose this property in the future, even though the controller may cause changes in $\mathcal{G}_{\sigma(x(t))}$. Intuitively, u_i should make the maximum distance between robot i and its neighbors nonincreasing.

Let $\mathcal{I}_i(x)$ denote the set of neighbor robots $j \in \mathcal{N}_i(x)$ that have maximum distance from robot i (generically $\mathcal{I}_i(x)$ is a singleton).

PROPOSITION 4.1. *Assume that for each i the control law u_i satisfies*

$$(\forall x) \quad \max_{j \in \mathcal{I}_i(x)} (x_i - x_j)^T u_i \leq 0. \tag{4.2}$$

If $\mathcal{G}_{\sigma(x^0)}$ is QSC and a solution $x(t)$ to (4.1) exists for all $t \geq 0$, then $\mathcal{G}_{\sigma(x(t))}$ is QSC for all $t \geq 0$.

Proof. Define

$$V(x) = \max_i \max_{j \in \mathcal{N}_i(x)} \|x_i - x_j\|^2.$$

Notice that $V(x) \leq r$, where r is the radius of the field of view of each robot. Also, define

$$\mathcal{I}(x) = \left\{(i,j) : \ V(x) = \|x_i - x_j\|^2, \ j \in \mathcal{N}_i(x)\right\},$$

the set of pairs of indices where the maximum is reached. By Lemma 2.2

$$\begin{aligned} D^+V(x(t)) &= 2 \max_{(i,j) \in \mathcal{I}(x)} \left[(x_i - x_j)^T u_i + (x_j - x_i)^T u_j\right] \\ &\leq 2 \max_{(i,j) \in \mathcal{I}(x)} (x_i - x_j)^T u_i + 2 \max_{(i,j) \in \mathcal{I}(x)} (x_j - x_i)^T u_j. \end{aligned}$$

It follows from (4.2) that

$$\max_{(i,j) \in \mathcal{I}(x)} (x_i - x_j)^T u_i \leq 0 \quad \text{and} \quad \max_{(i,j) \in \mathcal{I}(x)} (x_j - x_i)^T u_j \leq 0.$$

Hence $D^+V(x(t)) \leq 0$ for all $t \geq 0$, which means the already linked arcs will never be disconnected and therefore the conclusion follows. ∎

Next, we show that if the distributed control law u_i satisfies (4.2) as well as assumptions A1″ and A2″, then a solution $x(t)$ to (4.1) exists for all $t \geq 0$, and the robots rendezvous.

PROPOSITION 4.2. *Suppose $\mathcal{G}_{\sigma(x^0)}$ is QSC. If u_i satisfies* (4.2) *as well as* A1″ *and* A2″, *then the robots rendezvous.*

Proof. If $\mathcal{G}_{\sigma(x^0)}$ is fully connected, then $\mathcal{G}_{\sigma(x(t))}$ is fixed for all time $t \geq 0$ since no link will be dropped, by Proposition 4.1, and no link can be added. Then the conclusion follows from Corollary 3.9.

If instead $\mathcal{G}_{\sigma(x^0)}$ is not fully connected, then $\mathcal{G}_{\sigma(x(t))}$ is dynamic and switches for a finite number of times. To prove this, suppose by contradiction that for all $t \geq 0$, $\mathcal{G}_{\sigma(x(t))} = \mathcal{G}_{\sigma(x^0)}$. Then by Corollary 3.9, all the robots converge to a common location. So $\mathcal{G}_{\sigma(x(t))}$ will become fully connected at some time t, which contradicts the assumption that $\mathcal{G}_{\sigma(x(t))} = \mathcal{G}_{\sigma(x^0)}$ is not fully connected. Hence, there is a $t_1 \geq 0$ such that $\mathcal{G}_{\sigma(x(t_1))}$ has more links than $\mathcal{G}_{\sigma(x^0)}$ because no link will be dropped by Proposition 4.1. Repeating this argument a finite number of times eventually leads to the existence of t_i such that $\mathcal{G}_{\sigma(x(t_i))}$ is fully connected, and thus, it is fixed after t_i. Then the conclusion follows from Corollary 3.9 by treating $(t_i, x(t_i))$ as the initial condition. □

The control law given next is based on the algorithm first proposed in [1].

PROPOSITION 4.3. *A possible choice of u_i satisfying condition* (4.2) *as well as assumptions* A1″ *and* A2″ *is $u_i = e(0, y_{ij} : j \in \mathcal{N}_i(x))$, the Euclidean center of the set $\mathcal{Z} = \{0, y_{ij}, j \in \mathcal{N}_i(x)\}$.*

Proof. The Euclidean center of the set $\mathcal{Z}$ is the unique point w that minimizes the function $g(w) := \max_{z \in \mathcal{Z}} \|w - z\|$. Interpreted geometrically, $e(\cdot)$ is the center of the smallest m-sphere that contains the set of points $\{0, y_{ij}, j \in \mathcal{N}_i(x)\}$. Furthermore, it can be easily shown that it lies in the polytope $\widetilde{\mathcal{C}}_p^i = \text{co}\{0, y_{ij}, j \in \mathcal{N}_i(x)\}$ but not at its vertices if the polytope is not a singleton. Thus,

$$e\left(0, y_{ij} : j \in \mathcal{N}_i(x)\right) = \arg\min_{w \in \widetilde{\mathcal{C}}_p^i} \left(\max_{z \in \mathcal{Z}} \|w - z\| \right).$$

Then, by the maximum theorem [40], the function $e(\cdot)$ is continuous (but not locally Lipschitz by some other arguments), and hence u_i satisfies assumption A1″.

Next, $e(\cdot) \in \widetilde{\mathcal{C}}_p^i$ implies $e(\cdot) \in \mathcal{T}(0, \widetilde{\mathcal{C}}_p^i)$. Also, notice that $\mathcal{C}_p^i(x) = \text{co}\{x_i, x_j : \ j \in \mathcal{N}_i(x)\}$ is the translation of $\widetilde{\mathcal{C}}_p^i$ to the point x_i. Hence, $e(\cdot) \in \mathcal{T}(x_i, \mathcal{C}_p^i(x))$. In addition, if $\mathcal{C}_p^i(x)$ is not a singleton and x_i is its vertex, this means that $\widetilde{\mathcal{C}}_p^i$ is not a singleton and 0 is its vertex. Then by the fact that $e(\cdot)$ lies in $\widetilde{\mathcal{C}}_p^i$ but not at its vertices, it follows that $u_i = e(\cdot) \neq 0$. Thus u_i satisfies assumption A2″.

Finally, u_i satisfies (4.2). This can be seen from geometry. We show the case $m = 2$ for illustration. If $u_i = 0$, then it trivially satisfies (4.2). If $u_i \neq 0$, then the picture is as in Figure 4.1. The solid circle C_1 is the smallest circle enclosing the points 0 and $y_{ij}, j \in \mathcal{N}_i(x)$. The dotted circle C_2 is centered at the origin and goes through the intersection points between C_1 and its diameter, which is perpendicular to u_i. We know that if there are some y_{ij} in the closed shaded area, then one of them achieves the maximal distance from the origin among all $y_{ij}, j \in \mathcal{N}_i(x)$. On the other hand, there is at least one $j \in \mathcal{N}_i(x)$ such that y_{ij} is in the closed semicircle of C_1, since otherwise it is not the smallest circle. Hence, y_{ij} lies in the closed shaded area if $j \in \mathcal{I}_i(x)$. Moreover, the angle between u_i and such y_{ij} is less than $\pi/2$. This implies that $\max_{j \in \mathcal{I}_i(x)} (x_i - x_j)^T u_i \leq 0$. □

5. Proofs of the main results in section 3. Our proofs rely heavily on nonsmooth analysis involving the Dini derivative. They are partly inspired by a result of Narendra and Annaswamy [28], who show that with $\dot{V}(x,t) \leq 0$ uniform

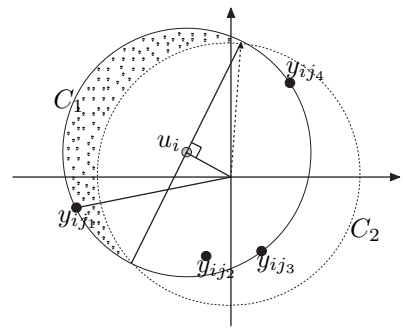


FIG. 4.1. *The smallest enclosing circle.*

asymptotic stability can be proved if there exists a positive T such that for all t, $V(x(t+T),t+T) - V(x(t),t) \leq -\gamma(\|x(t)\|) < 0$, where γ is a class $\mathcal{K}$ function. The difference here is that we deal with stability with respect to a *set*—the set of aggregate states where the subsystem states are all equal—rather than stability of an equilibrium point; an additional complication is that the natural V-functions are nondifferentiable.

Nagumo's theorem concerning set invariance is stated first, for later reference.

THEOREM 5.1 (see [3]). *Consider the system $\dot{y} = F(y)$, with $F : \mathbb{R}^l \to \mathbb{R}^l$, and let $\mathcal{Y} \subset \mathbb{R}^l$ be a closed convex set. Assume that, for each y^0 in $\mathcal{Y}$, there exists $\epsilon(y^0) > 0$ such that the system admits a unique solution $y(t, y^0)$ defined for all $t \in [0, \epsilon(y^0))$. Then,*

$$y^0 \in \mathcal{Y} \Longrightarrow \big(\forall t \in [0, \epsilon(y^0))\big)\ \ y(t, y^0) \in \mathcal{Y}$$

if and only if $F(y) \in \mathcal{T}(y, \mathcal{Y})$ for all $y \in \mathcal{Y}$.

Proof of Theorem 3.6. Let $\mathcal{A}$ be any compact convex set in $\mathcal{S}$ and consider any initial state $x^0 \in \mathcal{A}^n$ and any initial time t_0. For any piecewise constant switching signal $\sigma(t)$, let $x(t, t_0, x^0)$ be the solution of the switched interconnected system (3.2) with $x(t_0) = x^0$, and let $[t_0, t_0 + \epsilon(t_0, x^0))$ be its maximal interval of existence.

For any point $x \in \mathcal{A}^n$, it is obvious that $\mathcal{C}_p^i(x) \subset \mathcal{A}$ for all $i \in \mathcal{I}_0$ and $p \in \mathcal{P}$, by convexity of $\mathcal{A}$. Thus, by property (i) in Lemma 2.1,

$$f_p^i(x) \in \mathrm{ri}\left(\mathcal{T}(x_i, \mathcal{C}_p^i(x))\right) \subset \mathcal{T}(x_i, \mathcal{A})\ \forall i \in \mathcal{I}_0,\ \forall p \in \mathcal{P},$$

and by property (ii) in the same lemma,

$$g(t,x) := f_{\sigma(t)}(x) \in \mathcal{T}(x, \mathcal{A}^n)\ \forall t \in \mathbb{R},\ \forall x \in \mathcal{A}^n.$$

Set $y = (t, x)$ and construct the augmented system

$$\dot{y} = F(y) := \begin{bmatrix} 1 \\ g(y) \end{bmatrix}. \tag{5.1}$$

Since $g(t,x)$ admits a unique solution $x(t, t_0, x^0)$ defined for all $t \in [t_0, t_0 + \epsilon(t_0, x^0))$, it follows that for all $y^0 = (t_0, x^0) \in \mathbb{R} \times \mathcal{A}^n$, the augmented system (5.1) has a unique solution $y(t, y^0)$ defined on $[0, \epsilon(y^0))$. Moreover,

$$F(y) \in \mathcal{T}(t, \mathbb{R}) \times \mathcal{T}(x, \mathcal{A}^n) = \mathcal{T}(y, \mathbb{R} \times \mathcal{A}^n)\ \forall y \in \mathbb{R} \times \mathcal{A}^n.$$

Since $\mathbb{R} \times \mathcal{A}^n$ is closed and convex, by Theorem 5.1 it follows that

$$y^0 = (t_0, x^0) \in \mathbb{R} \times \mathcal{A}^n \Longrightarrow (\forall \tau \in [0, \epsilon(y^0)))\ y(\tau) \in \mathbb{R} \times \mathcal{A}^n. \tag{5.2}$$

The solution $y(\tau)$ to (5.1) with initial condition $y^0 = (t_0, x^0)$ is related to the solution $x(t)$ to $\dot{x} = g(t, x)$ with initial condition $x(t_0) = x^0$ as follows:

$$(\forall t \in [t_0, t_0 + \epsilon(t_0, x^0)))\ (t, x(t)) = y(t - t_0).$$

We thus rewrite condition (5.2) as

$$t_0 \in \mathbb{R} \text{ and } x^0 \in \mathcal{A}^n \Longrightarrow (\forall t \in [t_0, t_0 + \epsilon(t_0, x^0)))\ x(t) \in \mathcal{A}^n.$$

Since the set $\mathcal{A}^n$ is compact, it follows by Theorem 2.4 in [18] that, for all $x^0 \in \mathcal{A}^n$ and all t_0, $\epsilon(t_0, x^0) = \infty$ and the set $\mathcal{A}$ is positively invariant for the switched interconnected system (3.2) by definition 3.5. □

Now we need some additional notation. First, a hypercube in $\mathbb{R}^m$:

$$\mathcal{A}_r(z) = \{y \in \mathbb{R}^m : \|y - z\|_\infty \leq r\}.$$

Let $c > 0$ be large enough that $\mathcal{S}_c := \mathcal{S} \cap \mathcal{A}_c(0)$ is not empty. Now consider any $x = (x_1, \ldots, x_n)$, $x_i \in \mathcal{S}_c$. Each x_i lives in $\mathbb{R}^m$. Let $\mathcal{C}(x)$ denote the convex hull of the points $x_1, \ldots, x_n$; $\mathcal{C}(x)$ is a polytope in $\mathbb{R}^m$.

To simplify notation, we focus on the first axis in $\mathbb{R}^m$. Along this axis, let $a_1(x)$ and $b_1(x)$ denote the upper and lower ordinates of $\mathcal{C}(x)$, as in Figure 5.1. The set $\{y \in \mathcal{C}(x) : y_1 = a_1(x)\}$ is the first upper boundary of $\mathcal{C}(x)$. Finally, for small enough $r > 0$, define

$$\mathcal{H}_r(x) = \{y \in \mathcal{C}(x) : y_1 \leq a_1(x) - r\}.$$

The setup is summarized in Figure 5.1.

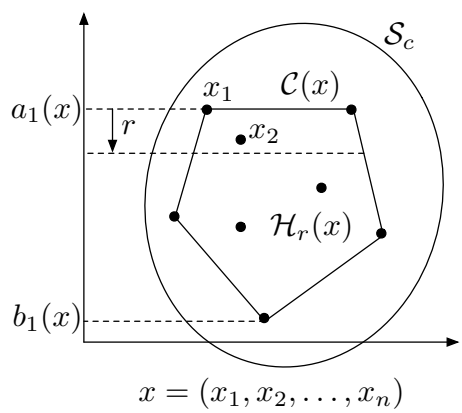


FIG. 5.1. *Illustration to define notation: $\mathcal{C}(x)$ is the convex hull of the points $x_1, \ldots, x_n$; $a_1(x), b_1(x)$ are its upper and lower ordinates; $\mathcal{H}_r(x)$ is the part of the convex hull below the line with ordinate $a_1(x) - r$.*

Now we need two technical lemmas for which we assume that the hypotheses of Theorem 3.8 hold. Due to space limitation, we have to omit the proofs and refer the reader to [21].

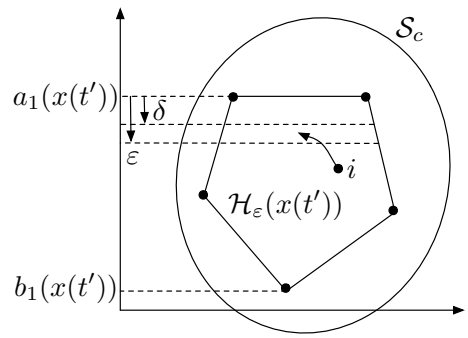


FIG. 5.2. *Illustration for Lemma* 5.2*: Agent i is in $\mathcal{H}_\varepsilon(x(t'))$ at time t_1, and it cannot get into the upper layer of width δ in the near future.*

The first lemma is illustrated in Figure 5.2.

LEMMA 5.2. *For every sufficiently large $c > 0$, there exists a class $\mathcal{KL}$ function $\gamma : [0, 2c] \times [0, \infty) \to [0, \infty)$ such that $\gamma(\triangle, 0) = \triangle$ and such that the following is true: For every $(t', x(t')) \in \mathbb{R} \times \mathcal{S}_c^n$, every $\varepsilon > 0$ sufficiently small, and every $T > 0$, if $x_i(t_1) \in \mathcal{H}_\varepsilon(x(t'))$ at $t_1 \geq t'$, then $x_i(t) \in \mathcal{H}_\delta(x(t'))$ for all $t \in [t_1, t_1 + T]$, where $\delta = \gamma(\varepsilon, T)$.*

The second lemma is illustrated in Figure 5.3.

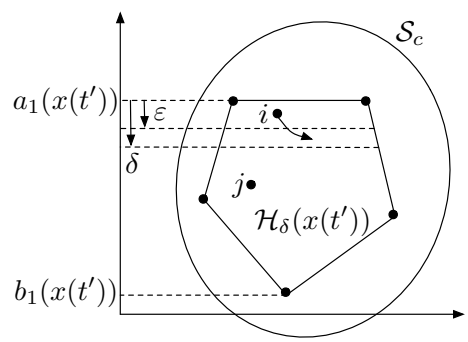


FIG. 5.3. *Illustration for Lemma* 5.3*: Agent i has the neighbor j in $\mathcal{H}_\delta(x(t'))$ and will consequently be pulled into $\mathcal{H}_\varepsilon(x(t'))$.*

LEMMA 5.3. *For every sufficiently large $c > 0$, there exists a class $\mathcal{K}$ function $\varphi : [0, 2c] \to [0, \infty)$ such that $\varphi(\triangle) < \triangle$ for $\triangle \neq 0$ and such that the following is true: For every $(t', x(t')) \in \mathbb{R} \times \mathcal{S}_c^n$ and every $\delta > 0$ sufficiently small, if there exist a pair (i, j) and a $t_1 \geq t'$ such that $j \in \mathcal{N}_i(t)$ and $x_j(t) \in \mathcal{H}_\delta(x(t'))$ for all $t \in [t_1, t_1 + \tau_D]$, then there exists a $t_2 \in [t', t_1 + \tau_D]$ such that $x_i(t_2) \in \mathcal{H}_\varepsilon(x(t'))$, where $\varepsilon = \varphi(\delta)$.*

Proof of Theorem 3.8. (Necessity.) To prove the contrapositive form, assume that $\mathcal{G}_{\sigma(t)}$ is not UQSC. That is, for every $T > 0$ there exists $t^* \geq 0$ such that $\mathcal{G}([t^*, t^* + T])$ is not QSC; i.e., it does not have a center. Then, in $\mathcal{G}([t^*, t^* + T])$ there are two nodes i^* and j^* such that for every node k either $k \not\rightarrow i^*$ or $k \not\rightarrow j^*$. Let $\mathcal{V}_1$ be the set of

nodes l such that $l \to i^*$ and let $\mathcal{V}_2$ be the set of nodes l such that $l \to j^*$. Obviously, $\mathcal{V}_1$ and $\mathcal{V}_2$ are disjoint. Moreover, for each node $i \in \mathcal{V}_1$ (resp., $\mathcal{V}_2$), the set of neighbors of agent i in $\mathcal{G}([t^*, t^* + T])$ is a subset of $\mathcal{V}_1$ (resp., $\mathcal{V}_2$). This implies that, for all $t \in [t^*, t^* + T]$ and for all $(i, j) \in \mathcal{V}_1 \times \mathcal{V}_2$,

$$\mathcal{N}_i(\sigma(t)) \subseteq \mathcal{V}_1 \quad \text{and} \quad \mathcal{N}_j(\sigma(t)) \subseteq \mathcal{V}_2.$$

Choose any $z_1, z_2 \in \mathcal{S}$ such that $z_1 \neq z_2$. Let $t_0 = t^*$ and pick any initial condition $x(t_0)$ such that

$$x_i(t_0) = \begin{cases} z_1 & \text{if } i \in \mathcal{V}_1, \\ z_2 & \text{if } i \in \mathcal{V}_2. \end{cases}$$

Then, by assumption A2, for all $t \in [t_0, t_0 + T]$,

$$x_i(t) = \begin{cases} z_1 & \forall i \in \mathcal{V}_1, \\ z_2 & \forall i \in \mathcal{V}_2. \end{cases}$$

Let $c = \max_i \|x_i(t_0)\|$ and let ε be a positive scalar smaller than $\|z_1 - z_2\|/2$. We have thus found $\varepsilon > 0$ and $c > 0$ such that, for all $T > 0$, there exists $t_0 = t^*$ such that

$$(\forall i)\ (\|x_i(t_0)\| \leq c) \wedge (x_i(t_0) \in \mathcal{S}),\ \text{but}\ (\forall \zeta \in \mathcal{S})(\exists t = t_0 + T)(\exists i)\ \|x_i(t) - \zeta\| > \epsilon.$$

Thus system (3.2) does not have the property of asymptotic state agreement on $\mathcal{S}$.

(Sufficiency.) Assume $\mathcal{G}_{\sigma(t)}$ is UQSC. By Theorem 3.7 the switched interconnected system (3.2) has the property of state agreement on $\mathcal{S}$, so it remains to show that $\forall \varepsilon > 0,\ \forall c > 0,\ \exists T^* > 0$ such that $\forall t_0 \geq 0$

$$(\forall i)\ x_i(t_0) \in \mathcal{S}_c \Longrightarrow (\exists \zeta \in \mathcal{S})(\forall t \geq t_0 + T^*)(\forall i)\ x_i(t) \in \mathcal{A}_\varepsilon(\zeta). \tag{5.3}$$

Let $\varepsilon > 0,\ c > 0$ be arbitrary. There exist a class $\mathcal{KL}$ function γ and a class $\mathcal{K}$ function φ satisfying the properties in Lemmas 5.2 and 5.3, respectively. For any given $t_0 \geq 0$ and $x^0 \in \mathcal{S}_c^n$, consider the solution $x(t)$ of (3.2) with $x(t_0) = x^0$ and the nonnegative function $V_j(x) := a_j(x) - b_j(x)$, $j = 1, \ldots, m$. Thus $V_j(x(t))$ equals the width in the jth direction of the convex hull of the agents at time t. By Theorem 3.6, for every $t \geq t' \geq t_0$, $x_i(t) \in \mathcal{C}(x(t')) \subset \mathcal{S}_c$ for all i. It follows that $V_j(x(t))$ is nonincreasing along the trajectory $x(t)$.

Since $\mathcal{G}_{\sigma(t)}$ is UQSC, there is a $T' > 0$ such that for each t the union digraph $\mathcal{G}([t, t + T'])$ is QSC. Let $T = T' + 2\tau_D$, where τ_D is the dwell time.

Claim. There exists a class $\mathcal{K}$ function η such that for every $t' \geq t_0$

$$V_1(x(t' + \bar{T})) - V_1(x(t')) \leq -\eta\left(V_1(x(t'))\right), \tag{5.4}$$

where $\bar{T} = 2nT$.

Let us postpone the proof of this claim and see how the theorem follows from the claim. From (5.4) we have

$$V_1\left(x(t_0 + k\bar{T})\right) \leq V_1\left(x(t_0)\right) - \eta\left(V_1(x(t_0))\right) - \cdots - \eta\left(V_1(x(t_0 + (k-1)\bar{T}))\right).$$

Notice that $x^0 \in \mathcal{S}_c^n(0)$ implies $V_1(x^0) \leq 2c$. In addition, considering the facts that η is a class $\mathcal{K}$ function and that $V_1(x(t))$ is nonincreasing, one obtains

$$V_1(x(t_0 + k\bar{T})) \leq 2c - k\eta(V_1(x(t_0 + k\bar{T}))).$$

This means there is a $T_1^* = k\bar{T} > 0$ (k large enough) such that $V_1(x(t)) < 2\varepsilon$ for all $t \geq t_0 + T_1^*$. For each $j = 2, \ldots, m$, by the same argument, there is a $T_j^* > 0$ such that $V_j(x(t)) < 2\varepsilon$ for all $t \geq t_0 + T_j^*$. Let $T^* = \max_j T_j^*$. Thus $V_j(x(t)) < 2\varepsilon$ for all $t \geq t_0 + T^*$ and all $j = 1, \ldots, m$. This in turn implies that there exists a $\zeta \in \mathcal{S}$ such that $x_i(t) \in \mathcal{A}_\varepsilon(\zeta)$ for all i and all $t \geq t_0 + T^*$. This proves (5.3).

Now we prove the claim. Inequality (5.4) says that the width, along the first axis, of the convex hull of the agents reduces measurably from time t' to time $t' + \bar{T}$. The proof is intricate and involves applying Lemmas 5.2 and 5.3 alternately.

We begin by constructing a family of parameters, $\varepsilon_1, \delta_1, \varepsilon_2, \ldots, \varepsilon_{n-1}, \delta_{n-1}, \varepsilon_n$. First, ε_1 is taken to be half the width at time t': $\varepsilon_1 = V_1(x(t'))/2$. Then δ_1 is produced by applying Lemma 5.2: $\delta_1 = \gamma(\varepsilon_1, \bar{T})$. Then ε_2 comes from Lemma 5.3, $\varepsilon_2 = \varphi(\delta_1)$, and δ_2 comes from Lemma 5.2, $\delta_2 = \gamma(\varepsilon_2, \bar{T})$. Continuing, we set

$$\begin{aligned}
\varepsilon_3 &= \varphi(\delta_2), \\
\delta_3 &= \gamma(\varepsilon_3, \bar{T}) \\
&\vdots \\
\varepsilon_{n-1} &= \varphi(\delta_{n-2}), \\
\delta_{n-1} &= \gamma(\varepsilon_{n-1}, \bar{T}), \\
\varepsilon_n &= \varphi(\delta_{n-1}).
\end{aligned}$$

Define $\bar{\gamma}(\cdot) := \gamma(\cdot, \bar{T})$. Then ε_n can be written as

$$\varepsilon_n = \eta\left(V_1(x(t'))\right),$$

where $\eta(\cdot) := \varphi \circ \bar{\gamma} \circ \cdots \circ \varphi \circ \bar{\gamma}(\cdot/2)$. It is a class $\mathcal{K}$ function since $\bar{\gamma}$ and φ both are. Since γ is class $\mathcal{KL}$ with the property $\gamma(\triangle, 0) = \triangle$ and $\bar{T} > 0$, it follows that $\delta_k < \varepsilon_k$. In addition, $\varepsilon_{k+1} < \delta_k$ because $\varphi(\triangle) < \triangle$ for $\triangle \neq 0$. Thus,

$$0 < \varepsilon_n < \delta_{n-1} < \cdots < \delta_1 < \varepsilon_1.$$

These parameters are used as in Figure 5.4.

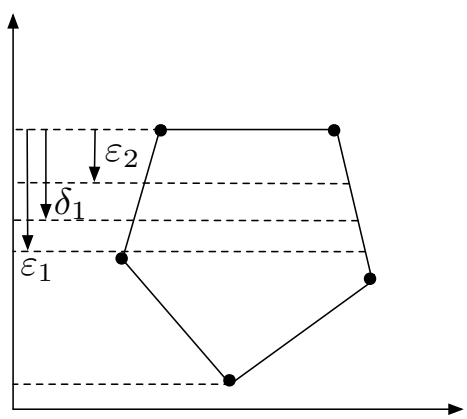


FIG. 5.4. *The parameters $\varepsilon_1, \delta_1, \varepsilon_2, \ldots, \varepsilon_{n-1}, \delta_{n-1}, \varepsilon_n$ with respect to the convex hull and the first axis.*

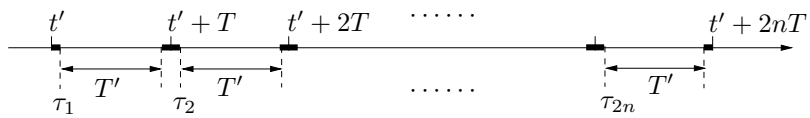


FIG. 5.5. *The time interval* $[t', t'+\bar{T}]$.

Let $\mathcal{V}_1$ and $\mathcal{V}_1^*$ be a partition of the node set $\mathcal{V}$ such that $i \in \mathcal{V}_1$ if $x_i(t') \in \mathcal{H}_{\varepsilon_1}$ and $i \in \mathcal{V}_1^*$ otherwise. Thus $\mathcal{V}_1$ is the set of agents located in the lower half of the convex hull in Figure 5.4 at time t'.

Next, we apply the two lemmas to construct a sequence of times at which certain events are known to occur. In what follows, hopefully without causing confusion, we use $\mathcal{H}_r$ to denote $\mathcal{H}_r(x(t'))$ for simplicity. As shown in Figure 5.5, let

$$\begin{aligned}\tau_1 &= t' + \tau_D,\\ \tau_2 &= t' + T + \tau_D\\ &\vdots\\ \tau_{2n} &= t' + (2n-1)T + \tau_D.\end{aligned}$$

For each $k = 1, \ldots, 2n$, the digraph $\mathcal{G}([\tau_k, \tau_k + T'])$ is QSC, and therefore it has a center, say c_k. Now c_k is either in $\mathcal{V}_1$ or in $\mathcal{V}_1^*$; thus at least n elements in $\{c_1, \ldots, c_{2n}\}$ lie in either $\mathcal{V}_1$ or $\mathcal{V}_1^*$. Assume without loss of generality that they lie in $\mathcal{V}_1$; thus there exist indices $1 \le k_1 < \cdots < k_n \le 2n$ such that $c_{k_i} \in \mathcal{V}_1$.

At time t', by definition, $\mathcal{H}_{\varepsilon_1}$ has at least one agent (see Figure 5.4). Moreover, by Lemma 5.2, for all i

$$x_i(t') \in \mathcal{H}_{\varepsilon_1} \Longrightarrow x_i(t) \in \mathcal{H}_{\delta_1} \ \forall t \in [t', t'+\bar{T}]. \tag{5.5}$$

Since $\mathcal{G}([\tau_{k_1}, \tau_{k_1} + T'])$ has a center c_{k_1} in $\mathcal{V}_1$, there exists a pair $(i,j) \in \mathcal{V}_1^* \times \mathcal{V}_1$ such that j is a neighbor of i in this digraph; otherwise there is no link from j to i for any $i \in \mathcal{V}_1^*$ and $j \in \mathcal{V}_1$, which contradicts the fact that the digraph has a center in $\mathcal{V}_1$. This further implies that there is a $\tau \in [\tau_{k_1}, \tau_{k_1} + T']$ such that $j \in \mathcal{N}_i(\tau)$. Since $\tau \in [\tau_{k_1}, \tau_{k_1} + T'] = [t' + (k_1 - 1)T + \tau_D, t' + k_1 T - \tau_D]$, it follows that $[\tau - \tau_D, \tau + \tau_D] \subset [t' + (k_1 - 1)T, t' + k_1 T]$. Since $\sigma(t) \in \mathcal{S}_{dwell}(\tau_D)$, there is an interval $[\bar{\tau}, \bar{\tau} + \tau_D]$, which contains τ and is a subinterval of $[t', t' + k_1 T]$, such that $j \in \mathcal{N}_i(t)$ for all $t \in [\bar{\tau}, \bar{\tau} + \tau_D]$. In addition, since $j \in \mathcal{V}_1$ or, what is the same, $x_j(t') \in \mathcal{H}_{\varepsilon_1}$, from (5.5) we know that $x_j(t) \in \mathcal{H}_{\delta_1}$ for all $t \in [t', t' + \bar{T}]$ (and of course for all $t \in [\bar{\tau}, \bar{\tau} + \tau_D]$). Thus, by Lemma 5.3, there exists $t_1 \in [t', \bar{\tau} + \tau_D] \subseteq [t', t' + k_1 T]$ such that $x_i(t_1) \in \mathcal{H}_{\varepsilon_2}$.

So we have shown on the one hand that the agents not in $\mathcal{H}_{\varepsilon_1}$ at t' are in $\mathcal{H}_{\varepsilon_2}$ at t_1. On the other hand, the agents in $\mathcal{H}_{\varepsilon_1}$ at t' remain in $\mathcal{H}_{\delta_1}$ at t_1 from (5.5), and therefore remain in $\mathcal{H}_{\varepsilon_2}$ at t_1 because $\mathcal{H}_{\delta_1} \subset \mathcal{H}_{\varepsilon_2}$. Hence, at time t_1, $\mathcal{H}_{\varepsilon_2}(x(t'))$ has at least two agents.

Let $\mathcal{V}_2$ and $\mathcal{V}_2^*$ be a partition of the node set $\mathcal{V}$ such that $i \in \mathcal{V}_2$ if $x_i(t_1) \in \mathcal{H}_{\varepsilon_2}$ and $i \in \mathcal{V}_2^*$ otherwise. Note that by (5.5)

$$k \in \mathcal{V}_1 \Longrightarrow x_k(t') \in \mathcal{H}_{\varepsilon_1} \underset{(5.5)}{\Longrightarrow} x_k(t_1) \in \mathcal{H}_{\delta_1} \subset \mathcal{H}_{\varepsilon_2} \Longrightarrow k \in \mathcal{V}_2,$$

so $\mathcal{V}_1 \subset \mathcal{V}_2$. In particular c_{k_2}, the center node of $\mathcal{G}([\tau_{k_2}, \tau_{k_2} + T'])$, is in $\mathcal{V}_2$ because it is in $\mathcal{V}_1$. Then we can apply the same argument to conclude that there are a

$t_2 \in [t_1, t' + k_2 T]$ and an i in $\mathcal{V}_2^*$ such that $x_i(t_2) \in \mathcal{H}_{\varepsilon_3}$ and therefore, $\mathcal{H}_{\varepsilon_3}$ has at least three agents at t_2.

Repeating this argument $n-1$ times leads to the result that there is a $t_{n-1} \in [t', t' + k_{n-1}T] \subset [t', t' + \bar{T}]$ such that $\mathcal{H}_{\varepsilon_n}$ has n agents at t_{n-1}. Hence,

$$V_1(x(t_{n-1})) \leq V_1(x(t')) - \varepsilon_n = V_1(x(t')) - \eta(V_1(x(t'))),$$

and (5.4) follows. □

6. Conclusions. In this paper we first studied the state agreement problem for a class of switched interconnected large-scale systems with a family of admissible vector fields. The interconnection structure is time varying and independent of the state. The key assumption about the vector fields, A2, generalizes Moreau's assumption in discrete time. Necessary and sufficient conditions, in terms of the interaction graph, are obtained to assure that the system achieves asymptotic state agreement. These results can be understood as connective stability, as in the framework of [36]. Achieving asymptotic state agreement of a large-scale interconnected system is robust with respect to either the coupling structure or parameter values. In addition, our results and analysis may be of independent interest in the field of switched systems.

Second, we studied the rendezvous problem in continuous time. The interconnection structure is defined in terms of the distances between agents and hence is state independent. We proved that the circumcenter control law is a solution to the problem.

The notion of state agreement in this paper is that the states of the subsystems are all equal and constant. This notion can potentially be generalized in the following two directions. First, state agreement could mean equality of all the trajectories of the subsystems. In other words, the trajectories of a collection of subsystems will follow, after some transient, the same path in time. This would be of interest in formation control of multiagent systems. Second, state agreement could mean equality of all the states after suitable state transformations. An example is a biochemical reaction network studied in [21].

In many state-agreement problems, the interaction graphs are bidirectional. For such cases, it is reasonable to conjecture that interconnected systems enjoy several special properties. For instance, results similar to those in Theorem 3.8 may be obtained with weaker assumptions on the smoothness of the vector fields.

Finally, we conjecture in the spirit of [2] that our result could be generalized by replacing $\mathcal{C}_p^i(x)$ in assumption A2 by a set-valued map satisfying suitable properties.

REFERENCES

[1] H. Ando, Y. Oasa, I. Suzuki, and M. Yamashita, *Distributed memoryless point convergence algorithm for mobile robots with limited visibility*, IEEE Trans. Robotics Automation, 15 (1999), pp. 818–828.

[2] D. Angeli and P. A. Bliman, *Stability of leaderless discrete-time multi-agent systems*, Math. Control Signals Systems, 18 (2006), pp. 293–322.

[3] J.-P. Aubin, *Viability Theory*, Birkhäuser Boston, Boston, 1991.

[4] R. W. Beard and V. Stepanyan, *Information consensus in distributed multiple vehicle coordinated control*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 2029–2034.

[5] I. V. Belykh, V. N. Belykh, and M. Hasler, *Blinking model and synchronization in small-world networks with a time-varying coupling*, Phys. D, 195 (2004), pp. 188–206.

[6] I. V. Belykh, V. N. Belykh, and M. Hasler, *Connection graph stability method for synchronized coupled chaotic systems*, Phys. D, 195 (2004), pp. 159–187.

[7] C. Berge and A. Ghouila-Houri, *Programming, Games, and Transportation Networks*, John Wiley and Sons, New York, 1965.
[8] F. H. Clarke, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
[9] J. Cortes, S. Martinez, and F. Bullo, *Robust rendezvous for mobile autonomous agents via proximity graphs in arbitrary dimensions*, IEEE Trans. Automat. Control, 51 (2006), pp. 1289–1298.
[10] J. M. Danskin, *The theory of max-min, with applications*, SIAM J. Appl. Math., 14 (1966), pp. 641–664.
[11] M. H. DeGroot, *Reach a consensus*, J. Amer. Statist. Assoc., 69 (1974), pp. 118–121.
[12] V. Gazi and K. M. Passino, *Stability analysis of swarms*, IEEE Trans. Automat. Control, 48 (2003), pp. 692–697.
[13] V. Gazi and K. M. Passino, *Stability analysis of social foraging swarms*, IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics, 34 (2004), pp. 539–557.
[14] J. Gunawardena, *Chemical Reaction Network Theory for In-Silico Biologists*, Technical report, Bauer Center for Genomics Research, Harvard University, Cambridge, MA, 2003.
[15] Y. Hatano and M. Mesbahi, *Agreement over random networks*, IEEE Trans. Automat. Control, 50 (2005), pp. 1867–1872.
[16] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.
[17] A. Jadbabaie, N. Motee, and M. Barahona, *On the stability of the Kuramoto model of coupled nonlinear oscillators*, in Proceedings of the 2004 American Control Conference, Boston, MA, 2004, pp. 4296–4301.
[18] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1996.
[19] J. Lin, A. S. Morse, and B. D. O. Anderson, *The multi-agent rendezvous problem. Part* 1: *The synchronous case*, SIAM J. Control Optim., submitted.
[20] J. Lin, A. S. Morse, and B. D. O. Anderson, *The multi-agent rendezvous problem. Part* 2: *The asynchronous case*, SIAM J. Control Optim., submitted.
[21] Z. Lin, *Coupled Dynamic Systems: From Structure Towards Stability and Stabilizability*, Ph.D. dissertation, University of Toronto, Toronto, 2005.
[22] Z. Lin, M. Broucke, and B. Francis, *Local control strategies for groups of mobile autonomous agents*, IEEE Trans. Automat. Control, 49 (2004), pp. 622–629.
[23] Z. Lin, B. Francis, and M. Maggiore, *Necessary and sufficient graphical conditions for formation control of unicycles*, IEEE Trans. Automat. Control, 50 (2005), pp. 121–127.
[24] Z. Lin, B. Francis, and M. Maggiore, *On the state agreement problem for multiple nonlinear dynamical systems*, in Proceedings of 16th IFAC World Congress, Prague, Czech Republic, 2005.
[25] Y. Moses and S. Rajsbaum, *A layered analysis of consensus*, SIAM J. Comput., 31 (2002), pp. 989–1021.
[26] L. Moreau, *Stability of Continuous-Time Distributed Consensus Algorithms*, 2004; available online at http://arxiv.org/abs/math/0409010.
[27] L. Moreau, *Stability of multiagent systems with time-dependent communication links*, IEEE Trans. Automat. Control, 50 (2005), pp. 169–182.
[28] K. S. Narendra and A. M. Annaswamy, *Persistent excitation in adaptive systems*, Internat. J. Control, 45 (1987), pp. 127–160.
[29] R. Olfati-Saber and R. M. Murray, *Consensus problems in networks of agents with switching topology and time-delays*, IEEE Trans. Automat. Control, 49 (2004), pp. 101–115.
[30] A. Pogromsky, G. Santoboni, and H. Nijmeijer, *Partial synchronization: From symmetry towards stability*, Phys. D, 172 (2002), pp. 65–87.
[31] W. Ren, R. W. Beard, and T. W. McLain, *Coordination variables and consensus building in multiple vehicle systems*, in Cooperative Control, Lecture Notes in Control and Inform. Sci. 309, V. Kumar, N. Leonard, and A. S. Morse, eds., Springer-Verlag, Berlin, 2005, pp. 171–188.
[32] T. Richardson, *Stable polygons of cyclic pursuit*, Ann. Math. Artif. Intell., 31 (2001), pp. 147–172.
[33] N. Rouche, P. Habets, and M. Laloy, *Stability Theory by Liapunov's Direct Method*, Springer-Verlag, New York, Heidelberg, 1977.
[34] R. Sepulchre, D. Paley, and N. Leonard, *Collective motion and oscillator synchronization*, in Cooperative Control, Lecture Notes in Control and Inform. Sci. 309, V. Kumar, N. Leonard, and A. S. Morse, eds., Springer-Verlag, Berlin, 2005, pp. 189–205.
[35] D. D. Siljak, *On stability of large-scale systems under structural perturbations*, IEEE Trans. on Systems, Man, and Cybernetics, 3 (1973), pp. 415–417.

[36] D. D. Siljak, *Decentralized Control of Complex Systems*, Academic Press, Boston, 1991.
[37] G. V. Smirnov, *Introduction to the Theory of Differential Inclusions*, AMS, Providence, RI, 2001.
[38] I. Stewart, M. Golubitsky, and M. Pivato, *Symmetry groupoids and patterns of synchrony in coupled cell networks*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 609–646.
[39] S. H. Strogatz, *From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators*, Phys. D, 143 (2000), pp. 1–20.
[40] R. K. Sundaram, *A First Course in Optimization Theory*, Cambridge University Press, Cambridge, UK, 1996.
[41] I. Suzuki and M. Yamashita, *Distributed anonymous mobile robots: Formation of geometric patterns*, SIAM J. Comput., 28 (1999), pp. 1347–1363.
[42] V. I. Vorotnikov, *On the coordinate synchronization problem for dynamical systems*, Differ. Equ., 40 (2004), pp. 14–22.
[43] C. W. Wu, *Synchronization in arrays of coupled nonlinear systems: Passivity, circle criterion, and observer design*, IEEE Trans. Circuits Syst. I: Fund. Theory Appl., 48 (2001), pp. 1257–1261.
[44] C. W. Wu, *Synchronization in coupled arrays of chaotic oscillators with nonreciprocal coupling*, IEEE Trans. Circuits Syst. I: Fund. Theory Appl., 50 (2003), pp. 294–297.
[45] C. W. Wu and L. O. Chua, *Synchronization in an array of linearly coupled dynamical systems*, IEEE Trans. Circuits Syst. I: Fund. Theory Appl., 42 (1995), pp. 430–447.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 308–331

QUICKEST DETECTION OF A MINIMUM OF TWO POISSON DISORDER TIMES*

ERHAN BAYRAKTAR† AND H. VINCENT POOR‡

Abstract. A multisource quickest detection problem is considered. Assume there are two independent Poisson processes X^1 and X^2 with disorder times θ_1 and θ_2, respectively; i.e., the intensities of X^1 and X^2 change at random unobservable times θ_1 and θ_2, respectively. θ_1 and θ_2 are independent of each other and are exponentially distributed. Define $\theta \triangleq \theta_1 \wedge \theta_2 = \min\{\theta_1, \theta_2\}$. For any stopping time τ that is measurable with respect to the filtration generated by the observations, define a penalty function of the form

$$R_\tau = \mathbb{P}(\tau < \theta) + c\mathbb{E}\left[(\tau - \theta)^+\right],$$

where $c > 0$ and $(\tau - \theta)^+$ is the positive part of $\tau - \theta$. It is of interest to find a stopping time τ that minimizes the above performance index. This performance criterion can be useful, e.g., in the following scenario: There are two assembly lines that produce products A and B, respectively. Assume that the malfunctioning (disorder) of the machines producing A and B are independent events. Later, the products A and B are to be put together to obtain another product C. A product manager who is worried about the quality of C will want to detect the minimum of the disorder times (as accurately as possible) in the assembly lines producing A and B. Another problem to which we can apply our framework is the Internet surveillance problem: A router receives data from, say, n channels. The channels are independent, and the disorder times of channels are $\theta_1, \ldots, \theta_n$. The router is said to be under attack at $\theta = \theta_1 \wedge \cdots \wedge \theta_n$. The administrator of the router is interested in detecting θ as quickly as possible. Since both observations X^1 and X^2 reveal information about the disorder time θ, even this simple problem is more involved than solving the disorder problems for X^1 and X^2 separately. This problem is formulated in terms of a three-dimensional sufficient statistic, and the corresponding optimal stopping problem is examined. The solution is characterized by iterating a suitable functional operator.

Key words. change detection, Poisson processes, optimal stopping

AMS subject classifications. 62L10, 62L15, 62C10, 60G40

DOI. 10.1137/050630933

1. Introduction.

Consider two independent Poisson processes $X^i = \{X_t^i : t \geq 0\}$ $i \in \{1, 2\}$ with the same arrival rate β. At some random unobservable times θ_1 and θ_2, with distributions

$$\mathbb{P}(\theta_i = 0) = \pi_i, \quad \mathbb{P}(\theta_i > t) = (1 - \pi_i)e^{-\lambda t} \text{ for } t \geq 0, \tag{1.1}$$

the arrival rates of the Poisson processes X^1 and X^2 change from β to α, respectively, i.e.,

$$X_t^i - \int_0^t h_i(s)ds, \quad t \geq 0,\ i = 1, 2, \tag{1.2}$$

are martingales in which

$$h_i(t) = [\beta 1_{\{s<\theta_i\}} + \alpha 1_{\{s\geq\theta_i\}}], \quad t \geq 0,\ i = 1, 2. \tag{1.3}$$

*Received by the editors May 7, 2005; accepted for publication (in revised form) November 11, 2006; published electronically April 13, 2007. This work was supported in part by the U.S. Army Pantheon Project and National Science Foundation under grant DMS-0604491.

http://www.siam.org/journals/sicon/46-1/63093.html

†Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (erhan@umich.edu).

‡School of Engineering and Applied Science, Princeton University, Princeton, NJ 08544 (poor@princeton.edu).

Here α and β are known positive constants. We seek a stopping rule τ that detects the instant $\theta = \theta_1 \wedge \theta_2$ of the first regime change as accurately as possible given the past and the present observations of the processes X^1 and X^2. More precisely, we wish to choose a stopping time τ of the history of the processes X^1 and X^2 that minimizes the following penalty function:

$$R_\tau = \mathbb{P}(\tau < \theta) + c\mathbb{E}\left[(\tau - \theta)^+\right]. \tag{1.4}$$

The first term in (1.4) penalizes the frequency of false alarms, and the second term penalizes the detection delay. The disorder time demarcates two regimes, and in each of these regimes the decision maker uses distinctly different strategies. Therefore, it is in the decision maker's interest to detect the disorder time as accurately as possible from its observations. Here, we are solving the case when a decision maker has two identical and independent sources to process. In section 9 we discuss how our analysis can be extended to nonidentical sources.

Quickest detection problems arise in a variety of applications such as seismology, machine monitoring, finance, health, and surveillance, among others (see, e.g., [1], [10], [7], [11], and [14]). Because Poisson processes are often used to model abrupt changes, Poisson disorder problems have potential applications to, e.g., to the effective control and prevention of infectious diseases, the quickest detection of quality and reliability problems in industrial processes, and the surveillance of Internet traffic to protect network servers from the attacks of malicious users. This is because the number of patients infected, the number of defected items produced, and the number of packets arriving at a network node are usually modeled by Poisson processes. In these examples the disorder time corresponds to the time when an outbreak occurs, when a machine in an assembly line breaks down, or when a router is under attack, respectively. The multisource quickest detection problem considered here can be applied to tackle these problems when there are multiple sources of information. For example, in the monitoring of industrial processes the minimum of disorder times represents the first time when one of many assembly lines in a plant breaks down during the production of a certain type of item. Let us be more specific: Assume that there are two assembly lines that produce products A and B, respectively. Assume also that the malfunctioning (disorder) of the machines producing A and B are independent events. Later, the products A and B are to be put together to obtain another product C. A product manager who is worried about the quality of C will want to detect the minimum of the disorder times (as accurately as possible) in the assembly lines producing A and B. The performance function (1.4) is an appropriate choice because the product manager will worry about the quality of the end product C, not of the individual pieces seperately. Another problem to which we can apply our framework is the Internet surveillance problem: A router receives data from, say, n channels. The channels are independent and the disorder times of the channels are $\theta_1, \ldots, \theta_n$. The router is said to be under attack at $\theta = \theta_1 \wedge \cdots \wedge \theta_n$. The administrator of the router is interested in detecting θ as quickly as possible.

The one-dimensional Poisson disorder problem, i.e., the problem of detecting θ_1 as accurately as possible given the observations from the Poisson process X^1, has recently been solved (see [2] and [3] and the references therein). The two-dimensional disorder problem we have introduced cannot be reduced to solving the corresponding one-dimensional disorder problems since both X^1 and X^2 reveal some information about θ whenever these processes jump. That is, if we take the minimum of the optimal stopping times that solve the one-dimensional Poisson disorder problems, then we obtain a stopping time that is a suboptimal solution to (1.4) (see Remark 4.1).

We will show that the quickest detection problem of (1.4) can be reduced to an optimal stopping problem for a three-dimensional piecewise-deterministic Markov process. Continuous-time Markov optimal stopping problems are typically solved by formulating them as free boundary problems associated with the infinitesimal generator of the Markov process. In this case, however the infinitesimal generator contains differential delay operators. Solving free boundary problems involving differential delay operators is a challenge even in the one-dimensional case, and the smooth fit principle is expected to fail (see [2] and [3] and the references therein). Instead as in [4] and [6] we work with an integral operator, iteration of which generates a monotonically increasing sequence of functions converging exponentially to the value function of the optimal stopping problem. That is, using the integral operator we reduce the problem to a sequence of deterministic optimization problems. This approach provides a new numerical method for calculating and characterizing the value function and the continuation region in addition to providing information about the shape and the location of the optimal continuation region. Using the structure of the paths of the piecewise-deterministic Markov process, we also provide a nontrivial bound on the optimal stopping time which can be used to obtain approximate stopping strategies.

The remainder of this paper is organized as follows: In sections 2 and 3, we restate the problem of interest under a suitable reference measure $\mathbb{P}_0$ that is equivalent to $\mathbb{P}$. Working under the reference measure $\mathbb{P}_0$ reduces the computations considerably, since under this measure the observations X^1 and X^2 are simple Poisson processes that are independent of the disorder times. Here we show that the quickest detection problem reduces to solving an optimal stopping problem for a three-dimensional statistic. In section 4, we analyze the path behavior of this sufficient statistic. In section 5, we provide a tight upper bound on the continuation region of the optimal stopping problem, which can be used to determine approximate detection rules besides helping us to determine the location and the shape of the continuation region. Here, we also show that the smallest optimal stopping time of the problem under consideration has finite expectation. In section 6, we convert the optimal stopping problem into sequences of deterministic optimal stopping problems using a suitably defined integral operator. In section 7, we construct optimal stopping times from sequences of stopping (alarm) times that sound before the processes X^1 and X^2 jump a certain number of times. In section 8 we discuss the structure of the optimal stopping regions. And finally, we discuss how to extend our approach to the case with more than two sources and to the case when the jump sizes are random and the jump size distribution changes at the time of disorder.

2. Problem description. Let us start with a probability space $(\Omega, \mathcal{F}, \mathbb{P}_0)$ that hosts two independent Poisson processes X^1 and X^2, both of which have rate β, as well as two independent random variables θ_1 and θ_2 independent of the Poisson processes with distributions

$$\mathbb{P}_0(\theta_i = 0) = \pi_i \quad \text{and} \quad \mathbb{P}_0(\theta_i > t) = (1 - \pi_i)e^{-\lambda_i t} \tag{2.1}$$

for $0 \leq t < \infty$, $i \in \{1, 2\}$, and for some known constants $\pi_i \in [0, 1)$ and $\lambda > 0$ for $i \in \{1, 2\}$. We denote by $\mathbb{F} = \{\mathcal{F}_t\}_{0 \leq t < \infty}$ the filtration generated by X^1 and X^2, i.e., $\mathcal{F}_t = \sigma(X^1_s, X^2_s, 0 \leq s \leq t)$, and denote by $\mathbb{G} = \{\mathcal{G}_t\}_{0 \leq t < \infty}$ the initial enlargement of $\mathbb{F}$ by θ_1 and θ_2, i.e., $\mathcal{G}_t \triangleq \sigma(\theta_1, \theta_2, X^2_s, X^2_s : 0 \leq s \leq t)$. The processes X^1 and X^2 satisfy (1.2) under a new probability measure $\mathbb{P}$, which is characterized by

$$\left.\frac{d\mathbb{P}}{d\mathbb{P}_0}\right|_{\mathcal{G}_t} \triangleq Z_t \triangleq Z^1_t Z^2_t, \tag{2.2}$$

where

$$Z_t^i \triangleq \exp\left(\int_0^t \log\left(\frac{h_i(s-)}{\beta}\right) dX_s^i - \int_0^t [h_i(s)-\beta]ds\right) \tag{2.3}$$

for $t \geq 0$ and $i \in \{1,2\}$ are exponential martingales (see, e.g., [5]). Under this new probability measure $\mathbb{P}$ of (2.2), θ_1 and θ_2 have the same distribution as they have under the measure $\mathbb{P}_0$, i.e., their distribution is given by (1.1). This holds because θ_1 and θ_2 are $\mathcal{G}_0$-measurable and $d\mathbb{P}/d\mathbb{P}_0|_{\mathcal{G}_0} = 1$, i.e., $\mathbb{P}$ and $\mathbb{P}_0$ coincide on $\mathcal{G}_0$. Under the new probability measure $\mathbb{P}$ the processes X^1 and X^2 have measurable intensities h_1 and h_2, respectively. That is to say that (1.2) holds. In other words, the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ describes the model posited in (1.1) and (1.2). Now, our problem is to find a quickest detection rule for the disorder times $\theta_1 \wedge \theta_2$, which is *adapted to* the history $\mathcal{F}$ generated by the observed processes X^1 and X^2 because the complete information (concerning θ_1 and θ_2) embodied in $\mathcal{G}$ is not available. We will achieve our goal by finding an $\mathcal{F}$-stopping time that minimizes (1.4).

In terms of the exponential likelihood processes

$$L_t^i \triangleq \left(\frac{\alpha}{\beta}\right)^{X_t^i} \exp(-(\alpha-\beta)t), \qquad t \geq 0,\ i \in \{1,2\}, \tag{2.4}$$

we can write

$$Z_t^i = 1_{\{\theta_i > t\}} + 1_{\{\theta_i \leq t\}} \frac{L_t^i}{L_{\theta_i}^i}. \tag{2.5}$$

Let us introduce the posterior probability process

$$\Pi_t \triangleq \mathbb{P}\left(\theta \leq t \middle| \mathcal{F}_t\right) = \frac{\mathbb{E}_0\left[Z_t 1_{\{\theta \leq t\}} \middle| \mathcal{F}_t\right]}{\mathbb{E}_0\left[Z_t \middle| \mathcal{F}_t\right]}, \tag{2.6}$$

where the second equality follows from the Bayes formula (see, e.g., [9]). Then it follows from (2.5) and (2.6) that

$$1 - \Pi_t = \frac{(1-\pi)e^{-2\lambda t}}{\mathbb{E}_0\left[Z_t \middle| \mathcal{F}_t\right]}, \quad \text{where} \tag{2.7}$$

$$\pi \triangleq 1 - (1-\pi_1)(1-\pi_2). \tag{2.8}$$

Let us now introduce the odds-ratio process

$$\Phi_t \triangleq \frac{\Pi_t}{1-\Pi_t}, \qquad 0 \leq t < \infty. \tag{2.9}$$

Then observe from (2.6) and (2.7) that

$$\mathbb{E}_0\left[Z_t 1_{\{\theta \leq t\}} | \mathcal{F}_t\right] = (1-\pi)e^{-\lambda t}\Phi_t, \tag{2.10}$$

$t \geq 0$. Now, we will write the penalty function of (1.4) in terms of the odds-ratio process:

$$\begin{aligned} \mathbb{E}\left[(\tau-\theta)^+\right] &= \mathbb{E}\left[\int_0^\infty 1_{\{\tau>t\}} 1_{\{\theta \leq t\}} dt\right] = \int_0^\infty \mathbb{E}_0\left[1_{\{\tau>t\}} \mathbb{E}_0\left[Z_t 1_{\{\theta\leq t\}} \middle| \mathcal{F}_t\right]\right] dt \\ &= (1-\pi)\mathbb{E}_0 \int_0^\tau e^{-2\lambda t}\Phi_t dt. \end{aligned} \tag{2.11}$$

Since $\{\tau < \theta\} \in \mathcal{G}_\theta$ we can write

$$\mathbb{P}(\tau < \theta) = \mathbb{E}_0\left[Z_\theta 1_{\{\tau<\theta\}}\right] = \mathbb{P}_0(\tau < \theta) = (1-\pi)\left(1 - \lambda\mathbb{E}_0\left[\int_0^\tau e^{-2\lambda t}dt\right]\right), \tag{2.12}$$

where the second equality follows since $Z_\theta = 1$ almost surely under P_0. Using (2.11) and (2.12) we can write the penalty function as

$$R_\tau(\pi_1, \pi_2) = 1 - \pi + c(1-\pi)\mathbb{E}_0\left[\int_0^\tau e^{-2\lambda t}\left(\Phi_t - \frac{\lambda}{c}\right)dt\right]. \tag{2.13}$$

On the other hand the following lemma obtains a representation for the odds-ratio process Φ.

LEMMA 2.1. *Let us denote*

$$\Phi_t^i \triangleq \frac{e^{\lambda t}}{1-\pi_i}\mathbb{E}_0\left[1_{\{\theta_i \le t\}}\frac{L_t^i}{L_{\theta_i}^i}\Big|\mathcal{F}_t^i\right] = \frac{\mathbb{P}(\theta_i \le t|\mathcal{F}_t)}{1-\mathbb{P}(\theta_i \le t|\mathcal{F}_t)} \tag{2.14}$$

for $t \ge 0$ and $i \in \{1,2\}$. Then we can write the odds-ratio process Φ as

$$\Phi_t = \Phi_t^1 + \Phi_t^2 + \Phi_t^1\Phi_t^2, \quad t \ge 0. \tag{2.15}$$

Proof. From (2.10)

$$\begin{aligned}
\Phi_t &= \frac{e^{2\lambda t}}{(1-\pi)}\mathbb{E}_0\left[Z_t 1_{\{\theta\le t\}}|\mathcal{F}_t\right] \\
&= \frac{e^{2\lambda t}}{(1-\pi)}\left\{\mathbb{P}_0(\theta_1 > t)\mathbb{E}_0\left[1_{\{\theta_1\le t\}}\frac{L_t^1}{L_{\theta_1}^1}\Big|\mathcal{F}_t^1\right] + \mathbb{P}(\theta_2 > t)\mathbb{E}_0\left[1_{\{\theta_2\le t\}}\frac{L_t^2}{L_{\theta_i}^2}\Big|\mathcal{F}_t^i\right]\right. \\
&\quad \left. + \mathbb{E}_0\left[1_{\{\theta_1\le t\}}\frac{L_t^1}{L_{\theta_1}^1}\Big|\mathcal{F}_t^1\right]\mathbb{E}_0\left[1_{\{\theta_2\le t\}}\frac{L_t^2}{L_{\theta_2}^2}\Big|\mathcal{F}_t^2\right]\right\}.
\end{aligned} \tag{2.16}$$

The second equality follows from (2.2), (2.5), and the independence of the sigma algebras $\mathcal{F}_t^1$ and $\mathcal{F}_t^2$. Now the claim follows from (2.1), (2.8), and (2.14). □

Using the fact that the likelihood ratio process L^i is the unique solution of the equation

$$dL_t^i = [(\alpha/\beta) - 1]L_{t-}^i(dX^i - \alpha\, dt), \quad L_0^i = 1 \tag{2.17}$$

(see, e.g., [13]) and by means of the chain-rule, we obtain

$$d\Phi_t^i = (\lambda + (\lambda - \alpha + \beta)\Phi_t^i)dt + [(\alpha/\beta) - 1]\Phi_t^i dX_t^i, \quad \Phi_0^i = \frac{\pi_i}{1-\pi_i} \tag{2.18}$$

for $t \ge 0$ and $i \in \{1,2\}$ (see [3]). If we let

$$\Phi_t^+ \triangleq \Phi_t^1 + \Phi_t^2, \;\; \Phi_t^\times \triangleq \Phi_t^1\Phi_t^2, \;\; t \ge 0, \tag{2.19}$$

then using a change of variable formula for jump processes gives

$$\begin{aligned}
d\Phi_t^\times &= [\lambda\Phi_t^+ + 2a\Phi_t^\times]dt + ((\alpha/\beta) - 1)\Phi_t^\times d(X_t^1 + X_t^2), \\
d\Phi_t^+ &= [2\lambda + a\Phi_t^+]dt + ((\alpha/\beta) - 1)[\Phi_t^1 dX_t^1 + \Phi_t^2 dX_t^2],
\end{aligned} \tag{2.20}$$

with $\Phi_0^\times = \pi_1\pi_2/[(1-\pi_1)(1-\pi_2)]$ and $\Phi_0^+ = \pi_1/(1-\pi_1) + \pi_2/(1-\pi_2)$, where $a \triangleq \lambda - \alpha + \beta$. Note that $X_t \triangleq X_t^1 + X_t^2$, $t \geq 0$, is a Poisson process with rate 2β under $\mathbb{P}_0$.

It is clear from (2.18) and (2.20) that

$$\Upsilon \triangleq (\Phi^\times, \Phi^+, \Phi^1) \tag{2.21}$$

is a piecewise-deterministic Markov process; therefore, the original change detection problem with penalty function (1.4) has been reformulated as (2.13) and (2.18)–(2.21), which is an optimal stopping problem for a two-dimensional Markov process driven by a three-dimensional piecewise-deterministic Markov process.

We will denote by $\mathcal{A}$ the infinitesimal generator of Υ. Its action on a smooth test function $f : \mathbb{B}_+^3 \to \mathbb{R}$ is given by

$$\begin{aligned}
&[\mathcal{A}f](\phi^\times, \phi^+, \phi^1) = D_{\phi^\times} f(\phi^\times, \phi^+, \phi^1)[\lambda\phi^+ + 2a\phi^+] + D_{\phi^+} f(\phi^\times, \phi^+, \phi^1)[2\lambda + a\phi^+] \\
&+ D_{\phi^1} f(\phi^\times, \phi^+, \phi^1)[\lambda + a\phi^1] + \beta\left[f\left(\frac{\alpha}{\beta}\,\phi^\times, \phi^+ + \left(\frac{\alpha}{\beta} - 1\right)\phi^1, \frac{\alpha}{\beta}\,\phi^1\right) - f(\phi^\times, \phi^+, \phi^1)\right] \\
&+ \beta\left[f\left(\frac{\alpha}{\beta}\,\phi^\times, \frac{\alpha}{\beta}\,\phi^+ - \left(\frac{\alpha}{\beta} - 1\right)\phi^1, \phi^1\right) - f(\phi^\times, \phi^+, \phi^1)\right].
\end{aligned} \tag{2.22}$$

Let us denote

$$\mathbb{B}_+^2 \triangleq \{(x,y) \in \mathbb{R}_+^2 : y \geq 2\sqrt{x}\} \quad \text{and} \tag{2.23}$$

$$\mathbb{B}_+^3 \triangleq \{(x,y,z) \in \mathbb{R}_+^3 : y \geq 2\sqrt{x}, y \geq z\}. \tag{2.24}$$

Now, for every $(\phi^\times, \phi^+, \phi^1) \in \mathbb{B}_+^3$, let us denote by $x(t, \phi^\times)$, $y(t, \phi^+)$, and $z(t, \phi^1)$, $t \in \mathbb{R}$, the solutions of

$$\begin{aligned}
&\frac{d}{dt}x(t,\phi^\times) = [\lambda y(t,\phi^+) + 2ax(t,\phi^\times)]dt, \quad x(0,\phi^\times) = \phi^\times, \\
&\frac{d}{dt}y(t,\phi^+) = [2\lambda + ay(t,\phi^+)]dt, \quad y(0,\phi^+) = \phi^+, \\
&\frac{d}{dt}z(t,\phi^1) = [\lambda + az(t,\phi^1)]dt, \quad z(0,\phi^1) = \phi^1.
\end{aligned} \tag{2.25}$$

The solutions of (2.25), when $a \neq 0$, are explicitly given by

$$\begin{aligned}
x(t,\phi^\times) &= \frac{\lambda^2}{a^2} + e^{2at}\left[\phi^\times - \frac{\lambda^2}{a^2}\right] + e^{2at}(1 - e^{-at})\frac{\lambda}{a}\left(\phi^+ + \frac{2\lambda}{a}\right), \\
y(t,\phi^+) &= -\frac{2\lambda}{a} + e^{at}\left(\phi^+ + \frac{2\lambda}{a}\right), \\
z(t,\phi^1) &= -\frac{\lambda}{a} + e^{at}\left(\phi^+ + \frac{\lambda}{a}\right).
\end{aligned} \tag{2.26}$$

Otherwise, $x(t,\phi^\times) = \phi^\times + \lambda t\phi^\times + \lambda^2 t^2$, $y(t,\phi^+) = \phi^+ + 2\lambda t$, and $z(t,\phi^1) = \phi^1 + \lambda t$. Note that the solution (x,y,z) of the system of equations in (2.25) satisfies the semigroup property, i.e., for every $s, t \in \mathbb{R}$,

$$\begin{aligned}
&x(t+s,\phi_0) = x(s,x(t,\phi_0)), \quad y(t+s,\phi_1) = y(s,y(t,\phi_1)), \\
&\text{and} \quad z(t+s,\phi_1) = z(s,z(t,\phi_1)).
\end{aligned} \tag{2.27}$$

Note from (2.18), (2.20), and (2.26) that

$$\Phi_t^{\times} = x(t-\sigma_n, \Phi_{\sigma_n}^{\times}), \qquad , \Phi_t^{+} = y(t-\sigma_n, \Phi_{\sigma_n}^{+}), \qquad \Phi_t^{1} = z(t-\sigma_n, \Phi_{\sigma_n}^{1}), \quad \sigma_n \le t < \sigma_{n+1},\ n \in \mathbb{N}, \tag{2.28}$$

and

$$\begin{aligned} \Phi_{\sigma_{n+1}}^{\times} &= \frac{\alpha}{\beta}\,\Phi_{\sigma_{n+1}-}^{\times}, \quad \Phi_{\sigma_{n+1}}^{1} = \frac{\alpha}{\beta}\,\Phi_{\sigma_{n+1}}^{1} 1_{\{X_{\sigma_{n+1}}^1 \neq X_{\sigma_{n+1}-}^1\}}, \quad \text{and} \\ \Phi_{\sigma_{n+1}}^{+} &= \left[\Phi_{\sigma_{n+1}-}^{+} + \left(\frac{\alpha}{\beta}-1\right)\Phi_{\sigma_{n+1}-}^{1}\right] 1_{\{X_{\sigma_{n+1}}^1 \neq X_{\sigma_{n+1}-}^1\}} \\ &\quad + \left[\frac{\alpha}{\beta}\,\Phi_{\sigma_{n+1}-}^{+} - \left(\frac{\alpha}{\beta}-1\right)\Phi_{\sigma_{n+1}-}^{1}\right] 1_{\{X_{\sigma_{n+1}}^2 \neq X_{\sigma_{n+1}-}^2\}}. \end{aligned} \tag{2.29}$$

Here, for any function h, $h(t-) \triangleq \lim_{s\uparrow t} h(t)$. Note that an observer watching Υ is able to tell whenever the processes X^1 and X^2 jump (see (2.18) and (2.20)), i.e., the filtration generated by Υ is the same as $\mathbb{F}$.

3. An optimal stopping problem. Let us denote the set of $\mathbb{F}$-stopping times by $\mathcal{S}$. The value function of the quickest detection problem

$$U(\pi_1, \pi_2) \triangleq \inf_{\tau \in \mathcal{S}} R_\tau(\pi_1, \pi_2) \tag{3.1}$$

can be written as

$$U(\pi_1, \pi_2) = (1-\pi)\left[1 + cV\left(\frac{\pi_1\pi_2}{1-\pi}, \frac{\pi_1+\pi_2-2\pi_1\pi_2}{1-\pi}, \frac{\pi_1}{1-\pi_1}\right)\right], \tag{3.2}$$

where V is the value function of the optimal stopping problem

$$V(\phi^{\times}, \phi^{+}, \phi^{1}) \triangleq \inf_{\tau \in \mathcal{S}} \mathbb{E}_0^{(\phi^{\times}, \phi^{+}, \phi^{1})}\left[\int_0^{\tau} e^{-\lambda t} h\left(\Phi_t^{\times}, \Phi_t^{+}\right) dt\right] \tag{3.3}$$

in which $(\phi^{\times}, \phi^{+}, \phi^{1}) \in \mathbb{D}_+^3$ and $h(x,y) \triangleq x + y - \lambda/c$. Here $\mathbb{E}_0^{(\phi^{\times}, \phi^{+}, \phi^{1})}$ is the expectation under $\mathbb{P}_0$ given that $\Phi_0^{\times} = \phi^{\times}$, $\Phi_0^{+} = \phi^{+}$, and $\Phi_0^{1} = \phi^{1}$.

It is clear from (3.3) that for both optimal stopping problems it is not optimal to stop before $(\Phi_t^{\times}, \Phi_t^{+})$, $t \ge 0$, leaves the *advantageous region* defined by

$$\mathbb{C}_0 \triangleq \{(\phi^{\times}, \phi^{+}) \in \mathbb{B}_+^2 : \phi^{\times} + \phi^{+} \le \lambda/c\}. \tag{3.4}$$

Let us also denote

$$\mathbb{C} \triangleq \{(\phi^{\times}, \phi^{+}, \phi^{1}) \in \mathbb{B}_+^3 : \phi^{\times} + \phi^{+} \le \lambda/c\}. \tag{3.5}$$

Also note that the only reason not to stop at the time of the first exit from the region $\mathbb{C}_0$ is the prospect of $(\Phi_t^{\times}, \Phi_t^{+})$, $t \ge 0$, returning to $\mathbb{C}_0$ at a future time.

Remark 3.1. It is reasonable to question our choice of statisitc, since it is clear that $(\Phi_t^1, \Phi_t^2)_{t\ge 0}$ contains all the information X has to offer. Our choice $(\Upsilon_t)_{t\ge 0}$, which is defined in (2.21), is motivated by the mere desire of having a concave value function U and a convex optimal stopping region. The concavity is due to the linearity of the function h (see Lemma 6.2 and its proof along with Lemma 6.1, (6.1), and Theorem 6.1).

If we had chosen to work with $(\Phi_t^1, \Phi_t^2)_{t\geq 0}$, then the relevant optimal stopping problem becomes

$$W(a,b) \triangleq \inf_{\tau\in\mathcal{S}} \mathbb{E}_0^{(a,b)} \left[\int_0^\tau e^{-\lambda t}\tilde{h}\left(\Phi_t^1, \Phi_t^2\right) dt\right] \tag{3.6}$$

in which

$$\tilde{h}(x,y) \triangleq x + y + xy, \quad (x,y) \in \mathbb{R}_+^2. \tag{3.7}$$

Since $h(\cdot,\cdot)$ is nonlinear the concavity of the value function, i.e., $W(\cdot,\cdot)$, is not concave. In fact, $W(x,y) = V(xy, x+y, x)$. The function V is concave, but W is not. So there is a trade-off between concavity and the dimension of the statistic to be used.

In what follows, for the sake of the simplicity of notation, when the meaning is clear, we will drop the superscripts of the expectation operators $\mathbb{E}_0^{(\phi^\times, \phi^+, \phi^1)}$.

4. Sample paths of $\Psi = (\Phi^\times, \Phi^+)$. It is illustrative to look at the sample paths of the sufficient statistic Ψ to understand the nature of the problem. Indeed, this way, for a certain parameter range, we will be able to identify the optimal stopping time without any further analysis.

From (2.26), we see that, if $a > 0$, then the paths of the processes $\Phi^\times$ and Φ^+ increase between the jumps, and otherwise the paths of the processes $\Phi^\times$ and Φ^+ mean-revert to the levels $2\lambda^2/a^2$ and $-2\lambda/a$, respectively. Also observe that $\Phi^\times$ and Φ^+ increase (decrease) with a jump if $\alpha \geq \beta$ ($\beta > \alpha$). See (2.20).

Case 1 ($\alpha \geq \beta$, $a > 0$). The following theorem follows from the description of the behavior of the paths above.

THEOREM 4.1. *If $\alpha > \beta$ and $a > 0$, then the stopping rule*

$$\tau_0 \triangleq \inf\{t \geq 0 : \Phi_t^\times + \Phi_t^+ \geq \lambda/c\} \tag{4.1}$$

is optimal for (3.3).

Proof. Under the hypothesis of the theorem and whenever a path of $(\Phi^\times, \Phi^+)$ leaves $\mathbb{C}_0$, it never returns. □

In section 5, we will identify another case (another range of parameters) in which the advantageous region $\mathbb{C}_0$ is the optimal continuation region and the stopping time τ_0 is optimal (see Cases 2B1a and 2B2a).

Remark 4.1. Let $\kappa_i \triangleq \inf\{t \geq 0 : \Phi_t^i \geq \lambda/c\}$. If $\alpha \geq \beta$ and $a > 0$, then κ_i is the optimal stopping time for the one-dimensional disorder problem with disorder time θ_i (see [3]). Let us define $\kappa \triangleq \kappa_1 \wedge \kappa_2$. Since with this choice of parameters $\Phi_\kappa^\times + \Phi_\kappa^+ > \lambda/c$, it follows that $\tau_0 < \kappa$ almost surely. Therefore, if we follow the rule dictated by the stopping time κ, then we pay an extra penalty for detection delay. This example illustrates that solving the two one-dimensional quickest detection problems separately in order to minimize the penalty function of (1.4) is suboptimal.

In what follows, we will consider the remaining cases: $\alpha \geq \beta$ and $a < 0$; $\alpha < \beta$.

5. Construction of a bound on the continuation region. In this section the purpose is to show that the continuation region of (3.3) is bounded. The construction of upper bounds is carried out in the next two theorems. These upper bounds are tight as the next theorem shows and might be used to construct useful approximations to the two optimal stopping times solving the problems defined in (3.3). We will carry out the analysis for $a = \lambda - \alpha + \beta \neq 0$. A similar analysis for this case can be similarly performed. As a result of Theorem 5.3 we are also able to conclude that the (smallest) optimal stopping time has a finite expectation.

The first two theorems in this section assume that an optimal stopping time of (3.3) exists and in particular the stopping time

$$\tau^*(a,b,c) \triangleq \inf\{t \geq 0 : V(\Upsilon_t) = 0, \qquad \Upsilon_0 = (a,b,c)\}, \tag{5.1}$$

is optimal. In section 7, we will verify that this assumption holds. We will denote by

$$\mathbf{\Gamma} \triangleq \{(a,b,c) \in \mathbb{B}_+^3 : v(a,b,c) = 0\}, \quad \mathbf{C} \triangleq \mathbb{B}_+^3 - \mathbf{\Gamma} \tag{5.2}$$

the optimal stopping region and optimal continuation region of (3.3), respectively.

THEOREM 5.1. *In this theorem the standing assumption is that $\alpha \geq \beta$ and that $a < 0$ (Case 2).*

Case 2A. Let us further assume that $\lambda/a^2 - 2/a \leq 1/c$ and denote

$$\mathbb{D}_0 \triangleq \left\{(x,y) \in \mathbb{B}_+^2 : x \cdot \left(\frac{1}{2(\lambda-a)}\right) + y \cdot \left(\frac{3\lambda - 2a}{2(\lambda-a)(2\lambda-a)}\right) + k > 0\right\} \tag{5.3}$$

in which

$$k \triangleq \frac{\lambda}{2a^2} - \frac{1}{a} - \frac{1}{2c} + \frac{\lambda^2}{2a^2(\lambda-a)} + \frac{1}{2\lambda - a} + 2\left(\frac{\lambda}{a} - \frac{\lambda^2}{a^2}\right). \tag{5.4}$$

Let $(\phi_0, \phi_1) \in \mathbb{D}_0 \cap (\mathbb{B}_+^2 - \mathbb{C}_0)$. Then for any $\phi_2 \leq \phi_1$, (ϕ_0, ϕ_1, ϕ_2) is in the stopping region of (3.3).

Case 2B. Assume that $\lambda/a^2 - 2/a \geq 1/c$ (standing assumption in the rest of the theorem). Consider the four different possible ranges of parameters.

Case 2B1. If $\lambda + a \leq 0$

- and if $-a/c - 1 \leq 0$ (*Case* 2B1a), then $\mathbb{C}$ in (3.5) is the optimal continuation region for (3.3);
- else if $-a/c - 1 > 0$ (*Case* 2B1b), then a superset of the continuation region can be constructed as follows. Let us introduce the line segment

$$C \triangleq \{(x,y) \in \mathbb{B}_+^2 : x + y = \lambda/c\} \tag{5.5}$$

and define

(5.6)
$$C_1 \triangleq \left\{(x,y) \in \mathbb{B}_+^2 : x = x(t,x^*), y = y(t,0) \text{ for } t \in [0,t^*]\right\} \bigcup \left\{(x,y) \in C : x < \frac{\lambda}{c} + \frac{2\lambda}{\lambda-a}\left(1 + \frac{a}{c}\right),\ y > -\frac{2\lambda}{\lambda-a}\left(1 + \frac{a}{c}\right)\right\}.$$

Here, t^* is the solution of $y(-t^*, -\frac{2\lambda}{\lambda-a}(1+\frac{a}{c})) = 0$, and $x^* = x(-t^*, \frac{\lambda}{c} + \frac{2\lambda}{\lambda-a}(1+\frac{a}{c}))$. The curve C_1 separates $\mathbb{R}_+^2$ into two connected regions. Let us denote the region that lies above the curve C_1 by

$$\mathbb{D}_1 \triangleq \{(x,y) \in \mathbb{B}_+^2 : \text{there exists a positive number } \tilde{y}(x) < y \text{ such that } (x, \tilde{y}(x)) \in C_1\}. \tag{5.7}$$

Then $[(\mathbb{B}_+^2 - \mathbb{D}_1) \times \mathbb{R}_+] \cap \mathbb{B}_+^3$ is an upper bound on the continuation region of (3.3).

Case 2B2. If $\lambda + a > 0$

- and if $-a/c - 1 < 0$ (*Case* 2B2a), then $\mathbb{C}$ in (3.5) is the optimal continuation region for (3.3);
- else if $-a/c - 1 > 0$, then $[(\mathbb{B}_+^2 - \mathbb{D}_1) \times \mathbb{R}_+] \cap \mathbb{B}_+^3$ is an upper bound on the continuation region of (3.3).

Note that all the supersets of the continuation we constructed are bounded subsets of $\mathbb{R}_+^3$.

Proof. Note that

$$\Phi_t^+ \geq y(t, \phi_1), \quad \Phi_t^\times \geq x(t, \phi_0), \quad t \geq 0, \tag{5.8}$$

almost surely if $\Phi_0^+ = \phi_1$ and $\Phi_0^\times = \phi_0$. This is because $\Phi^\times$ and Φ^+ increase with jumps.

From this observation we obtain the following inequality:

$$\inf_{\tau \in \mathcal{S}} \mathbb{E}_0 \left[\int_0^\tau e^{-2\lambda s} h(\Phi_s^\times, \Phi_s^+) ds \right] \geq \inf_{\tau \in \mathcal{S}} \mathbb{E}_0 \left[\int_0^\tau e^{-2\lambda s} h(x(s, \phi_0), y(s, \phi_1)) ds \right] \tag{5.9}$$

$$= \inf_{t \in [0, \infty]} \int_0^t e^{-2\lambda s} h(x(s, \phi_0), y(s, \phi_1)) ds. \tag{5.10}$$

Note that if for a given (ϕ_0, ϕ_1) the expression in (5.10) is equal to zero, then the infimum on the left-hand side of (5.9) is attained by setting $\tau = 0$. In what follows we will find a subset of the stopping region of the optimal stopping problem using this argument.

Case 2A ($\lambda/a^2 - 2/a \leq 1/c$). In this case the mean reversion level of the path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$, $(\phi_0, \phi_1) \in \mathbb{B}_+^2$, namely, $(\lambda^2/a^2, -2\lambda/a)$, is inside the region $\mathbb{C}_0$ which is defined in (3.4). In this case, for any $(\phi_0, \phi_1) \in \mathbb{B}_+^2 - \mathbb{C}_0$ the minimizer $t_{\text{opt}}(\phi_0, \phi_1)$ of the expression in (5.10) is either 0 or ∞ by the following argument: For any $(\phi_0, \phi_1) \in \mathbb{B}_+^2 - \mathbb{C}_0$ the path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ is in the advantageous region $\mathbb{C}_0$ all the time except for possibly a finite duration. Therefore if

$$\int_0^\infty e^{-2\lambda s} h(x(s, \phi_0), y(s, \phi_1)) ds < 0, \tag{5.11}$$

then in order to minimize (5.10) it is never optimal to stop. On the other hand if (5.11) is positive, then it is not worth taking the journey into the advantageous region, and it is optimal to stop immediately in order to minimize (5.10).

We shall find the pairs (ϕ_0, ϕ_1) for which $t_{\text{opt}} = 0$. Using (2.26) we can write

$$\int_0^\infty e^{-2\lambda s} h(x(s, \phi_0), y(s, \phi_1)) ds = \phi_0 \left(-\frac{1}{\alpha - \beta} \right) + \phi_1 \left(\frac{a}{(\alpha - \beta)^2} \right) + k, \tag{5.12}$$

where k is given by (5.4). Note that if $(\phi_0, \phi_1) \in \mathbb{D}_0 \cap (\mathbb{B}_+^2 - \mathbb{C}_0)$, then by (5.9) and (5.10) we can see that the infimum in (5.10) is equal to 0. Therefore $[((\mathbb{B}_+^2 - \mathbb{D}_0) \cup \mathbb{C}_0) \times \mathbb{R}_+] \cap \mathbb{B}_+^3$ is a superset of the optimal continuation region of (3.3).

Case 2B ($\lambda/a^2 - 2/a \geq 1/c$). In this case the mean reversion level of $t \to (x(t, \phi_0), y(t, \phi_1))$ is outside $\mathbb{C}_0$. Therefore, the minimizer of (5.10) is $t_{\text{opt}}(\phi_0, \phi_1) \in \{0, t_c(\phi_0, \phi_1), \infty\}$, where $t_c(\phi_0, \phi_1)$ is the exit time of the path $(x(t, \phi_0), y(t, \phi_1))$ from $\mathbb{C}_0$. The derivative

$$\frac{d}{dt}[x(t, \phi_0) + y(t, \phi_1)] = (\lambda + a) y(t, \phi_1) + 2a x(t, \phi_1) + 2\lambda \tag{5.13}$$

vanishes if $(x(t,\phi_0), y(t,\phi_1))$ meets the line segment

$$L = \{(x,y) \in \mathbb{B}_+^2 : (\lambda + a)y + 2ax + 2\lambda = 0\}. \tag{5.14}$$

Note that the mean reversion level belongs to L, i.e.,

$$\left(\frac{\lambda^2}{a^2}, -\frac{2\lambda}{a}\right) \in L. \tag{5.15}$$

Case 2B1 ($\lambda + a < 0$). (In addition to $\alpha > \beta$, $a < 0$ and $\lambda/a^2 - 2/a \geq 1/c$.) In this case the line L is decreasing (as a function of x).

Case 2B1a ($-a/c - 2 < 0$). (In addition to $\alpha > \beta$, $a < 0$, $\lambda/a^2 - 2/a \geq 1/c$, and $\lambda + a < 0$.) In this case the line segment C in (5.5) lies entirely below L. Assume that a path $(x(\cdot,\phi_0), y(\cdot,\phi_1))$ originating at $(\phi_0,\phi_1) \in \mathbb{B}_+^2 - \mathbb{C}_0$ enters $\mathbb{C}_0$ at time $t_0 > 0$. This path must leave $\mathbb{C}_0$ at time $t_1 < \infty$ since the mean reversion level $(\lambda^2/a^2, -\lambda/a) \notin \mathbb{C}_0$. This implies that, for any $t \in (t_0, t_1)$, $x(t,\phi_0) + y(t,\phi_1) < \lambda/c$ and $x(t_0,\phi_0) + y(t_0,\phi_1) = \lambda/c$. This yields a contradiction, because $\lambda + a < 0$ together with (5.13) implies that $t \to x(t,\phi_0) + y(t,\phi_1)$ is increasing below the line segment L. Therefore the minimizer $t_{\text{opt}}(\phi_0,\phi_1)$ of (5.10) is equal to 0 if $(\phi_0,\phi_1) \notin \mathbb{C}_0$, and it is equal to $t_c(\phi_0,\phi_1)$ if $(\phi_0,\phi_1) \in \mathbb{C}_0$. From (5.9) we can conclude that $\mathbb{C}$ is equal to the optimal continuation region of (3.3).

Case 2B1b ($-a/c - 1 > 0$). In this case the line segments C and L intersect at $I = (x^I, y^I) \triangleq (\frac{\lambda}{c} + \frac{2\lambda}{\lambda - a}(1 + \frac{a}{c}), -\frac{2\lambda}{\lambda - a}(1 + \frac{a}{c}))$. By running the paths backward in time, we can find x^* such that

$$(x^*, 0) = \left(x\left(-t^*, x^I\right), y\left(-t^*, y^I\right)\right). \tag{5.16}$$

By the semigroup property (2.27), we have

$$\begin{aligned} x(t^*, x^*) &= x\left(t^*, x\left(-t^*, x^I\right)\right) = x\left(t^* + (-t^*), x^I\right) \\ &= x\left(0, x^I\right) = x^I. \end{aligned} \tag{5.17}$$

Similarly, $y(t^*, 0) = y^I$. The function $t \to x(t,x^*) + y(t,0)$ is decreasing on $(0,t^*)$ and increasing on (t^*,∞). It follows that the path $t \to (x(t,x^*), y(t,0))$ is tangential to C at I and lies above the region $\mathbb{C}_0$.

We will now show that if a path $(x(\cdot,\phi_0), y(\cdot,\phi_1))$ originates in $\mathbb{D}_1$, then it stays in $\mathbb{D}_1$. Let us first consider a pair $(\phi_0,\phi_1) \in \mathbb{D}_1$ such that $\phi_1 < -2\lambda/a$. Consider the curve

$$P \triangleq \left\{(x,y) \in \mathbb{B}_+^2 : x = x(t,x^*), y = y(t,0) \text{ for } t \in [0,\infty)\right\}. \tag{5.18}$$

The following remark will be useful in completing the proof.

Remark 5.1. The semigroup property in (2.27) implies that two distinct curves $(x(\cdot,\phi_0^a), y(\cdot,\phi_1^a))$ and $(x(\cdot,\phi_0^b), y(\cdot,\phi_1^b))$ do not intersect. If

$$(x(t^a,\phi_0^a), y(t^a,\phi_1^a)) = (x(t^b,\phi_0^b), y(t^b,\phi_1^b)) = (\phi_0,\phi_1) \tag{5.19}$$

for some $t^a, t^b \in \mathbb{R}$, then (2.27) implies that

$$\begin{aligned} (x(t,\phi_0^a), y(t,\phi_1^a)) &= (x(t^a + (t - t^a), \phi_0^a), y(t^a + (t - t^a), \phi_1^a)) \\ &= (x(t - t^a, \phi_0), y(t - t^a, \phi_1)) \\ &= (x(t^b + (t - t^a), \phi_0^b), y(t^b + (t - t^a), \phi_1^b)) \\ &= (x(t^b - t^a + t, \phi_0^b), y(t^b - t^a + t, \phi_1^b)) \quad \text{for all} \quad t \in \mathbb{R}, \end{aligned} \tag{5.20}$$

i.e., the two curves are identical after a reparametrization.

If the point (ϕ_0, ϕ_1) lies above P and we recall that P lies above $\mathbb{C}_0$, then by Remark 5.1 the path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ will lie above $\mathbb{C}_0$. If the point (ϕ_0, ϕ_1) lies between P and $\mathbb{C}_0$, then the path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ will lie below the line segment L. This observation together with the fact that $\lambda + a < 0$ implies (using (5.13)) that the function $t \to x(t, \phi_0) + y(t, \phi_1)$ is increasing. Therefore the path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ cannot intersect $\mathbb{C}_0$.

Now let us consider a pair $(\phi_0, \phi_1) \in \mathbb{D}_1$ such that $\phi_1 > -2\lambda/a$. If (ϕ_0, ϕ_1) lies above L, then the function $t \to x(t, \phi_0) + y(t, \phi_1)$ is decreasing and its range is $[\phi_0 + \phi_1, 2\lambda/a(\lambda/a - 1))$, which is always above λ/c, and therefore the path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ does not enter $\mathbb{C}_0$. If the point (ϕ_0, ϕ_1) lies below L, then $t \to x(t, \phi_0) + y(t, \phi_1)$ is increasing. This monotonicity implies that the path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ cannot visit $\mathbb{C}_0$. If $\phi_1 = -2\lambda/a$, then $y(t, \phi_1) = -2\lambda/a$ for all $t \geq 0$. $x(t, \phi_0)$ increases or decreases depending on whether $(\phi_0, -2\lambda/a)$ is below or above L. Therefore if $(\phi_0, -2\lambda/a) \notin \mathbb{C}_0$, then $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ never visits $\mathbb{C}_0$. These arguments show that if a path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ originates in $\mathbb{D}_1$, then it stays in $\mathbb{D}_1$. Therefore if $(\phi_0, \phi_1) \in \mathbb{D}_1$, then the infimum in (5.10) is equal to 0 (by (5.9) and (5.10)). Therefore $[(\mathbb{B}^2_+ - \mathbb{D}_1) \times \mathbb{R}_+] \cap \mathbb{B}^3_+$ is a superset of the optimal continuation region of (3.3).

Case 2B2 ($\lambda + a > 0$). In this case L is increasing (as a function of x). The function $t \to x(t, \phi_0) + y(t, \phi_1)$ is increasing if (ϕ_0, ϕ_1) lies above L, and it is decreasing otherwise.

Case 2B2a ($-a/c - 1 \leq 0$). In this case the line segments L and C do not intersect. Let us first consider a pair $(\phi_0, \phi_1) \in \mathbb{B}^2_+$ such that $\phi_1 < -2\lambda/a$. If $(\phi_0, \phi_1) \notin \mathbb{C}_0$ lies above the line segment L, then $t \to x(t, \phi_0) + y(t, \phi_1)$ is increasing and the path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ cannot enter $\mathbb{C}_0$. Consider the curve

$$\tilde{P} \triangleq \left\{ (x, y) \in \mathbb{B}^2_+ : x = x\left(t, -\frac{\lambda}{a}\right), y = y(t, 0) \text{ for } t \in [0, \infty) \right\}, \tag{5.21}$$

which starts at the intersection of L with the x-axis. The semigroup property Remark 5.1 implies that no path starting to the right of $\tilde{P}$ intersects $\tilde{P}$ and therefore lies to the right of the region $\mathbb{C}_0$. Therefore, if (ϕ_0, ϕ_1) is below the line segment L, then the path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ never visits the advantageous region $\mathbb{C}_0$. (Note that if the path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ meets the line L at time $t_L(\phi_0, \phi_1)$, then $t \to (x(t, \phi_0) + y(t, \phi_1))$ is decreasing (increasing) on $[0, t_L]$ ($[t_L, \infty)$).

Now let us consider a point $(\phi_0, \phi_1) \in \mathbb{B}^2_+ - \mathbb{C}_0$ such that $\phi_1 > -2\lambda/a$. Then $t \to (x(t, \phi_0) + y(t, \phi_1))$ is increasing on $[0, t_L(\phi_0, \phi_1)]$ and is decreasing on $(t_L(\phi_0, \phi_1), \infty)$ (it decreases to $-2\lambda/a + \lambda^2/a^2 > \lambda/c$). And the monotonicity of $t \to x(t, \phi_0) + y(t, \phi_1)$ on $[0, t_L(\phi_0, \phi_1)]$ implies that $x(t, \phi_0) + y(t, \phi_1) > \lambda/c$ for $t \in [0, t_L(\phi_0, \phi_1)]$. If $\phi_1 = -2\lambda/a$, then $y(t, \phi_1) = -2\lambda/a$ for all $t \geq 0$. $x(t, \phi_0)$ increases (decreases) depending on whether $(\phi_0, -2\lambda/a)$ is above or below L. These arguments show that if a path $(x(\cdot, \phi_0), y(\cdot, \phi_1))$ originates in $\mathbb{B}^2_+ - \mathbb{C}_0$, then it stays in $\mathbb{B}^2_+ - \mathbb{C}_0$. Therefore the minimizer $t_{\text{opt}}(\phi_0, \phi_1)$ of (5.10) for any $\phi_0, \phi_1 \in \mathbb{B}^2_+ - \mathbb{C}_0$ is equal to zero. Now using (5.9) and (5.10) the optimal continuation region of (3.3) is equal to $\mathbb{C}$.

Case 2B2b ($-a/c - 1 > 0$). In this case the line segments C and L intersect at $I = (x^I, y^I)$. Arguments similar to those of Case 2B1b show that $[(\mathbb{B}^2_+ - \mathbb{D}_1) \times \mathbb{R}_+] \cap \mathbb{B}^3_+$, in which $\mathbb{D}_1$ is defined in (5.7), is a superset of the optimal continuation region of (3.3). □

THEOREM 5.2. *Let us assume that* $\alpha < \beta$ *(*Case 3, $\alpha < \beta$*) and define*

$$\mathbb{D}_2 \triangleq \left\{ (x, y) \in \mathbb{B}^2_+ : x + y \geq \frac{\lambda + 2\beta}{c} \right\}. \tag{5.22}$$

Then $[(\mathbb{B}^2_+ - \mathbb{D}_1) \times \mathbb{R}_+] \cap \mathbb{B}^3_+$, *which is a bounded region in* $\mathbb{R}^3_+$, *is an upper bound on the continuation region of* (3.3).

Proof. Note that in this case $a > 0$. The paths of the processes $\Phi^\times$ and Φ^+ increase between the jumps and decrease with a jump. If $\tau \in \mathcal{S}$, then there is a constant $t \geq 0$ such that $\tau \wedge \sigma_1 = t \wedge \sigma_1$ almost surely. Hence we can write

(5.23)
$$\begin{aligned}
\mathbb{E}_0\left[\int_0^\tau e^{-\lambda s} h(\Psi_s) ds\right] &= \mathbb{E}_0\left[\int_0^{\tau\wedge\sigma_1} e^{-\lambda s} h(\Psi_s) ds\right] + \mathbb{E}_0\left[1_{\{\tau \geq \sigma_1\}} \int_{\sigma_1}^\tau e^{-\lambda s} h(\Psi_s) ds\right] \\
&= \mathbb{E}_0\left[\int_0^{t\wedge\sigma_1} e^{-\lambda s} h(\Psi_s) ds\right] + \mathbb{E}_0\left[1_{\{t \geq \sigma_1\}} \int_{\sigma_1}^\tau e^{-\lambda s} h(\Psi_s) ds\right] \\
&\geq \mathbb{E}_0\left[\int_0^{t\wedge\sigma_1} e^{-\lambda s} h(\Psi_s) ds\right] - \frac{1}{c}\mathbb{E}_0\left[1_{\{t \geq \sigma_1\}} e^{-\lambda\sigma_1}\right] \\
&= \int_0^t e^{-(\lambda+2\beta)s}\left[h(x(s,\phi_0), y(s,\phi_1)) - \frac{2\beta}{c}\right] ds
\end{aligned}$$

using also the fact that σ_1 has exponential distribution with rate 2β. From (5.23) it follows that if $x(s,\phi_0) + y(s,\phi_1) - (\lambda + 2\beta)/c > 0$, then $\mathbb{E}_0\left[\int_0^\tau e^{-\lambda s} h(\Psi_s) ds\right] > 0$ for every stopping time $\tau \neq 0$, $\tau \in \mathcal{S}$. Since the paths $x(t,\phi_0)$ and $y(t,\phi_1)$ are increasing, we can conclude that stopping immediately is optimal for (3.3). That is, $\tau = 0$ is optimal for (3.3) if $(\phi_0, \phi_1) \in \mathbb{D}_2$ and $\phi_2 \leq \phi_1$, in which $\mathbb{D}_2$ is as in (5.22). □

Theorems 5.1 and 5.2 can be used to determine approximate detection rules besides helping us to determine the location and the shape of the continuation region. As we have seen in Cases 2B1a and 2B2a, these approximate rules turn out to be tight. The next theorem is essential in proving the fact that the smallest optimal stopping time of (3.3) has a finite expectation.

THEOREM 5.3. *Let* τ_D *be the exit time of the process* Υ *from a bounded region* $D \subset \mathbb{B}^3_+$. *Then* $E_0^{\phi^\times,\phi^+,\phi^1}[\tau_D] < \infty$ *for every* $(\phi^\times, \phi^+, \phi^1) \in \mathbb{B}^3_+$. *Hence* τ^* *defined in* (5.1) *has a finite expectation.*

Proof. Let $f(\phi^\times, \phi^+, \phi^1) \triangleq \phi^\times + \phi^+$. Then it follows from (2.22) that

(5.24)
$$\begin{aligned}
[\mathcal{A}f](\phi^\times, \phi^+, \phi^1) &= \lambda\phi^+ + 2a\phi^\times + 2\lambda + a\phi^+ \\
&\quad + \beta\left[\frac{\alpha}{\beta}\phi^+ + \phi^\times + \left(\frac{\alpha}{\beta} - 1\right)\phi^1 - \phi^\times - \phi^+\right] \\
&\quad + \beta\left[\frac{\alpha}{\beta}\phi^+ + \frac{\alpha}{\beta}\phi^\times - \left(\frac{\alpha}{\beta} - 1\right)\phi^1 - \phi^\times - \phi^+\right] = 2\lambda(\phi^\times + \phi^+ + 1) \geq 2\lambda
\end{aligned}$$

for every $(\phi^\times, \phi^+, \phi^1) \in \mathbb{B}^3_+$. Since f is bounded on D and $\tau_D \wedge t$ is a bounded $\mathbb{F}$-stopping time, we have

$$\mathbb{E}_0\left[f(\Upsilon_{\tau_D\wedge t})\right] = f(\Upsilon_0) + \mathbb{E}_0\left[\int_0^{\tau_D\wedge t} [Af](\Upsilon_s) ds\right] \geq 2\lambda E_0[\tau_D \wedge t]. \tag{5.25}$$

On the other hand

$$\mathbb{E}_0\left[f(\Upsilon_{\tau_D\wedge t})\right] \leq \frac{\alpha}{\beta}\xi$$

in which $\xi = \min\{a \in \mathbb{R}_+ : \text{for any } (x,y,z) \in D, \max(x,y,z) \leq a\} < \infty$. An application of the monotone convergence theorem implies that $E_0[\tau_D] < \infty$. □

The results of this section can be used to determine approximate detection rules besides helping us to determine the location and the shape of the continuation region. As we have seen in Cases 2B1a and 2B2a, these approximate rules turn out to be tight.

6. Optimal stopping with time horizon σ_n. In this section, we will first approximate the optimal stopping problem (3.3) by a sequence of optimal stopping problems. Let us denote

$$V_n(a,b,c) \triangleq \inf_{\tau\in\mathcal{S}} \mathbb{E}_0^{a,b,c}\left[\int_0^{\tau\wedge\sigma_n} e^{-\lambda t} h\left(\Phi_t^{\times}, \Phi_t^{+}\right) dt\right] \tag{6.1}$$

for all $(a,b,c)\in\mathbb{B}_+^3$ and $n\in\mathbb{N}$. Here, σ_n is the nth jump time of the process X.

Observe that $(V_n)_{n\in\mathbb{N}}$ is a decreasing sequence and that each of its members satisfies $-1/c < V_n < 0$. Therefore the pointwise limit $\lim_n V_n$ exists. It can be shown that more is true using the fact that the function h is bounded from below and σ_n is a sum of independent exponential random variables.

LEMMA 6.1. *For any $(a,b,c)\in\mathbb{B}_+^3$,*

$$0 \le V_n(a,b,c) - V(a,b,c) \le \frac{1}{c}\left(\frac{2\beta}{2\beta+\lambda}\right)^n. \tag{6.2}$$

Proof. For any $\tau\in\mathcal{S}$,

$$\begin{aligned}\mathbb{E}_0\left[\int_0^{\tau} e^{-\lambda t} h\left(\Phi_t^{\times}, \Phi_t^{+}\right) dt\right] &= \mathbb{E}_0\left[\int_0^{\tau\wedge\sigma_n} e^{-\lambda t} h\left(\Phi_t^{\times}, \Phi_t^{+}\right) dt\right] \\ &\quad + \mathbb{E}_0\left[1_{\{\tau\ge\sigma_n\}}\int_{\sigma_n}^{\tau} e^{-\lambda t} h(\Phi_t^{\times}, \Phi_t^{+}) dt\right].\end{aligned} \tag{6.3}$$

The first term on the right-hand side of (6.3) is greater than V_n. Since $h(\cdot,\cdot) > -\lambda/c$ we can show that the second term is greater than

$$-\frac{\lambda}{c}\mathbb{E}_0^{\phi_0,\phi_1}\left[1_{\{\tau\ge\sigma_n\}}\int_{\sigma_n}^{\tau} e^{-\lambda s} ds\right] \ge -\frac{1}{c}\mathbb{E}_0^{\phi_0,\phi_1}\left[e^{-\lambda\sigma_n}\right] \ge -\frac{1}{c}\left(\frac{2\beta}{\lambda+2\beta}\right)^n. \tag{6.4}$$

To show the last inequality we have used the fact that σ_n is a sum of n independent and identically distributed exponential random variables with rate 2β. Now, the proof of the lemma follows immediately. □

As in [4] and [6], to calculate the value functions V_n iteratively we introduce the functional operators J, J_t. These operators are defined through their actions on bounded functions $g:\mathbb{B}_+^3\to\mathbb{R}$ as follows:

(6.5)

$$[Jg](t,a,b,c) \triangleq \mathbb{E}_0^{a,b,c}\left[\int_0^{t\wedge\sigma_1} e^{-\lambda s} h(\Phi_s^{\times}, \Phi_s^{+}) ds + 1_{\{t\ge\sigma_1\}} e^{-\lambda\sigma_1} g(\Phi_{\sigma_1}^{\times}, \Phi_{\sigma_1}^{+}, \Phi_{\sigma_1}^{1})\right], \text{ and}$$

$$[J_t g](a,b,c) \triangleq \inf_{s\in[t,\infty]} [Jg](s,a,b,c), \quad t\in[0,\infty].$$

Observe that

(6.6)
$$\begin{aligned}
&\mathbb{E}_0\left[1_{\{t\geq\sigma_1\}}e^{-\lambda\sigma_1}g(\Phi^{\times}_{\sigma_1},\Phi^{+}_{\sigma_1},\Phi^{1}_{\sigma_1})\right]\\
&=\mathbb{E}_0\Bigg[\left(g\left(\frac{\alpha}{\beta}\Phi^{\times}_{\sigma_1-},\Phi^{+}_{\sigma_1-}+\left(\frac{\alpha}{\beta}-1\right)\Phi^{1}_{\sigma_1-},\frac{\alpha}{\beta}\Phi^{1}_{\sigma_1-}\right)1_{\{X^1_{\sigma_1}\neq X^1_{\sigma_1-}\}}\\
&\quad+g\left(\frac{\alpha}{\beta}\Phi^{\times}_{\sigma_1-},\frac{\alpha}{\beta}\Phi^{+}_{\sigma_1-}-\left(\frac{\alpha}{\beta}-1\right)\Phi^{1}_{\sigma_1-},\Phi^{1}_{\sigma_1-}\right)1_{\{X^2_{\sigma_1}\neq X^2_{\sigma_1-}\}}\Bigg)1_{\{t\geq\sigma_1\}}e^{-\lambda\sigma_1}\Bigg]\\
&=\frac{1}{2}\int_0^t 2\beta e^{-(\lambda+2\beta)s}g\left(\frac{\alpha}{\beta}\,x(s,a),y(s,b)+\left(\frac{\alpha}{\beta}-1\right)z(s,c),\frac{\alpha}{\beta}\,z(s,c)\right)ds\\
&\quad+\frac{1}{2}\int_0^t 2\beta e^{-(\lambda+2\beta)s}g\left(\frac{\alpha}{\beta}\,x(s,a),\frac{\alpha}{\beta}\,y(s,b)-\left(\frac{\alpha}{\beta}-1\right)z(s,c),z(s,c)\right)ds.
\end{aligned}$$

To derive (6.6) we used the fact that σ_1 has exponential distribution with rate 2β; the dynamics in (2.20); and the fact that conditioned on the event in which there is a jump, it has $1/2$ probability of coming from X^1.

Using (6.6) and Fubini's theorem we can write

$$[Jg](t,a,b,c)=\int_0^t e^{-(\lambda+2\beta)s}(h+\beta\cdot g\circ(F_1+F_2))(x(s,a),y(s,b),z(s,c))ds, \tag{6.7}$$

where

$$F_i(a,b,c)=\left(\frac{\alpha}{\beta}\,a,\left(\frac{\alpha}{\beta}\right)^{i-1}b+(-1)^i\left(\frac{\alpha}{\beta}-1\right)c,\left(\frac{\alpha}{\beta}\right)^{2-i}c\right),\quad i\in\{1,2\}. \tag{6.8}$$

Using (2.26) it can be shown that

$$\lim_{t\to\infty}[Jg](t,a,b,c)=[Jg](\infty,a,b,c)<\infty. \tag{6.9}$$

LEMMA 6.2. *For every bounded function f, the mapping J_0f is bounded. If f is a concave function, then J_0f is also a concave function. If $f_1\leq f_2$, then $J_0f_1\leq J_0f_2$.*

Proof. The third assertion of the lemma directly follows from the representation (6.7). The first assertion holds since h is bounded from below and $J_0f(a,b,)\leq Jf(0,a,b,c)=0$. The second assertion follows from the linearity of the functions $x(t,\cdot)$, $y(t,\cdot)$, $h(\cdot,\cdot)$, $F_1(\cdot,\cdot,\cdot)$, and $F_2(\cdot,\cdot,\cdot)$. □

Using Lemma 6.2 we can prove the following corollary.

COROLLARY 6.1. *Let us define the sequences of function $(v_n)_{n\in\mathbb{N}}$ by*

$$v_0\triangleq 0,\quad v_n\triangleq J_0v_{n-1}. \tag{6.10}$$

Then every $n\in\mathbb{N}$, v_n is bounded and concave, and $v_{n+1}\leq v_n$. Therefore, the limit $v=\lim_n v_n$ exists and is bounded and concave. Moreover, v_n for all $n\in\mathbb{N}$ and v are increasing in each of their arguments.

Proof. The proof of the first part directly follows from Lemma 6.2. That v_n for all $n\in\mathbb{N}$ and v are increasing in each of their arguments follows from the fact that these functions are bounded from above and below and that they are concave. □

We will need the following lemma to give a characterization of the stopping times of the filtration $\mathbb{F}$ (see [5]).

LEMMA 6.3. *For every $\tau\in\mathcal{S}$, there are $\mathcal{F}_{\sigma_n}$-measurable random variables $\xi_n:\Omega\to\infty$ such that $\tau\wedge\sigma_{n+1}=(\sigma_n+\xi_n)\wedge\sigma_{n+1}$ $\mathbb{P}_0$ almost surely on $\{\tau\geq\sigma_n\}$.*

The main theorem of this section can be proven by induction using Lemma 6.3 and the strong Markov property.

THEOREM 6.1. *For every $n \in \mathbb{N}$, v_n defined in Corollary* 6.1 *is equal to V_n. For $\varepsilon \geq 0$, let us denote*

$$r_n^\varepsilon(a,b,c) \triangleq \inf\{t \in (0,\infty] : [Jv_n](t,(a,b,c)) \leq [J_0 v_n](a,b,c) + \varepsilon\}. \tag{6.11}$$

And let us define a sequence of stopping times by $S_1^\varepsilon \triangleq r_0^\varepsilon(\Upsilon_0) \wedge \sigma_1$ and

$$S_{n+1}^\varepsilon \triangleq \begin{cases} r_n^{\varepsilon/2}(\Upsilon_0) & \text{if } \sigma_1 \geq r_n^{\varepsilon/2}(\Upsilon_0), \\ \sigma_1 + S_n^{\varepsilon/2} \circ \theta_{\sigma_1} & \text{otherwise.} \end{cases} \tag{6.12}$$

Here θ_s is the shift operator on Ω, i.e., $X_t \circ \theta_s = X_{s+t}$. Then S_n^ε is an ε-optimal stopping time of (6.1)*, i.e.,*

$$\mathbb{E}_0^{a,b,c}\left[\int_0^{S_n^\varepsilon} e^{-\lambda t} h(\Psi_t)dt\right] \leq v_n(a,b,c) + \varepsilon \tag{6.13}$$

in which $\Psi_t = (\Phi_t^\times, \Phi_t^+)$, $t \geq 0$.

Proof. See the appendix. □

Theorem 6.1 shows that the value function V_n of the optimal stopping problem defined in (6.1) and the function v_n introduced in Corollary 6.1 by an iterative application of the operator J_0 are equal. This implies that the value function of the optimal stopping problem of (6.1) can be found by solving a sequence of deterministic minimization problems.

7. Optimal stopping time.

THEOREM 7.1. *τ^* defined in* (5.1) *is the smallest optimal stopping time for* (3.3).

This theorem shows that Γ defined in (5.2) is indeed an optimal stopping region. We will divide the proof of this theorem into several lemmas.

The following dynamic programming principle can be proven by the special representation of the stopping times of a jump process (Lemma 6.3) and the strong Markov property.

LEMMA 7.1. *For any bounded function $g : \mathbb{B}_+^3 \to \mathbb{R}$ we have*

$$[J_t g](a,b,c) = [Jg](t,a,b,c) + e^{-(\lambda+2\beta)t}[J_0 g](x(t,a), y(t,b), z(t,c)). \tag{7.1}$$

Let us denote

$$r_n(a,b,c) \triangleq r_n^0(a,b,c), \tag{7.2}$$

which is well defined because of (6.9) and the continuity of the function $t \to [Jf](t,a,b,c)$, $t \geq 0$. (See (6.11) for the definition of r_n^0.) Let us also denote

$$r(a,b,c) \triangleq \inf\{t \geq 0 : [JV](t,a,b,c) = J_0 V(a,b,c)\}. \tag{7.3}$$

COROLLARY 7.1. *The functions r_n and r defined by* (7.2) *and* (7.3)*, respectively, satisfy*

$$\begin{aligned} r_n(a,b,c) &= \{t \geq 0 : v_{n+1}(x(t,a), y(t,b), z(t,c)) = 0\}, \\ r(a,b,c) &= \{t \geq 0 : v(x(t,a), y(t,b), z(t,c)) = 0\}, \end{aligned} \tag{7.4}$$

with the convention that $\inf \emptyset = 0$. *Together with* (7.4), *Corollary* 6.1 *implies that* $r_n(a,b,c) \uparrow r(a,b,c)$ *as* $n \uparrow \infty$.

Proof. Suppose that $r_n(a,b,c) < \infty$. Then from (6.11) it follows that

$$\begin{aligned}
&[Jv_n](r_n(a,b,c),a,b,c) = [J_0 v_n](a,b,c) = [J_{r_n(a,b,c)} v_n](a,b,c) \\
&= [Jv_n](r_n(a,b,c),a,b,c) \\
&\quad + e^{-(\lambda+2\beta)r_n(a,b,c)} v_{n+1}(x(r_n(a,b,c),a), y(r_n(a,b,c),b), z(r_n(a,b,c),c)).
\end{aligned} \tag{7.5}$$

Here the second equality follows from the definition of the operator J_t in (6.5), and the third equation follows from Lemma 7.1 and the fact that $J_0 v_n = v_{n+1}$. From (7.5) it follows that $v_{n+1}(x(r_n(a,b,c),a),\ y(r_n(a,b,c),b), z(r_n(a,b,c),c)) = 0$. For $t \in (0, r_n(a,b,c))$ we have $[Jv_n](t,a,b,c) > [J_0 v_n](a,b,c) = [J_{r_n(a,b,c)} v_n](a,b,c) = J_t v_n(a,b,c)$, since the function $s \to [J_s v_n](a,b,c)$ is nondecreasing. Using Lemma 7.1 we can write

$$[J_0 v_n](a,b,c) = [J_t v_n](a,b,c) = [Jv_n](t,a,b,c) + e^{-(\lambda+2\beta)} v_{n+1}(x(t,a), y(t,b), z(t,c)), \tag{7.6}$$

which implies that $v_{n+1}(x(t,a), y(t,b), z(t,c)) < 0$ for all $t < r_n(a,b,c)$.

Now, suppose that $r_n(a,b,c) = \infty$. Then $v_{n+1}(x(t,a), y(t,b), z(t,c)) < 0$ for every $t \in (0,\infty)$ which can be shown using the same arguments as above. Therefore $\{t > 0 : v_{n+1}(x(t,a), y(t,b), z(t,c)) = 0\} = \emptyset$, and (7.4) holds.

The proof for the representation of r can be proven using the same line of argument and the fact that $J_0 V = V$. The fact that $J_0 V = V$ can be proven by the dominated convergence theorem, since the sequences $(v_n(a,b,c))_{n\geq 0}$ and $([Jv_n](t,a,b,c))_{n\geq 0}$ are decreasing and since v_n is a bounded function for all $n \in \mathbb{N}$. □

In the next lemma we construct optimal stopping times for the family of problems introduced in (6.1).

LEMMA 7.2. *Let us denote* $S_n \triangleq S_n^0$, *where* S_n^ε *is defined in Theorem* 6.1 *for* $\varepsilon \geq 0$. *Then the sequence* $(S_n)_{n\in\mathbb{N}}$ *is an almost surely increasing sequence. Moreover* $S_n < \tau^*$ *almost surely for all* n.

Proof. Since $r_1 > 0$, using Corollary 7.1 we can write

$$S_2 - S_1 = \left\{ \begin{array}{ll} r_1 - r_0 & \text{if } \sigma_1 > r_1 \\ \sigma_1 - r_0 + S_1 \circ \theta_{\sigma_1} & \text{if } r_0 < \sigma_1 \leq r_1 \\ S_1 \circ \theta_{\sigma_1} & \text{if } \sigma_1 \leq r_0 \end{array} \right\} > 0. \tag{7.7}$$

Now, let us assume that $S_n - S_{n-1} > 0$ almost surely. From Lemma 7.1 we have that $r_n > r_{n-1}$. Using this fact and the induction hypothesis we can write

$$S_{n+1} - S_n = \left\{ \begin{array}{ll} r_n - r_{n-1} & \text{if } \sigma_1 > r_n \\ \sigma_1 - r_{n-1} + S_n \circ \theta_{\sigma_1} & \text{if } r_{n-1} < \sigma_1 \leq r_n \\ (S_n - S_{n-1}) \circ \theta_{\sigma_1} & \text{if } \sigma_1 \leq r_{n-1} \end{array} \right\} > 0, \tag{7.8}$$

which proves the first assertion of the lemma.

From Corollary 7.1 and the definition of τ^* it follows that $\tau^* \wedge \sigma_1 = r \wedge \sigma_1$. Therefore $\tau^* \wedge \sigma_1 > r_0 \wedge \sigma_1 = S_1$ since $r_0 < r$. Now, we will assume that $S_n < \tau^*$ and show that $S_{n+1} < \tau^*$. On $\{\sigma_1 \leq r_n\}$ we have that

$$S_{n+1} = \sigma_1 + S_n \circ \theta_{\sigma_1} < \sigma_1 + \tau^* \circ \theta_{\sigma_1}. \tag{7.9}$$

Since $\tau^* \wedge \sigma_1 = r \wedge \sigma_1$ and $r > r_n$, if $\sigma_1 \leq r_n$, then $\tau^* \wedge \sigma_1 = \sigma_1$. Because τ^* is a hitting time, on the set $\{\sigma_1 \leq r_n\} \subset \{\sigma_1 \leq \tau^*\}$ the following holds:

$$S_{n+1} \leq \sigma_1 + \tau^* \circ \theta_{\sigma_1} = \tau^*.$$

On the other hand if $\sigma_1 > r_n$, then $\tau^* \wedge \sigma_1 = r \wedge \sigma_1 > r_n$. Therefore on $\{\sigma_1 > r_n\}$, $S_{n+1} = r_n < \tau^*$. This concludes the proof of the second assertion. □

LEMMA 7.3. *Let us denote* $\Psi_t = (\Phi_t^{\times}, \Phi_t^{+})$, $t \geq 0$. *If* $S^* \triangleq \lim_n S_n$, *then* $S^* = \tau^*$ *almost surely. Moreover,* τ^* *is an optimal stopping time, i.e.,*

$$V(a,b,c) = \mathbb{E}_0^{a,b,c}\left[\int_0^{S^*} e^{-\lambda s} h(\Psi_s) ds\right].$$

Proof. The limit $S^* \triangleq \lim_n S_n$ exists since $(S_n)_{n\in\mathbb{N}}$ is increasing and $S_n \leq \tau^* < \infty$ (as a corollary of Theorem 5.3) for all n. Let us show that S^* is optimal.

$$\begin{aligned} \mathbb{E}_0\left[\lim_n \int_0^{S_n} e^{-\lambda t} h(\Psi_t) dt\right] &\leq \liminf_n \mathbb{E}_0\left[\int_0^{S_n} e^{-\lambda t} h(\Psi_t) dt\right] \\ &= \lim_n V_n(a,b,c) = V(a,b,c). \end{aligned} \tag{7.10}$$

The first inequality follows from Fatou's lemma, which we can apply since

$$\int_0^{S_n} e^{-\lambda t} h(\Psi_t) dt \geq \int_0^{\infty} e^{-\lambda t} h(\Psi_t) dt \geq -\frac{\sqrt{2}}{c} \quad \text{almost surely.}$$

The first equality in (7.10) follows from Theorem 6.1. Now it can be seen from (7.10) that S^* is an optimal stopping time. Taking the limit of (6.12) as $n \to \infty$ and using Corollary 7.1, we conclude that $\tau^* = S^*$. □

Proof of Theorem 7.1. The proof of the optimality of τ^* follows directly from Lemma 7.3. We will show that τ^* is the smallest optimal stopping time.

Given any $\mathbb{F}$-stopping time $\tau < \tau^*$, let us define

$$\tilde{\tau} \triangleq \tau + \tau^* \circ \theta_\tau. \tag{7.11}$$

Then the stopping time $\tilde{\tau}$ satisfies

(7.12)

$$\begin{aligned} \mathbb{E}_0^{a,b,c}\left[\int_0^{\tilde{\tau}} e^{-\lambda s} h(\Psi_s) ds\right] &= \mathbb{E}_0^{a,b,c}\left[\int_0^{\tau} e^{-\lambda s} h(\Psi_s) ds + \int_\tau^{\tilde{\tau}} e^{-\lambda s} h(\Psi_s) ds\right] \\ &= \mathbb{E}_0^{a,b,c}\left[\int_0^{\tau} e^{-\lambda s} h(\Psi_s) ds + e^{-\lambda\tau} \int_0^{\tau^*\circ\theta_\tau} e^{-\lambda s} h(\Psi_{s+\tau}) ds\right] \\ &= \mathbb{E}_0^{a,b,c}\left[\int_0^{\tau} e^{-\lambda s} h(\Psi_s) ds + e^{-\lambda\tau} V(\Upsilon_\tau)\right] \\ &< \mathbb{E}_0^{a,b,c}\left[\int_0^{\tau} e^{-\lambda s} h(\Psi_s) ds\right]. \end{aligned}$$

Here the third equality follows from the strong Markov property of the process Υ, and the inequality follows since $V(\Upsilon_\tau) < 0$. Equation (7.12) shows that any optimal stopping time $\tau < \tau^*$ cannot be optimal. □

8. Structure of the continuation and stopping regions. Let us recall (5.2) and denote

$$\mathbf{\Gamma_n} \triangleq \{(a,b,c) \in \mathbb{B}_+^3 : v_n(a,b,c) = 0\}, \quad \mathbf{C}_n \triangleq \mathbb{B}_+^3 - \mathbf{\Gamma_n}. \tag{8.1}$$

We have shown in Theorem 7.1 that the Γ of (5.2) is the optimal stopping region for (3.3) and the first hitting time τ^* of Υ to this set is optimal. On the other hand although Γ_n is an optimal stopping region for (6.1), the description of the optimal stopping times S_n^0 (see (6.12)) is more involved. These optimal stopping times are not hitting times of the sets Γ_n. S_n^0 prescribes to stop if Υ hits Γ_n before it jumps. Otherwise if there is a jump before Υ reaches Γ_n, then S_n^0 prescribes to stop when the process hits Γ_{n-1} before the next jump, and so on.

Theorem 6.1 shows that V_n of (6.1) and the functions v_n introduced in Corollary 6.1 are equal. Therefore, their respective limits V and v are also equal. Recall that V^n converges to V uniformly and the convergence rate is exponential (see Lemma 6.1). Since $(v_n)_{n\in\mathbb{N}}$ is a decreasing sequence with limit v, the stopping regions in (8.1) are nested and satisfy $\Gamma \subset \cdots \subset \Gamma_n \subset \Gamma_{n-1} \subset \cdots \Gamma_1$ and $\Gamma = \cap_{n=1}^{\infty} \Gamma_n$.

By Corollary 6.1 we know that each v_n is concave and bounded, which also implies that the limit v is concave and bounded. This in turn implies that the stopping regions Γ_n and Γ are convex and closed. Since we show in section 5 that the continuation region is bounded, it can readily be shown that the stopping regions Γ_n and Γ are the epigraphs of some mappings γ_n and γ which are convex and strictly decreasing and the numbers $x_n \triangleq \inf\{y \in \mathbb{R}_+ : \gamma_n(y) = 0\}$ and $x \triangleq \inf\{y \in \mathbb{R}_+ : \gamma(x) = 0\}$ are finite.

9. Extensions.

9.1. Nonidentical sources. Consider two independent Poisson processes X^1 and X^2 with arrival rates β_1 and β_2, respectively. At some random unobservable times θ_1 and θ_2, with distributions

$$\mathbb{P}(\theta_i = 0) = \pi_i, \quad \mathbb{P}(\theta_i > t) = (1-\pi_i)e^{-\lambda_i t} \text{ for } t \geq 0, \tag{9.1}$$

the arrival rates of the Poisson processes X^1 and X^2 change from β_i to α_i, respectively, i.e.,

$$X_t^i - \int_0^t h_i(s)ds, \quad t \geq 0,\, i = 1,2, \tag{9.2}$$

are martingales, in which

$$h_i(t) = [\beta_i 1_{\{s<\theta_i\}} + \alpha_i 1_{\{s\geq\theta_i\}}], \quad t \geq 0,\, i = 1,2. \tag{9.3}$$

Here α_1, α_2, β_1, and β_2 are known positive constants. Then the dynamics of $\Phi^\times$ defined in (2.19) becomes

$$d\Phi_t^\times = [\lambda_2\Phi_t^1 + \lambda_1\Phi_t^2 + (a_1+a_2)\Phi_t^+]dt + \Phi_t^\times\left[((\alpha_1/\beta_1)-1)dX_t^1 + ((\alpha_2/\beta_2)-1)dX_t^2\right] \tag{9.4}$$

in which $a_i = \lambda_i - \alpha_i + \beta_i$, $i \in \{1,2\}$. Let us introduce

$$x(t,\phi_0) = e^{(a_1+a_2)t}\phi_0 + \int_0^t e^{(a_1+a_2)(t-u)}(\lambda_2 y(u,\phi_1) + \lambda_1 z(u,\phi_2))du \quad \text{in which} \tag{9.5}$$

$$y(t,\phi_1) = -\frac{\lambda_1}{a_1} + e^{a_1 t}\left(\phi_1 + \frac{\lambda}{a}\right), \quad z(t,\phi_2) = -\frac{\lambda_2}{a_2} + e^{a_2 t}\left(\phi_2 + \frac{\lambda_2}{a_2}\right).$$

Then $\Phi_t^{\times}$, Φ_t^1, and Φ_t^2, $t \geq 0$, can be written as

$$\Phi_t^{\times} = x(t-\sigma_n, \Phi_{\sigma_n}^{\times}), \qquad \Phi_t^1 = y(t-\sigma_n, \Phi_{\sigma_n}^{+}), \qquad \Phi_t^2 = z(t-\sigma_n, \Phi_{\sigma_n}^1), \quad \sigma_n \leq t < \sigma_{n+1},\ n \in \mathbb{N}, \tag{9.6}$$

and

$$\begin{aligned} \Phi_{\sigma_{n+1}}^{\times} &= \left(\frac{\alpha_1}{\beta_1} 1_{\{X^1_{\sigma_{n+1}} \neq X^1_{\sigma_{n+1-}}\}} + \frac{\alpha_2}{\beta_2} 1_{\{X^2_{\sigma_{n+1}} \neq X^2_{\sigma_{n+1-}}\}}\right) \Phi_{\sigma_{n+1-}}^{\times}, \\ \Phi_{\sigma_{n+1}}^1 &= \frac{\alpha_1}{\beta_1} 1_{\{X^1_{\sigma_{n+1}} \neq X^1_{\sigma_{n+1-}}\}} \Phi_{\sigma_{(n+1)-}}^1, \quad \Phi_{\sigma_{n+1}}^2 = \frac{\alpha_2}{\beta_2} 1_{\{X^2_{\sigma_{n+1}} \neq X^2_{\sigma_{n+1-}}\}} \Phi_{\sigma_{(n+1)-}}^1. \end{aligned} \tag{9.7}$$

Choosing $\Upsilon_t = \left(\Phi_t^{\times}, \Phi_t^1, \Phi_t^2\right)$, $t \geq 0$, as the Markovian statistic to work with, we can extend our analysis to deal with nonidentical sources.

9.2. When there are more than two sources. We have solved a two-source quickest detection problem in which the aim is to detect the minimum of two disorder times. Our approach can easily be generalized to problems including several dimensions. To clarify how this generalization works, let us show what the sufficient statistics are when there are three independent sources. Assume that the observations come from the independent sources X^1, X^2, and X^3. Let Φ_t be the odds ratio defined in (2.9). Then

$$\Phi_t = \Phi_t^1 + \Phi_t^2 + \Phi_t^3 + \Phi_t^1\Phi_t^2 + \Phi_t^1\Phi_t^3 + \Phi_t^2\Phi_t^3 + \Phi_t^1\Phi_t^2\Phi_t^3 \tag{9.8}$$

in which Φ^i, $i \in \{1,2,3\}$, is defined as in (2.14). Let us denote $\Phi_t^{(i,j)} \triangleq \Phi_t^i\Phi_t^j$, $i,j \in \{1,2,3\}$, and $\Phi_t^{(x)} \triangleq \Phi_t^1\Phi_t^2\Phi_t^3$, $t \geq 0$. The dynamics of these processes can be written as

$$\begin{aligned} d\Phi_t^{(i,j)} &= [\lambda(\Phi_t^i + \Phi_t^j) + 2(\lambda - \alpha + \beta)\Phi_t^{(i,j)}]dt + \left(\frac{\alpha}{\beta} - 1\right)\Phi_t^{(i,j)} d(X_t^i + X_t^j), \\ d\Phi_t^{(x)} &= \left[\lambda\left(\Phi_t^{(1,2)} + \Phi_t^{(1,3)} + \Phi_t^{(2,3)}\right) + 3(\lambda - \alpha + \beta)\Phi_t^{(x)}\right] dt \\ &\quad + \left(\frac{\alpha}{\beta} - 1\right)\Phi_t^{(x)} d(X_t^1 + X_t^2 + X_t^3). \end{aligned} \tag{9.9}$$

We can see from (2.13) and (9.9) that $\Upsilon \triangleq (\Phi^1, \Phi^2, \Phi^3, \Phi_t^{(1,2)}, \Phi_t^{(1,3)}, \Phi_t^{(2,3)}, \Phi^{(x)})$ is a seven-dimensional Markovian statistic whose natural filtration is equal to the filtration generated by X^1, X^2, and X^3. From this one can see that the results of sections 6 and 7 can be extended to the three-dimensional case since these results rely only on the fact that the sufficient statistic Υ is a strong Markov process. The boundedness of the continuation region can also be shown as in section 5 since these results can be derived from the sample path properties of the sufficient statistic.

As a result, our results are applicable for decision making with large-scale distributed networks of information sources. In the future, using the techniques developed here, we would like to solve a multisource detection problem where the observations come from correlated sources. We also would like to extend our results and develop change detection algorithms that can be applied effectively to multiple source data that involves both continuous and discrete event phenomena.

9.3. When the jump sizes of the observations are random. Consider two independent *compound* Poisson processes $X^i = \{X_t^i : t \geq 0\}$, $i \in \{1,2\}$, where

$$X_t^i = X_0^i + \sum_{j=1}^{N_t^i} Y_j^i \tag{9.10}$$

in which N^i, $i \in \{1,2\}$, are two independent Poisson processes whose common rate $\beta > 0$ changes to α at some random unobservable times θ_i, $i \in \{1,2\}$, respectively. The random variables $Y_j^i \in \mathbb{R}^d$, $i \in \{1,2\}$, which are also termed as "marks," are independent and identically distributed with a common distribution, ν, which is called as the "mark distribution." At the change time θ_i the mark distribution of the process X^i changes from ν to μ. We will assume that μ is absolutely continuous with respect to ν and denote the Radon–Nikodym derivative by $r(y) \triangleq \frac{d\mu}{d\nu}(y),\ y \in \mathbb{R}^d$. In this case L_t^i in (2.4) becomes

$$L_t^i = e^{-(\alpha-\beta)t} \prod_{k=1}^{N_t^i} \left(\frac{\alpha}{\beta}\, r(Y_k^i) \right). \tag{9.11}$$

The likelihood ratio process L^i is the unique solution of the stochastic differential equation (see, e.g., [8])

$$dL_t^i = L_t^i \left(-(\alpha-\beta)dt + \int_{y\in\mathbb{R}^d} \left(\frac{\alpha}{\beta}\, r(y) - 1 \right) p(dtdy) \right), \quad L_0^i = 1, \tag{9.12}$$

where p is a random measure that is defined as

$$p^i((0,t] \times A) \triangleq \sum_{k=1}^{\infty} 1_{\{\sigma_k^i \leq t\}} 1_{\{Y_k^i \in A\}}, \quad t \geq 0, \tag{9.13}$$

and for any A that is a Borel measurable subset of $\mathbb{R}^d$. Here σ_k^i is the kth jump time of the process X^i. Now using the change of variable formula for semimartingales (see, e.g., [12]), we can write

$$d\Phi_t^i = (\lambda + (\lambda - \alpha + \beta)\Phi_t^i)dt + \Phi_{t-}^i \int_{y\in\mathbb{R}^d} \left(\frac{\alpha}{\beta}\, r(y) - 1 \right) p^i(dtdy), \quad \Phi_0^i = \frac{\pi_i}{1-\pi_i} \tag{9.14}$$

for $t \geq 0$ and $i \in \{1,2\}$. Note that $\Phi_{\sigma_n}^i = \frac{\alpha}{\beta}\, r(Y_n)\Phi_{\sigma_n -}^i$ at the nth jump time of the process X^i. Using a change of variable formula for semimartingales, the dynamics of $\Phi^\times$ and Φ^+ in (2.19) can be written as

$$\begin{aligned} d\Phi_t^\times &= [\lambda\Phi_t^+ + a\Phi_t^\times]dt + \Phi_{t-}^\times \int_{y\in\mathbb{R}^d} \left(\frac{\alpha}{\beta}\, r(y) - 1 \right) (p^1 + p^2)(dtdy), \\ d\Phi_t^+ &= [2\lambda + a\Phi_t^+]dt + \Phi_{t-}^1 \int_{y\in\mathbb{R}^d} \left(\frac{\alpha}{\beta}\, r(y) - 1 \right) p^1(dtdy) \\ &\quad + \Phi_{t-}^2 \int_{y\in\mathbb{R}^d} \left(\frac{\alpha}{\beta}\, r(y) - 1 \right) p^2(dtdy), \end{aligned} \tag{9.15}$$

with initial conditions $\Phi_0^\times = \pi_1\pi_2/[(1-\pi_1)(1-\pi_2)]$ and $\Phi_0^+ = \pi_1/(1-\pi_1) + \pi_2/(1-\pi_2)$.

The bounds on the continuation region constructed for the simple Poisson disorder problem in section 5 can also be shown to bound the continuation region of the compound Poisson disorder problem. On the other hand the results in sections 6 and 7 can be shown to hold. The only change will be the form of the operator J in (6.7). But this new operator can be shown to share the same properties as its counterpart for the unmarked case.

10. Appendix.

Proof of Theorem 6.1. We will prove only that $V_n = v_n$ and S_n^ε is an ε-optimal stopping time of (6.1).

The proof will be carried out in three steps.

(i) First we will show that $V_n \geq v_n$. To establish this fact, it is enough to show that for any stopping time $\tau \in \mathcal{S}$

$$\mathbb{E}_0^{a,b,c}\left[\int_0^{\tau\wedge\sigma_n} e^{-\lambda t}h(\Psi_t)dt\right] \geq v_n(a,b,c). \tag{10.1}$$

In order to prove (10.1) we will show that

$$\begin{aligned}&\mathbb{E}_0\left[\int_0^{\tau\wedge\sigma_n} e^{-\lambda t}h(\Psi_t)dt\right]\\ &\quad\geq \mathbb{E}_0\left[\int_0^{\tau\wedge\sigma_{n-k+1}} e^{-\lambda t}h(\Psi_t)dt + 1_{\{\tau\geq\sigma_{n-k+1}\}}e^{-\lambda\sigma_{n-k+1}}v_{k-1}(\Upsilon_{\sigma_{n-k+1}})\right]\end{aligned} \tag{10.2}$$

for $k \in \{1,2,\ldots,n+1\}$. Note that (10.1) follows from (10.2) if we set $k = n+1$. In what follows we will show (10.2) by induction.

When $k = 1$, (10.2) is satisfied since $v_0 = 0$. Assume that (10.2) holds for $1 \leq k \leq n+1$. Let us denote the right-hand side of (10.2) by ρ_{k-1}. We can write $\rho_{k-1} = \rho_{k-1}^1 + \rho_{k-1}^2$, where

$$\begin{aligned}\rho_{k-1}^1 &\triangleq \mathbb{E}_0\left[\int_0^{\tau\wedge\sigma_{n-k}} e^{-\lambda t}h(\Psi_t)dt\right] \quad \text{and}\\ \rho_{k-1}^2 &\triangleq \mathbb{E}_0\left[1_{\{\tau\geq\sigma_{n-k}\}}\left(\int_{\sigma_{n-k}}^{\tau\wedge\sigma_{n-k+1}} e^{-\lambda t}h(\Psi_t)dt\right.\right.\\ &\qquad\left.\left.+\,1_{\{\tau\geq\sigma_{n-k+1}\}}e^{-\lambda\sigma_{n-k+1}}v_{k-1}(\Upsilon_{\sigma_{n-k+1}})\right)\right].\end{aligned} \tag{10.3}$$

Now by Lemma 6.3, there exists an $\mathcal{F}_{\sigma_{n-k}}$-measurable random variable ξ_{n-k} such that

$$\tau\wedge\sigma_{n-k+1} = (\sigma_{n-k}+\xi_{n-k})\wedge\sigma_{n-k+1} \quad \text{almost surely on } \{\tau\geq\sigma_{n-k}\}. \tag{10.4}$$

Equation (10.4) together with the strong Markov property of Υ (with respect to the filtration $\mathbb{F}$) implies that

$$\rho_{k-1}^2 = \mathbb{E}_0\left[1_{\{\tau\geq\sigma_{n-k}\}}e^{-\lambda\sigma_{n-k}}f_{k-1}(\xi_{n-k},\Upsilon_{\sigma_{n-k}})\right] \tag{10.5}$$

in which

$$\begin{aligned}f_{k-1}(r,(a,b,c)) &\triangleq \mathbb{E}_0^{a,b,c}\left[\int_0^{r\wedge\sigma_1} e^{-\lambda t}h(\Psi_t)dt + 1_{\{r\geq\sigma_1\}}e^{-\lambda\sigma_1}v_{k-1}(\Upsilon_{\sigma_1})\right]\\ &= Jv_{k-1}(r,(a,b,c)) \geq J_0v_{k-1}(a,b,c) = v_k(a,b,c)\end{aligned} \tag{10.6}$$

in which the second equality and the first inequality follow from (6.5) and the last equality follows from (6.10). Therefore

$$\rho_{k-1}^2 \geq \mathbb{E}_0\left[1_{\{\tau \geq \sigma_{n-k}\}} e^{-\lambda \sigma_{n-k}} v_k(\Upsilon_{\sigma_{n-k}})\right]. \tag{10.7}$$

Now using (10.2), (10.3), and (10.7) we obtain that (10.2) holds when k is replaced by $k+1$. At this point we have proved by induction that (10.2) holds for $k = 1, 2, \ldots, n+1$.

(ii) The converse of (i), $V_n \leq v_n$, follows from (6.13) since $S_n^\varepsilon \leq \sigma_n$ by construction (see (6.12)).

(iii) What is left to prove is (6.13). If $n = 1$, then the left-hand side of (6.13) becomes

$$\begin{aligned}\mathbb{E}_0^{a,b,c}\left[\int_0^{r_0^\varepsilon(a,b,c)\wedge\sigma_1} e^{-\lambda t} h(\Psi_t)dt\right] &= Jv_0(r_0^\varepsilon(a,b,c),a,b,c) \\ &\leq J_0 v_0(a,b,c) + \varepsilon = v_1(a,b,c) + \varepsilon.\end{aligned} \tag{10.8}$$

Now, suppose that (6.13) holds for all $\varepsilon > 0$ for some n. Using the fact that $S_{n+1}^\varepsilon \wedge \sigma_1 = r_n^{\varepsilon/2} \wedge \sigma_1$ almost surely and the strong Markov property of Υ, we can write

$$\begin{aligned}&\mathbb{E}_0\left[\int_0^{S_{n+1}^\varepsilon} e^{-\lambda t} h(\Psi_t)dt\right] \\ &\quad= \mathbb{E}_0\left[\int_0^{S_{n+1}^\varepsilon\wedge\sigma_1} e^{-\lambda t} h(\Psi_t)dt + 1_{\{S_{n+1}^\varepsilon \geq \sigma_1\}} \int_{\sigma_1}^{S_{n+1}^\varepsilon} e^{-\lambda t} h(\Psi_t)dt\right] \\ &\quad= \mathbb{E}_0\left[\int_0^{r_n^{\varepsilon/2}(a,b,c)\wedge\sigma_1} e^{-\lambda t} h(\Psi_t)dt\right] + \mathbb{E}_0\left[1_{\{r_n^{\varepsilon/2}(a,b,c)\geq\sigma_1\}} e^{-\lambda\sigma_1} g_n(\Upsilon_{\sigma_1})\right]\end{aligned} \tag{10.9}$$

in which

$$g_n(a,b,c) \triangleq \mathbb{E}_0^{a,b,c}\left[\int_0^{S_n^{\varepsilon/2}} e^{-\lambda t} h(\Psi_t)dt\right] \leq v_n(a,b,c) + \varepsilon/2. \tag{10.10}$$

The inequality in (10.10) follows from the induction hypothesis. Using (10.10) we can write (10.9) as

$$\begin{aligned}&\mathbb{E}_0^{a,b,c}\left[\int_0^{S_{n+1}^\varepsilon} e^{-\lambda t} h(\Psi_t)dt\right] \\ &\quad\leq \mathbb{E}_0^{a,b,c}\left[\int_0^{r_n^{\varepsilon/2}(a,b,c)\wedge\sigma_1} e^{-\lambda t} h(\Psi_t)dt + 1_{\{r_n^{\varepsilon/2}(a,b,c)\geq\sigma_1\}} e^{-\lambda\sigma_1} v_n(\Upsilon_{\sigma_1})\right] + \varepsilon/2 \\ &\quad= Jv_n(r_n^{\varepsilon/2}(a,b,c),a,b,c) + \varepsilon/2 \leq v_{n+1}(a,b,c) + \varepsilon.\end{aligned} \tag{10.11}$$

This proves (6.13) when n is replaced by $n+1$. □

Acknowledgments. We are grateful to the two referees for their detailed comments that helped us improve the manuscript. We also would like to thank Semih Sezer for insightful comments.

REFERENCES

[1] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*, Prentice–Hall, Englewood Cliffs, NJ, 1993.

[2] E. Bayraktar and S. Dayanik, *Poisson disorder problem with exponential penalty for delay*, Math. Oper. Res., 31 (2006), pp. 217–233.

[3] E. Bayraktar, S. Dayanik, and I. Karatzas, *The standard Poisson disorder problem revisited*, Stochastic Process. Appl., 115 (2005), pp. 1437–1450.

[4] E. Bayraktar, S. Dayanik, and I. Karatzas, *Adaptive poisson disorder problem*, Ann. Appl. Prob., 16 (2006), pp. 1190–1261.

[5] P. Brémaud, *Point Processes and Queues*, Springer-Verlag, Berlin, 1981.

[6] M. H. A. Davis, *Markov Models and Optimization*, Chapman and Hall, London, 1993.

[7] P. Dube and R. Mazumdar, *A Framework for Quickest Detection of Traffic Anomalies in Networks*, preprint, Purdue University, West Lafayette, IN, 2001. Available online at http://citeseer.ist.psu.edu/506551.html

[8] J. Jacod and A. Shiryaev, *Limit Theorem for Stochastic Processes*, Springer-Verlag, Berlin, 1987.

[9] R. S. Lipster and A. N. Shiryaev, *Statistics of Random Processes*, Springer-Verlag, Berlin, 2001.

[10] S. N. Neftci, *Optimal prediction of cyclical downturns*, J. Econom. Dynam. Control, 4 (1982), pp. 225–241.

[11] H. V. Poor, *Quickest detection with exponential penalty for delay*, Ann. Statist., 26 (1998), pp. 2179–2205.

[12] P. Protter, *Stochastic Integration and Differential Equations*, Springer-Verlag, Berlin, 1990.

[13] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, Springer-Verlag, Berlin, 1999.

[14] C. Sonesson and D. Bock, *A review and discussion of prospective statistical surveillance in public health*, J. Roy. Statist. Soc. A, 166 (2003), pp. 5–21.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 332–355

UNIQUENESS OF CONSTRAINED VISCOSITY SOLUTIONS IN HYBRID CONTROL SYSTEMS*

MINYI HUANG†

Abstract. We study constrained viscosity solutions with an unbounded growth for a class of first order Hamilton–Jacobi–Bellman equations arising in hybrid control systems. To deal with the boundary constraint and rapid growth of the solutions, we construct a particular set of test functions and under very mild conditions establish a comparison theorem which gives the estimate of distance between the subsolution and the supersolution. The comparison theorem implies uniqueness of the constrained viscosity solution if its existence is ensured; and under some additional assumptions we give an existence result by showing that the value function is a constrained viscosity solution. We then apply the obtained uniqueness results to an optimal scheduling problem and finally to stochastic manufacturing systems.

Key words. hybrid control systems, optimal control, HJB equations, constrained viscosity solutions, fluid models, manufacturing systems

AMS subject classifications. 93E20, 93E03, 49L25, 49L20

DOI. 10.1137/050635845

1. Introduction.

This paper is concerned with the analysis of a class of first order Hamilton–Jacobi–Bellman (HJB) equations with discrete transitions and state constraints. Such equations arise naturally in the optimal control of stochastic systems with random structural changes in dynamics, which are modeled as Markovian jumps. These systems involve both continuum and discrete components in their evolution and are referred to as hybrid systems, and they have been investigated from a wide range of backgrounds including production planning subject to random machine breakdown and repair [20] and the control of fluid queueing models for communication networks [19, 7, 17], among others [6, 9, 14, 24, 28]. Due to physical limitation, a typical feature for many control models is that the system state is restricted to a certain set; for instance, the level for buffers must be maintained nonnegative [20, 25]. To deal with the resulting HJB equation, one needs to take into account both the discrete transitions and the state space constraints and to adopt the notion of appropriately defined constrained viscosity solutions first introduced in [22].

Specifically, Soner studied an optimal control problem and introduced first order constrained viscosity solutions in [22], where the deterministic state trajectory is restricted to a given subset of $\mathbb{R}^n$, and in a companion work [23] along this line, viscosity solutions were analyzed for controlled piecewise deterministic Markov processes [6] defined on a subset of $\mathbb{R}$, which leads to an integral HJB equation. Later on, the result in [22] was generalized in [18] by identifying weaker sufficient conditions for ensuring continuity of the value function and in [12] by an additional boundary characterization of the subsolution via a so-called inward Hamiltonian reflecting boundary constraints.

*Received by the editors July 12, 2005; accepted for publication (in revised form) December 27, 2006; published electronically April 13, 2007. This work was partially supported by the Australian Research Council.

http://www.siam.org/journals/sicon/46-1/63584.html

†Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, Victoria 3010, Australia. Current address: Department of Information Engineering, Research School of Information Sciences and Engineering, The Australian National University, Canberra, ACT 0200, Australia (minyi.huang@rsise.anu.edu.au).

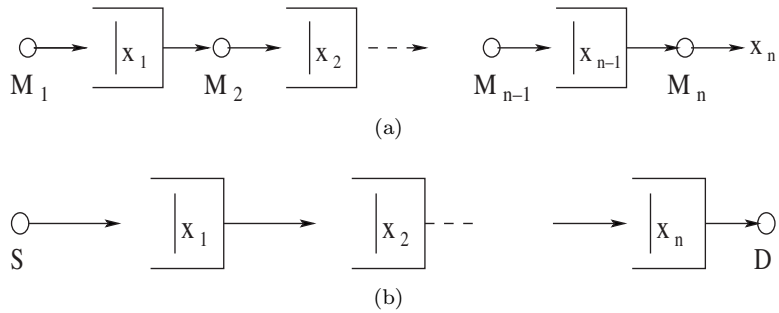


FIG. 1. (a) *A manufacturing system with n machines (M_i, $1 \le i \le n$) and $n-1$ buffers;* (b) *a multihop communication network where the source S and destination D are connected by n buffers.*

For HJB equations involving finite state Markov chains, viscosity solutions have been studied in [20, 8, 26, 27]. In particular, the authors in [26, 27] considered controlled random transitions but there were no state constraints.

Although viscosity solutions with state constraints are of importance and have their primary motivation in optimal control, in many application problems, the existing results face limitation. Notably, in the sequence of work [22, 18, 12] considering first order HJB equations for deterministic systems, uniqueness is obtained for uniformly continuous and bounded solutions. In [15] and [5], bounded continuous solutions were analyzed on a bounded domain. Also, in a singular perturbation control problem with partial state constraints [1], uniqueness and existence theorems were established with bounded continuous solutions. For illustrating the limitation of those previous results, we consider the optimal control of a single buffer fluid model with controlled input and output and, as a well-motivated practice, introduce a linear holding cost for a positive buffer level (see section 6 for details). This readily leads to unbounded value functions, and existing results for constrained viscosity solutions are difficult to apply.

In this work we study uniqueness of constrained viscosity solutions for a class of stochastic hybrid systems. We concentrate our attention to two concrete types of domains for the state variable. The particular structure of the state space has adequate generality and is frequently encountered in a wide range of application problems arising in manufacturing systems and communication networks [20, 25, 19] (see Figure 1 for illustration), though a generalization of the state space to other forms is possible. In introducing our solution notion, we generalize the definition of constrained viscosity solutions for standard HJB equations of deterministic models to a coupled HJB equation system. Resulting from the state space constraints, this definition leads to specifying the viscosity sub/supersolution in two different regions, respectively, i.e., characterizing the subsolution in a smaller region—the interior of the constrained state space. Such a differentiation by two regions is important for developing a solution framework for uniqueness analysis. We prove uniqueness of the solution within the class of functions satisfying a polynomial growth and local Hölder continuity. In establishing the comparison result in this paper, a crucial step is to obtain suitable test functions involved in the definition of constrained viscosity solutions. Towards this end, we construct the auxiliary function Φ by first dominating the sub/supersolution growth by an exponential function and then introducing a pair of perturbation parameters (τ, ε) [see (11)] such that the resulting maxima (w.r.t. x) can be tuned to the interior of the state space to generate desired test functions for the subsolution.

The proof of the comparison theorem depends on generalizing typical techniques for deterministic systems [22, 2].

For proving existence results in the general model, a quite difficult step is to show continuity of the value function. The key idea in our analysis is to truncate a small time interval by the jump time of the Markov chain so that locally the system dynamics act like a time-invariant model. This resulting feature enables us to use a certain time-shifting technique to construct auxiliary admissible controls for cost estimates. In particular, using a recursive estimation procedure, we obtain Hölder continuity of the value function, and we mention that as a byproduct this method can be used to strengthen some existing continuity results in the literature for state constrained optimal control problems.

Our work differs from most existing analysis on constrained viscosity solutions for deterministic systems in that we need to deal with a system of coupled equations and the solution growth is rapid. Our solution notion for the coupled HJB equation system and the uniqueness results provide a unified analytical basis for the optimal control of this class of hybrid systems. In particular, our uniqueness results are applicable to classical stochastic manufacturing models (see, e.g., [20]), where to our best knowledge the existing work has not provided uniqueness results for the coupled HJB equations when nonnegative buffer level constraints are imposed.

The organization of the paper is as follows. In section 2, we first describe the optimal control problem for the hybrid system and introduce the notion of constrained viscosity solutions. The comparison result and uniqueness theorem are stated in section 3. The proof of the comparison theorem is technical and postponed to section 4. For the general hybrid system model, section 5 first shows Hölder continuity of the value function under some technical conditions and proves that it is the unique constrained viscosity solution. In section 6, we study an optimal data traffic scheduling problem and prove the existence and uniqueness of constrained viscosity solutions by applying the result in section 3. In section 7, we further apply the results in section 3 to a well-studied stochastic manufacturing system, which complements existence theorems in the manufacturing literature [20]. Finally, a few concluding remarks are presented in section 8.

2. The HJB equation and constrained viscosity solutions. Consider a hybrid control system described by the following differential equation:

$$\frac{dX(t)}{dt} = F(X(t), \theta(t), u(t)), \qquad t \geq 0, \tag{1}$$

with initial condition $X(0) \in \bar{Q}$. Here X and θ are called the state and mode variables, respectively. The trajectory of X on $[0, \infty)$ is required to be in $\bar{Q}$, which is a closed subset of $\mathbb{R}^n$ with a nonempty interior Q. Moreover, θ is a continuous time Markov chain with state space $\Theta = \{1, 2, \ldots, m\}$ and transition probability rate matrix $\Pi_\theta = (\pi_{ij})_{m\times m}$, which is also called the generator. It is assumed that, with probability one, the trajectory of θ is right continuous with left limit. Given $\theta(t) = k$, the control $u(t)$ takes values from a compact set $U_k \subset \mathbb{R}^d$. Let $\mathcal{F}_t$ denote the σ-algebra generated by the Markov chain θ up to time t, i.e., $\mathcal{F}_t = \sigma(\theta(s), s \leq t)$. Associated with $X(0) = x$ and $\theta(0) = k$, the admissible control set is written as $\mathcal{U}_{x,k}$ consisting of all controls $u(\cdot)$ satisfying $u(t) \in U_{\theta(t)}$ and adapted to $\mathcal{F}_t$ such that $P\{X(t) \in \bar{Q},\ \forall\ t \geq 0\} = 1$. We make the convention that for all $(x, k) \in \bar{Q} \times \Theta$, $\mathcal{U}_{x,k}$ is nonempty and that the state process $X(t)$ associated with an admissible control is uniquely determined on $[0, \infty)$ with exception on a null set of samples. Given initial condition $(x, k) \in \bar{Q} \times \Theta$

at $t = 0$, let the cost function be given by

$$\begin{aligned} v(x,k) &= \inf_{u\in\mathcal{U}_{x,k}} J(x,k,u) \\ &\triangleq \inf_{u\in\mathcal{U}_{x,k}} E\Big[\int_0^\infty e^{-\rho t}L(X(t),\theta(t),u(t))dt|X(0)=x,\theta(0)=k\Big], \end{aligned} \tag{2}$$

where $\rho > 0$ is a discount factor and L is the cost integrand before discount.

To facilitate the subsequent analysis, we set some convention on notation. We may alternatively denote $X(t)$ as X_t with a real-valued subscript $t \geq 0$, and the same convention holds for $u(t)$ and $\theta(t)$, etc. The letter u may stand for a value in U_k for a certain $k \in \Theta$ or a control adapted to $\mathcal{F}_t$; the specific interpretation should be clear from the context. Throughout the paper, for a real-valued vector y, $|y|$ denotes its Euclidean norm.

For any function $\varphi : \Theta \to \mathbb{R}$, we define the map

$$[\Pi_\theta\varphi(\cdot)](i) = \sum_{j\neq i}\pi_{ij}[\varphi(j)-\varphi(i)], \tag{3}$$

where $\pi_{ii} + \sum_{j\neq i}\pi_{ij} = 0$.

We assume for any given $k \in \Theta$, both $F(x,k,u)$ and $L(x,k,u)$ are continuous in $(x,u) \in \bar{Q}\times U_k$. A formal application of dynamic programming leads to the following equation system:

$$\rho v(x,k) = \inf_{u\in U_k}\Big[v_x^T(x,k)F(x,k,u) + [\Pi_\theta v(x,\cdot)](k) + L(x,k,u)\Big], \tag{4}$$

where $(x,k) \in \bar{Q}\times\Theta$ and the superscript $(\cdot)^T$ denotes the transpose of a vector or matrix. Note that due to the action of the generator, (4) gives a system of m coupled equations. For convenience of exposition, we simply refer to (4) as the HJB equation for the underlying optimal control problem. Write

$$\tilde{H}(x,k,v_x(x,k),v(x,\cdot),u) = v_x^T(x,k)F(x,k,u) + [\Pi_\theta v(x,\cdot)](k) + L(x,k,u).$$

Then the HJB equation (4) may be written in the compact form:

$$\begin{aligned} \rho v(x,k) &= \inf_{u\in U_k}\tilde{H}(x,k,v_x(x,k),v(x,\cdot),u) \\ &\triangleq H(x,k,v_x(x,k),v(x,\cdot)), \qquad (x,k)\in\bar{Q}\times\Theta, \end{aligned} \tag{5}$$

where the dot entry in (5) indicates that for each fixed k, the term H depends on the whole vector $[v(x,1),\ldots,v(x,m)]$.

DEFINITION 1. *Let $\underline{v}(x,k)$, $\overline{v}(x,k)$, and $v(x,k)$ be functions from $\bar{Q}\times\Theta$ to $\mathbb{R}$, each being continuous in x for all $k \in \Theta$.*

(i) *$\underline{v}(x,k)$ is a viscosity subsolution to (5) on $Q\times\Theta$ if for any $k_0 \in \Theta$ and any function $\phi \in C^1(\bar{Q})$, we have*

$$\rho\underline{v}(x_0,k_0) - H(x_0,k_0,\phi_x(x_0),\underline{v}(x_0,\cdot)) \leq 0 \tag{6}$$

at x_0, whenever $\underline{v}(x,k_0)-\phi(x)$ attains a local maximum at $x = x_0 \in Q$.

(ii) *$\overline{v}(x,k)$ is a viscosity supersolution to (5) on $\bar{Q}\times\Theta$ if for any $k_0 \in \Theta$ and $\phi \in C^1(\bar{Q})$, we have*

$$\rho\overline{v}(x_0,k_0) - H(x_0,k_0,\phi_x(x_0),\overline{v}(x_0,\cdot)) \geq 0 \tag{7}$$

at x_0, whenever $\overline{v}(x,k_0)-\phi(x)$ attains a local minimum at $x = x_0 \in \bar{Q}$.

(iii) $v(k,x)$ *is called a constrained viscosity solution on* $\bar{Q}\times\Theta$ *to* (5) *if it is both a viscosity subsolution on* $Q\times\Theta$ *and a viscosity supersolution on* $\bar{Q}\times\Theta$.

In the definition of the viscosity supersolution, the minima x_0 may lie on the boundary of $\bar{Q}$. The function ϕ involved in either (i) or (ii) in Definition 1 is called the test function.

Denote by $C_p(\bar{Q}\times\Theta)$ the set of functions $g(x,k)$ from $\bar{Q}\times\Theta$ to $\mathbb{R}$, which are continuous in $x\in\bar{Q}$ for any given $k\in\Theta$ and have a polynomial growth rate, i.e., for any $g\in C_p(\bar{Q}\times\Theta)$, one can find positive constants C and b, depending on that particular function, such that $|g(x,k)|\le C(1+|x|^b)$ for all $(x,k)\in\bar{Q}\times\Theta$. For $a_1,a_2\in\mathbb{R}$, denote $a_1\vee a_2=\max\{a_1,a_2\}$, and $a_1\wedge a_2=\min\{a_1,a_2\}$. Given $\gamma\in(0,1]$ and $g\in C_p(\bar{Q}\times\Theta)$, define

$$Hol(g,\gamma,R)\stackrel{\triangle}{=}\sup_{k\in\Theta}\sup_{|x|\vee|x'|\le R}\frac{|g(k,x')-g(k,x)|}{|x'-x|^{\gamma}},$$

where $x,x'\in\bar{Q}$ and $0<R<\infty$. The value $Hol(g,\gamma,R)\le\infty$ is called the local Hölder constant for g associated with $R>0$, where γ is the Hölder exponent. For the case $\gamma=1$, $Hol(g,1,R)$ reduces to the local Lipschitz constant and is denoted as $Lip(g,R)$.

Define $C^{loc}_{p,Hol}(\bar{Q}\times\Theta)$ as the class of functions $g\in C_p(\bar{Q}\times\Theta)$ satisfying local Hölder continuity; i.e., there exists a Hölder exponent $\gamma\in(0,1]$ such that $Hol(g,\gamma,R)<\infty$ for all $R>0$. Furthermore, we define $C^{loc}_{p,Lip}(\bar{Q}\times\Theta)$ as the class of functions $g\in C_p(\bar{Q}\times\Theta)$ satisfying local Lipschitz continuity in x; i.e., $Lip(g,R)<\infty$ for all $R>0$.

Obviously, $C^{loc}_{p,Lip}(\bar{Q}\times\Theta)\subset C^{loc}_{p,Hol}(\bar{Q}\times\Theta)$. In addition, if $g\in C^{loc}_{p,Hol}(\bar{Q}\times\Theta)$ with Hölder exponent γ_2 and $0<\gamma_1<\gamma_2\le 1$, g is also locally Hölder continuous with exponent γ_1.

In establishing our main results, we concentrate on two types of structures for $\bar{Q}$.

Case (i). For state constraint in a subspace:

$$\bar{Q}_a\stackrel{\triangle}{=}[0,\infty)^{n-1}\times(-\infty,\infty),$$

where the integer $n\ge 2$. The interior of the set is $Q_a=(0,\infty)^{n-1}\times(-\infty,\infty)$.

Case (ii). For state constraint in full space:

$$\bar{Q}_b\stackrel{\triangle}{=}[0,\infty)^{n},$$

where $n\ge 1$. The interior of the set is $Q_b=(0,\infty)^n$.

Corresponding to $\bar{Q}_a$ and $\bar{Q}_b$, the state variable x is restricted to the *positive orthant* of $\mathbb{R}^n$ or its $n-1$-dimensional subspace. Indeed, cases (i) and (ii) can cover fairly general application models as shown in Figure 1, and they are also applicable to systems with more complicated buffer interconnection; see, e.g., [21]. It is worth noting that in the manufacturing fluid model given by Figure 1(a), the first $n-1$ entries in x correspond to buffer levels and must be positive; the last entry x_n, which denotes the inventory level of the final product, however, can be negative and interpreted as backlog. Although our technique developed in this paper may be extended to deal with other forms of $\bar{Q}$, we do not intend to treat the most general form.

3. The comparison theorem and uniqueness of solutions. The objective of this section is to establish a comparison result which plays an important role in proving uniqueness. Existence analysis will be presented for the general model in section 5 and for more concrete models in sections 6 and 7.

Let $L_i(x,k,u)$, $i=1,2$, be two functions with $u \in U_k$ and $(x,k) \in \bar{Q} \times \Theta$. Replacing $L(x,k,u)$ by $L_i(x,k,u)$ in the original HJB equation (4), we write two new equations:

$$\rho v(x,k) = H_1(x,k,v_x(x,k),v(x,\cdot)), \tag{8}$$

$$\rho v(x,k) = H_2(x,k,v_x(x,k),v(x,\cdot)), \tag{9}$$

where $(x,k) \in \bar{Q} \times \Theta$ and the construction for H_i, $i=1,2$, is obvious.

3.1. Main results. We make the following assumptions.

(A1) For any given $k \in \Theta$, $F(x,k,u)$ and $L_i(x,k,u)$, $i=1,2$, are continuous in $(x,u) \in \bar{Q} \times U_k$.

(A1′) For any given $k \in \Theta$, $F(x,k,u)$ and $L(x,k,u)$ are continuous in $(x,u) \in \bar{Q} \times U_k$.

(A2) $F_{\max} \stackrel{\triangle}{=} \sup_{(x,k)\in\bar{Q}\times\Theta} \sup_{u\in U_k} |F(x,k,u)| < \infty$.

Under (A1) and (A1′), we have the following equicontinuity in x on compact sets. Let φ stand for F, L, or L_i. For a given compact subset $B_{\bar{Q}}$ of $\bar{Q}$, when $x, x' \in B_{\bar{Q}}$ and $|x-x'| \to 0$, we have

$$|\varphi(x,k,u) - \varphi(x',k,u)| \to 0 \tag{10}$$

with a vanishing rate not depending on (k,u).

THEOREM 2. *Let $\bar{Q}$ be either $\bar{Q}_a$ or $\bar{Q}_b$, and suppose* (A1)–(A2) *hold. If v_1, $v_2 \in C^{loc}_{p,Hol}(\bar{Q}\times\Theta)$ are, respectively, a viscosity subsolution to* (8) *on $Q\times\Theta$ and a viscosity supersolution to* (9) *on $\bar{Q}\times\Theta$, then the inequality holds:*

$$\sup_{\bar{Q}\times\Theta} [v_1(x,k) - v_2(x,k)] \le \rho^{-1} \sup_{\bar{Q}\times\Theta} \sup_{U_k} [L_1(x,k,u) - L_2(x,k,u)].$$

Theorem 2 is the so-called comparison theorem, and it immediately implies the following uniqueness theorem.

THEOREM 3. *Let $\bar{Q}$ be either $\bar{Q}_a$ or $\bar{Q}_b$, and suppose* (A1′) *and* (A2) *hold. If $v \in C^{loc}_{p,Hol}(\bar{Q}\times\Theta)$ is a constrained viscosity solution to* (5)*, then it is unique in the function class $C^{loc}_{p,Hol}(\bar{Q}\times\Theta)$.*

3.2. Some preliminary lemmas. To prove Theorem 2, we need to establish a sequence of preliminary results. The basic approach is to introduce a suitable comparison function Φ for the construction of smooth test functions ϕ to generate the local minima and maxima and then to apply the definition of viscosity sub/supersolutions. A key technique will be developed such that the obtained maxima for $v_1 - \phi$, as specified during the proof of Theorem 2, do not occur at the boundary of $\bar{Q}$, which is crucial for subsequently applying the definition of viscosity subsolutions.

Let v_1 and v_2 be the viscosity sub/supersolution, respectively. For both Case (i) $\bar{Q} = \bar{Q}_a = [0,\infty)^{n-1} \times (-\infty,\infty)$ and Case (ii) $\bar{Q} = \bar{Q}_b = [0,\infty)^n$, we use the same function $\Phi(x,y,k)$ constructed as follows. Denote $\mathbf{1}_n = (1,1,\dots,1)^T$, and for $v_1, v_2 \in C^{loc}_{p,Hol}(\bar{Q}\times\Theta)$, let

$$\Phi(x,y,k) = v_1(x,k) - v_2(y,k) - \left|\frac{x-y}{\varepsilon} - \tau\mathbf{1}_n\right|^2 - \alpha[\zeta(x)+\zeta(y)], \qquad x,y \in \bar{Q}, \tag{11}$$

where $\zeta(x) = \exp(\beta\sqrt{|x|^2+1})$, with $\beta = \rho F_{\max}^{-1}$, and $\varepsilon, \tau, \alpha$ are all parameters chosen within the interval $(0,1]$ throughout sections 3 and 4.

The construction of Φ is based on the methods in [20, 22, 11, 5]; however, with the simultaneous appearance of state constraints and rapid growth, it is necessary to predominate v_1 and v_2 by the exponential term $\zeta(x)$ and subsequently insert the small perturbation term $\tau\mathbf{1}_n$, the magnitude of which can be adjusted independently. This differs from the technique in [22, 5]. During the maximization of Φ, τ causes a useful asymmetry between x and y in producing the increment of Φ. Such an effect is further amplified by reducing ε provided that τ is fixed first, and this ensures that x can be tuned to the interior of $\bar{Q}$ leading to desired test functions.

Since both v_1 and v_2 have a polynomial growth rate, it is clear that there exists $(\hat{x}, \hat{y}, \hat{k}) \in \bar{Q} \times \bar{Q} \times \Theta \stackrel{\triangle}{=} \Gamma$ such that

$$\Phi(\hat{x}, \hat{y}, \hat{k}) = \sup_{(x,y,k)\in\Gamma} \Phi(x, y, k), \tag{12}$$

where the values of $\hat{x}$, $\hat{y}$, and $\hat{k}$ depend on ε, τ and α. However, for a given $\alpha \in (0, 1]$, we may obtain a uniform bound for $|\hat{x}|$ and $|\hat{y}|$ when the value of ε and τ varies on $(0, 1]$.

LEMMA 4. *Suppose $\bar{Q} = \bar{Q}_a$ or $\bar{Q}_b$. Let $v_1, v_2 \in C^{loc}_{p,Hol}(\bar{Q} \times \Theta)$ be given in* (11) *and $(\hat{x}, \hat{y}, \hat{k})$ be obtained from* (12). *Then there exists a positive constant, depending only on α and denoted as C_α, such that*

$$|\hat{x}| \vee |\hat{y}| \le C_\alpha.$$

Proof. It suffices to analyze for $\bar{Q} = \bar{Q}_a$. Since $\Phi(\hat{x}, \hat{y}, \hat{k}) \ge \Phi(0, 0, \hat{k})$, it follows that

$$\begin{aligned} &v_1(\hat{x}, \hat{k}) - v_2(\hat{y}, \hat{k}) - \left|\frac{\hat{x} - \hat{y}}{\varepsilon} - \tau\mathbf{1}_n\right|^2 - \alpha[\zeta(\hat{x}) + \zeta(\hat{y})] \\ &\ge v_1(0, \hat{k}) - v_2(0, \hat{k}) - n\tau^2 - \alpha[\zeta(0) + \zeta(0)], \end{aligned}$$

which gives

$$\begin{aligned} &\alpha[\zeta(\hat{x}) + \zeta(\hat{y})] + \left|\frac{\hat{x} - \hat{y}}{\varepsilon} - \tau\mathbf{1}_n\right|^2 \\ &\le v_1(\hat{x}, \hat{k}) - v_1(0, \hat{k}) - v_2(\hat{y}, \hat{k}) + v_2(0, \hat{k}) + n\tau^2 + \alpha[\zeta(0) + \zeta(0)]. \end{aligned}$$

Without loss of generality, assume $C_0 > 0$ and $b_0 > 0$ have been found such that $|v_1(x, k)| \vee |v_2(x, k)| \le C_0(1 + |x|^{b_0})$, for $(x, k) \in \bar{Q} \times \Theta$. Since

$$\alpha[\zeta(\hat{x}) + \zeta(\hat{y})] \le C_0(4 + |\hat{x}|^{b_0} + |\hat{y}|^{b_0}) + n + [\zeta(0) + \zeta(0)],$$

there exists $C_\alpha > 0$, depending on α but not on ε and τ, such that $|\hat{x}| \vee |\hat{y}| \le C_\alpha$. □

Notice that the selection of C_α implicitly depends on the associated parameters C_0 and b_0. However, for convenience of presentation, in our analysis we simply say it depends only on α, since v_1 and v_2 are assumed to be picked out from $C^{loc}_{p,Hol}(\bar{Q} \times \Theta)$ and hence fixed.

LEMMA 5. *Suppose $\bar{Q} = \bar{Q}_a$ or $\bar{Q}_b$ and fix $\alpha \in (0, 1]$. For $(\hat{x}, \hat{y}, \hat{k})$ given in* (12), *the following properties hold:* (i) $\sup_{\varepsilon\in(0,1]} \varepsilon^{-1}|\hat{x} - \hat{y}| = O(1)$, *where the right-hand side is independent of τ, and* (ii) $\lim_{\varepsilon\to 0+} |\hat{x} - \hat{y}| = 0$ *uniformly w.r.t. τ.*

Proof. It is adequate to consider $\bar{Q} = \bar{Q}_a$. Since $2\Phi(\hat{x}, \hat{y}, \hat{k}) \geq \Phi(\hat{x}, \hat{x}, \hat{k}) + \Phi(\hat{y}, \hat{y}, \hat{k})$, we get

$$\begin{aligned} &2v_1(\hat{x}, \hat{k}) - 2v_2(\hat{y}, \hat{k}) - 2\left|\frac{\hat{x} - \hat{y}}{\varepsilon} - \tau \mathbf{1}_n\right|^2 - 2\alpha[\zeta(\hat{x}) + \zeta(\hat{y})] \\ &\geq v_1(\hat{x}, \hat{k}) - v_2(\hat{x}, \hat{k}) - n\tau^2 - 2\alpha\zeta(\hat{x}) + v_1(\hat{y}, \hat{k}) - v_2(\hat{y}, \hat{k}) - n\tau^2 - 2\alpha\zeta(\hat{y}). \end{aligned}$$

Suppose v_1, v_2 have exponent $\gamma_1, \gamma_2 \in (0, 1]$, respectively, for local Hölder continuity. Hence

$$\begin{aligned} \left|\frac{\hat{x} - \hat{y}}{\varepsilon} - \tau \mathbf{1}_n\right|^2 &\leq \frac{1}{2}[v_1(\hat{x}, \hat{k}) - v_1(\hat{y}, \hat{k}) + v_2(\hat{x}, \hat{k}) - v_2(\hat{y}, \hat{k})] + n\tau^2 \\ &\leq \frac{1}{2}[Hol(v_1, \gamma_1, C_\alpha)|\hat{x} - \hat{y}|^{\gamma_1} + Hol(v_2, \gamma_2, C_\alpha)|\hat{x} - \hat{y}|^{\gamma_2}] + n\tau^2 \\ &\leq (1 + C_\alpha)[Hol(v_1, \gamma_1, C_\alpha) + Hol(v_2, \gamma_2, C_\alpha)] + n\tau^2, \end{aligned} \tag{13}$$

since $\hat{x} \vee \hat{y} \leq C_\alpha$ by Lemma 4.

By use of the triangular inequality for norms, for $\tau \in (0, 1]$, we get

$$\begin{aligned} \left|\frac{\hat{x} - \hat{y}}{\varepsilon}\right| &\leq \left|\frac{\hat{x} - \hat{y}}{\varepsilon} - \tau \mathbf{1}_n\right| + |\tau \mathbf{1}_n| \\ &\leq \sqrt{(1 + C_\alpha)[Hol(v_1, \gamma_1, C_\alpha) + Hol(v_2, \gamma_2, C_\alpha)] + n\tau^2} + n\tau \\ &\leq \sqrt{(1 + C_\alpha)[Hol(v_1, \gamma_1, C_\alpha) + Hol(v_2, \gamma_2, C_\alpha)] + n} + n, \end{aligned}$$

which implies assertion (i) and subsequently (ii). This completes the proof. □

The proof of Lemmas 4 and 5 adopts the techniques in [11, 20] dealing with unbounded viscosity solutions for first order HJB equations. The next lemma is essential for deriving the comparison result in section 4.

LEMMA 6. *Let* $(\hat{x}, \hat{y}, \hat{k})$ *be given by* (12) *and* $\hat{x} = [\hat{x}_1, \ldots, \hat{x}_n]^T$, $\hat{y} = [\hat{y}_1, \ldots, \hat{y}_n]^T$. *For given* $\tau, \alpha \in (0, 1]$, *if* $\varepsilon > 0$ *is sufficiently small, we have* (i) $\hat{x}_i > \hat{y}_i$ *for* $1 \leq i \leq n - 1$, *if* $\bar{Q} = \bar{Q}_a$, *or* (ii) $\hat{x}_i > \hat{y}_i$ *for* $1 \leq i \leq n$, *if* $\bar{Q} = \bar{Q}_b$, *which further implies* $\hat{x} \in Q$ *for both cases.*

Proof. We give only the proof for assertion (i). The proof for assertion (ii) can be handled similarly. The proof is quite technical, and we break it into three steps.

Step 1. Let α and τ be given with $\bar{Q} = \bar{Q}_a$. We assume assertion (i) is invalid, and hence there exists a sequence $\varepsilon_i \downarrow 0$, $i \geq 1$, such that there is at least one (denoted as the n_ith) coordinate component satisfying

$$\hat{x}_{n_i}^{(i)} - \hat{y}_{n_i}^{(i)} \leq 0, \tag{14}$$

where $(\hat{x}^{(i)}, \hat{y}^{(i)}, \hat{k}^{(i)})$ is determined from (12) by taking $\varepsilon = \varepsilon_i$. Obviously, each n_i is picked out from the index set $\{1, 2, \ldots, n - 1\}$.

If necessary, we may take a subsequence $S_J \triangleq \{\varepsilon_{i_j}, j \geq 1\}$ such that both the coordinate index n_i and $\hat{k}^{(i)}$ take constant values along S_J. In all of the following we base the analysis on the subsequence S_J; however, to simplify the notation we simply represent S_J using the sequence $\{\varepsilon_i, i \geq 1\}$ and without loss of generality take $n_i \equiv 1$ and $\hat{k}^{(i)} = \hat{k}$.

Hence, we rewrite (14) as

$$\hat{x}_1^{(i)} - \hat{y}_1^{(i)} \leq 0, \qquad i \geq 1. \tag{15}$$

By definition, we have

$$\Phi(\hat{x}^{(i)},\hat{y}^{(i)},\hat{k}) = \sup_{\bar{Q}\times\bar{Q}\times\Theta} \Phi(x,y,k) \geq \Phi(\hat{x}^{(i)}+\chi_\delta,\hat{y}^{(i)},\hat{k}) \tag{16}$$

for any $0<\delta\leq 1$, where we denote the vector $\chi_\delta=(\delta,0,\ldots,0)^T$. Since $\hat{x}^{(i)}\in\bar{Q}$, it is clear that $\hat{x}^{(i)}+\chi_\delta\in\bar{Q}$.

Step 2. Now we show that (15) together with (16) leads to a contradiction. By (16), we have

$$\begin{aligned}
&\Phi(\hat{x}^{(i)},\hat{y}^{(i)},\hat{k})\\
&= v_1(\hat{x}^{(i)},\hat{k}) - v_2(\hat{y}^{(i)},\hat{k}) - \left|\frac{\hat{x}^{(i)}-\hat{y}^{(i)}}{\varepsilon_i}-\tau\mathbf{1}_n\right|^2 - \alpha[\zeta(\hat{x}^{(i)})+\zeta(\hat{y}^{(i)})]\\
&\geq v_1(\hat{x}^{(i)}+\chi_\delta,\hat{k}) - v_2(\hat{y}^{(i)},\hat{k}) - \left|\frac{\hat{x}^{(i)}+\chi_\delta-\hat{y}^{(i)}}{\varepsilon_i}-\tau\mathbf{1}_n\right|^2 - \alpha[\zeta(\hat{x}^{(i)}+\chi_\delta)+\zeta(\hat{y}^{(i)})],
\end{aligned}$$

which readily yields

$$\begin{aligned}
T_i \triangleq& \left|\frac{\hat{x}_1^{(i)}-\hat{y}_1^{(i)}}{\varepsilon_i}-\tau\right|^2 - \left|\frac{\hat{x}_1^{(i)}+\delta-\hat{y}_1^{(i)}}{\varepsilon_i}-\tau\right|^2\\
\leq& v_1(\hat{x}^{(i)},\hat{k}) - v_1(\hat{x}^{(i)}+\chi_\delta,\hat{k}) + \alpha\zeta(\hat{x}^{(i)}+\chi_\delta) - \alpha\zeta(\hat{x}^{(i)}).
\end{aligned} \tag{17}$$

Obviously, for $\delta\in(0,1]$ we have $|\hat{x}^{(i)}|\vee|\hat{x}^{(i)}+\chi_\delta|\leq C_\alpha+1$ by Lemma 4. Denote the constant $D_{\zeta,\alpha}=\sup_{|x|\leq C_\alpha+1}|\zeta'(x)|$.

By (15) and then using the local Hölder and local Lipschitz continuity of v_1 and ζ, respectively, it is easy to check that

$$T_i = -\frac{\delta^2}{\varepsilon_i^2} + \frac{2\delta}{\varepsilon_i}\left|\frac{|\hat{x}_1^{(i)}-\hat{y}_1^{(i)}|}{\varepsilon_i}+\tau\right| \leq \delta^{\gamma_1} Hol(v_1,\gamma_1,C_\alpha+1) + \delta\alpha D_{\zeta,\alpha},$$

where $\gamma_1\in(0,1]$ is the Hölder exponent for v_1, and therefore

$$\frac{2\delta\tau}{\varepsilon_i} \leq \frac{\delta^2}{\varepsilon_i^2} + \delta^{\gamma_1} Hol(v_1,\gamma_1,C_\alpha+1) + \delta\alpha D_{\zeta,\alpha}. \tag{18}$$

Since (18) holds for all $0<\delta\leq 1$, for the case with subscript index i, we take $\delta=\varepsilon_i^{\frac{2}{2-\gamma_1}}$ to obtain

$$2\tau\varepsilon_i^{\frac{-\gamma_1}{2-\gamma_1}} \leq 1 + Hol(v_1,\gamma_1,C_\alpha+1) + \varepsilon_i^{\frac{2(1-\gamma_1)}{2-\gamma_1}}\alpha D_{\xi,\alpha}. \tag{19}$$

Letting $i\to\infty$, since $\tau>0$ is fixed, (19) leads to

$$Hol(v_1,\gamma_1,C_\alpha+1)\geq\infty,$$

which is a contradiction since $v_1\in C^{loc}_{p,Hol}(\bar{Q}\times\Theta)$ with the exponent γ_1.

Step 3. Combining Steps 1 and 2 above, we see that the initial assumption that (i) is invalid does not hold. Hence assertion (i) is proven. Since $\hat{y}_i\geq 0$ for $1\leq i\leq n-1$, it follows that $\hat{x}_i>0$ for $i\leq n-1$, and consequently $\hat{x}\in Q$. □

Notice that in order to derive the contradiction in Step 2 of the proof, it is necessary to take τ as an independent variable such that its magnitude may be controlled separately.

4. Proof of Theorem 2. We give only the proof for Case (i) $\bar{Q} = \bar{Q}_a$, and Case (ii) $\bar{Q} = \bar{Q}_b$ can be treated without further difficulty. Let $(\hat{x}, \hat{y}, \hat{k})$ be obtained from (12). For given τ and α, by Lemma 6 we can pick a sufficiently small $\varepsilon_{\tau,\alpha}$ depending on the pair (τ, α) such that for all $0 < \varepsilon \le \varepsilon_{\tau,\alpha}$, its associated $(\hat{x}, \hat{y})$ is in the set $Q \times \bar{Q}$. In the following analysis we assume $\varepsilon \le \varepsilon_{\tau,\alpha}$ is always satisfied. In particular, $\hat{x}$ is in the open set Q.

Let $\phi_1(x) = v_2(\hat{y}) + |\frac{x-\hat{y}}{\varepsilon} - \tau \mathbf{1}_n|^2 + \alpha[\zeta(x) + \zeta(\hat{y})]$ and $\phi_2(y) = v_1(\hat{x}) - |\frac{\hat{x}-y}{\varepsilon} - \tau \mathbf{1}_n|^2 - \alpha[\zeta(\hat{x}) + \zeta(y)]$ be two test functions. Then on $\bar{Q}$, $v_1(x) - \phi_1(x)$ attains its maximum at $\hat{x} \in Q$, and $v_2(y) - \phi_2(y)$ attains its minimum at $\hat{y} \in \bar{Q}$. Hence we apply Definition 1 for viscosity sub/supersolutions to get

$$\rho v_1(\hat{x}, \hat{k}) - \inf_{u \in U_{\hat{k}}} \left\{ \left[2\left(\frac{\hat{x} - \hat{y}}{\varepsilon} - \tau \mathbf{1}_n \right) + \alpha \zeta'(\hat{x}) \right]^T F(\hat{x}, \hat{k}, u) + \Pi_\theta[v_1(\hat{x}, \cdot)](\hat{k}) + L_1(\hat{x}, \hat{k}, u) \right\} \le 0, \tag{20}$$

$$\rho v_2(\hat{y}, \hat{k}) - \inf_{u \in U_{\hat{k}}} \left\{ \left[2\left(\frac{\hat{x} - \hat{y}}{\varepsilon} - \tau \mathbf{1}_n \right) + \alpha \zeta'(\hat{y}) \right]^T F(\hat{y}, \hat{k}, u) + \Pi_\theta[v_2(\hat{y}, \cdot)](\hat{k}) + L_2(\hat{y}, \hat{k}, u) \right\} \ge 0. \tag{21}$$

The pair of inequalities (20) and (21) yields

$$\begin{aligned}
&\rho v_1(\hat{x}, \hat{k}) - \rho v_2(\hat{y}, \hat{k}) \\
&\le \inf_{u \in U_{\hat{k}}} \left\{ \left[2\left(\frac{\hat{x} - \hat{y}}{\varepsilon} - \tau \mathbf{1}_n \right) + \alpha \zeta'(\hat{x}) \right]^T F(\hat{x}, \hat{k}, u) + \Pi_\theta[v_1(\hat{x}, \cdot)](\hat{k}) + L_1(\hat{x}, \hat{k}, u) \right\} \\
&\quad - \inf_{u \in U_{\hat{k}}} \left\{ \left[2\left(\frac{\hat{x} - \hat{y}}{\varepsilon} - \tau \mathbf{1}_n \right) - \alpha \zeta'(\hat{y}) \right]^T F(\hat{y}, \hat{k}, u) + \Pi_\theta[v_2(\hat{y}, \cdot)](\hat{k}) + L_2(\hat{y}, \hat{k}, u) \right\} \\
&\le \sup_{u \in U_{\hat{k}}} 2 \left| \frac{\hat{x} - \hat{y}}{\varepsilon} - \tau \mathbf{1}_n \right| \cdot |F(\hat{x}, \hat{k}, u) - F(\hat{y}, \hat{k}, u)| \\
&\quad + \sup_{u \in U_{\hat{k}}} \alpha |F^T(\hat{x}, \hat{k}, u) \zeta'(\hat{x}) + F^T(\hat{y}, \hat{k}, u) \zeta'(\hat{y})| \\
&\quad + \sup_{u \in U_{\hat{k}}} [L_1(\hat{x}, \hat{k}, u) - L_2(\hat{y}, \hat{k}, u)] + \left\{ \Pi_\theta[v_1(\hat{x}, \cdot)](\hat{k}) - \Pi_\theta[v_2(\hat{y}, \cdot)](\hat{k}) \right\} \\
&\triangleq A_1(\varepsilon, \hat{x}, \hat{y}, \hat{k}) + A_2(\hat{x}, \hat{y}, \hat{k}) + A_3(\hat{x}, \hat{y}, \hat{k}) + A_4(\hat{x}, \hat{y}, \hat{k}).
\end{aligned} \tag{22}$$

Let α and τ be fixed first. Now in (22) we take a sequence $\varepsilon_i \downarrow 0$ with the associated $(\hat{x}_i, \hat{y}_i, \hat{k}_i)$ determined by (12). Here the subscript $i \ge 1$ in $\hat{x}_i$ is used to label the sequence and should not be confused as the index of a coordinate component. Since $|\hat{x}_i| \vee |\hat{y}_i| \le C_\alpha$ for all $i \ge 1$ by Lemma 4, there exists a subsequence denoted by

$S_{x,y,k} = \{(\hat{x}_{i_j}, \hat{y}_{i_j}, \hat{k}_{i_j}), j \geq 1\}$, which converges to a limit (x^*, x^*, k^*) in view of the fact that $\lim_{\varepsilon_i \to 0+} |\hat{x}_i - \hat{y}_i| = 0$ by Lemma 5. From (13) in the proof of Lemma 5, it is seen that

$$\left| \frac{\hat{x} - \hat{y}}{\varepsilon} - \tau \mathbf{1}_n \right| \leq \sqrt{(1 + C_\alpha)[Hol(v_1, \gamma_1, C_\alpha) + Hol(v_2, \gamma_2, C_\alpha)] + n\tau^2}, \tag{23}$$

where the fixed parameter γ_i is the Hölder exponent of $v_i \in C^{loc}_{p,Hol}(\bar{Q} \times \Theta)$, $i = 1, 2$. We combine (23) with the uniform continuity of $F(x, k, u)$ in x for $|x| \leq C_\alpha$ (see (10)) to get

$$\lim_{j \to \infty} A_1(\varepsilon_{i_j}, \hat{x}_{i_j}, \hat{y}_{i_j}, \hat{k}_{i_j}) = 0.$$

Using the continuity of F, L_1, and L_2 with respect to x, it can be checked that both $A_2(\hat{x}, \hat{y}, \hat{k})$ and $A_3(\hat{x}, \hat{y}, \hat{k})$ are continuous in the arguments $(\hat{x}, \hat{y})$. Then we have

$$\begin{aligned} \lim_{j \to \infty} (A_2 + A_3)(\hat{x}_{i_j}, \hat{y}_{i_j}, \hat{k}_{i_j}) = 2\alpha & \sup_{u \in U_{k^*}} |\zeta'(x^*) F(x^*, k^*, u)| \\ & + \sup_{u \in U_{k^*}} [L_1(x^*, k^*, u) - L_2(x^*, k^*, u)]. \end{aligned}$$

Now it readily follows from (22) that

$$\begin{aligned} \rho v_1(x^*, k^*) - \rho v_2(x^*, k^*) \leq 2\alpha & \sup_{u \in U_{k^*}} |\zeta'(x^*) F(x^*, k^*, u)| \\ & + \sup_{u \in U_{k^*}} [L_1(x^*, k^*, u) - L_2(x^*, k^*, u)] + A_4(x^*, x^*, k^*). \end{aligned} \tag{24}$$

On the other hand, for any $(x, x, k) \in \bar{Q} \times \bar{Q} \times \Theta$ and the set of parameters $(\varepsilon_{i_j}, \tau, \alpha)$, we have $\Phi(x, x, k) \leq \Phi(\hat{x}_{i_j}, \hat{y}_{i_j}, \hat{k}_{i_j})$, i.e.,

$$\begin{aligned} & v_1(x, k) - v_2(x, k) - n\tau^2 - 2\alpha\zeta(x) \\ & \leq v_1(\hat{x}_{i_j}, \hat{k}_{i_j}) - v_2(\hat{y}_{i_j}, \hat{k}_{i_j}) - \left| \frac{\hat{x}_{i_j} - \hat{y}_{i_j}}{\varepsilon_{i_j}} - \tau \mathbf{1}_n \right|^2 - \alpha(\zeta(\hat{x}_{i_j}) + \zeta(\hat{y}_{i_j})) \\ & \leq v_1(\hat{x}_{i_j}, \hat{k}_{i_j}) - v_2(\hat{y}_{i_j}, \hat{k}_{i_j}) - \alpha(\zeta(\hat{x}_{i_j}) + \zeta(\hat{y}_{i_j})). \end{aligned} \tag{25}$$

Taking $j \to \infty$ in (25) and invoking (24), we get

$$\begin{aligned} v_1(x, k) - v_2(x, k) & \leq v_1(x^*, k^*) - v_2(x^*, k^*) + n\tau^2 + 2\alpha\zeta(x) - 2\alpha\zeta(x^*) \\ & \leq 2\alpha\rho^{-1} \sup_{u \in U_{k^*}} |F(x^*, k^*, u)| \cdot |\zeta'(x^*)| \\ & \quad + \rho^{-1} \sup_{u \in U_{k^*}} [L_1(x^*, k^*, u) - L_2(x^*, k^*, u)] \end{aligned} \tag{26}$$

$$+ \rho^{-1} A_4(x^*, x^*, k^*) + n\tau^2 + 2\alpha\zeta(x) - 2\alpha\zeta(x^*). \tag{27}$$

By setting $x = x^*$ on both sides of (26), we have

$$v_1(x^*, k) - v_2(x^*, k) \leq v_1(x^*, k^*) - v_2(x^*, k^*) + n\tau^2$$

for all $k \in \Theta$, which gives

$$\begin{aligned} A_4(x^*, x^*, k^*) &= \sum_{k \neq k^*} \pi_{k^*k}[v_1(x^*, k) - v_1(x^*, k^*)] - \sum_{k \neq k^*} \pi_{k^*k}[v_2(x^*, k) - v_2(x^*, k^*)] \\ &= \sum_{k \neq k^*} \pi_{k^*k}\Big\{[v_1(x^*, k) - v_2(x^*, k)] - [v_1(x^*, k^*) - v_2(x^*, k^*)]\Big\} \\ &\leq n\tau^2 \sum_{k \neq k^*} \pi_{k^*k} = n\tau^2 |\pi_{k^*k^*}|. \end{aligned}$$

By use of the expression for $\zeta(x)$, it can be shown that

$$\rho^{-1} \sup_{k \in \Theta} \sup_{u \in U_k} |F(x, k, u)| \cdot |\zeta'(x)| \leq \zeta(x)$$

for all $x \in \mathbb{R}$, and hence it follows from (27) that

$$\begin{aligned} v_1(x, k) - v_2(x, k) &\leq n\tau^2 + 2\alpha\zeta(x) + \rho^{-1} \sup_{u \in U_{k^*}} [L_1(x^*, k^*, u) - L_2(x^*, k^*, u)] \\ &\quad + \rho^{-1} n\tau^2 |\pi_{k^*k^*}| \\ &\leq n\tau^2 + 2\alpha\zeta(x) + \rho^{-1} \sup_{\bar{Q} \times \Theta} \sup_{u \in U_k} [L_1(x, k, u) - L_2(x, k, u)] \\ &\quad + \rho^{-1} n\tau^2 \max_k |\pi_{kk}|. \end{aligned}$$

Taking $\tau \to 0+$ and then $\alpha \to 0+$, we get

$$v_1(x, k) - v_2(x, k) \leq \rho^{-1} \sup_{\bar{Q} \times \Theta} \sup_{u \in U_k} [L_1(x, k, u) - L_2(x, k, u)],$$

which completes the proof. □

5. The value function as a constrained viscosity solution. In this section we give an existence result by showing that the value function v associated with (1) and (2) gives a constrained viscosity solution. Under Definition 1, we first need to show that $v(x, k)$ is continuous in x, which is rather technical with the state constraints involved. To this end, we need some restrictions on the control set and the cost integrand in this general model. Here we take the state space to be $\bar{Q}_b$, and the case for $\bar{Q}_a$ can be treated analogously.

5.1. Hölder continuity of the value function and existence theorem. For deterministic systems, there has been a fair amount of work on continuity of infinite horizon value functions with state constraints, and usually only uniform continuity is proven; see [2] and references therein. By assuming a sufficiently large discount factor, Lipschitz continuity was obtained in [16, 12]. The proof in [12] made use of the viscosity sub/supersolution properties after showing that the value function is continuous and is the unique viscosity solution, and this method was extended to prove Hölder regularity in a state constrained diffusion model [13]. Here we take a different approach to obtain Hölder continuity by recursive upper bound estimates. Unlike [12, 16], our method does not involve the HJB equation and there is no restriction on the discount factor.

THEOREM 7. *Suppose $\bar{Q} = \bar{Q}_b$, and* (A1′)–(A2) *hold. In addition, we assume that:*

(i) *each U_k, $k \in \Theta$, is equal to the same compact set $U \subset \mathbb{R}^m$;*

(ii) *there exist positive constants* $K_i < \infty$, $i = 1, 2, 3$, *such that*

$$|F(x,u,k) - F(y,u,k)| \le K_1|x-y|, \qquad \forall\ x, y \in \mathbb{R}^n,\ u \in U,\ k \in \Theta,$$
$$\sup_{x\in\mathbb{R}^n, u\in U, k\in\Theta} |F(x,u,k)| \le K_2, \qquad \sup_{x\in\mathbb{R}^n, u\in U, k\in\Theta} |L(x,u,k)| \le K_3;$$

(iii) *there exist a continuous function* $h : \partial\bar{Q} \to U$ *and constant* $\beta_1 > 0$ *such that*

$$F_i(x, h(x), k) \ge \beta_1 \tag{28}$$

for $x \in \partial\bar{Q}$, $k \in \Theta$ *and each* $i \in \{1, \dots, n\}$, *where* $\partial\bar{Q}$ *denotes the boundary of* $\bar{Q}$, *and* F_i *is the* i*th component of* F.

Then for the value function v *defined in* (2), *we have the assertions:*

(a) v *is bounded and Hölder continuous on* $\bar{Q}$ *(w.r.t.* x*), and*

(b) v *is a unique constrained viscosity solution to* (5) *within the function class* $C^{loc}_{p,Hol}(\bar{Q} \times \Theta)$.

Remark. The proof of continuity relies on a modifying procedure, which consists of taking the control $u = h(\hat{x})$ for a short period when hitting $\hat{x} \in \partial\bar{Q}$ and switching back to a shifted version of the original control. Condition (i) ensures the admissibility of the modified control.

Remark. Condition (iii) is based on the idea of controllability on boundary initially due to Soner [22]; also see, e.g., [2, Chapter 5]. It means the state trajectory can be lifted inward at the boundary points and may be relaxed to other forms. For illustration, consider the example $\bar{Q} = [0, \infty) \times [0, \infty)$ and fix $r > 0$. Then in addition to (28) being restricted on $x \in \partial\bar{Q} \cap \{x, |x| \le r\}$, we may relax (iii) by only requiring $F_1(x, h(x), k) \ge \beta_1$ if $x = (0, x_2)$, where $x_2 \ge r$, and a similar requirement for F_2 in the case $x = (x_1, 0)$, $x_1 > r$.

Remark. We establish uniqueness in the class $C^{loc}_{p,Hol}(\bar{Q} \times \Theta)$, although v is bounded.

For the value function v and $r > 0$, define

$$\nu(r) = \sup\{|v(x,k) - v(y,k)| : |x-y| < r, \text{ and } x, y \in \bar{Q},\ k \in \Theta\}. \tag{29}$$

Before proving Theorem 7, we give the following lemma on Hölder continuity. The proof is based on recursive estimation by gradually approaching the origin with small intervals for r. As an interesting byproduct for deterministic problems, Lemma 8 implies that the uniform continuity results in [22, 2] may be strengthened to Hölder regularity.

LEMMA 8. *For* $\nu : (0, \infty) \to \mathbb{R}$ *defined in* (29), *suppose* v *is bounded and there exist constants* $C > 0$, $0 < \alpha < 1$, *and* $D > 1$ *such that*

$$\nu(r) \le Cr + \alpha\nu(Dr) \tag{30}$$

for all $r > 0$, *and let* ε_0 *be selected to satisfy either case* (i) $0 < \varepsilon_0 < -(\ln\alpha)/(\ln D)$ *and* $\varepsilon_0 \le 1$ *or case* (ii) $\varepsilon_0 = -(\ln\alpha)/(\ln D)$ *provided that* $-(\ln\alpha)/(\ln D) < 1$. *Then* v *is Hölder continuous on* $\bar{Q}$ *with exponent* ε_0, *i.e.,*

$$\sup_{k\in\Theta, x\ne y} \frac{|v(x,k) - v(y,k)|}{|x-y|^{\varepsilon_0}} < \infty.$$

Proof. Note that for sufficiently small $\varepsilon_0 > 0$, it always satisfies case (i). It is obvious that $\nu(r)$ monotonically increases with $r > 0$ and is bounded since v is

bounded. For the estimate below, it suffices to restrict r to the interval $(0,1]$. We denote $\Psi(r) = \nu(r)r^{-\varepsilon_0}$, and it follows from (30) that

$$\Psi(r) \le Cr^{1-\varepsilon_0} + \alpha D^{\varepsilon_0}\Psi(Dr). \tag{31}$$

We take the decomposition $(0,1] = \cup_{i=0}^{\infty} I_i$, where $I_i = (D^{-(i+1)}, D^{-i}]$. It is clear that $\sup_{r\in I_0}\Psi(r) < \infty$ since v is bounded. For $r \in (D^{-(i+2)}, D^{-(i+1)}]$ and $i \ge 0$, we have $Dr \in (D^{-(i+1)}, D^{-i}]$, which in conjunction with (31) gives

$$\sup_{r\in I_i}\Psi(r) \le CD^{-(1-\varepsilon_0)i} + \alpha D^{\varepsilon_0}\sup_{r\in I_{i-1}}\Psi(r), \qquad i \ge 1. \tag{32}$$

By iterating (32), it follows that

$$\sup_{r\in I_k}\Psi(r) \le C\sum_{i=0}^{k-1} D^{-(1-\varepsilon_0)(k-i)}(\alpha D^{\varepsilon_0})^i + (\alpha D^{\varepsilon_0})^k \sup_{r\in I_0}\Psi(r). \tag{33}$$

Denote $S_k = C\sum_{i=0}^{k-1} D^{-(1-\varepsilon_0)(k-i)}(\alpha D^{\varepsilon_0})^i$. For case (i), we have $\alpha D^{\varepsilon_0} < 1$ and $S_k \le C\sum_{i=0}^{\infty}(\alpha D^{\varepsilon_0})^i = C(1-\alpha D^{\varepsilon_0})^{-1}$. For case (ii), we have $\alpha D^{\varepsilon_0} = 1$, $1-\varepsilon_0 > 0$, and therefore $S_k = C\sum_{i=0}^{k-1} D^{-(1-\varepsilon_0)(k-i)} < C(D^{1-\varepsilon_0}-1)^{-1}$.

Combining cases (i) and (ii), we see that the right-hand side of (33) is bounded by a constant independent of k. Hence we conclude that

$$\sup_{k\in\Theta, 0<|x-y|\le 1}\frac{|v(x,k)-v(y,k)|}{|x-y|^{\varepsilon_0}} \le \sup_{r\in(0,1]}\Psi(r) < \infty,$$

for ε_0 determined by either case (i) or case (ii), which implies the Hölder continuity of v. □

Remark. If $\alpha D < 1$ holds, (30) implies Lipschitz continuity of v since we may take $\varepsilon_0 = 1$ for case (i).

5.2. Proof of Theorem 7. We begin by proving assertion (a), which is broken into two steps.

Step 1. Let $(z,k) \in \bar{Q}\times\Theta$ be the initial condition at $t=0$ and τ_k the first jump time of $\theta(t)$ starting from $k \in \Theta$. If k is an absorbing state of $\theta(t)$, we simply have $\tau_k \equiv \infty$. We write $\mathcal{U}_{z,k}$ as $\mathcal{U}_z$ since all $U_k = U$. Following the same method as in [22, 2], we first show that there exist a small $t^* > 0$ and a constant $C_1 > 0$ such that for all $(z,k) \in \bar{Q}\times\Theta$ and u adapted to $\mathcal{F}_t = \sigma(\theta(s), s \le t)$, there is $\bar{u} \in \mathcal{U}_z$ such that

$$|J_{t^*\wedge\tau_k}(z,k,u) - J_{t^*\wedge\tau_k}(z,k,\bar{u})| \le C_1 \sup_{0\le t\le t^*\wedge\tau_k} d(X(t,z,k,u),\bar{Q}), \tag{34}$$

where $J_{t^*\wedge\tau_k} = \int_0^{t^*\wedge\tau_k} e^{-\rho t}L(X,u,\theta)(t)dt$ with the initial condition (z,k) at $t=0$, $X(t,z,k,u)$ is the state at time t associated with the initial condition (z,k) and control u, and $d(X(t,z,k,u),\bar{Q})$ denotes the distance between the state and $\bar{Q}$ on that particular sample ω.

For proving (34), we need to determine two constants $t^*, \kappa > 0$ below. Before proceeding to do so, we set $t_0 = \tau_{z,k,u} \wedge t^*$, where we define $\tau_{z,k,u} = \inf\{0 \le t \le t^*, X(t,z,k,u) \in \partial\bar{Q}\}$, if $X(t,z,k,u)$ reaches $\partial\bar{Q}$ before t^*, or $\tau_{z,k,u} = t^*$, if $X(t,z,k,u) \in Q$ for all $t \le t^*$, and $\varepsilon = \sup_{0\le t\le t^*\wedge\tau_k} d(X(t,z,k,u),\bar{Q})$. Let u be any control adapted to $\mathcal{F}_t$. We construct the new control

$$\hat{u}(t) = u(t)1_{[0,t_0)} + h(X(t_0))1_{[t_0,t_0+\kappa\varepsilon]} + u(t-k\varepsilon)1_{(t_0+\kappa\varepsilon,\infty)}, \tag{35}$$

which is adapted to $\mathcal{F}_t$. Below we will show that $X(t,z,k,\hat{u}) \in \bar{Q}$ for all $t \le t^* \wedge \tau_k$ after t^* and κ are appropriately chosen; by repeating this construction procedure on successive small intervals covering $[0,\infty)$, we obtain $\bar{u} \in \mathcal{U}_z$ and $\bar{u} \equiv \hat{u}$ on $[0, t^* \wedge \tau_k]$. Once this is done, the nonemptiness of $\mathcal{U}_z$ and $\sup_{\bar{Q}\times U\times\Theta} |L(x,u,k)| < \infty$ implies that v is bounded on $\bar{Q}\times\Theta$.

By uniform continuity of F (w.r.t. x), there exists $\delta > 0$ such that $F_i(z, h(z_0), k) \ge \beta_1/2$ provided that $|z - z_0| \le \delta$ and $z_0 \in \partial\bar{Q}$.

We first make the restriction $t^* < \delta/(2K_2)$. If $d(z, \partial\bar{Q}) \ge \delta/2$, then $t_0 = t^*$ and $X(t,z,k,u) \in Q$ for $t \le t^*$. Now it suffices to consider the case $t_0 = \tau_{z,k,u} < t^* < \delta/(2K_2)$. It can be checked that $|X(t,z,k,\hat{u}) - X(t_0,z,k,u)| \le \delta$, and therefore $F_i(X(t,z,k,\hat{u}), h(X(t_0,z,k,u)), \theta_t) \ge \beta_1/2$, $1 \le i \le n$, for $t \le t^*$. It is obvious that $X(t,z,k,\hat{u}) \in \bar{Q}$ for all $t \le t^* \wedge (t_0 + \kappa\varepsilon)$ by the construction of $\hat{u}$; for the case $t^* \wedge \tau_k \le (t_0 + \kappa\varepsilon)$, we immediately have $X(t,z,k,\hat{u}) \in \bar{Q}$ for $t \le t^* \wedge \tau_k$.

If $t^* \wedge \tau_k > (t_0 + \kappa\varepsilon)$, we apply a similar method as in [2, pp. 272–274] to show that $X(t,z,k,\hat{u}) \in \bar{Q}$ for all $t \le t^* \wedge \tau_k$. Indeed, for $t_0 + \kappa\varepsilon \le t \le t^* \wedge \tau_k$, we may write

$$\begin{aligned}
\hat{X}_t^{(i)} &\ge X_{t_0}^{(i)} + \frac{\beta_1}{2}\kappa\varepsilon + \int_{t_0+\kappa\varepsilon}^{t} F_i(\hat{X}_s, \hat{u}_s, \theta_s)ds \\
&= X_{t_0}^{(i)} + \frac{\beta_1}{2}\kappa\varepsilon + \int_{t_0+\kappa\varepsilon}^{t} F_i(X_{s-\kappa\varepsilon}, \hat{u}_s, \theta_s)ds \\
&\quad + \int_{t_0+\kappa\varepsilon}^{t} F_i(\hat{X}_s, \hat{u}_s, \theta_s)ds - \int_{t_0+\kappa\varepsilon}^{t} F_i(X_{s-\kappa\varepsilon}, \hat{u}_s, \theta_s)ds, \qquad (36)
\end{aligned}$$

where $\hat{X}_t = X(t,z,k,\hat{u})$, $X_t = X(t,z,k,u)$, and we use the superscript i in $\hat{X}_t$, X_t to denote the ith component in the vector. Recalling the construction of $\hat{u}$ for $t_0 + \kappa\varepsilon \le t \le t^* \wedge \tau_k$, we have

$$\begin{aligned}
X_{t_0}^{(i)} + \int_{t_0+\kappa\varepsilon}^{t} F_i(X_{s-\kappa\varepsilon}, \hat{u}_s, \theta_s)ds &= X_{t_0}^{(i)} + \int_{t_0}^{t-\kappa\varepsilon} F_i(X_s, u_s, \theta_{s+\kappa\varepsilon})ds \\
&= X_{t_0}^{(i)} + \int_{t_0}^{t-\kappa\varepsilon} F_i(X_s, u_s, \theta_s)ds \\
&= X_{t-\kappa\varepsilon}^{(i)} \ge -\varepsilon, \qquad (37)
\end{aligned}$$

where the inequality in (37) holds by the definition of ε. On the other hand, by the Lipschitz continuity of F_i, we have

$$\begin{aligned}
&\left| \int_{t_0+\kappa\varepsilon}^{t} F_i(\hat{X}_s, \hat{u}_s, \theta_s)ds - \int_{t_0+\kappa\varepsilon}^{t} F_i(X_{s-\kappa\varepsilon}, \hat{u}_s, \theta_s)ds \right| \\
&\le K_1 \int_{t_0+\kappa\varepsilon}^{t} |\hat{X}_s - X_{s-\kappa\varepsilon}|ds \\
&\le K_1 |\hat{X}_{t_0+\kappa\varepsilon} - X_{t_0}| \int_{t_0+\kappa\varepsilon}^{t} e^{K_1(s-t_0-\kappa\varepsilon)}ds \qquad (38) \\
&\le |\hat{X}_{t_0+\kappa\varepsilon} - X_{t_0}|(e^{K_1(t-t_0-\kappa\varepsilon)} - 1) \le \kappa\varepsilon K_2(e^{K_1(t-t_0-\kappa\varepsilon)} - 1), \qquad (39)
\end{aligned}$$

where (38) is obtained by estimating $|\hat{X}_s - X_{s-\kappa\varepsilon}|$ via Gronwall inequality.

Hence for $t_0 + \kappa\varepsilon \le t \le t^* \wedge \tau_k$, it follows from (36), (37), and (39) that

$$\hat{X}_t^{(i)} \ge \frac{\beta_1}{2}\kappa\varepsilon - \varepsilon - \kappa\varepsilon K_2(e^{K_1(t-t_0-\kappa\varepsilon)} - 1). \qquad (40)$$

We conclude that if we take $t^* = \min\{\frac{1}{K_1}\ln(\frac{\beta_1}{4K_2}+1), \frac{\delta}{3K_2}\}$ and $\kappa = 4/\beta_1$, then $\hat{X}_t = X(t,z,k,\hat{u}) \in \bar{Q}$ for all $t \le t^* \wedge \tau_k$. This completes the construction of $\hat{u}$ and subsequently that of $\bar{u} \in \mathcal{U}_z$. The inequality (34) is obtained by use of the boundedness of L and simple integral estimates as in [2].

Step 2. Now we proceed to prove continuity of the value function. Let t^* be determined as above and $|z-y| < r$, where $z, y \in \bar{Q}$. For any $\delta_1 > 0$, by the optimality principle we may find $u \in \mathcal{U}_z$ such that

$$E[J_{t^*\wedge\tau_k}(z,k,u) + e^{-\rho(t^*\wedge\tau_k)}v(X(t^*\wedge\tau_k,z,k,u),\theta_{t^*\wedge\tau_k})] \le v(z,k)+\delta_1.$$

Based on u we construct $\bar{u} \in \mathcal{U}_y$ by use of (35). By basic estimates similar to those in [2, pp. 274–275] we can show $|X(t^*\wedge\tau_k,z,k,u) - X(t^*\wedge\tau_k,y,k,\bar{u})| \le C_2 r$ and $|J_{t^*\wedge\tau_k}(z,k,u) - J_{t^*\wedge\tau_k}(y,k,\bar{u})| \le C_3 r$ for constants $C_2 > 1, C_3 > 0$. Subsequently, we get

$$\begin{aligned} v(y,k) - v(z,k) \le\ & E[J_{t^*\wedge\tau_k}(y,k,\bar{u}) + e^{-\rho(t^*\wedge\tau_k)}v(X(t^*\wedge\tau_k,y,k,\bar{u}),\theta_{t^*\wedge\tau_k})] \\ & - E[J_{t^*\wedge\tau_k}(z,k,u) + e^{-\rho(t^*\wedge\tau_k)}v(X(t^*\wedge\tau_k,z,k,u),\theta_{t^*\wedge\tau_k})] + \delta_1. \end{aligned}$$

By arbitrariness of $\delta_1 > 0$, it follows that $\nu(r) \le C_3 r + Ee^{-\rho(t^*\wedge\tau_k)}\nu(C_2 r)$. If k is an absorbing state, we have $0 < Ee^{-\rho(t^*\wedge\tau_k)} = e^{-\rho t^*} \stackrel{\triangle}{=} \alpha_1 < 1$; otherwise, τ_k is exponentially distributed with the density function $\lambda_k e^{-\lambda_k t}$ on $[0,\infty)$, where $\lambda_k = -\pi_{kk} > 0$, and we have $Ee^{-\rho(t^*\wedge\tau_k)} = \lambda_k/(\lambda_k+\rho) + \rho e^{-(\lambda_k+\rho)t^*}/(\lambda_k+\rho) \le 1-\rho(1-e^{-\rho t^*})/(\lambda^*+\rho) \stackrel{\triangle}{=} \alpha_2 < 1$, where $\lambda^* = \max_{k\in\Theta}\{|\pi_{kk}|\}$. Hence we obtain

$$\nu(r) \le C_3 r + \alpha\nu(C_2 r), \tag{41}$$

where $\alpha = \max\{\alpha_1,\alpha_2\} < 1$ and $C_2 > 1$. This leads to Hölder continuity of v by Lemma 8.

For proving assertion (b), the verification of the constrained viscosity solution property is similar to the state unconstrained case, and we omit the details here. Uniqueness of the constrained viscosity solution follows from Theorem 3. □

Remark. For brevity, we only give a detailed proof of existence in Theorem 11 which deals with a composite mode variable, and the steps there can be adapted to this theorem in a straightforward manner to verify the constrained viscosity solution property of the value function.

Remark. For the estimation in section 5.2, it is necessary to apply truncation by the jump time τ_k; otherwise the derivation for (37) and (38) is invalid. Also note that F in the dynamics and the cost integrand L are restricted to be bounded. With a more general growth condition in x for F and L, the corresponding ODE estimates will be more challenging.

6. An optimal scheduling problem. As an application of the results in section 3, we consider a fluid buffer control problem for data traffic relay arising in communication networks; relevant background information can be found in the wireless application work [10] and references therein. Suppose a relay buffer is deployed to connect a source and a destination; see Figure 2. The incoming and outgoing links are described by two continuous time independent finite state Markov chains $y(t)$ and $z(t)$, indicating a certain channel quality. Suppose that $y(t)$ and $z(t)$ have state spaces $S_y = \{1,\dots,m_1\}$, $S_z = \{1,\dots,m_2\}$ and transition probability rate matrices $\Pi_y = (p_{ij})_{m_1\times m_1}, \Pi_z = (q_{ij})_{m_2\times m_2}$, respectively.

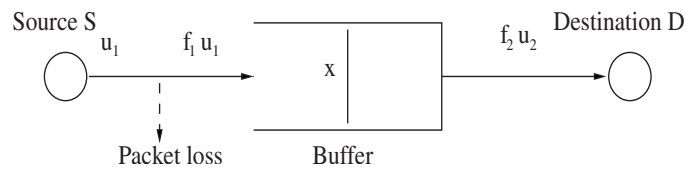


FIG. 2. *The fluid buffer model.*

Let $X \geq 0$ denote the buffer level (number of data packets), and let u_i, $i = 1, 2$, be the transmission rate (packets per second) at the incoming and outgoing links, respectively. Write the buffer level dynamics in the form:

$$\frac{dX(t)}{dt} = [u_1 f_1(y, u_1) - u_2 f_2(z, u_2)](t) \stackrel{\triangle}{=} F(y, z, u)(t), \qquad t \geq 0, \tag{42}$$

subject to $X \geq 0$. Here f_i, $i = 1, 2$, is the success probability of transmission given the link state y or z and rate u_i, and $u = [u_1, u_2]^T$. Notice that the buffer level decrease rate is only a fraction of u_2 since a packet which fails to reach the destination is not immediately deleted and will stay for retransmission. Furthermore, for limiting interference, at a given time it is allowed to transmit at only one link, either the incoming or the outgoing link [10].

We define the discounted *utility* function as

$$J_{ut}(x, i, j, u) = E\bigg[\int_0^\infty e^{-\rho t}[F_1(y, u_1) + F_2(z, u_2) - \lambda X](t)dt | X(0) = x, y(0) = i, z(0) = j\bigg], \tag{43}$$

where $\rho > 0$, $\lambda > 0$, $F_1(y, u_1) \stackrel{\triangle}{=} u_1 f_1(y, u_1)$, and $F_2(z, u_1) \stackrel{\triangle}{=} u_2 f_2(z, u_2)$. The term λX corresponds to a linear holding cost for the buffer level. The function $J^0(y, u_1, z, u_2) = F_1 + F_2$ is naturally interpreted as the instantaneous aggregate utility of the buffer in successfully transporting infinitesimal traffic volume by one hop—operating in either the receiving or the transmitting mode.

Let $U_i = [\underline{R}_i, \bar{R}_i]$, $i = 1, 2$, where $0 < \underline{R}_i < \bar{R}_i < \infty$. The control at a given time is denoted as $u = (u_1, u_2)^T \in U \stackrel{\triangle}{=} (U_1 \times \{0\}) \cup (\{0\} \times U_2)$. Define the σ-algebra $\mathcal{F}_t = \sigma(y(s), z(s), s \leq t)$.

The objective for the optimal scheduling problem is to maximize J_{ut} or, equivalently, to minimize $-J_{ut}$. Specializing the general formulation in section 2 to the current setting, we denote the admissible control set $\mathcal{U}_{x,i,j}$ with the initial condition (x, i, j) for $(X(t), y(t), z(t))$. Let $v(x, i, j)$ denote the value function for minimizing $J(x, i, j, u) \stackrel{\triangle}{=} -J_{ut}(x, i, j, u)$, where $x \in [0, \infty), i \in S_y$, and $j \in S_z$, and write $L(x, i, j, u) = -F_1(i, u_1) - F_2(j, u_2) + \lambda x$. The following assumption is used throughout this section.

(A3) f_1 (resp., f_2) is a function mapping $S_y \times U_1$(resp., $S_z \times U_2$) $\to [0, 1]$ and is continuous in u_1 (resp., u_2).

6.1. Existence and uniqueness of viscosity solutions. For a function $\varphi(x, i, j)$ continuous in $x \in [0, \infty)$, define the operator

$$[\Pi\varphi(x, \cdot, \cdot)](i, j) = \sum_{i' \neq i} p_{ii'}[\varphi(x, i', j) - \varphi(x, i, j)] + \sum_{j \neq j'} q_{jj'}[\varphi(x, i, j') - \varphi(x, i, j)],$$

where $\Pi_y = (p_{ij})_{m_1 \times m_1}$ and $\Pi_z = (q_{ij})_{m_2 \times m_2}$. For the value function v, we write the HJB equation in the compact form:

$$\begin{aligned}\rho v(x,i,j) &= [\Pi v(x,\cdot,\cdot)](i,j) + \inf_{u\in U}\Big[v_x(x,i,j)(F_1(i,u_1) - F_2(j,u_2)) + L(x,i,j,u)\Big] \\ &= H(x,i,j,v_x(x,i,j), v(x,\cdot,\cdot)). \qquad (44)\end{aligned}$$

Notice that after introducing a new set of indices for the joint Markov chain (y,z) with its associated transition probability rate matrix, (44) can be written in the standard form in section 2. The details for such a conversion are omitted here. Before proving that the value function v is a constrained viscosity solution to (44), we show that v is continuous in x.

LEMMA 9. *Let $0 \le \hat{x} < x < \infty$ be given and $(y(0), z(0)) = (i,j) \in S_y \times S_z$ be fixed. For any $u \in \mathcal{U}_{x,i,j}$, there exists $\hat{u} \in \mathcal{U}_{\hat{x},i,j}$ such that*

(i) $\sup_{t\ge 0} |\hat{X}(t) - X(t)| \le |\hat{x} - x|$, *and*

(ii) *with probability one, we have*

$$\Big| \int_0^t \big\{[F_1(y,\hat{u}_1) - F_1(y,u_1)] + [F_2(z,u_2) - F_2(z,\hat{u}_2)]\big\}(s)ds \Big| \le 2|\hat{x} - x|$$

for all $t > 0$, where $X(t)$ and $\hat{X}(t)$ are, respectively, the solution associated with the control u, $\hat{u}$ and the initial condition x, $\hat{x}$.

Proof. For $u \in \mathcal{U}_{x,i,j}$, let $X(t,\hat{x},u)$ denote the state at time t with the initial condition $\hat{x} \ge 0$ and control u. Let $\tau_1 = \inf\{t \ge 0 | X(t,\hat{x},u) = 0\}$ and $\tau_1 = \infty$ on $\{X(t,\hat{x},u) > 0,\ \forall\ t \ge 0\}$. Denote $\delta = |x - \hat{x}|/(\bar{R}_1 + \bar{R}_2)$. We construct the control $u^{(1)}$ as follows:

$$u^{(1)}(t) = \begin{cases} u(t) & \text{for } t < \tau_1, \\ [\bar{R}_1, 0]^T & \text{for } t \in [\tau_1, \tau_1 + \delta), \\ u(t) & \text{for } t \ge \tau_1 + \delta. \end{cases} \qquad (45)$$

Suppose τ_k and $u^{(k)}$, $k \ge 1$, have been constructed. Define $\tau_{k+1} = \inf\{t \ge \tau_k + \delta | X(t,\hat{x},u^{(k)}) = 0\}$ on $\{\tau_k < \infty\}$, $\tau_{k+1} = \infty$ on $\{\tau_k = \infty\} \cup \{\tau_k < \infty$, and $X(t,\hat{x},u^{(k)}) > 0,\ \forall\ t \ge \tau_k + \delta\}$; define $u^{(k+1)}$ by setting (u,τ_1) as $(u^{(k)},\tau_{k+1})$ on the right-hand side of (45). This procedure may be terminated if the stopping time τ_k at a certain stage k equals ∞ with probability one. Let $\hat{u}(t) = u^{(k)}(t)$ for $t \le \tau_{k+1}$, and it can be shown that this gives a well-defined control on $[0,\infty)$ and $\hat{u} \in \mathcal{U}_{\hat{x},i,j}$.

In (46) below, $X(t)$ and $\hat{X}(t)$ are associated with u and $\hat{u}$, respectively. By the construction of $\hat{u}$, it is easy to check that $\hat{X}(t) - X(t) \ge -|\hat{x} - x|$ for all $t \ge 0$. Now we show that for all $t \ge 0$, $\hat{X}(t) - X(t) \le |\hat{x} - x|$, which obviously holds for $t \le \tau_1$. Suppose $t \in [\tau_k, \tau_{k+1})$. Since $0 = \hat{X}(\tau_k) \le X(\tau_k)$, we have

$$\begin{aligned}\hat{X}(t) - X(t) &= \hat{X}(\tau_k) + \int_{\tau_k}^t (\hat{F}_1 - \hat{F}_2)(s)ds - X(\tau_k) - \int_{\tau_k}^t (F_1 - F_2)(s)ds \\ &\le \int_{\tau_k}^{t\wedge(\tau_k+\delta)} (\hat{F}_1 - \hat{F}_2)(s)ds - \int_{\tau_k}^{t\wedge(\tau_k+\delta)} (F_1 - F_2)(s)ds \\ &= \int_{\tau_k}^{t\wedge(\tau_k+\delta)} (\hat{F}_1 - F_1)(s)ds + \int_{\tau_k}^{t\wedge(\tau_k+\delta)} (F_2 - \hat{F}_2)(s)ds \\ &\le \bar{R}_1\delta + \bar{R}_2\delta = |\hat{x} - x|, \qquad (46)\end{aligned}$$

where we denote $F_1 = F_1(y, u_1)$, $\hat{F}_1 = F_1(y, \hat{u}_1)$, $F_2 = F_2(z, u_2)$, etc. Hence $\sup_{t\geq 0} |\hat{X}(t) - X(t)| \leq |\hat{x} - x|$, and (i) follows. On the other hand, we have

$$\hat{X}(t) - X(t) = \hat{x} - x + \int_0^t [(\hat{F}_1 - F_1) + (F_2 - \hat{F}_2)](s)ds. \tag{47}$$

By use of (47) and (i) we get

$$\sup_{t\geq 0} \Big| \int_0^t [(\hat{F}_1 - F_1) + (F_2 - \hat{F}_2)](s)ds \Big| \leq \sup_{t\geq 0} |\hat{X}(t) - X(t)| + |\hat{x} - x| \leq 2|\hat{x} - x|,$$

and (ii) follows. □

LEMMA 10. *The value function $v(x, i, j)$ is Lipschitz continuous with respect to $x \in [0, \infty)$.*

Proof. Take $0 \leq \hat{x} < x$. We need to estimate $|v(\hat{x}, i, j) - v(x, i, j)|$. For any $\varepsilon > 0$, there exists $u_\varepsilon \in \mathcal{U}_{x,i,j}$ such that $v(x, i, j) \leq J(x, i, j, u_\varepsilon) \leq v(x, i, j) + \varepsilon$. Based on u_ε, we construct $\hat{u}_\varepsilon \in \mathcal{U}_{\hat{x},i,j}$ satisfying (i) and (ii) in Lemma 9. Using the same set of notation as in (46) and noticing $\hat{F}_1 - F_1 \geq 0$, $F_2 - \hat{F}_2 \geq 0$, we have

$$\begin{aligned}
&\Big| \int_0^\infty e^{-\rho t}(\hat{F}_1 + \hat{F}_2)dt - \int_0^\infty e^{-\rho t}(F_1 + F_2)dt \Big| \\
= &\Big| \int_0^\infty e^{-\rho t}(\hat{F}_1 - F_1)dt + \int_0^\infty e^{-\rho t}(\hat{F}_2 - F_2)dt \Big| \\
\leq &\Big| \int_0^\infty e^{-\rho t}(\hat{F}_1 - F_1)dt + \int_0^\infty e^{-\rho t}(F_2 - \hat{F}_2)dt \Big| \\
\leq &\Big| \int_0^\infty [(\hat{F}_1 - F_1) + (F_2 - \hat{F}_2)]dt \Big| \leq 2|\hat{x} - x|, \qquad (48)
\end{aligned}$$

where the last inequality follows from Lemma 9(ii).

By (48) and Lemma 9(i), we can check that

$$\begin{aligned}
|J(\hat{x}, i, j, \hat{u}_\varepsilon) - J(x, i, j, u_\varepsilon)| &\leq 2|\hat{x} - x| + E\int_0^\infty e^{-\rho t}\lambda|\hat{X}(t) - X(t)|dt \\
&\leq (2 + \lambda/\rho)|\hat{x} - x|.
\end{aligned}$$

Hence $v(\hat{x}, i, j) \leq v(x, i, j) + \varepsilon + (2 + \lambda/\rho)|\hat{x} - x|$. On the other hand, suppose $\hat{u}_\varepsilon$ has been found such that $J(\hat{x}, i, j, \hat{u}_\varepsilon) \leq v(\hat{x}, i, j) + \varepsilon$; then obviously $\hat{u}_\varepsilon \in \mathcal{U}_{x,i,j}$, and we can verify that $J(x, i, j, \hat{u}_\varepsilon) \leq J(\hat{x}, i, j, \hat{u}_\varepsilon) + (\lambda/\rho)|\hat{x} - x|$ and hence $v(x, i, j) \leq v(\hat{x}, i, j) + \varepsilon + (\lambda/\rho)|\hat{x} - x|$.

Thus $|v(\hat{x}, i, j) - v(x, i, j)| \leq (2 + \lambda/\rho)|\hat{x} - x| + \varepsilon$. Since $\varepsilon > 0$ is arbitrary, we get $|v(\hat{x}, i, j) - v(x, i, j)| \leq (2 + \lambda/\rho)|\hat{x} - x|$, and the lemma follows. □

THEOREM 11. *The value function $v : [0, \infty) \times S_y \times S_z \to \mathbb{R}$ is a unique constrained viscosity solution to* (44) *in the function class $C^{loc}_{p,Lip}([0, \infty) \times S_y \times S_z)$.*

Proof. See the appendix. □

7. Application to stochastic manufacturing systems. In this section we consider production rate control involving n machines in a tandem queue with $n - 1$ buffers between neighboring machines. The associated optimal control problem has been well studied in the stochastic manufacturing literature; see [20, 21]. Let the system model be given as

$$\frac{dX(t)}{dt} = (Au + Bz)(t), \qquad t \geq 0, \tag{49}$$

where $X \in \mathbb{R}^n$, $u \in \mathbb{R}^n_+$, and $z \in \mathbb{R}_+$, and

$$A = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & & & \vdots & \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -1 \end{bmatrix}.$$

Here all upper subdiagonal entries in A are -1. The state space for X is $\bar{Q} \stackrel{\triangle}{=} [0,\infty)^{n-1} \times (-\infty,\infty)$. Notice that the last component in X is the inventory level of the final product, which may be negative and accordingly interpreted as backlog. The first $n-1$ entries in X denote the buffer levels and hence are nonnegative. The variable z denotes a finite state Markov chain describing the random demanding rate. The cost function to be minimized is of the form

$$J(x,k,z,u) = E\bigg[\int_0^\infty e^{-\rho t} L(X(t),u(t))dt | X(0)=x, k(0)=k, z(0)=z\bigg],$$

where (x,k,z) is the initial condition. Here $k(t) \in \mathbb{R}^n$ is vector Markov process with discrete values describing the machine capacity. Let the state space and generator for (k,z) be denoted by $\mathcal{C} \times \mathcal{D}$ and Π, respectively. For the initial condition (x,k,z), the admissible control set $\mathcal{U}_{x,k,z}$ consists of controls such that (i) $u(t)$ is adapted to $\mathcal{F}_t = \sigma(k(s),z(s),s \le t)$, (ii) $0 \le u(t) \le k(t)$ (holding entrywise), and (iii) $X(t) \in \bar{Q}$ at all times $t \ge 0$. We also assume

$$|L(x,u) - L(x',u')| \le C(1+|x|^d + |x'|^d)(|x-x'| + |u-u'|),$$

where $d > 0$ is a constant. For a given mode $k(t) = k = (k_1,\dots,k_n) \in \mathcal{C}$, let the machine capacity region be denoted by $U_k = \{u = (u_1,\dots,u_n)^T | 0 \le u_i \le k_i, i = 1,\dots,n\}$. Let $v(x,k,z)$ be the value function associated with the cost $J(x,k,z,u)$ and the admissible control set $\mathcal{U}_{x,k,z}$. The interested reader is referred to [20, Chapter 4] for a detailed account of this class of problems.

We write the HJB equation

$$\rho v(x,k,z) = \inf_{u \in U_k} \Big[v_x(x,k,z)(Au+Bz) + [\Pi v(x,\cdot,\cdot)](k,z) + L(x,u)\Big], \tag{50}$$

where $(x,k,z) \in \bar{Q} \times \mathcal{C} \times \mathcal{D}$ and $\Pi v(x,\cdot,\cdot)$ is determined in an obvious manner. Set

$$\tilde{H}(x,k,z,v_x(x,k,z),v(x,\cdot,\cdot),u) = v_x(x,k,z)(Au+Bz) + [\Pi v(x,\cdot,\cdot)](k,z) + L(x,u).$$

Then (50) may be written in the compact form:

$$\begin{aligned} \rho v(x,k,z) &= \inf_{u \in U_k} \tilde{H}(x,k,z,v_x(x,k,z),v(x,\cdot,\cdot),u) \\ &\stackrel{\triangle}{=} H(x,k,z,v_x(x,k,z),v(x,\cdot,\cdot)), \quad (x,k,z) \in \bar{Q} \times \mathcal{C} \times \mathcal{D}. \end{aligned} \tag{51}$$

Now we apply the results in section 3 and characterize the value function as the unique constrained viscosity solution to (51).

THEOREM 12. *The value function $v : \bar{Q} \times \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ is the unique constrained viscosity solution to the HJB equation* (51) *in the function class $C^{loc}_{p,Lip}(\bar{Q} \times \mathcal{C} \times \mathcal{D})$.*

Proof. The continuity and growth estimates have been given in [20, Chapter 4]. The viscosity sub/supersolution properties for v under Definition 1 can be verified

by a similar method as in proving Theorem 11, and uniqueness follows from Theorem 3. □

It is worthwhile to note that within our solution notion, for x on the boundary of $\bar{Q}$, the right-hand side of (50) is calculated by minimizing over U_k, and the state space constraint is not explicitly involved, which differs from [20, pp. 65–71] in dealing with state constraints. The key reason here is that the viscosity subsolution property is specified only on Q, and by use of this slightly weaker specification, we can still establish uniqueness on $\bar{Q}$ owing to the continuity of the solution.

8. Concluding remarks. In this paper we study optimal control of a class of stochastic hybrid systems with state space constraints. The notion of constrained viscosity solutions is introduced. We establish a comparison theorem for the subsolution and supersolution, and under some mild conditions for the general model, the value function is characterized as the unique constrained viscosity solution to the HJB equation. The uniqueness result obtained in the general setting is further applied to a communication buffer model and a standard manufacturing system.

For future research, it is of interest to generalize the state constrained viscosity solution analysis to systems with switch cost. To gain some motivation, we consider the fluid communication buffer model in section 6. Intuitively, a high buffer level will produce a high holding cost, and on the other hand, a very low buffer level limits the controller in choosing a more beneficial action. Hence, with a certain combination of values for the buffer level x and mode variable (y, z), the control may switch rapidly between positive u_1 and positive u_2 in order to attain or approximate the optimal cost. This leads to the so-called chattering effect, which is undesirable in practical applications. We note that this kind of chattering may also occur in manufacturing systems where the machine's operation switches between the production of multiple products [20]. It is of interest to develop numerical methods to identify the critical buffer levels where chattering may occur. Furthermore, for chattering avoidance, an effective means is to introduce a switch cost, and then one needs to deal with quasi-variational inequalities [3, 4] instead of a usual HJB equation. A detailed study of optimization and numerical computation of these hybrid systems with both state space constraints and switch cost will be reported in future work.

Appendix. Proof of Theorem 11.

It is obvious that $v \in C^{loc}_{p,Lip}([0,\infty) \times S_y \times S_z)$. It suffices to show that v is a constrained viscosity solution, and uniqueness follows from Theorem 3 since $C^{loc}_{p,Lip}([0,\infty) \times S_y \times S_z) \subset C^{loc}_{p,Hol}([0,\infty) \times S_y \times S_z)$.

We give the proof by carrying out elementary estimates. Denote $\bar{Q} = [0,\infty)$ and $Q = (0,\infty)$. After suitably labeling, we may denote the joint process $(y(t), z(t))$ by an equivalent integer-valued Markov chain $\theta(t)$ with state space $\mathcal{P} = \{1, 2, \ldots, m\}$ containing $m = m_1 \times m_2$ entries, and let the associated generator for $\theta(s)$ be $\Pi_\theta = (\pi_{ij})_{m\times m}$. All of our estimates below may easily translate into a form in terms of the process (y, z), and we omit the details. First, we show v is a subsolution on $Q \times \mathcal{P}$. The functions $v(x,k)$, $L(x,k,u)$, and $F(k,u)$, $k \geq 1$ (instead of $v(x,i,j)$, etc.), are used in an obvious manner. For any given $k_0 \in \mathcal{P}$, suppose $v(x,k_0) - \phi(x)$ attains a local maximum at $x_0 \in Q$ in a neighborhood $N_{x_0} \subset Q$, where $\phi \in C^1(\bar{Q})$. Without loss of generality, we assume $v(x_0,k_0) = \phi(x_0)$, since otherwise $\phi(x)$ may be replaced by $\phi(x) - \phi(x_0) + v(x_0,k_0)$. It is easy to check that $v(x,k_0) \leq v(x_0,k_0) - \phi(x_0) + \phi(x)$ for all $x \in N_{x_0}$.

For a given initial state $x_0 \in Q$ and any $\bar{u} \in U$, there exist a sufficiently small interval $[0, \delta]$ and an admissible control $\tilde{u}$ defined on $[0, \infty)$ such that $\tilde{u}(t) \equiv \bar{u}$ on $[0, \delta]$ and $x(t) \in N_{x_0}$ for all $0 < t \leq \delta$. For the given x_0, $\delta > 0$ may be selected independently of the control. Let τ be the first jump time of $\theta(t)$ with initial state $k_0 \in \mathcal{P}$. If k_0 is nonabsorbing, τ has an exponential probability density function $|\pi_{k_0 k_0}| e^{\pi_{k_0 k_0} t}$, $t \geq 0$. The estimates below are applicable to both nonabsorbing and absorbing k_0. By the dynamic programming principle, for $h \in (0, \delta)$, we have

$$\begin{aligned}
\phi(x_0) = v(x_0, k_0) &\leq E \int_0^h e^{-\rho s} L(X, \theta, \tilde{u})(s) ds + E e^{-\rho h} v(X(h), \theta(h)) \\
&\leq E \int_0^h e^{-\rho s} L(X, \theta, \tilde{u})(s) ds + E e^{-\rho h} v(X(h), \theta(h)) 1_{(h<\tau)} \\
&\quad + E e^{-\rho h} v(X(h), \theta(h)) 1_{(h \geq \tau)} \\
&\leq E \int_0^h e^{-\rho s} L(X, \theta, \tilde{u})(s) ds + E e^{-\rho h} v(x_0, k_0) 1_{(h<\tau)} \\
&\quad + E e^{-\rho h} [\phi(X(h)) - \phi(x_0)] 1_{(h<\tau)} + E e^{-\rho h} v(X(h), \theta(h)) 1_{(h \geq \tau)} \\
&\triangleq I_1 + I_2 + I_3 + I_4. \qquad \text{(A.1)}
\end{aligned}$$

It is easy to obtain the estimates

$$\begin{aligned}
I_1 &= L(x_0, k_0, \bar{u}) h + o(h), \\
I_2 &= [1 - \rho h + o(h)] v(x_0, k_0) e^{\pi_{k_0 k_0} h} \\
&= v(x_0, k_0) - \rho v(x_0, k_0) h + v(x_0, k_0) \pi_{k_0 k_0} h + o(h), \\
I_3 &= E e^{-\rho h} [\phi(X(h)) - \phi(x_0)] - E e^{-\rho h} [\phi(X(h)) - \phi(x_0)] 1_{(h \geq \tau)} \\
&= \phi_x(x_0) F(k_0, \bar{u}) h + o(h) + O\left(\left(E |\phi(X(h)) - \phi(x_0)|^2 \cdot E |1_{(h \geq \tau)}|^2 \right)^{\frac{1}{2}} \right) \\
&= \phi_x(x_0) F(k_0, \bar{u}) h + o(h), \\
I_4 &= E e^{-\rho h} v(x_0, \theta(h)) 1_{(h \geq \tau)} + E e^{-\rho h} [v(X(h), \theta(h)) - v(x_0, \theta(h))] 1_{(h \geq \tau)} \\
&= E e^{-\rho h} v(x_0, \theta(h)) 1_{(h \geq \tau)} \\
&\quad + O\left(\left(E |v(X(h), \theta(h)) - v(x_0, \theta(h))|^2 \cdot E |1_{(h \geq \tau)}|^2 \right)^{\frac{1}{2}} \right) \\
&= E e^{-\rho h} v(x_0, \theta(h)) 1_{(h \geq \tau)} + o(h) \qquad \text{(A.2)} \\
&= h \sum_{k \neq k_0} \pi_{k_0 k} v(x_0, k) + o(h),
\end{aligned}$$

where (A.2) is obtained by the continuity of v with respect to x. Recalling $v(x_0, k_0) = \phi(x_0)$, we get

$$\begin{aligned}
0 &\leq I_1 + I_2 + I_3 + I_4 - \phi(x_0) \\
&= L(x_0, k_0, \bar{u}) h - \rho v(x_0, k_0) h + v(x_0, k_0) h \pi_{k_0 k_0} \\
&\quad + \phi_x(x_0) F(k_0, \bar{u}) h + h \sum_{k \neq k_0} \pi_{k_0 k} v(x_0, k) + o(h) \\
&= -\rho v(x_0, k_0) h + \phi_x(x_0) F(k_0, \bar{u}) h + h \sum_{k \neq k_0} \pi_{k_0 k} [v(x_0, k) - v(x_0, k_0)] \\
&\quad + L(x_0, k_0, \bar{u}) h + o(h)
\end{aligned}$$

since $\pi_{k_0k_0} + \sum_{k\neq k_0} \pi_{k_0,k} = 0$. Letting $h \to 0$, we get the desired inequality for the viscosity subsolution since $\bar{u}$ is arbitrary.

Now we show v is also a viscosity supersolution. Suppose there exists a neighborhood N_{x_0} such that $v(x,k_0) - \phi(x)$ attains a local minimum at $x_0 \in N_{x_0} \cap \bar{Q}$ for a given $k_0 \in \mathcal{P}$; for any given $\varepsilon > 0$, we can find a sequence of admissible controls $u^{(i)}$, $i \geq 1$, such that

$$\begin{aligned} v(x_0,k_0) + \frac{\varepsilon}{i} &\geq E\int_0^{\frac{1}{i}} e^{-\rho s} L(X,\theta,u^{(i)})(s)ds + Ee^{-\frac{\rho}{i}} v(X(\tfrac{1}{i}),\theta(\tfrac{1}{i})) \\ &\triangleq I_1 + I_2 + I_3 + I_4, \end{aligned} \tag{A.3}$$

where we express the right-hand side by use of the same set of notation I_i, $1 \leq i \leq 4$, as in (A.1) with $\tilde{u}$ replaced by $u^{(i)}$. Now we give the estimates as follows:

$$\begin{aligned} I_1 + I_3 &= E\int_0^{\frac{1}{i}} e^{-\rho s} L(x_0,k_0,u^{(i)})(s)ds + Ee^{-\frac{\rho}{i}}[\phi(X(\tfrac{1}{i})) - \phi(x_0)] + o\left(\frac{1}{i}\right) \\ &= E\int_0^{\frac{1}{i}} L(x_0,k_0,u^{(i)})(s)ds + E[\phi(X(\tfrac{1}{i})) - \phi(x_0)] + o\left(\frac{1}{i}\right) \\ &= E\int_0^{\frac{1}{i}} [L(x_0,k_0,u^{(i)}) + \phi_x(X)F(k_0,u^{(i)})](s)ds + o\left(\frac{1}{i}\right) \\ &= E\int_0^{\frac{1}{i}} [L(x_0,k_0,u^{(i)}) + \phi_x(x_0)F(k_0,u^{(i)})](s)ds + o\left(\frac{1}{i}\right) \\ &\geq \frac{1}{i}\inf_{u\in U}[L(x_0,k_0,u^{(i)}) + \phi_x(x_0)F(k_0,u^{(i)})] + o\left(\frac{1}{i}\right), \end{aligned} \tag{A.4}$$

$$I_2 + I_4 = v(x_0,k_0) - \frac{\rho}{i}v(x_0,k_0) + \frac{1}{i}\sum_{k\neq k_0} \pi_{k_0k}[v(x_0,k) - v(x_0,k_0)] + o\left(\frac{1}{i}\right), \tag{A.5}$$

where the higher order term $o(\frac{1}{i})$ is derived via basic estimates using the dynamics of $X(t)$ and the Markov chain $\theta(t)$ and holds uniformly with respect to ε. Taking $i \to \infty$, it follows from (A.3)–(A.5) that

$$\rho v(x_0,k_0) + \varepsilon \geq \inf_{u\in U}\Big\{\phi_x(x_0)F(k_0,\bar{u}) + \sum_{k\neq k_0} \pi_{k_0k}[v(x_0,k) - v(x_0,k_0)] + L(x_0,k_0,\bar{u})\Big\}.$$

Since $\varepsilon > 0$ is arbitrary, it follows that v is a viscosity supersolution on $\bar{Q} \times \mathcal{P}$. □

Acknowledgments. I thank the referees and the Associate Editor for their valuable comments and suggestions, which have helped improve the quality of the paper, and thank the Corresponding Editor Professor T. E. Duncan for his efforts in handling the review process. The support of Professor R. J. Evans, Dr. S. Dey, and Dr. G. N. Nair of Department of Electrical and Electronic Engineering, The University of Melbourne, is gratefully acknowledged.

REFERENCES

[1] F. Bagagiolo and M. Bardi, *Singular perturbation of a finite horizon problem with state-space constraints*, SIAM J. Control Optim., 36 (1998), pp. 2040–2060.

[2] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[3] A. Bensoussan and J. L. Lions, *Impulse Control and Quasi-Variational Inequalities*, Bordes, Paris, 1984.
[4] I. Capuzzo-Dolcetta and L. C. Evans, *Optimal switching for ordinary differential equations*, SIAM J. Control Optim., 22 (1984), pp. 143–161.
[5] I. Capuzzo-Dolcetta and P.-L. Lions, *Hamilton-Jacobi equations with state constraints*, Trans. Amer. Math. Soc., 318 (1990), pp. 643–683.
[6] M. H. A. Davis, *Markov Models and Optimization*, Chapman and Hall/CRC, London, 1993.
[7] A. I. Elwalid and D. Mitra, *Effective bandwidth of general Markovian traffic sources and admission control of high speed networks*, IEEE/ACM Trans. Networking, 1 (1993), pp. 329–343.
[8] W. H. Fleming and Q. Zhang, *Risk-sensitive production planning of a stochastic manufacturing system*, SIAM J. Control Optim., 36 (1998), pp. 1147–1170.
[9] J. P. Hespanha, *Stochastic hybrid systems: application to communication networks*, in Hybrid Systems: Computation and Control, R. Alur and G. J. Pappas, eds., Lect. Notes Comput. Sci. 2993, Springer, Berlin, 2004, pp. 387–401.
[10] M. Huang and S. Dey, *Joint rate/power adaptation and dynamic buffer management in wireless data relay networks*, in Proceedings of the American Control Conference, Minneapolis, MN, 2006, pp. 6097–6102.
[11] H. Ishii, *Uniqueness of unbounded viscosity solutions of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 721–748.
[12] H. Ishii and S. Koike, *A new formulation of state constraint problems for first-order PDEs*, SIAM J. Control Optim., 34 (1996), pp. 554–571.
[13] H. Ishii and P. Loreti, *A class of stochastic optimal control problems with state constraint*, Indiana Univ. Math. J., 51 (2002), pp. 1167–1196.
[14] Y. Ji and H. J. Chizeck, *Controllability, stabilizability, and continuous-time Markovian jump linear quadratic control*, IEEE Trans. Automat. Control, 35 (1990), pp. 777–788.
[15] P. Loreti, *Some properties of constrained viscosity solutions of Hamilton-Jacobi-Bellman equations*, SIAM J. Control Optim., 25 (1987), pp. 1244–1252.
[16] P. Loreti and M. E. Tessitore, *Approximation and regularity results on constrained viscosity solutions of Hamilton-Jacobi-Bellman equations*, J. Math. Syst. Estim. Control, 4 (1994), pp. 467–483.
[17] D. Mitra, *Stochastic theory of a fluid model of multiple failure-susceptible producers and consumers coupled by a buffer*, Adv. Appl. Prob., 20 (1988), pp. 646–676.
[18] M. Motta, *On nonlinear optimal control problems with state constraints*, SIAM J. Control Optim., 33 (1995), pp. 1411–1424.
[19] S. Rajagopal, V. G. Kulkarni, and S. Stidham, Jr., *Optimal flow control of a stochastic fluid-flow system*, IEEE J. Sel. Areas Comm., 13 (1995), pp. 1219–1228.
[20] S. P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhäuser, Boston, 1994.
[21] S. P. Sethi and X. Y. Zhou, *Stochastic dynamic job shops and hierarchical production planning*, IEEE Trans. Automat. Control, 39 (1994), pp. 2061–2076.
[22] H. M. Soner, *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.
[23] H. M. Soner, *Optimal control with state-space constraint* II, SIAM J. Control Optim., 24 (1986), pp. 1110–1122.
[24] D. D. Sworder, *Feedback control of a class of linear systems with jump parameters*, IEEE Trans. Automat. Control, 14 (1969), pp. 9–14.
[25] M. H. Veatch and L. M. Wein, *Optimal control of a two-station tandem production/inventory system*, Oper. Res., 42 (1994), pp. 337–350.
[26] M. Xiao and T. Başar, *Optimal control of piecewise deterministic nonlinear systems with controlled transitions: viscosity solutions, their existence and uniqueness*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 4712–4717.
[27] M. Xiao and T. Başar, *Viscosity solutions of two classes of coupled HJB equations*, J. Inequal. Appl., 6 (2001), pp. 519–545.
[28] M. W. Wonham, *Random differential equations in control theory*, in Probabilistic Methods in Applied Mathematics, Vol II, A.T. Bharucha, ed., Academic Press, New York, 1971, pp. 131–213.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 1, pp. 356–378

NONLINEAR DIFFUSION GOVERNED BY MCKEAN–VLASOV EQUATION ON HILBERT SPACE AND OPTIMAL CONTROL*

N. U. AHMED†

Abstract. This paper deals with optimal feedback control of measure-valued solutions of nonlinear diffusion governed by McKean–Vlasov equations. Questions of existence, uniqueness, and regularity properties of measure-valued solutions are addressed. A class of feedback controls furnished with a weak topology is introduced, and some important topological properties of the attainable set corresponding to these controls are presented. We consider several typical control problems with objective functionals which are functions of measures and prove the existence of optimal controls.

Key words. McKean–Vlasov equation, stochastic control, attainable set, optimal control

AMS subject classifications. 47H20, 49L25, 49J20, 93E20

DOI. 10.1137/050645944

1. Introduction. Let H be a separable Hilbert space, let $\mathcal{P}(H)$ the space of probability measures defined on $\mathcal{B}(H)$ (the Borel field of subsets of H), and let $C^k(H)$ the space of k-times Fréchet differentiable functions on H. In this paper we consider the control problem associated to a measure-valued function μ_t described by the following McKean–Vlasov equation (written in weak form) in $\mathcal{P}(H)$:

$$\begin{cases} \frac{d}{dt}\langle \mu_t, \varphi\rangle = \langle \mu_t, L^u(\mu_t)\varphi\rangle \ t \in [0,T], \\ \mu_0 = \nu \in \mathcal{P}(H) \quad \text{initial data}, \end{cases} \tag{1}$$

where $\varphi \in C(H)$ is a smooth test function, $u : I \times H \to E$ is a control law, and $\{L^u(\mu) : \mu \in \mathcal{P}\}$ is a family of second-order differential operators on $C^2(H)$ defined by

$$L^u(\mu)\varphi(x) = \frac{1}{2}\mathrm{Tr}(QD^2\varphi) + (A^*D\varphi, x) + (f(x,\mu), D\varphi) + (g(x,u), D\varphi), \tag{2}$$

where $A : D(A) \subset H \longrightarrow H$ is the infinitesimal generator of a C_0 semigroup of bounded linear operators $S(t), t \geq 0$ on H; $f : H \times \mathcal{P}(H) \longrightarrow H$ and $g : H \times E \longrightarrow H$ are suitable maps where E is a separable Banach space. The operator D^k denotes the Fréchet derivative of order $k = 1, 2, \ldots$. We shall make more precise assumptions on f and g later.

Under suitable conditions as shown in section 3, for each given initial data $\nu \in \mathcal{P}(H)$ and control u, (1) has a unique measure-valued solution $\{\mu_t^u : t \in [0,T]\}$. One of the control problems studied in this paper can be stated as follows: Find a feedback control law u so that the corresponding cost functional J defined by

$$J(u) \equiv \Phi(\mu_T^u) \tag{3}$$

is minimum, where Φ is a given real-valued function on $\mathcal{P}(H)$.

Three physically motivated examples are presented in section 5 illustrating the practical relevance of the theory developed in this paper. In general, motivations

*Received by the editors November 23, 2005; accepted for publication (in revised form) December 27, 2006; published electronically April 13, 2007.
http://www.siam.org/journals/sicon/46-1/64594.html

†School of Information Technology and Engineering and Department of Mathematics, University of Ottawa, Ottawa, Ontario, K1N 6N5, Canada (ahmed@site.uottawa.ca).

for studying such control problems are twofold. First, control problems related to measure-valued functions are very interesting both mathematically and practically, and there are very few papers in the literature focusing on this area. For finite dimensional systems see [2]. Second, the above problem arises quite naturally in the study of stochastic control for nonlinear diffusion processes described by the following Ito stochastic equation in H:

$$(4)\quad \begin{cases} dX_t = AX_t dt + f(X_t, \mu_t)dt + g(X_t, u(t, X_t))dt + \sqrt{Q}dW \quad t \in [0, T], \\ \mu_t = \text{ probability law of } X_t, \\ X_0 = \xi \text{ has initial distribution } \nu, \end{cases}$$

where W is a H-valued cylindrical Wiener process, and ξ is a given H-valued random variable. For instance, one may be interested in finding a feedback control u to minimize the quantity

$$(5)\qquad E(C(X_T^u)),$$

where X_T^u denotes the value at time T of the solution process of (4) corresponding to a feedback control law u, and C is a given cost function. Using the Yosida approximation of the operator A and Ito's formula, it is easy to verify that the probability distribution of the solution process X^u satisfies the controlled McKean–Vlasov equation (1), and the cost functional (5) is a special case of (3) with $\Phi(\mu) = \langle \mu, C\rangle$. Another interesting situation is where the terminal probability distribution is required to approximate as closely as possible to a desired (target) probability measure μ_d. In this case Φ may be chosen as

$$\Phi(\mu) \equiv d_P(\mu, \mu_d),$$

where d_P is the Prohorov metric on $\mathcal{P}(H)$ or any other metric compatible with the weak topology. Since A is generally an unbounded operator, (1) needs delicate treatment. This can be done using the Ornstein–Uhlenbeck semigroup generated by the Gaussian process Y, which is given by the solution of the linear SDE

$$(6)\quad \begin{aligned} dY(t) &= AY(t)dt + \sqrt{Q}dW(t), \quad t \in I, \\ Y(0) &\equiv Y_0 = x. \end{aligned}$$

Let $\varphi \in BC(H)$ (space of bounded continuous real-valued functions on H) and define the transition (operator) semigroup $P_t, t \geq 0$, through the conditional expectation,

$$(P_t\varphi)(x) \equiv E\{\varphi(Y_t)|Y_0 = x\}, \quad t \geq 0.$$

It is clear that this is a contraction semigroup on $B(H)$ (bounded Borel measurable functions with the supnorm topology) but not strongly continuous. This is because A is generally an unbounded operator. It is known from a result of Da Prato and Zabczyk [3, 5] that if A is the generator of an exponentially stable C_0 semigroup $S(t), t \geq 0$ on H, and Q is a nuclear operator in H (or more generally $\sup_{t\geq 0} \mathrm{Tr} Q_t < \infty$, where $Q_t \equiv \int_0^t S(r)QS^*(r)dr$), then there exists an invariant measure μ^* associated with the operator P_t in the sense that

$$\mu^*(P_t\varphi) = \mu^*(\varphi) \;\forall\; \varphi \in B(H) \text{ and } t \geq 0,$$

where $\mu(\psi) \equiv \int_H \psi(x)\mu(dx)$. Using this invariant measure one can construct the Hilbert space $L_2(H, \mu^*) \supset B(H)$. Then the semigroup P_t admits a continuous extension to the Hilbert space $L_2(H, \mu^*)$ which we shall continue to denote by the same

symbol $P_t, t \geq 0$. The corresponding infinitesimal generator is the closed extension of the original generator $\mathcal{A}$ of the Markov process $\{Y\}$, where

$$\mathcal{A}\varphi(x) \equiv (1/2)\mathrm{Tr}(D^2\varphi(x)Q) + (A^*D\varphi(x), x). \tag{7}$$

We denote the closed extension by $\mathcal{A}$ itself. We introduce two other operators associated with (1) as follows:

$$F(\mu)\varphi(x) \equiv \langle f(x,\mu), D\varphi(x)\rangle_H, \quad G(u)\varphi(x) \equiv \langle g(x,u), D\varphi(x)\rangle_H \tag{8}$$

for $x \in H$ and $\varphi \in C^1(H)$. Then (1) can be written in the weak form as follows:

$$\begin{aligned} \frac{d}{dt}\mu_t(\varphi) &= \mu_t(\mathcal{A}\varphi) + \mu_t(F(\mu_t)\varphi) + \mu_t(G(u)\varphi), \\ \mu_0(\varphi) &= \nu(\varphi). \end{aligned} \tag{9}$$

Using the adjoint of the Ornstein–Uhlenbeck semigroup and the variation of constants formula, we can formulate the problem (9) as an integral equation

$$\mu_t = P_t^*\nu + \int_0^t P_{t-s}^* F^*(\mu_s)\mu_s ds + \int_0^t P_{t-s}^* G^*(u_s)\mu_s ds, \quad t \in I \tag{10}$$

on the space $M_{\lambda^2}(H)$. A solution of this equation (if one exists) is the mild solution of (9). We use this integral equation throughout the paper.

The McKean–Vlasov equation has been studied extensively in the literature starting from McKean himself [14], Dawson [6], Gärtner [11], Dawson and Gärtner [7], Léonard [13], Funaki [10], Ahmed and Ding [1, 2], and others (see the references therein). For linear diffusions in finite dimensional spaces, control theory is well developed as seen in Fleming [8] and Fleming and Soner [9]. Recently Mahmudov and McKibben [16] have considered a class of second-order evolution equations of the McKean–Vlasov type on Hilbert space proving the existence of solutions including some controllability results. Optimal control problems for nonlinear diffusions of the McKean–Vlasov type seem to have been treated only for finite dimensional systems [2]. To the knowledge of the author infinite dimensional problems seem to be untouched. Here we consider optimal control problems for nonlinear diffusions of the McKean–Vlasov type on infinite dimensional Hilbert spaces.

The rest of the paper is organized as follows. Basic notations and function spaces are introduced in section 2. In section 3, questions of the existence of solutions and their regularity properties are addressed. In section 4, some necessary topological properties of attainable sets are presented, and questions of existence of optimal controls for several control problems are studied.

2. Some basic notations. Define $\lambda(x) \equiv 1 + |x|_H = (1 + |x|)$ and introduce the family of Banach spaces

$$C_\rho(H) \equiv \left\{ \varphi \in C(H) : \| \varphi \|_\rho \equiv \sup_{x\in H} \frac{|\varphi(x)|}{\lambda^2(x)} + \sup_{x\neq y} \frac{|\varphi(x)-\varphi(y)|}{|x-y|} < \infty \right\}, \tag{11}$$

$$C_{\lambda^k}(H) \equiv \left\{ \varphi \in C(H) : \| \varphi \|_{\lambda^k} \equiv \sup_{x\in H} \frac{|\varphi(x)|}{\lambda^k(x)} < \infty \right\}, \quad k = 1, 2, \ldots, \tag{12}$$

and

$$C_\lambda^1(H) \equiv \left\{ \varphi \in C(H) : \| \varphi \|_{C_\lambda^1} \equiv \sup_{x\in H} \frac{|\varphi(x)|}{\lambda(x)} + \sup_{x\in H} \frac{|D\varphi(x)|_H}{\lambda(x)} < \infty \right\}. \tag{13}$$

It is easy to see that the embeddings $C^1_\lambda \hookrightarrow C_\lambda \hookrightarrow C_{\lambda^2}$ are continuous and that the embedding $C_\lambda \hookrightarrow C_{\lambda^2}$ is continuous and dense.

For $1 \le p < \infty$, let $M^s_{\lambda^p}(H)$ denote the Banach space of signed Borel measures on $\mathcal{B}(H)$ such that

$$\| \mu \|^p_{\lambda^p} \equiv \int_H \lambda^p(x)|\mu|(dx) < \infty,$$

and set $M_{\lambda^2}(H) \equiv M^s_{\lambda^2}(H) \cap \mathcal{P}(H)$, where $|\mu|(C), C \in \mathcal{B}(H)$, denotes the total variation of μ over C, $|\mu| = \mu_1 + \mu_2$ and $\mu = \mu_1 - \mu_2$ is the Jordan decomposition of μ, with μ_1, μ_2 being bounded positive measures. We furnish $M_{\lambda^2}(H)$ with the following metric topologies:

$$\text{(M1)}: \qquad \rho(\mu,\nu) \equiv \sup\{\langle \mu - \nu, \varphi\rangle, \varphi \in C_\rho(H) \text{ and } \| \varphi \|_\rho \le 1\}; \tag{14}$$

and

$$\text{(M2)}: \rho^*(\mu,\nu) \equiv \sup\{\langle \mu - \nu, \varphi\rangle, \varphi \in C_{\lambda^2}(H) \text{ and } \| \varphi \|_{C_{\lambda^2}(H)} \le 1\}. \tag{15}$$

Note that

$$\rho(\mu,\nu) \le \rho^*(\mu,\nu) \quad \forall \quad \mu,\nu \in M_{\lambda^2}(H).$$

Since solutions of (9) are expected to be functions of time taking values from the space of measures $M_{\lambda^2}(H)$, we need the topological spaces $C(I,(M_{\lambda^2}(H),\rho))$ and $C(I,(M_{\lambda^2}(H),\rho^*))$. It is easy to see that these are metric spaces with respect to the metric topologies: $d(\mu,\vartheta) \equiv \sup\{\rho(\mu_t,\vartheta_t), t \in I\}$ and $d^*(\mu,\vartheta) \equiv \sup\{\rho^*(\mu_t,\vartheta_t), t \in I\}$. Later in the paper we will have to consider a family of such metric spaces indexed by the end point of the interval $[0,\tau], 0 < \tau \le T$. We shall denote such spaces by

$$(C_\tau, d) \equiv C([0,\tau],(M_{\lambda^2}(H),\rho)), \quad (C_\tau, d^*) \equiv C([0,\tau],(M_{\lambda^2}(H),\rho^*)),$$

with d, d^* restricted to the time interval $[0,\tau]$. So in this notation the original spaces are denoted by (C_T, d) and (C_T, d^*), respectively. These function spaces were used in [1, 2].

3. Existence and regularity of solutions. First we introduce the class of admissible controls. Let E be a separable Banach space and denote $B_\lambda(H,E)$ to be the class of Borel measurable maps from H to E. Furnished with the norm topology

$$|u|_\lambda \equiv \sup\left\{\frac{|u(x)|_E}{\lambda(x)}, x \in H\right\},$$

this is a Banach space. Let $B(I, B_\lambda(H,E))$ denote the space of bounded Borel measurable functions from the interval I to the Banach space $B_\lambda(H,E)$. Again, furnished with the sup norm topology

$$\| u \|_\lambda \equiv \sup\big\{|u(t,\cdot)|_\lambda \equiv |u_t|_\lambda, t \in I\big\},$$

$B(I, B_\lambda(H,E))$ is also a Banach space. Note that $B(I, B_\lambda(H,E))$ is isometrically isomorphic with the space $B_\lambda(I \times H, E)$ of Borel measurable functions on $I \times H$ with values in E and furnished with the norm topology given by

$$\| u \|_{B_\lambda(I\times H,E)} \equiv \sup\left\{\frac{|u(t,x)|_E}{\lambda(x)}, (t,x) \in I \times H\right\}.$$

Later, for purposes of optimal controls, we will introduce a weaker topology on the Banach space $B_\lambda(I \times H, E)$.

Assumptions. The basic assumptions used in the paper are the following.

Assumption 1. A is the generator of an exponentially stable C_0 semigroup of operators $S(t), t \geq 0$, on H, and Q is a positive nuclear operator so that $\sup_{t\geq 0} \mathrm{Tr} Q_t < \infty$, where $Q_t \equiv \int_0^t S(r)QS^*(r)dr$ is the covariance operator of the Orstein–Uhlenbeck process Y.

Assumption 2. For all $t \geq 0$, $\mathrm{Im}S(t) \subset \mathrm{Im}Q_t^{1/2}$, and there exists a constant $c > 0$ and $\alpha \in [0, 1)$ so that $\parallel B(t) \parallel \equiv \parallel Q_t^{-1/2} S(t) \parallel \leq c/t^\alpha, t > 0$.

Assumption 3. $f : H \times (M_{\lambda^2}(H), \rho) \longrightarrow H$ satisfying the following properties: There exist $\ell, k > 0$ such that

(i): $|f(x,\mu)|_H \leq \ell(1 + |x| + \parallel \mu \parallel_\lambda),\ \forall\ x \in H, \mu \in (M_{\lambda^2}(H), \rho)$,

(ii): $|f(x,\mu) - f(y,\nu)| \leq k\big(|x-y| + \rho(\mu,\nu)\big)\ \forall x, y \in H, \mu, \nu \in (M_{\lambda^2}(H), \rho)$.

Assumption 4. $g : H \times E \longrightarrow H$ satisfying the following properties: There exist $L, K > 0$ such that

(i): $|g(x,u)| \leq L(1 + |x| + |u|_E),\ \forall\ x \in H, u \in E$,

(ii): $|g(x,u) - g(y,v)| \leq K(|x-y| + |u-v|_E),\ \forall\ x, y \in H$ and $u, v \in E$.

These are natural variations used in [1, 2] of standard assumptions used in the study of stochastic differential equations; see Da Prato and Zabczyk [3, 4, 5]. The first part of Assumption 2 is equivalent to null controllability of the linear system $\dot{x} = Ax + \sqrt{Q}u$ and the strong Feller property of the semigroup $P_t, t \geq 0$. The second part can be relaxed by simply requiring the function to be locally integrable. The first part of the assumption guarantees the existence of a unique invariant measure μ^* for the transition semigroup $P_t, t \geq 0$. For more details see Da Prato and Zabczyk [5, section 7.2]. Now we are prepared to prove the existence of a solution to (9). First we need an a priori bound.

LEMMA 3.1. *Suppose Assumptions* 1–4 *hold and let* μ *be a solution of* (9) *(weak sense) corresponding to a control* $u \in B_\lambda(I \times H, E)$ *and the initial state* $\nu \in M_{\lambda^2}(H)$. *Then for each* $t \in I$, $\mu_t \in M_{\lambda^2}(H)$, *and there exists a finite number* $R > 0$ *such that* $\parallel \mu_t \parallel_{M_{\lambda^2}(H)} \leq R$ *for all* $t \in I$.

Proof. First we replace the unbounded operator A by its Yosida approximation $A_n \equiv nAR(n, A)$, where $R(n, A)$ is the resolvent of the operator A corresponding to $n \in \rho(A) \cap N$, with $\rho(A)$ being the resolvent set of A. Then let μ^n be the corresponding weak solution of $(9)_n (\equiv (16))$, which is given by

$$\begin{aligned} \frac{d}{dt}\mu_t^n(\varphi) &= \mu_t^n(\mathcal{A}_n\varphi) + \mu_t^n(F(\mu_t^n)\varphi) + \mu_t^n(G(u)\varphi), \\ \mu_0^n(\varphi) &= \nu(\varphi), \end{aligned} \tag{16}$$

where $\mathcal{A}_n$ is the operator $\mathcal{A}$ given by the expression (7) with A replaced by A_n. Let $\{e_i\} \subset D(A)$ be a complete orthonormal basis of H. Take $\varphi_i(x) \equiv (x, e_i)^2$ and substitute in (17). By straightforward computation, one can verify that

$$\begin{aligned} \sum_{i\geq 1} \mu_t^n(\varphi_i) &= \sum_{i\geq 1} (Q_t^n e_i, e_i) \equiv \mathrm{Tr}(Q_t^n), \\ \sum_{i\geq 1} \mu_t^n(\mathcal{A}_n\varphi_i) &= \sum_{i\geq 1} (Qe_i, e_i) + 2\int_H \sum_{i\geq 1} \{(e_i, x)(e_i, A_n x)\}\mu_t^n(dx) \\ &= \mathrm{Tr}Q + \int_H (A_n x, x), \mu_t^n(dx), \end{aligned}$$

$$\sum_{i\geq 1}\mu_t^n(F(\mu_t^n)\varphi_i) = 2\int_H \sum_{i\geq 1}(f(x,\mu_t^n),e_i)(x,e_i)\mu_t^n(dx),$$
$$= 2\int_H (f(x,\mu_t^n),x)\mu_t^n(dx),$$

$$\sum_{i\geq 1}\mu_t^n(G(u)\varphi_i) = 2\int_H \sum_{i\geq 1}(g(x,u_t),e_i)(x,e_i)\mu_t^n(dx)$$
$$= 2\int_H (g(x,u_t),x)\mu_t^n(dx).$$

Substituting these in (17), we obtain the following identity:

$$\frac{d}{dt}\mathrm{Tr}(Q_t^n) = \mathrm{Tr}Q + \int_H (A_n x,x)\mu_t^n(dx)$$
$$+2\int_H (f(x,\mu_t^n),x)\mu_t^n(dx) + 2\int_H (g(x,u),x)\mu_t^n(dx),$$
$$\mathrm{Tr}Q_0^n = \mathrm{Tr}Q^\nu, \tag{17}$$

where Q^ν is given by

$$(Q^\nu h,h) \equiv \int_H (h,x)^2\nu(dx). \tag{18}$$

Since A is dissipative, it is not difficult to verify that its Yosida approximation A_n is also dissipative. Integrating (17) and employing the dissipativity property of A_n we obtain the following inequality:

$$\mathrm{Tr}Q_t^n \leq \mathrm{Tr}Q^\nu + t\ \mathrm{Tr}Q + 2\int_0^t\int_H (f(x,\mu_s^n),x)\mu_s^n(dx)ds$$
$$+2\int_0^t\int_H (g(x,u_s),x)\mu_s^n(dx)ds. \tag{19}$$

Using Assumption 3(i), one can easily verify that

$$2\left|\int_H (f(x,\mu_s^n),x)\mu_s^n(dx)\right| \leq 2\ell(1+3\mathrm{Tr}(Q_s^n)), \quad s\in I. \tag{20}$$

Since $u\in B_\lambda(I\times H,E)$ there exists a finite positive number r_u such that

$$|u_s(x)|_E \leq r_u\lambda(x) = r_u(1+|x|),\ \forall\ x\in H, s\in I.$$

Using this and Assumption 4(i), one can again verify that

$$2\left|\int_H (g(x,u_s),x)\mu_s^n(dx)\right| \leq L(1+r_u)(1+3\mathrm{Tr}Q_s^n). \tag{21}$$

Defining $a_1 \equiv 2\ell + L(1+r_u)$, $a_2 \equiv 6\ell + 3L(1+r_u)$, it follows from (19)–(21) that

$$\mathrm{Tr}Q_t^n \leq \mathrm{Tr}Q^\nu + t\mathrm{Tr}Q + a_1 t + a_2\int_0^t \mathrm{Tr}Q_s^n\ ds. \tag{22}$$

Hence by Gronwall's lemma we obtain

$$\mathrm{Tr}Q_t^n \le a_3 \exp a_2 t, \quad t \in I, \tag{23}$$

where $a_3 \equiv \mathrm{Tr}Q^\nu + T(a_1 + \mathrm{Tr}Q)$. Since the right-hand side of this expression is independent of n we have $\sup_{n\ge1} \mathrm{Tr}(Q_t^n) < \infty$. In other words,

$$\sup_{n\ge1} \int_H |x|_H^2 \mu_t^n(dx) < \infty. \tag{24}$$

Thus if $\mu_t^u, t \in I$, is any solution of (9) corresponding to any admissible control u, we have

$$\begin{aligned} \| \mu_t^u \|_{\lambda^2} &\equiv \int_H \lambda^2(x)\mu_t^u(dx) \le 2(1 + \mathrm{Tr}Q_t^u) \\ &\le 2(1 + a_3 \exp a_2 T) \equiv R < \infty, \ \forall\, t \in I = [0,T], \end{aligned} \tag{25}$$

where the constant R depends on the data $\{T, u, Q^\nu, Q\}$. This completes the proof. □

Corollary 3.2. *Suppose the assumptions of Lemma* 3.1 *hold, and let* $\Lambda \equiv \{\mu^n\} \subset (C_T, d^*)$ *be the family of probability measure-valued functions with* μ^n *solving the McKean–Vlasov equation* $(9)_n = (16)$ *over the time interval* $[0,T]$. *Then each* t*-section of* Λ *denoted by* $\Lambda(t)$ *is a relatively weakly compact subset of* $M_{\lambda^2}(H)$.

Proof. By Lemma 3.1 the set Λ is a bounded subset of (C_T, d^*). Thus each t-section $\Lambda(t) \equiv \{\mu_t^n : \mu^n \in \Lambda\}$ is a bounded subset of $M_{\lambda^2}(H)$. We have also seen that for each $t \in I$, $\sup_n \mathrm{Tr}Q_t^n < \infty$ for all positive integers n. Thus the family of covariance operators $\{Q_t^n\} (\in \mathcal{L}_N^+(H)$ the space of positive nuclear operators) are nuclear uniformly with respect to n. This means that $\Lambda(t)$ is tight and hence a relatively weakly compact subset of $M_{\lambda^2}(H)$. □

Now we are prepared to prove the existence and uniqueness of the solution of (9). We prove that it has a mild solution in the sense that the integral equation (10) has a solution. In our paper [1, Ahmed and Ding] the stochastic method, involving the Ito stochastic differential equation, was used to prove the existence of the solution. Here we use a direct analytic approach.

Theorem 3.3. *Suppose the assumptions of Lemma* 3.1 *hold. Then for every initial data* $\nu \in M_{\lambda^2}(H)$ *and control* $u \in B_\lambda(I \times H, E)$, (10) *has a unique solution* $\mu \equiv \mu^u \in C(I, (M_{\lambda^2}(H), \rho^*))$.

Proof. We use the integral equation (10) to prove existence. Let $\gamma \in (C_T, d^*)$, $u \in B_\lambda(I \times H, E)$ be fixed but arbitrary and consider the integral equation

$$\mu_t = P_t^*\nu + \int_0^t P_{t-s}^* F^*(\gamma_s)\mu_s ds + \int_0^t P_{t-s}^* G^*(u_s)\mu_s ds, \quad t \in I. \tag{26}$$

This is a linear integral equation on $M_{\lambda^2}(H)$, and it is relatively easy to verify that this equation has a unique solution $\mu \in (C_T, d^*)$. We present a brief outline of this. Define the operator Γ by

$$(\Gamma\mu)_t \equiv P_t^*\nu + \int_0^t P_{t-s}^* F^*(\gamma_s)\mu_s ds + \int_0^t P_{t-s}^* G^*(u_s)\mu_s ds, \quad t \in I. \tag{27}$$

We show that Γ has a unique fixed point in (C_T, d^*). Let $\mu, \vartheta \in (C_T, d^*)$; then

$$(\Gamma\mu)_t - (\Gamma\vartheta)_t = \int_0^t P^*_{t-s}\big(F^*(\gamma_s)(\mu_s - \vartheta_s)\big)ds + \int_0^t P^*_{t-s}\big(G^*(u_s)(\mu_s - \vartheta_s)\big)ds, \quad t \in I. \tag{28}$$

It is known from a result of Ahmed and Ding [1, Lemma 1(b), p. 79] that

$$P_t : C_\lambda(H) \longrightarrow C^1_\lambda(H)$$

and that there exists a constant $b > 0$ such that

$$\| P_t \|_{\mathcal{L}(C_\lambda(H), C^1_\lambda(H)} \leq b/t^\alpha, \quad t > 0,$$

where $\alpha \in [0, 1)$ is as given in Assumption 2. Thus $P^*_t : (C^1_\lambda(H))^* \longrightarrow (C_\lambda(H))^*$, and

$$\| P^*_t \|_{\mathcal{L}((C^1_\lambda(H))^*, (C_\lambda(H))^*)} \leq b/t^\alpha$$

also. Since $C_\lambda(H)$ is dense in $C_{\lambda^2}(H)$, we have

$$\| \Gamma\mu_t - \Gamma\vartheta_t \|_{(C_\lambda(H))^*} = \rho^*(\Gamma\mu_t, \Gamma\vartheta_t).$$

Using this and the preceding estimate it follows from (28) that

$$\rho^*(\Gamma\mu_t, \Gamma\vartheta_t) \leq \int_0^t (b/(t-s)^\alpha) \ \| (F^*(\gamma_s)(\mu_s - \vartheta_s) \|_{(C^1_\lambda)^*} \, ds + \int_0^t (b/(t-s)^\alpha) \ \| (G^*(u_s)(\mu_s - \vartheta_s) \|_{(C^1_\lambda)^*} \, ds. \tag{29}$$

Following similar steps as in [1] and using the fact that $\rho(\mu, \vartheta) \leq \rho^*(\mu, \vartheta)$ we find that

$$|\langle F^*(\gamma_s)\mu_s - F^*(\gamma_s)\vartheta_s, \varphi\rangle| \leq b_2 \rho^*(\mu_s, \vartheta_s) \| \varphi \|_{C^1_\lambda(H)}, \tag{30}$$

where $b_2 \equiv \ell(1 + R)$, and $R \equiv \sup\{|\gamma_s|_\lambda, s \in I\}$. Since $\gamma \in (C_T, d^*)$, $R < \infty$. Hence we conclude that

$$\| (F^*(\gamma_s)(\mu_s - \vartheta_s) \|_{(C^1_\lambda)^*} \ \leq \ b_2 \ \rho^*(\mu_s, \vartheta_s). \tag{31}$$

Similarly, considering the second term in (29) we find that

$$|\langle G^*(u_s)\mu_s - G^*(u_s)\vartheta_s, \varphi\rangle| = \left| \int_H (g(x, u_s), D\varphi)(\mu_s - \vartheta_s)(dx) \right| \leq L(1 + r_u)\rho^*(\mu_s, \vartheta_s)|\varphi|_{C^1_\lambda(H)}. \tag{32}$$

Hence we have

$$|G^*(u_s)\mu_s - G^*(u_s)\vartheta_s|_{(C^i_\lambda)^*} \ \leq \ b_3 \ \rho^*(\mu_s, \vartheta_s), \tag{33}$$

where $b_3 \equiv L(1 + r_u)$. Defining $b_4 \equiv b(b_2 + b_3)$ it follows from (29), (31), and (33) that

$$\rho^*(\Gamma\mu_t, \Gamma\vartheta_t) \ \leq \ b_4 \ \int_0^t (1/(t-s)^\alpha)\rho^*(\mu_s, \vartheta_s)ds, \quad t \in I. \tag{34}$$

By successive substitution of (34) into itself, after n steps one finds that, for $t \in I$,

$$\begin{aligned} &\rho^*(\Gamma^n \mu_t, \Gamma^n \vartheta_t) \\ &\leq b_4^n \prod_{m=1}^{n-1} \beta(1-\alpha, m(1-\alpha)) \int_0^t (t-s)^{n(1-\alpha)-1} \rho^*(\mu_s, \vartheta_s) ds, \end{aligned} \tag{35}$$

where $\beta(p,q)$ denotes the standard beta function. Using this expression, one can deduce that

$$d^*(\Gamma^n \mu, \Gamma^n \vartheta) \leq \ \eta_n \ d^*(\mu, \vartheta), \tag{36}$$

where η_n is given by the product

$$\eta_n \equiv \big(b_4^n T^{n(1-\alpha)}/n(1-\alpha)\big) \prod_{m=1}^{n-1} \beta(1-\alpha, m(1-\alpha)).$$

For sufficiently large n, we have $\eta_n < 1$, and hence the operator Γ^n is a contraction on (C_T, d^*). Thus by the Banach fixed point theorem Γ^n and hence Γ has a unique fixed point. This proves the existence of a unique mild solution of the linear integral equation (26) for every given $\gamma \in (C_T, d^*)$ and $u \in B_\lambda(I \times H, E)$. Now we are prepared to prove the existence of the solution of the nonlinear integral equation (10). For fixed initial data $\nu \in M_{\lambda^2}(H)$ and $u \in B_\lambda(I \times H, E)$, consider the map

$$\gamma \longrightarrow \mu^\gamma \equiv \Psi(\gamma)$$

from (C_T, d^*) to itself. This is the solution map determined by (26). Clearly it follows from the proof given above that this map is uniquely defined. For the proof of the existence of the solution of the nonlinear problem (10) it suffices to prove the existence of a fixed point of the operator Ψ. Let $\gamma, \theta \in (C_T, d^*)$, and let $\mu^\gamma \equiv \Psi(\gamma), \mu^\theta \equiv \Psi(\theta)$ denote the unique solutions of the following integral equations:

$$\mu_t^\gamma = P_t^* \nu + \int_0^t P_{t-s}^* F^*(\gamma_s) \mu_s^\gamma ds + \int_0^t P_{t-s}^* G^*(u_s) \mu_s^\gamma ds, \quad t \in I, \tag{37}$$

$$\mu_t^\theta = P_t^* \nu + \int_0^t P_{t-s}^* F^*(\theta_s) \mu_s^\theta ds + \int_0^t P_{t-s}^* G^*(u_s) \mu_s^\theta ds, \quad t \in I. \tag{38}$$

Since $\gamma, \mu^\gamma, \theta, \mu^\theta \in (C_T, d^*)$ there exists a positive number R (not necessarily the same R as used previously) such that

$$\sup\{\| \gamma_t \|_{M_{\lambda^2}}, \| \mu_t^\gamma \|_{M_{\lambda^2}}, \| \theta_t \|_{M_{\lambda^2}}, \| \mu_t^\theta \|_{M_{\lambda^2}}, t \in I\} \leq R.$$

Subtracting (38) from (37) we obtain

$$\begin{aligned} \mu_t^\gamma - \mu_t^\theta &= \int_0^t P_{t-s}^* \big(F^*(\gamma_s)\mu_s^\gamma - F^*(\theta_s)\mu_s^\theta\big) ds \\ &\quad + \int_0^t P_{t-s}^* G^*(u_s)(\mu_s^\gamma - \mu_s^\theta) ds, \quad t \in I. \end{aligned} \tag{39}$$

Using (39) and carrying out similar computations as in the first part of the proof, we obtain

$$\begin{aligned} \| \big(F^*(\gamma_s)\mu_s^\gamma - F^*(\theta_s)\mu_s^\theta\big) \|_{(C_\lambda^1)^*} &\leq \ell(1 + |\gamma_s|_\lambda)\rho^*(\mu_s^\gamma, \mu_s^\theta) + k|\mu_s^\theta|_\lambda \rho^*(\gamma_s, \theta_s) \\ &\leq \ell(1+R)\rho^*(\mu_s^\gamma, \mu_s^\theta) + kR\rho^*(\gamma_s, \theta_s), \end{aligned} \tag{40}$$

$$\| \left(G^*(u_s)[\mu_s^\gamma - \mu_s^\theta]\right) \|_{(C_\lambda^1)^*} \leq L(1 + |u_s|_\lambda)\rho^*(\mu_s^\gamma, \mu_s^\theta), \tag{41}$$

and hence

$$\begin{aligned} \rho^*(\mu_t^\gamma, \mu_t^\theta) \leq & \int_0^t (C_1/(t-s)^\alpha)\rho^*(\gamma_s, \theta_s)ds \\ & + \int_0^t (C_2/(t-s)^\alpha)\rho^*(\mu_s^\gamma, \mu_s^\theta)ds \end{aligned} \tag{42}$$

for $t > 0$, where $C_1 = bKR$ and $C_2 = b\{\ell(1+R) + L(1+ \| u \|_\lambda)\}$. For any $\tau \in (0, T]$ and any $\varrho, \sigma \in (C_T, d^*)$, define

$$d_\tau^*(\varrho, \sigma) \equiv \sup\{\rho^*(\varrho_s, \sigma_s), 0 \leq s \leq \tau\}.$$

Using this notation, one can readily deduce from (42) that

$$d_\tau^*(\mu^\gamma, \mu^\theta) \leq C_3(\tau)d_\tau^*(\gamma, \theta) + C_4(\tau)d_\tau^*(\mu^\gamma, \mu^\theta), \tag{43}$$

where

$$C_3(\tau) = (C_1/(1-\alpha))\tau^{1-\alpha}, \quad C_4(\tau) = (C_2/(1-\alpha))\tau^{1-\alpha}.$$

Since $\alpha \in [0, 1)$, it is clear that C_3 and C_4 are positive, increasing, and continuous functions of $\tau \in I$ starting from $C_3(0) = C_4(0) = 0$. Hence there exists a $\tau_1 \in (0, T]$ such that

$$C_3(\tau_1) < (1/2), \quad C_4(\tau_1) < (1/2).$$

For such a choice of τ_1, there exists a $\kappa \in (0, 1)$ such that (43) reduces to

$$d_{\tau_1}^*(\Psi(\gamma), \Psi(\theta)) \equiv d_{\tau_1}^*(\mu^\gamma, \mu^\theta) \ \leq \ \kappa \ d_{\tau_1}^*(\gamma, \theta). \tag{44}$$

Thus the map Ψ is a contraction on the metric space $(C_{\tau_1}, d_{\tau_1}^*)$, and hence it has a unique fixed point $\mu^o \in (C_{\tau_1}, d_{\tau_1}^*)$. Clearly $\mu_{\tau_1}^o \in M_{\lambda^2}(H)$. Starting with $\nu = \mu_{\tau_1}^o$ and continuing this process with the integral equation

$$\mu_t = P_{t-\tau_1}^* \mu_{\tau_1}^o + \int_{\tau_1}^t P_{t-s}^* F^*(\mu_s)\mu_s ds + \int_{\tau_1}^t P_{t-s}^* G^*(u_s)\mu_s ds$$

for $t \in (\tau_1, T]$, again we can find a nonempty interval $(\tau_1, \tau_2]$ on which the above equation has a unique solution. If $\tau_2 \geq T$, the process terminates; otherwise, it is continued. Since I is a compact interval, the process terminates in a finite number of steps. Piecing together the solutions constructed on each of the subintervals as indicated above, we obtain a unique solution μ^o defined for the entire interval I. Thus we have proved the existence of a unique solution of our original problem. We may denote this solution by $\mu^o \equiv \mu^u$ to indicate its dependence on the control. This completes the proof. □

Remark. Under an additional assumption we can prove that $w - \lim_{t \downarrow 0} \mu_t^o = \nu$. Indeed, suppose the assumptions of Theorem 3.3 hold and ν is absolutely continuous with respect to the invariant measure μ^* with the Radon–Nikodým derivative $d\nu/d\mu^* = h^* \in L_2(H, \mu^*)$. Then the solution of (10) satisfies the property

$$w - \lim_{t \downarrow 0} \mu_t^o = \nu.$$

4. Optimal control. First we introduce a weak topology on the Banach space $B_\lambda(I \times H, E)$. We recall our assumption that H is a separable Hilbert space and E is any separable Banach space. A sequence $\{u^n\}$ is said to converge weakly to an element $u \in B_\lambda(I \times H, E)$ if and only if for every $\eta \in B_\lambda(I \times H, E^*)$ and $\vartheta \in (C_T, d^*)$ we have

$$\int_{I \times H} \langle u^n(t,x) - u(t,x), \eta(t,x) \rangle \vartheta_t(dx)dt \longrightarrow 0.$$

Note that any weakly convergent sequence in the sense of the topology introduced above has a unique limit. Let U be a closed convex subset of E and r a finite positive number. Consider the set

$$\mathcal{U}_0 \equiv \Big\{ u \in B_\lambda(I \times H, E) : u(t,x) \in U \ \forall \ (t,x) \in I \times H, \\ \text{and} \ \ \| \ u \ \|_\lambda \leq r < \infty \Big\}.$$

We assume that it is furnished with the relative weak topology. Since U is a closed convex set, it follows from the Hahn–Banach theorem that $\mathcal{U}_0$ is weakly (sequentially) closed. For admissible controls we choose a weakly compact subset $\mathcal{U}_{ad}$ of the set $\mathcal{U}_0$.

For control problems considered in this paper we restrict ourselves to system (9) with g given by

$$g(x,u) = g_1(x) + g_0(x)u,$$

where $g_1 : H \longrightarrow H$ and $g_0(x) \in \mathcal{L}(E,H)$, with

$$\sup_{x \in H} \{ \| \ g_0(x) \ \|_{\mathcal{L}(E,H)} \} \leq K_0 < \infty.$$

Since g_1 can be included in f, without loss of generality we may assume that $g_1 = 0$. In this case the operator $G(u)$ is replaced by

$$G_0(u)\varphi = (g_0 u, D\varphi)_H = \langle u, g_0^* D\varphi \rangle_{E,E^*}.$$

In other words the system is linear in control. Let $\mathcal{L}_k(E,H)$ denote the class of compact operators from E to H, a subset of the space of bounded linear operators $\mathcal{L}(E,H)$. We introduce the following additional assumption on the operator-valued function g_0:

Assumption 5. $g_0 : H \longrightarrow \mathcal{L}_k(E,H)$ is bounded Borel measurable satisfying

$$\sup\{ \| \ g_0(x) \ \|_{\mathcal{L}(E,H)}, x \in H \} \leq K_0.$$

Control problems. Throughout this section we consider the system (9) replaced by the following system:

$$\begin{aligned} \frac{d}{dt}\mu_t(\varphi) &= \mu_t(\mathcal{A}\varphi) + \mu_t(F(\mu_t)\varphi) + \mu_t(G_0(u)\varphi), \\ \mu_0(\varphi) &= \nu(\varphi), \end{aligned} \tag{45}$$

where G_0 takes the place of the operator G. We are interested in the following control problems.

Problem 1. Find a control $u \in \mathcal{U}_{ad}$ that minimizes the functional

$$J(u) \equiv \Psi(\mu_T^u),$$

where Ψ is a real-valued function on $M_{\lambda^2}(H)$ and μ^u is the mild solution of system equation (45) corresponding to the control u.

Problem 2. Find a control $u \in \mathcal{U}_{ad}$ that maximizes the functional

$$J(u) \equiv F(\mu_{t_1}^u(\varphi_1), \mu_{t_2}^u(\varphi), \ldots, \mu_{t_m}^u(\varphi_m)),$$

where $F : R^m \longrightarrow \overline{R}$, $\{t_i, i = 1, 2, \ldots, m\}$ are distinct points from the interval I, and $\{\varphi_i, i = 1, 2, \ldots, m\}$ are elements of $BC(H)$, the space of bounded continuous functions on H.

Problem 3. Find a control $u \in \mathcal{U}_{ad}$ that minimizes the functional

$$J(u) \equiv \int_{I \times H} \big\{\ell_0(t, x, \mu_t^u) + |u_t(x)|_E\big\}\mu_t^u(dx)dt, \tag{46}$$

where μ^u is the mild solution of the system equation (45) corresponding to the control $u \in \mathcal{U}_{ad}$. Here $\ell_0 : I \times H \times M_{\lambda^2}(H) \longrightarrow R$ is the cost integrand.

In order to prove the existence of optimal controls we must prove that the attainable sets are closed. Denote by Ξ the family of (probability) measure-valued functions which are solutions of (45) corresponding to the admissible set of controls $\mathcal{U}_{ad}$. This is denoted by

$$\Xi \equiv \{\mu^u \in (C_T, d^*), \mu^u \text{ solves } (45) : u \in \mathcal{U}_{ad}\}.$$

Similarly, for each $t \in I$, define the attainable set as being the set of states in $M_{\lambda^2}(H)$ described by the system (45) at time t as u describes the set $\mathcal{U}_{ad}$. This is given by the t-section

$$\Xi(t) \equiv \{\mu_t^u : \mu^u \in \Xi\}$$

of the set Ξ. For attainable sets we have the following result.

THEOREM 4.1. *Consider the system* (45) *with the admissible controls* $\mathcal{U}_{ad}$ *as introduced above, and suppose that the assumptions of Theorem* 3.3 *hold and that the operator-valued function* g_0 *satisfies the hypothesis* 5. *Then for each* $t \in I$, *the attainable set* $\Xi(t)$ *is a weakly compact subset of* $M_{\lambda^2}(H)$.

Proof. We prove that the set $\Xi(t)$ is bounded and a relatively (weakly) compact subset of $M_{\lambda^2}(H)$. The proof is then concluded by demonstrating that it is also weakly closed. Considering the question of boundedness, it follows from Lemma 3.1 that

$$\sup\{\mathrm{Tr}Q_t^u, t \in I\} \leq a_3 \exp a_2 T, \tag{47}$$

where

$$a_3 = \mathrm{Tr}Q^\nu + T(a_1 + \mathrm{Tr}Q), \quad a_1 = 2\ell + L(1 + r_u), \quad a_2 = 6\ell + 3\ell(1 + r_u)$$

and

$$r_u = \| u \|_\lambda \equiv \sup\{|u(t, x)|_E/\lambda(x) : (t, x) \in I \times H\}.$$

By our choice of the set of admissible controls we have

$$\sup\{r_u, u \in \mathcal{U}_{ad}\} \leq r.$$

Hence all of the parameters $\{a_1, a_2, a_3\}$ are bounded above by a finite positive number, and therefore there exists a finite positive number $\tilde{R}$ such that

$$\sup_{u\in\mathcal{U}_{ad}}\sup_{t\in I}\{\mathrm{Tr}Q_t^u\}\leq \tilde{R}. \tag{48}$$

From this estimate two conclusions can be drawn. Since

$$\parallel \mu_t^u \parallel_{M_{\lambda^2}(H)}\leq 2(1+\mathrm{Tr}Q_t^u)$$

it follows from (48) that

$$\sup_{u\in\mathcal{U}_{ad}}\sup_{t\in I}\parallel \mu_t^u \parallel_{M_{\lambda^2}(H)}\leq 2(1+\tilde{R}).$$

This shows that the set Ξ and each of its t-sections $\Xi(t)$ are bounded. Since, for each $t\in I$,

$$\sup_{u\in\mathcal{U}_{ad}}\mathrm{Tr}(Q_t^u)\leq \tilde{R},$$

the family of covariance operators $\{Q_t^u, u\in\mathcal{U}_{ad}\}$ is a compact subset of the space $\mathcal{L}_N^+(H)$ (the space of positive, symmetric, nuclear operators on H). Then it follows from a well-known result on weak compactness of a subset of the space of probability measures on Hilbert spaces [12, Theorem 2, p. 377] that the set $\Xi(t)$ is relatively weakly compact. For compactness we must prove that it is weakly closed. Let $\sigma^n\in\Xi(t)$ with weak limit σ^o. We show that $\sigma^o\in\Xi(t)$. Since $\sigma^n\in\Xi(t)$, there exists an admissible control $u^n\in\mathcal{U}_{ad}$ and the corresponding (unique) mild solution $\mu^n\in(C_T,d^*)$ of Equation (10) with G replaced by G_0 (see Equation (45)) such that $\sigma^n=\mu_t^n$. Since $\mathcal{U}_{ad}$ is weakly compact, there exists a subsequence of the sequence $\{u_n\}$, relabeled as the original sequence, and an element $u^o\in\mathcal{U}_{ad}$ so that $u^n\xrightarrow{\tau_w}u^o$. Let $\mu^o\in(C_T,d^*)$ denote the solution of (10) corresponding to the control u^o. We show that $\mu_t^n\xrightarrow{\rho^*}\mu_t^o$. It follows from (10) that

$$\begin{aligned}\langle(\mu_t^n-\mu_t^o),\varphi\rangle &= \int_0^t\langle\big[F^*(\mu_s^n)\mu_s^n-F^*(\mu_s^o)\mu_s^o\big],P_{t-s}\varphi\rangle ds\\ &\quad+\int_0^t\langle\big[G_0^*(u_s^n)\mu_s^n-G_0^*(u_s^o)\mu_s^o\big],P_{t-s}\varphi\rangle ds.\end{aligned} \tag{49}$$

We split this into several parts as follows:

$$\begin{aligned}\langle(\mu_t^n-\mu_t^o),\varphi\rangle &= \int_0^t\int_H\langle(f(x,\mu_\theta^n)-f(x,\mu_\theta^o),DP_{t-\theta}\varphi\rangle\mu_\theta^n(dx)d\theta\\ &\quad+\int_0^t\int_H\langle f(x,\mu_\theta^o),DP_{t-\theta}\varphi\rangle(\mu_\theta^n-\mu_\theta^o)(dx)d\theta\\ &\quad+\int_0^t\int_H\langle g_0u_\theta^n,DP_{t-\theta}\varphi\rangle(\mu_\theta^n-\mu_\theta^o)(dx)d\theta\\ &\quad+\int_0^t\int_H\langle u_\theta^n-u_\theta^o,g_0^*DP_{t-\theta}\varphi\rangle\mu_\theta^o(dx)d\theta.\end{aligned} \tag{50}$$

Following similar computations as in the first part of the proof of Theorem 3.3, it follows from (50) that for each $t\in I$ we have

$$\rho^*(\mu_t^n,\mu_t^o)\leq\int_0^t(C/(t-\theta)^\alpha)\rho^*(\mu_\theta^n,\mu_\theta^o)d\theta+E_n(t), \tag{51}$$

where the constant C is dependent only on the parameters $\{\ell, k, r, \tilde{R}, b, K_0, T\}$, and the function E_n is given by

$$E_n(t) \equiv \sup\{e_n(t,\varphi) : \varphi \in B_1(C_\lambda(H))\},$$

with e_n given by

$$e_n(t,\varphi) = \int_0^t \int_H \langle u_\theta^n - u_\theta^o, g_0^* DP_{t-\theta}\varphi\rangle_{E,E^*} \mu_\theta^o(dx)d\theta. \tag{52}$$

It is well known that a bounded linear operator is compact if and only if its adjoint is compact. Thus $g_0^*(x)$ is compact for each $x \in H$, and hence $g_0^* : H \longrightarrow \mathcal{L}_k(H, E^*)$. Since u^n converges weakly to u^o, $g_0^*(x)$ is compact, and $P_t : C_\lambda(H) \longrightarrow C_\lambda^1(H)$, it follows from (52) that

$$e_n(t,\varphi) \longrightarrow 0 \quad \text{uniformly with respect to} \quad \varphi \in B_1(C_\lambda(H)).$$

Hence for any $t \in I$ we have

$$\lim_{n\to\infty} E_n(t) = 0. \tag{53}$$

Using (53) in (51) one can easily verify that for any $t \in I$

$$\lim_{n\to\infty} d_t^*(\mu^n, \mu^o) = 0. \tag{54}$$

Thus, in particular, $\mu_t^n \xrightarrow{w} \mu_t^o$. This, combined with the uniqueness of the weak limit, implies the identity $\sigma^o = \mu_t^o$ and therefore $\sigma^o \in \Xi(t)$ and so $\Xi(t)$ is weakly closed. Thus for each $t \in I$ the attainable set $\Xi(t)$ is a weakly compact subset of $M_{\lambda^2}(H)$. This completes our proof. □

The question of continuity of the map $u \longrightarrow \mu^u$ is very important in the study of control problems. As a corollary of the previous result we have the following result.

COROLLARY 4.2. *For the system* (45)*, the solution map $u \longrightarrow \mu^u$ is sequentially continuous with respect to the weak topology on $B_\lambda(I \times H, E)$ and the metric topology on $C(I, M_{\lambda^2}(H), \rho^*)$).*

Proof (outline). Let $u^n \in B_\lambda(I \times H, E)$, and suppose it converges to $u^o \in B_\lambda(I \times H, E)$ in the weak topology with μ^n and μ^o being the corresponding solutions of (45). Then following similar steps as in the preceding theorem one arrives at the same expression (51) leading to the conclusion (54). Since this is valid for every $t \in I$ and $d^* = d_T^*$, we have

$$\lim_{n\to\infty} d^*(\mu^n, \mu^o) \longrightarrow 0.$$

This completes the outline of the proof. □

Now we consider the following terminal control problem.

Problem 1. Find $u \in \mathcal{U}_{ad}$ that minimizes the functional

$$J(u) \equiv \Psi(\mu_T^u).$$

THEOREM 4.3. *Consider the system* (45) *and suppose the assumptions of Theorem* 4.1 *hold. Let Ψ be a weakly lower semicontinuous function on $M_{\lambda^2}(H)$ and*

$$\inf\{\Psi(\mu), \mu \in \Xi(T)\} > -\infty.$$

Then there exists an optimal control that solves Problem 1.

Proof. By hypothesis $\inf\{\Psi(\mu), \mu \in \Xi(T)\} \equiv m_0 > -\infty$. Let $\{\nu^n\}$, from the attainable set $\Xi(T)$, be a minimizing sequence for Ψ. Since this set is compact in the weak topology, there exists a subsequence of the sequence $\{\nu^n\}$, relabeled as the original sequence, and an element $\nu^o \in \Xi(T)$ so that $\nu^n \xrightarrow{w} \nu^o$. Now it follows from lower semicontinuity of Ψ and the minimizing property of the sequence $\{\nu^n\}$ that

$$\Psi(\nu^o) \leq \underline{\lim}_{n\to\infty}\Psi(\nu^n) = \lim_{n\to\infty} \Psi(\nu^n) = m_0.$$

Since $\nu^o \in \Xi(T)$, it is evident that $\Psi(\nu^o) \geq m_0$. Combining the above inequalities we obtain $\Psi(\nu^o) = m_0$. Thus Ψ attains its minimum on $\Xi(T)$. Since $\Xi(T)$ is the attainable set, there exists a control $u^o \in \mathcal{U}_{ad}$ so that the corresponding solution μ^{u^o} has the terminal value $\mu_T^{u^o} = \nu^o$. This proves the existence of an optimal control. □

Examples. Some simple examples of Ψ are the following.

Example 1. $\Psi(\mu) = \int_H V(x)\mu(dx)$ for $V \in C_{\lambda^2}(H)$.

Example 2. $\Psi(\mu) \equiv \mu(D)$, where D is any open set in H.

Example 3. $\Psi(\mu) \equiv \rho^*(\mu_d, \mu)$, where μ_d is the target measure. Clearly this is a continuous function.

Another control problem of similar nature is the following.

Problem 2. Find a control that maximizes the functional

$$J(u) \equiv F(\mu_{t_1}^u(\varphi_1), \mu_{t_2}^u(\varphi_2), \ldots, \mu_{t_m}^u(\varphi_m)),$$

where $\{t_1, t_2, \ldots, t_m\}$ are distinct points from the set I, $\{\varphi_i, i = 1, 2, \ldots m\}$ are elements of $BC(H)$, and

$$\mu(\varphi) \equiv \int_H \varphi(x)\mu(dx), \quad \mu \in M_{\lambda^2}(H).$$

THEOREM 4.4. *Consider the system* (45) *with the control problem* 2*, and suppose the assumptions of Theorem* 4.1 *hold. Let* $F : R^m \longrightarrow \overline{R}$ *be an upper semicontinuous function bounded on bounded sets and* $\{\varphi_i\} \in BC(H)$*. Then Problem* 2 *has a solution.*

Proof. Upper semicontinuity of F on R^m implies upper semicontinuity of J on $\mathcal{U}_{ad}$ in the weak topology. This follows from similar arguments as in the proof of Theorem 4.3 for a single index T. Since F is bounded on bounded sets of R^m and $\sup\{|\mu_t^u(\varphi_i)|, u \in \mathcal{U}_{ad}\} < \infty$ for each $t \in I$, we have

$$\sup\{J(u), u \in \mathcal{U}_{ad}\} < \infty.$$

The existence of optimal control now follows from weak compactness of the set of admissible controls $\mathcal{U}_{ad}$ and upper semicontinuity of J. □

One physical interpretation of Problem 2 is that it demands a control that maximizes the probability of visiting (or hitting) certain sites at specified points of time. Considering nonnegative functions $\{\varphi_i\}$, the sites are determined by the supports of the functions $\{\varphi_i\}$.

Now we consider the Lagrange problem 3.

THEOREM 4.5. *Consider the system* (45) *and the Lagrange problem* 3*, and suppose the assumptions of Theorem* 4.1 *hold with the exception that* E *is now a separable Hilbert space. Further suppose* $\ell_0 : I \times H \times M_{\lambda^2}(H) \longrightarrow R$ *is a nonnegative map satisfying the following conditions: There exists a positive number* L_0 *such that*

(C1): $|\ell_0(t, x, \nu) - \ell_0(t, x, \vartheta)| \leq L_0\ \rho^*(\nu, \vartheta)\ \forall \nu, \vartheta \in M_{\lambda^2}(H), (t, x) \in I \times H$;

(C2): $|\ell_0(t, x, \nu)| \leq L_0\ \left(\lambda^2(x) + |\nu|_\lambda^2\right)\ \forall\ (t, x) \in I \times H,$ *and* $\nu \in M_{\lambda^2}(H)$;

(C3): *$\nu \longrightarrow \ell_0(t,x,\nu)$ is lower semicontinuous on $(M_{\lambda^2}(H),\rho^*)$ for all $(t,x) \in I \times H$.*

Then there exists an optimal control for the Lagrange problem 3.

Proof. Since $\mathcal{U}_{ad}$ is weakly compact and $J(u) \geq 0$ for all $u \in \mathcal{U}_{ad}$, it suffices to prove that the functional $u \longrightarrow J(u)$, as given by (46), is lower semicontinuous. Let $\{u^n\}$ be a sequence of controls from $\mathcal{U}_{ad}$ which converges in the weak topology to an element $u^o \in \mathcal{U}_{ad}$, and let $\{\mu^n\}$ and μ^o denote the corresponding (mild) solutions of (45) associated with the controls $\{u^n\}$ and u^o, respectively. It follows from Corollary 4.2 that, along a subsequence if necessary, $\mu^n \xrightarrow{d^*} \mu^o$. Since $\{\mu^n\}$ is a convergent sequence and $\mu^o \in (C_T, d^*)$, there exists a finite positive number R (not necessarily the same as the previous ones) such that

$$\sup\{|\mu_t^n|_{\lambda^2}, |\mu_t^o|_{\lambda^2}, t \in I\} \leq R.$$

Consider the objective functional (46) decomposed into two parts $J(u) \equiv J_1(u)+J_2(u)$ evaluated at u^o. Focusing on the first part, we can rewrite this as follows:

$$\begin{aligned} J_1(u^o) &= \int_{I\times H} \ell_0(t,x,\mu_t^o)(\mu_t^o - \mu_t^n)(dx)dt \\ &\quad + \int_{I\times H} \big(\ell_0(t,x,\mu_t^o) - \ell_0(t,x,\mu_t^n)\big)\mu_t^n(dx)dt + J_1(u^n). \end{aligned} \tag{55}$$

Denoting the first term on the right by $J_{1,1}$ and the second by $J_{1,2}$, it is easy to verify that

$$|J_{1,1}| \leq L_0(1+R^2)\int_I \rho^*(\mu_t^n,\mu_t^o)\ dt, \tag{56}$$

$$|J_{1,2}| \leq L_0 R \int_I \rho^*(\mu_t^o,\mu_t^n)\ dt. \tag{57}$$

By Corollary 4.2, weak convergence of u^n to u^o implies d^* convergence of μ^n to μ^o. Thus $\rho^*(\mu_t^n,\mu_t^o) \longrightarrow 0$ for each $t \in I$. Further, as seen in Theorem 4.1, the set Ξ is bounded, and hence there exists a finite positive number $\hat{R}$ such that $\rho^*(\mu_t^n,\mu_t^o) \leq 2\hat{R}$ for all $n \in N$ and for all $t \in I$. Hence by the Lebesgue dominated convergence theorem

$$\int_I \rho^*(\mu_t^n,\mu_t^o)dt \longrightarrow 0 \quad \text{as } n \to \infty. \tag{58}$$

Clearly it follows from (55)–(58) that

$$J_1(u^o) \leq \liminf_{n\to\infty} J_1(u^n). \tag{59}$$

Now consider the part J_2 evaluated at u^o given by

$$J_2(u^o) \equiv \int_{I\times H} |u_t^o(x)|_E \ \ \mu_t^o(dx)dt. \tag{60}$$

By the Riesz representation theorem for Hilbert spaces, there exists a Borel measurable function $v_t^o(x)$ such that $|v_t^o(x)|_E = 1$ for all $(t,x) \in I \times H$. In fact one can choose

$$v_t^o(x) = \begin{cases} u_t^o(x)/|u_t^o(x)| & \text{if } |u_t^o(x)|_E \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly this is a Borel measurable function since u^o is. Thus J_2 can be written as

$$\begin{aligned} J_2(u^o) &\equiv \int_{I\times H} \langle u_t^o(x), v_t^o(x)\rangle_E \ \ \mu_t^o(dx)dt \\ &= \int_{I\times H} \langle u_t^o(x) - u_t^n(x), v_t^o(x)\rangle_E \ \ \mu_t^o(dx)dt \\ &\quad + \int_{I\times H} \langle u_t^n(x), v_t^o(x)\rangle_E \ \ (\mu_t^o - \mu_t^n)(dx)dt \\ &\quad + \int_{I\times H} \langle u_t^n(x), v_t^o(x)\rangle_E \ \ \mu_t^n(dx) \\ &\equiv J_{2,1} + J_{2,2} + J_{2,3}. \end{aligned} \tag{61}$$

Since $u^n \xrightarrow{\tau_w} u^o$, the first term converges to zero as $n \to \infty$. For the second term, it follows from boundedness of the set $\mathcal{U}_{ad}$ that

$$|J_{2,2}| \le r \int_I \rho^*(\mu_t^o, \mu_t^n)dt. \tag{62}$$

Hence again by the dominated convergence theorem we have

$$\int_I \rho^*(\mu_t^n, \mu_t^o)dt \longrightarrow 0 \quad \text{as } n \to \infty. \tag{63}$$

For the third term we have

$$|J_{2,3}| \le J_2(u^n). \tag{64}$$

Combining (60)–(63) we find that

$$J_2(u^o) \le \liminf_{n\to\infty} J_2(u^n). \tag{65}$$

Thus it follows from (59) and (65) that

$$J(u^o) \le \liminf_{n\to\infty} J(u^n),$$

and hence the functional J given by (46) is weakly lower semicontinuous on $\mathcal{U}_{ad}$. This completes the proof. □

Another control problem of significant interest is to find a control that maximizes the probability of following a moving (set-valued) target as closely as possible. Let $\Upsilon(t) \subset H$ denote a set-valued function (target). The problem is to find a control that maximizes the objective functional given by

$$J(u) \equiv \int_I w(t)\ \mu_t^u(\Upsilon(t))dt, \tag{66}$$

where w is a nonnegative weighting function.

THEOREM 4.6. *Consider the system* (45)*, and let* $t \to \Upsilon(t)$ *be a nonempty set-valued function with closed values in* H *and continuous in the Hausdorff metric, and let* w *be a continuous nonnegative real-valued function. Then there exists a control* $u^o \in \mathcal{U}_{ad}$ *at which* $J(u)$ *given by* (66) *attains its maximum.*

Proof. Clearly

$$0 \le J(u) \le \int_I w(t)dt < \infty.$$

Thus it suffices to prove that J is upper semicontinuous. Let $\{u^n\} \in \mathcal{U}_{ad}$, and suppose $u^n \xrightarrow{\tau_w} u^o$. Since $\mathcal{U}_{ad}$ is τ_w compact, $u^o \in \mathcal{U}_{ad}$. Let $\{\mu^n, \mu^o\}$ denote the corresponding solutions of (45). Then by Corollary 4.2, along a subsequence if necessary, $\mu^n \xrightarrow{d^*} \mu^o$. Since $\Upsilon(t)$ is closed-valued and $\mu_t^n \xrightarrow{w} \mu_t^o$ it follows from [15, Theorem 6.1, p. 40] that

$$\limsup_{n\to\infty} \mu_t^n(\Upsilon(t)) \le \mu_t^o(\Upsilon(t)) \ \forall\ t \in I.$$

Since w is nonnegative, it follows from the above inequality that

$$\limsup_{n\to\infty} w(t)\mu_t^n(\Upsilon(t)) \le w(t)\mu^o(\Upsilon(t)) \ \forall\ t \in I$$

also. Hence

$$\begin{aligned} \limsup_{n\to\infty} \int_I \{w(t)\mu_t^n(\Upsilon(t))\}dt &\le \int_I \limsup_{n\to\infty}\{w(t)\mu_t^n(\Upsilon(t))\}dt \\ &\le \int_I \{w(t)\mu_t^o(\Upsilon(t))\}dt. \end{aligned} \tag{67}$$

Clearly it follows from the definition of J (see (66)) and the inequality (67) that

$$\limsup_{n\to\infty} J(u^n) \le J(u^o),$$

proving upper semicontinuity of J. Hence there exists an admissible control maximizing J. This completes the proof. □

Remark 1. Theorem 4.6 holds under much more general conditions than those stated. For example, $w \in L_1^+(I)$ and $t \longrightarrow \Upsilon(t)$ a measurable multifunction with closed values.

Remark 2. Assuming continuity of the multifunction $t \longrightarrow \Upsilon(t)$ from I to $2^H \setminus \emptyset$ we can also admit J of the form

$$J(u) \equiv \int_I \mu_t^u(\Upsilon(t))\ \rho(dt)$$

provided ρ is a countably additive bounded positive measure on I having bounded total variation. For example, ρ may be given by a weighted sum of a finite number of Dirac measures.

Time optimal control is another interesting topic. Let $\mu_d \in M_{\lambda^2}(H)$ be a target measure, and suppose

$$\mu_d \in \bigcup_{0\le t<\infty} \Xi(t);$$

that is, the system is controllable in finite time.

Problem 4. The problem is to find a control $u^o \in \mathcal{U}_{ad}$ such that $\mu_d = \mu_{\tau^o}^{u^o}$ and $\mu_t^u \ne \mu_d$ for any $t < \tau^o$ and any $u \in \mathcal{U}_{ad}$. In other words, u^o is the time optimal control and τ^o is the optimal time.

THEOREM 4.7. *Consider the system* (45)*, suppose the assumptions of Theorem* 4.1 *hold, let* $\mu_d(\ne \nu) \in M_{\lambda^2}(H)$*, and suppose the system is controllable in finite time. Then Problem* 4 *has a solution.*

Proof. Define the function $\mathcal{T}$ given by

$$\mathcal{T}(t) \equiv \rho^*(\mu_d, \Xi(t)), \quad t \geq 0. \tag{68}$$

This is a well-defined continuous function since the solutions are elements of (C_T, d^*) for every finite $T > 0$. By assumption the system is controllable in finite time implying the existence of a finite $T > 0$ such that $\mu_d \in \bigcup_{t\in[0,T]} \Xi(t) \neq \emptyset$. This implies that the set $R_c \equiv \{t \geq 0 : \mathcal{T}(t) = 0\}$ is nonempty and $\inf R_c \equiv \tau^o < \infty$. The proof now follows from the facts that $\mathcal{T}$ is continuous and the attainable sets $\Xi(t), t \geq 0$, are closed. Thus there exists a control u^o so that the corresponding solution of (45) hits the target μ_d at time τ^o, and there exists no other control that drives the system to the target earlier. This completes the proof. □

5. Some examples. For illustration we present a few examples. We have already mentioned the general case in the introduction which says that any stochastic differential equation in which the coefficients are dependent on the probability law of the process itself gives rise to McKean–Vlasov diffusion. This phenomenon is common in physical sciences dealing with the dynamics of charge density waves (CDW), mean-field dynamics of soft spins (SDW), chemical reactions, population biology, power flow in mobile communication network, etc. Another example is Kushner's equation arising in the study of nonlinear filtering. Given the history of observation, the conditional probability law is governed by an equation of the McKean–Vlasov type. Here we present some simple examples that arise naturally from applications.

Example 1 (mobile communication). In the study of stochastic power control of a wireless network, Olama, Djouadi, and Charalambos [17] have proposed a stochastic differential equation model that describes the dynamics of power flow between mobile transreceivers and a base station. Considering radial distance r separating the transmitter from the base station (receiver) as the spatial coordinate, the logarithm of power denoted by $\{X(t,r)\}$ is a stochastic process governed by a differential equation of the form

$$dX(t,r) = \beta(t,r)(\overline{X(t,r)} - X(t,r))dt + \delta(t,r)dw(t) \tag{69}$$

for $r \in D \subset (0,R)$ and $t \geq 0$, where $\overline{X(t,r)}$ denotes the mean of the process X, w is the standard scalar Brownian motion, and the variables β and δ are certain given functions of space and time which we may assume to be bounded measurable. The authors call this mean-reverting SDE. Numerical results presented by the authors seem to indicate that the model is a good approximation of the power distribution actually measured on the site. Since the noise level may also depend on the geographical position, we propose to replace the stochastic term by a space-time Brownian motion and add a diffusion term representing dissipation of power in the environment, giving

$$\begin{aligned} \partial_t X(t,r) = {} & \partial/\partial r(e(r)\partial X/\partial r)dt + \beta(t,r)(\overline{X(t,r)} - X(t,r))dt \\ & + dW(t,r), \end{aligned} \tag{70}$$

where $e(r)$ is a strictly positive function representing the diffusivity property of the medium. We use a homogeneous Dirichlet boundary condition of the form

$$X(t,0) = 0, X(t,R) = 0, \tag{71}$$

and add a term u to represent the (log of) power transmitted which we may consider as the control variable. Thus the controlled version of (70) is given by

$$\begin{aligned} \partial_t Y(t,r) = {} & \partial/\partial r(e(r)\partial Y/\partial r)dt + \beta(t,r)(\overline{Y(t,r)} - Y(t,r))dt \\ & + u(t,r)dt + dW(t,r), \end{aligned} \tag{72}$$

with a homogeneous Dirichlet boundary condition. We introduce the Hilbert space $H \equiv L_2(D)$ and define the operator A by

$$D(A) \equiv H_0^1(D) \cap H^2(D), \tag{73}$$
$$(A\varphi)(r) \equiv \partial/\partial r\big(e(r)\partial\varphi/\partial r\big) \text{ for } \varphi \in D(A).$$

The operator $f : R_+ \times \mathcal{P}(H) \times H \longrightarrow H$ is given by

$$f(t,\mu,y) \equiv \beta(t,\cdot)\bigg(\int_H \xi\mu(d\xi) - y(\cdot)\bigg), \tag{74}$$

and the control operator is given by $Bu = u$ (B identity operator in H). We introduce the H-valued Brownian motion $W(t) \equiv \{W(t,r), r \in D\}$ with the covariance given by

$$E\{(W(t),h)(W(\tau),g)\} = t \wedge \tau(Qh,g),$$

where

$$(Qh,g) \equiv \int_{D\times D} q(r,s)h(r)g(s)drds,$$

with q being a positive symmetric Hilbert–Schmidt kernel belonging to $L_2(D \times D)$. Defining $y(t) \equiv Y(t,\cdot)$, using cylindrical Brownian motion, and denoting it by the same symbol W, we can rewrite the basic equation in our abstract form on the Hilbert space H as follows:

$$dy = Aydt + f(t,\mu_t,y)dt + Budt + \sqrt{Q}dW, \quad y(0) = y_0, \tag{75}$$
$$\mu_t = \mathcal{L}y(t), \quad t \geq 0. \tag{76}$$

Assuming $e(r) > \gamma > 0$, it is easy to verify that A generates an exponentially stable C_0 semigroup on H. The rest of the assumptions of our existence theorem are obviously satisfied since f is linear, the operator B is bounded in H, and $\sqrt{Q}$ is nuclear. An interesting control problem for the mobile station is to find a control law that maximizes the power delivery to the base station, at any given time, say, T. One may formulate the problem as follows. For the control space E choose $E = L_2(D) = H$. Since the transmitter power is limited, we may choose a closed bounded convex set $U \subset E$, and for admissible feedback policies we use the set

$$\mathcal{U}_{ad} \equiv \{u \in B_\lambda(I \times H, H) : u(t,y) \in U \ \forall\ (t,y) \in I \times H\}.$$

Then choose any $\varphi \in L_2(D)$ having support that contains the mobile, and choose a control from the admissible class $\mathcal{U}_{ad}$ that maximizes the functional

$$u \longrightarrow E\exp\bigg\{\int_D Y^u(T,r)\varphi(r)dr\bigg\} \equiv \int_H e^{(y,\varphi)}\mu_T^u(dy).$$

For example, let $\Gamma(t), t \geq 0$, be a closed convex set with values in $2^D \setminus \emptyset$ representing the closed neighborhood of the location of the mobile, and choose $\varphi(r) \equiv \chi_{\Gamma(t)}(r)$. In this case the problem looks like

$$u \longrightarrow E\exp\bigg(\int_{\Gamma(t)} Y^u(t,r)dr\bigg) = \int_H e^{(y,\varphi)}\mu_t^u(dy) \longrightarrow \sup.$$

This is a special case of Theorem 4.4 and hence the existence of optimal control follows from it.

Example 2 (Brussellator). In the study of chemical kinetics, certain trimolecular reactions are known to exhibit periodic behavior. This was discovered by a group of scientists from Brussell, and in their honor the name Brussellator was adopted. In the well-stirred case, this is described by a pair of stochastic ordinary differential equations containing the mean-field coupling:

$$dx_1 = (a - (b+1)x_1 + x_1^2 x_2)dt + D_1(\overline{x}_1 - x_1)dt + \sigma_1(x_1)dw_1, \tag{77}$$

$$dx_2 = (bx_1 - x_1^2 x_2)dt + D_2(\overline{x}_2 - x_2)dt + \sigma_2(x_2)dw_2, \tag{78}$$

where the vector $\{x_1, x_2\}$ denotes the concentration of the two chemicals (product of reaction). This model is obtained from the McKean–Vlasov limit of a model due to Dawson (see [6] and the references in [18]). For a complete justification of the model see [18]. Under certain assumptions on the parameters $\{a, b, D_1, D_2\}$ it was proved by Scheutzow [18, Theorem 3.4, p. 446] that the system has a solution whose law $\mu_t, t \geq 0$, is supported on $R_+ \times R_+$ and that this law is periodic in time. A self-organizing property is commonly observed in biological species. It is an amazing fact that this property has been observed in some chemical reactions also. This may have a profound significance on the origin of biological species.

The model described above is based on the assumption that the concentration of the chemicals is uniformly distributed throughout the volume of the reactor. This requires controlling the mixing by a stirrer and possibly by a heat source controlling the temperature. Thus in fact the concentration of the chemicals should be considered as functions of time and position in the reactor volume $\Omega \subset R^3$. Including a multiplicative control, and assuming uniform diffusivity, the PDE version of this, describing the evolution of concentration distribution in the reactor Ω, is given by

$$dx = Axdt + f(x, \mu)dt + G_0(x)udt + \sigma dW, \tag{79}$$

where the operator A is given by the matrix of Laplacians

$$A \equiv \begin{pmatrix} \Delta & 0 \\ 0 & \Delta \end{pmatrix}$$

subject to a homogeneous Dirichlet (zero) boundary condion on the boundary $\partial\Omega$. The drift f is easily identified from (77)–(78), the operator σ is identified as

$$\sigma \equiv \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix},$$

and W is the space-time Brownian motion. For the state space we choose $H \equiv L_2(\Omega) \times L_2(\Omega)$, for the control space $E \equiv L_p(\Omega), p \in [1, \infty)$, with G_0 a bounded operator-valued function on H with values in $\mathcal{L}(E, H)$. More specifically we assume that $\sup\{\| G_0(x) \|_{\mathcal{L}(E,H)}, x \in H\} < \infty$. This provides a multiplicative and possibly reactive control. Since σ is allowed to depend on the state and f has quadratic nonlinearity, strictly speaking our theory fails to cover this case. However, due to conservation of mass, the concentration of chemicals can never be unbounded. Thus by appropriate truncation we may replace f by $f_r(x) \equiv f(\rho_r(x))$, where ρ_r is the retract of the ball $B_r \subset H$ of radius r around the origin ($\infty > r \geq$ the total mass of the reactor including its contents). For this f_r with σ independent of the state and

$\sqrt{\sigma}\sigma^*$ positive nuclear our results hold. For example, let $\Gamma \subset H$ be a target set of concentration, and suppose it is required to reach this target at time T with maximum probability. The problem is to find a control law that maximizes the functional $J(u) \equiv \mu_T^u(\Gamma)$. Under the assumptions used for admissible controls (section 4), if Γ is a nonempty closed convex set, there is an optimal control that solves the problem.

Example 3 (charge density waves/spin density waves). The discrete mean-field version of the Fukuyama–Lee model for CDW as presented by Bonilla [19] is given by a system of coupled stochastic ordinary differential equations:

$$d\Theta_i = \left(E - \alpha \sin(\Theta_i - \beta) - K\left(\Theta_i - (1/N)\sum_1^N \Theta_j\right)\right) dt + \sqrt{\gamma} dw_i, \tag{80}$$

$$i = 1, 2, \ldots, N, \tag{81}$$

where Θ_i is the phase at the site i, K is the stiffnes of coupling of CDW, and E is the applied electric field which can be used as the control. Again the interaction is of the mean-field type. The parameters $\{\alpha, \beta, \gamma\}$ are fixed, α representing impurity potential and β the pinning angle, and $\{w_i\}$ are independent standard Brownian motions. Again the McKean–Vlasov limit ($N \to \infty$) of this equation is an equation of the form (9) describing the temporal evolution of phase density or the corresponding measure. The operators are identified as follows:

$$\mathcal{A}\varphi \equiv (1/2)\gamma D^2\varphi, \quad (F(\mu)\varphi)(\theta) = f(\theta, \bar{\theta})D\varphi, \quad G(u)\varphi \equiv uD\varphi,$$
$$\bar{\theta} \equiv \int_R \theta\mu(d\theta), \quad \mu \equiv \mathcal{L}(\theta), \quad f(y,z) \equiv -\alpha\sin(y-\beta) - K(y-z),$$

where $D\varphi$ and $D^2\varphi$ denote the first and second partials of φ with respect to θ, respectively. Note that this is a scalar McKean–Vlasov equation with $H = E = R$. For an excellent physical interpretation of this model the reader is referred to [19]. It is known that by slowly increasing the electric field $u = E$ to a threshold value, one can break free the CDW from impurities and force sliding, thereby causing flow of current. It is claimed by physicists that CDW has great industrial potential, with the prospect of developing switches, capacitors, detectors, superconductors, etc. Since all of our assumptions hold, without going into details we simply mention that our results on optimal control apply to this problem. For example, let Γ_0 denote the region of phase space in which the CDW behaves as an insulator and Γ_c the region in which it acts as a (super)conductor. Suppose at time zero $\mu_0(\Gamma_0) = 1$ and that the set

$$\{t \geq 0, u \in \mathcal{U}_{ad} : \mu_t^u(\Gamma_c) = 1\} \neq \emptyset.$$

The problem is to find a driving force that minimizes the transit time from the set Γ_0 to the set Γ_c. In other words, find a control policy that minimizes the cost functional

$$J(u) \equiv \inf\{t \geq 0 : \mu_t^u(\Gamma_c) = 1\}.$$

This is a time optimal control as treated in section 4.

Remark. In this paper we have been concerned with the questions of the existence of optimal controls. We have not attempted to develop necessary conditions of optimality. This remains an open problem for the future.

REFERENCES

[1] N. U. Ahmed and X. Ding, *A semilinear McKean-Vlasov stochastic evolution equation in Hilbert space*, Stochastic Process. Appl., 60 (1995), pp. 65–85.
[2] N. U. Ahmed and X. Ding, *Controlled McKean-Vlasov equations*, Commun. Appl. Anal., 5 (2001), pp. 183–206.
[3] G. Da Prato and J. Zabczyk, *Regular densities of invariant measures for nonlinear stochastic equations*, J. Funct. Anal., 130 (1995), pp. 427–449.
[4] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Encyclopedia Math. Appl. 44, Cambridge University Press, London, 1992.
[5] G. Da Prato and J. Zabczyk, *Ergodicity for Infinite Dimensional Systems*, London Math. Soc. Lecture Note Ser. 229, Cambridge University Press, London, 1996.
[6] D. A. Dawson, *Critical dynamics and fluctuations for a mean-field model of cooperative behavior*, J. Stat. Phys., 31 (1983), pp. 29–85.
[7] D. A. Dawson and J. Gärtner, *Large Deviations, Free Energy Functional and Quasi-Potential for a Mean Field Model of Interacting Diffusions*, Mem. Amer. Math. Soc. 398, Providence, RI, 1989.
[8] W. H. Fleming, *Nonlinear semigroup for controlled partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 286–301.
[9] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Appl. Math. (N.Y.) 25, Springer-Verlag, New York, 1993.
[10] T. Funaki, *A certain class of diffusion processes associated with nonlinear parabolic equations*, Probab. Theory Related Fields, 67 (1984), pp. 331–348.
[11] J. Gärtner, *On the Mckean-Vlasov Limit for Interacting Diffusions* I, II, Akademie Der Wissenschaften Der DDR, Karl–Weierstrass–Institute Für Mathematik, Berlin, 1986.
[12] I. I. Gihman and A. V. Skorohod, *The Theory of Stochastic Processes* I, Springer-Verlag, New York, 1974 (translated from the Russian by S Kotz).
[13] C. Léonard, *Une loi des grands nombres pour des systèmes de diffusions avec interaction à coefficients nonbornés*, Ann. Inst. H. Poincaré, 22 (1984), pp. 237–262.
[14] H. P. McKean, *A class of Markov processes associated with nonlinear parabolic equations*, Proc. Natl. Acad. Sci. USA, 56 (1966), pp. 1907–1911.
[15] K. R. Parthasarathy, *Probabiliy Measures on Metric Spaces*, Academic, New York and London, 1967.
[16] N. I. Mahmudov and M. A. McKibben, *Abstract second order damped McKean-Vlasov stochastic evolution equations*, Stoch. Anal. Appl., 24 (2006), pp. 303–328.
[17] M. M. Olama, S. M. Djouadi, and C. D. Charalambos, *Stochastic Power Control for Time-varying Long-Term Fading Wireless Networks*, University of Cyprus, Nicosia, Cyprus, preprint 2006/2007.
[18] M. Scheutzow, *Periodic behavior of the stochastic Brusselator in the mean-field limit*, Probab. Theory Related Fields, 72 (1986), pp. 425–462.
[19] L. L. Bonilla, *Stable nonequilibrium probability densities and phase transitions for mean-field models in the thermodynamic limit*, J. Stat. Phys., 46 (1987), pp. 659–678.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 379–394

NULL CONTROLLABILITY OF SOME SYSTEMS OF TWO PARABOLIC EQUATIONS WITH ONE CONTROL FORCE*

SERGIO GUERRERO†

Abstract. In this paper we establish some exact controllability results for systems of two parabolic equations. First, we prove the existence of insensitizing controls for the L^2 norm of the gradient of solutions of linear heat equations. Then, in the worst situation where null controllability for a system of two parabolic equations can hold, we prove this result for some general couplings.

Key words. Carleman inequalities, system of parabolic equations, controllability

AMS subject classifications. 35K40, 93B05

DOI. 10.1137/060653135

1. Introduction.

Let $\Omega \subset \mathbb{R}^N$ $(N \geq 1)$ be a bounded connected open set whose boundary $\partial\Omega$ is regular enough. Let $T > 0$, and let $\omega \subset \Omega$ be a (small) nonempty open subset which will usually be referred to as a *control domain*. We will use the notation $Q = \Omega \times (0,T)$ and $\Sigma = \partial\Omega \times (0,T)$.

The main objective of this paper is to establish some new controllability results for coupled parabolic equations.

The first main result of this paper concerns insensitizing controls. More precisely, we want to insensitize a functional associated with a state system, which is a linear parabolic equation. Let us introduce an open set $\mathcal{O} \subset \Omega$, which is called the *observatory* (or observation open set).

In order to state our problem, we introduce the following system:

$$\begin{cases} y_t - \Delta y + ay + B \cdot \nabla y = v1_\omega + f & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y|_{t=0} = y^0 + \tau\widehat{y}^0 & \text{in } \Omega. \end{cases} \tag{1}$$

Here, v is the control, $y^0 \in L^2(\Omega)$, and $a \in \mathbf{R}$ and $B \in \mathbf{R}^N$ are constants. Furthermore, we suppose that $\widehat{y}^0$ is unknown with $\|\widehat{y}^0\|_{L^2(\Omega)} = 1$ and that τ is a small unknown real number. Then, the interpretation of system (1) is that y is the temperature of a body, v is a localized heat source, (where we have access to the body) to be chosen, f is another heat source, and the initial state of the body is partially unknown.

In general, the functional J_τ we would like to insensitize (which is called *sentinel*) has to be differentiable. In this framework, the task is to find a control v such that the influence of the unknown data $\tau\widehat{y}^0$ is not perceptible for J_τ (see (3) below).

In the literature, the usual functional is given by the L^2 norm of the state (see [4], [6], or [19], for instance). Here, we are interested in insensitizing the L^2 norm of

*Received by the editors February 27, 2006; accepted for publication (in revised form) October 18, 2006; published electronically April 17, 2007.
http://www.siam.org/journals/sicon/46-2/65313.html

†Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Boîte Corrier 187, 75035 Cedex 05 Paris, France (guerrero@ann.jussieu.fr).

the gradient of the state (solution of (1)). Thus, let us introduce the functional

$$J_\tau(y) = \iint_{\mathcal{O}\times(0,T)} |\nabla y|^2 \, dx\, dt, \tag{2}$$

where y is the solution of (1).

Our objective is to find a control v such that the presence of the unknown data is imperceptible for J_τ, that is to say, such that

$$\frac{\partial J_\tau}{\partial \tau}(y)|_{\tau=0} = 0 \quad \forall\, \widehat{y}^0 \in L^2(\Omega) \quad \text{such that } \|\widehat{y}^0\|_{L^2(\Omega)} = 1. \tag{3}$$

If this holds, we will say that the control v insensitizes the functional J_τ.

Usually, insensitizing problems are formulated in an equivalent way as a controllability problems of a cascade system (see, for instance, [17] and [4] for a rigorous deduction of this fact). Indeed, if we consider the adjoint state of (1) (or apply the Lagrange principle), it is very easy to see that condition (3) is equivalent to $w|_{t=0} \equiv 0$ in Ω, where w together with z fulfills

$$\begin{cases} z_t - \Delta z + az + B\cdot\nabla z = v1_\omega + f & \text{in } Q, \\ -w_t - \Delta w + aw - B\cdot\nabla w = \nabla\cdot(\nabla z 1_{\mathcal{O}}) & \text{in } Q, \\ z = 0,\ w = 0 & \text{on } \Sigma, \\ z|_{t=0} = y^0,\ w|_{t=T} = 0 & \text{in } \Omega. \end{cases} \tag{4}$$

Here, we have denoted $z \equiv y|_{\tau=0}$. Assume for a moment that $y^0 \in L^2(\Omega)$ and $f,\, v \in L^2(Q)$. Then, it is not difficult to prove that there exists a unique solution (z, w) of (4) which belongs to $L^2(0,T; H_0^1(\Omega))^2$ and depends continuously on (y^0, f, v) in $L^2(\Omega)\times L^2(Q)^2$.

To our best knowledge, the first time this kind of problem was addressed was in [18] for second and fourth order parabolic equations of the heat kind and for the Navier–Stokes system. As we said above, all results around this subject concern the functional $\tilde{J}_\tau(y) = \|y\|^2_{L^2(\mathcal{O}\times(0,T))}$ with y a solution of a parabolic system. In [4], the authors prove the existence of ε-insensitizing controls (i. e., such that $|J'_\tau(y)|_{\tau=0}| \le \varepsilon$) for solutions of a semilinear heat system with C^1 and globally Lipschitz nonlinearities. In [6], the author proved the existence of insensitizing controls for the same system. For an extension of this result to more general nonlinearities, see [5] and the references therein.

As we shall see in the statement of Theorem 1 below, we will take $y^0 \equiv 0$. For a justification of this fact and a possible choice of more general initial conditions, see [6].

Throughout this paper we will suppose that $\omega \cap \mathcal{O} \neq \emptyset$. This is a condition that has always been imposed in the literature where insensitizing controls are concerned. Recently, for the (simpler) situation of looking for an ε-insensitizing control and the functional $\tilde{J}_\tau$, it has been demonstrated that this condition is not necessary for solutions of linear heat equations (see [7]).

The controllability result for system (4) is given in the following theorem.

THEOREM 1. *Let $m > 3$ be a real number and $y^0 \equiv 0$. Then, there exists a constant $K_0 > 0$ depending on $\Omega, \omega, \mathcal{O}, T, a$, and B such that for any $f \in L^2(Q)$ satisfying $\|e^{K_0/t^m} f\|_{L^2(Q)} < +\infty$, there exists a control v such that the corresponding solution (w, z) of (4) satisfies $w|_{t=0} \equiv 0$ in Ω.*

COROLLARY 2. *There exists insensitizing controls v of the functional J_τ given by* (2).

Remark 1. The same result holds when a and B are functions which depends only on the time variable t and are in $L^\infty(0,T)$. The proof of this fact is direct from that of Theorem 1.

Let us briefly explain the difficulties a controllability result for system (4) possesses. For this, we introduce the associated adjoint system:

$$\begin{cases} -\psi_t - \Delta\psi + a\psi - B\cdot\nabla\psi = \nabla\cdot((\nabla\varphi)1_{\mathcal{O}}) & \text{in } Q, \\ \varphi_t - \Delta\varphi + a\varphi + B\cdot\nabla\varphi = 0 & \text{in } Q, \\ \psi = 0,\ \varphi = 0 & \text{on } \Sigma, \\ \psi|_{t=T} = 0,\ \varphi|_{t=0} = \varphi^0 & \text{in } \Omega. \end{cases} \tag{5}$$

It is by now a classical fact that the null controllability result we want to prove for system (4) is equivalent to the following *observability inequality*:

$$\iint_Q e^{-K_0/t^m}|\varphi|^2\,dx\,dt \le C\iint_{\omega\times(0,T)}|\psi|^2\,dx\,dt, \tag{6}$$

where m is some positive number and C and K_0 are two positive constants depending on $\Omega, \omega, \mathcal{O}, T$, a, and B but independent of φ^0 (see, for instance, [14] or [11]). The main idea one usually follows in order to prove (6) is a combination of observability inequalities for ψ and φ (as solutions of heat equations) and trying to eliminate the local term (concentrated in $\omega\times(0,T)$) concerning φ. The great difficulty one encounters when trying this for system (5) is that no local estimate of the kind

$$\iint_{\widetilde{\omega}\times(0,T)}|\varphi|^2\,dx\,dt \le C\iint_{\omega\times(0,T)}|\Delta\varphi|^2\,dx\,dt, \quad \widetilde{\omega}\subset\omega,$$

can be obtained using local arguments (observe that ω can be taken as small as we want, so we can always suppose that $\overline{\omega}\cap\partial\Omega = \emptyset$).

This means that we have to find another way to locally relate φ and ψ. The idea we follow here is to first obtain an observability inequality of the kind

$$\iint_Q e^{-K_1/t^m}|\varphi|^2\,dx\,dt \le C\iint_{\omega\times(0,T)}|\Delta\varphi|^2\,dx\,dt. \tag{7}$$

The reason why an estimate like (7) is not easy to prove relies on the fact that no boundary conditions are known for $\Delta\varphi$. More details about this are given in subsection 2.1, below.

Remark 2. Is this result true when the coefficients a and B depend on the space variable? We observe here that, in this situation, not even the following unique continuation property is known:

$$\psi = 0 \ \text{ in } \omega\times(0,T) \Rightarrow \varphi,\psi \equiv 0 \ \text{ in } \Omega\times(0,T).$$

As an extension of the result stated in Theorem 1, some insensibilization properties have recently been demonstrated for the more complicated situation of a system

of the Stokes kind. More precisely, we consider the functional $\tilde{J}_\tau$, with y the solution of

$$\begin{cases} y_t - \Delta y + ay + B \cdot \nabla y + \nabla p = v1_\omega + f & \text{in } Q, \\ \nabla \cdot y = 0 & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y|_{t=0} = \tau \widehat{y}^0 & \text{in } \Omega, \end{cases}$$

with a and B constants. In [13], the existence of controls insensitizing $\tilde{J}_\tau$ is established. See also [13] for an extension of this to more general functionals and further controllability results for coupled Stokes-like systems.

The second and main objective of this paper is to extend the previous controllability result to more intrinsic coupled parabolic systems. The question is, Which coupling do we need to be able to control the whole system with only one control force?

As far as the controllability of strongly coupled parabolic equations is concerned, in the literature the local exact controllability of phase field systems was proved in [3], while the global version was later proved in [1]. For more general coupled parabolic equations with only one control force, some results have been given in [2] and [12].

In this paper, we will concentrate on studying the null controllability of systems of two parabolic equations, where the coupling terms are first order space derivatives in one equation and second order space derivatives in the other. In this situation, we will be again interested in controlling only one of the two equations while driving both states to zero at $t = T$.

We consider, for instance, the case where we control the lower order coupling term equation. We set the following control coupled system:

$$\begin{cases} w_t - \Delta w + cw + E \cdot \nabla w = P_2(t,x;D)(z\,\theta_2) & \text{in } Q, \\ z_t - \Delta z + hz + K \cdot \nabla z = P_1(t,x;D)(w\,\theta_1) + v1_\omega & \text{in } Q, \\ w = z = 0 & \text{on } \Sigma, \\ w|_{t=0} = w^0, \quad z|_{t=0} = z^0 & \text{in } \Omega, \end{cases} \tag{8}$$

where c, E, h, and K are constants and $P_i(t,x;D)$ $(i = 1, 2)$ is a partial differential operator in the space variables of order i such that

$$P_i(t,x;D)u = \sum_{|\beta| \le i} m_{i,\beta}(t,x)\partial_x^\beta u, \quad m_{i,\beta} \in L^\infty(0,T;W^{2/i,\infty}(\Omega)) \tag{9}$$

(that is to say, $m_{2,\beta}$, $\partial_x(m_{2,\beta})$, $m_{1,\beta}$, $\partial_x(m_{1,\beta})$, $\partial_x^2(m_{1,\beta}) \in L^\infty(Q)$). In (8), $\theta_i \in C^2(\overline{\Omega})$ $(1 \le i \le 2)$. We assume that there exists a nonempty open set $\omega_2 \subset \omega$ and a constant $C > 0$ such that $|\theta_2| \ge C > 0$ in ω_2. Observe that, in particular, one can take θ_1 and θ_2 to have a support as small as we want (one can also take $\theta_1 \equiv \theta_2 \equiv 1$ in Ω, which is the best possible situation).

Also for system (8) we have the existence and uniqueness of solution (w, z). For instance, if $v \in L^2(Q)$ and $(w^0, z^0) \in L^2(\Omega)^2$, then $(w, z) \in L^2(0,T;H_0^1(\Omega))^2$, which depends continuously on $(v, w^0, z^0) \in L^2(Q) \times L^2(\Omega)^2$.

Our objective here is to drive both w and z to zero at time T by means of the control v. Accordingly, we consider the corresponding adjoint system:

$$\begin{cases} -\varphi_t - \Delta\varphi + c\varphi - E\cdot\nabla\varphi = (P_1^*(t,x;D)\psi)\theta_1 & \text{in } Q,\\ -\psi_t - \Delta\psi + h\psi - K\cdot\nabla\psi = (P_2^*(t,x;D)\varphi)\theta_2 & \text{in } Q,\\ \varphi = \psi = 0 & \text{on } \Sigma,\\ \varphi|_{t=T} = \varphi^0, \quad \psi|_{t=T} = \psi^0 & \text{in } \Omega, \end{cases} \tag{10}$$

where P_1^* and P_2^* are the adjoint operators of P_1 and P_2, respectively.

It is very easy to prove the classical fact that the previous controllability property is equivalent to the following observability inequality:

$$\|\varphi|_{t=0}\|^2_{L^2(\Omega)} + \|\psi|_{t=0}\|^2_{L^2(\Omega)} \le C \iint_{\omega\times(0,T)} |\psi|^2\,dx\,dt,$$

with $C = C(\Omega,\omega,T) > 0$ independent of (φ^0,ψ^0).

In order to achieve this, we need the following properties to hold for the differential operator P_2:

$$m_{2,\beta} \text{ are constant} \tag{11}$$

and

$$\|u\|_{H^2(\Omega)} \le C\|P_2^*u\|_{L^2(\Omega)} \quad \forall u \in H^2(\Omega)\cap H_0^1(\Omega), \tag{12}$$

for some $C = C(\Omega) > 0$.

Observe that no boundary condition for $P_2^*\varphi$ is demanded.

THEOREM 3. *Assume that conditions* (11)–(12) *hold. Then, there exists a control* v *such that the solution of* (8) *satisfies* $w|_{t=T} \equiv z|_{t=T} \equiv 0$ *in* Ω.

Remark 3. Other boundary conditions can be considered in system (8). For instance, if one imposes Neumann boundary conditions, Theorem 3 also holds when we impose

$$\|u\|_{H^2(\Omega)} \le C\|P_2^*u\|_{L^2(\Omega)} \quad \forall u \in H^2(\Omega),\ \frac{\partial u}{\partial n}|_\Sigma = 0 \tag{13}$$

instead of (12).

In general, if one imposes $Bw|_\Sigma = 0$ as a boundary condition for w in (8), Theorem 3 holds if

$$\|u\|_{H^2(\Omega)} \le C\|P_2^*u\|_{L^2(\Omega)} \quad \forall u \in H^2(\Omega),\ Bu|_\Sigma = 0. \tag{14}$$

Remark 4. One can extend the result stated in Theorem 3 to the case where c, E h, and K are functions which depend on time and belong to $L^\infty(0,T)$, no matter which boundary conditions are considered.

Remark 5. Instead of the operator P_2, one could have also considered an operator L containing a first order time derivative. Indeed, let

$$Lu = \sum_{|\gamma|\le 1} \ell_\gamma\,\partial_t^\gamma u, \quad \ell_\gamma \text{ constants.}$$

Then, Theorem 3 holds for L instead of P_2. Observe that the proof in this situation is simpler, since the boundary conditions satisfied by φ are also satisfied by $L^*\varphi$.

Remark 6. A combination of Theorem 3, Remark 3, and Remark 5 yields that when we consider the differential operator

$$Qu = \sum_{|\gamma|\leq 1, |\beta|\leq 2} (\ell_\gamma \, \partial_t^\gamma + m_\beta \partial_x^\beta) u, \quad \ell_\gamma, m_\beta \text{ constants}$$

instead of P_2, the result stated in Theorem 3 holds as long as

$$\|u\|_{H^2(\Omega)} \leq C \|Q^* u\|_{L^2(\Omega)} \quad C = C(\Omega) > 0,$$

for any $u \in H^2(\Omega)$ satisfying the same boundary conditions as w in (8).

Finally, we consider the situation where we control the higher order coupling term equation. Thus, let us introduce the following system:

$$\begin{cases} \mathbf{p}_t - \Delta \mathbf{p} + c\mathbf{p} + E \cdot \nabla \mathbf{p} = \overrightarrow{P_2}(t,x;D)(q\,\theta_4) + \mathbf{v}1_\omega & \text{in } Q, \\ q_t - \Delta q + hq + K \cdot \nabla q = P_1(t,x;D)(\mathbf{p}\,\theta_3) & \text{in } Q, \\ \mathbf{p} = q = 0 & \text{on } \Sigma, \\ \mathbf{p}|_{t=0} = \mathbf{p}^0, \quad q|_{t=0} = q^0 & \text{in } \Omega. \end{cases} \tag{15}$$

Here, $\mathbf{p}$ is a vector-valued function and $\overrightarrow{P_2}$ a vectorial differential operator of order 2 in space such that each component is given by (9) for $i = 2$. On the other hand, q is a scalar-valued function and P_1 is a divergence-type operator, that is to say, $P_1\mathbf{f} \in \mathbf{R}$ for $\mathbf{f}$ a vector-valued function. Finally, $\theta_i \in C^2(\overline{\Omega})$ $(3 \leq i \leq 4)$ and we assume the existence of a nonempty open subset $\omega_3 \subset \omega$ and a positive constant C such that $\theta_3 \geq C > 0$ in ω_3.

For $\mathbf{v} \in L^2(Q)^N$ and $(\mathbf{p}^0, q^0) \in L^2(\Omega)^{N+1}$, there exists a unique solution $(\mathbf{p}, q) \in L^2(0,T; H_0^1(\Omega))^{N+1}$ which depends continuously on $(\mathbf{v}, \mathbf{p^0}, q^0) \in L^2(Q)^N \times L^2(\Omega)^{N+1}$.

Observe that now we are controlling the first equation. Obviously, the adjoint system associated with (15) is again (10). In order to establish the corresponding null controllability result, this time we need to impose the following conditions on the operator P_1:

$$m_{1,\beta} \text{ are constant} \tag{16}$$

and

$$\|u\|_{L^2(0,T;H^1(\Omega))} \leq C \|\overrightarrow{P_1^*} u\|_{L^2(Q)^N} \quad \forall u \in H_0^1(\Omega). \tag{17}$$

The corresponding result in this situation is presented in the following theorem.

THEOREM 4. *Assume conditions* (16)–(17) *are satisfied. Then, there exists a control* v *such that the solution of* (15) *satisfies* $\mathbf{p}|_{t=T} \equiv q|_{t=T} \equiv 0$ *in* Ω.

Once Theorem 3 is demonstrated, one can follow the same ideas in order to prove Theorem 4 by just adapting the corresponding arguments.

For the sake of completeness, we present a system for which Theorem 4 applies (for simplicity, we take $\theta_3 \equiv \theta_4 \equiv 1$):

$$\begin{cases} \mathbf{p}_t - \Delta \mathbf{p} + c\mathbf{p} + E \cdot \nabla \mathbf{p} = (P_{2,1}, \dots, P_{2,N})q, + \mathbf{v}1_\omega & \text{in } Q, \\ q_t - \Delta q + hq + K \cdot \nabla q = \nabla \cdot \mathbf{p} & \text{in } Q, \\ \mathbf{p} = q = 0 & \text{on } \Sigma, \\ \mathbf{p}|_{t=0} = \mathbf{p}^0, \quad q|_{t=0} = q^0 & \text{in } \Omega, \end{cases}$$

with $P_{2,j}$ differential operators of order 2 in the x variable satisfying (9).

This paper is organized as follows. In section 2, we prove Theorem 1. In subsection 2.1 we prove new Carleman-type estimates, which will be crucial for this proof, and in subsection 2.2 we combine some results and conclude its proof. Finally, in section 3 we prove Theorem 3.

2. Insensitizing controls for the functional J_τ. As we saw in the introduction, we can restrict ourselves to proving the null controllability of the coupled system (4), that is to say, Theorem 1.

As usual, in order to prove this result we concentrate on the corresponding adjoint system:

$$\begin{cases} -\varphi_t - \Delta\psi + a\psi - B\cdot\nabla\psi = \nabla\cdot((\nabla\varphi)1_{\mathcal{O}}) & \text{in } Q,\\ \varphi_t - \Delta\varphi + a\varphi + B\cdot\nabla\varphi = 0 & \text{in } Q,\\ \psi = 0,\ \varphi = 0 & \text{on } \Sigma,\\ \psi|_{t=T} = 0,\ \varphi|_{t=0} = \varphi^0 & \text{in } \Omega, \end{cases} \tag{18}$$

where $\varphi^0 \in L^2(\Omega)$. As explained in the introduction, in the framework of controllability it is a classical fact that the null controllability property for system (4) is equivalent to the following observability inequality:

$$\iint_Q e^{-K_0/t^m}|\psi|^2\,dx\,dt \le C\iint_{\omega\times(0,T)}|\psi|^2\,dx\,dt, \tag{19}$$

for certain positive constants K_0 and C which are independent of ψ^0, and for some positive m.

For the proof of (19), we will follow a classical approach consisting of obtaining a suitable weighted-like estimate (the so-called *Carleman estimate*) similar to the observability inequality. For a systematic use of this kind of estimate see, for instance, [14] or [11].

In order to establish this Carleman inequality, we need to define some weight functions:

$$\begin{gathered} \alpha_m(x,t) = \frac{\exp\{\frac{k(m+1)}{m}\lambda\|\eta^0\|_\infty\} - \exp\{\lambda(k\|\eta^0\|_\infty + \eta^0(x))\}}{t^m(T-t)^m},\\ \alpha_m^*(t) = \max_{x\in\overline{\Omega}}\alpha_m(x,t) = \alpha_m|_{\partial\Omega}(x,t), \quad \xi_m(x,t) = \frac{e^{\lambda(k\|\eta^0\|_\infty+\eta^0(x))}}{t^m(T-t)^m},\\ \xi_m^*(t) = \min_{x\in\overline{\Omega}}\xi_m(x,t) = \xi_m|_{\partial\Omega}(x,t), \end{gathered} \tag{20}$$

where $m > 3$ and $k > m$ are fixed. Here, $\eta^0 \in C^2(\overline{\Omega})$ satisfies

$$|\nabla\eta^0| \ge C > 0 \ \text{ in } \Omega\setminus\overline{\omega_0}, \quad \eta^0 > 0 \ \text{ in } \Omega, \quad \text{and} \quad \eta^0 \equiv 0 \ \text{ on } \partial\Omega, \tag{21}$$

with $\emptyset \ne \omega_0 \subset \omega\cap\mathcal{O}$ an open set. The proof of the existence of such a function η^0 is given in [11]. The weights (20) were first considered in [10] in order to obtain Carleman estimates for the three-dimensional micropolar fluid model.

Accordingly, we define $I(s,\lambda;\cdot)$ as follows:

$$
\begin{aligned}
I(s,\lambda;g) := {} & s^{-1}\iint_Q e^{-2s\alpha_m}\xi_m^{-1}\left(|g_t|^2+|\Delta g|^2\right)\,dx\,dt \\
& + \ s\lambda^2\iint_Q e^{-2s\alpha_m}\xi_m|\nabla g|^2\,dx\,dt + s^3\lambda^4\iint_Q e^{-2s\alpha_m}\xi_m^3|g|^2\,dx\,dt.
\end{aligned}
\tag{22}
$$

Furthermore, we denote by $I_w(s,\lambda,\cdot)$ the terms in the expression of $I(s,\lambda,\cdot)$ concerning the $L^2(Q)$ and $L^2(0,T;H_0^1(\Omega))$ norms (that is, the integrals appearing in the second line of (22)).

With this notation, we can prove the following result.

PROPOSITION 5. *There exists a positive constant C which depends on Ω, ω, and T such that*

$$
\begin{aligned}
& I_w(\Delta\varphi) + s^2\lambda^4\iint_Q e^{-3s\alpha_m}\xi_m^2(s^2\lambda^2\xi_m^2|\psi|^2+|\nabla\psi|^2)\,dx\,dt \\
& \qquad \le C(1+T^2)s^7\lambda^8\iint_{\omega\times(0,T)} e^{-2s\alpha_m}\xi_m^7|\psi|^2\,dx\,dt,
\end{aligned}
\tag{23}
$$

for any $\lambda \ge C$ and $s \ge C(T^{2m}+T^m)$.

Remark 7. From the Carleman inequality (23), one can readily deduce the observability inequality (19). Indeed, it suffices to combine the fact that $\psi|_{t=T} \equiv 0$ with the dissipation of $\|\nabla\varphi(t)\|_{L^2(\Omega)}$ as t goes to T (see, for instance, [11]). As a consequence, the proof of Theorem 1 is achieved.

The proof of Proposition 5 is divided into two steps, which correspond to subsections 2.1 and 2.2. The first, and more important, step deals with the equation satisfied by φ (which is independent of ψ). In the second step, we combine both equations in order to conclude the desired inequality (23).

Before starting with this, we recall a Carleman estimate which will be essential in our proof.

This estimate concerns energy solutions of heat equations with nonhomogeneous Neumann boundary conditions.

LEMMA 6. *Let $u^0 \in L^2(\Omega)$, $f_1 \in L^2(Q)$, $f_2 \in L^2(Q)^N$, and $f_3 \in L^2(\Sigma)$. Then there exists a constant $C(\Omega,\omega_0) > 0$ such that the solution $u \in L^2(0,T;H^1(\Omega)) \cap L^\infty(0,T;L^2(\Omega))$ of*

$$
\begin{cases}
u_t - \Delta u = f_1 + \nabla\cdot f_2 & \text{in } Q,\\
\dfrac{\partial u}{\partial n} + f_2.n = f_3 & \text{on } \Sigma,\\
u|_{t=T} = u^0 & \text{in } \Omega
\end{cases}
$$

satisfies

$$
\begin{aligned}
I_w(u) \le C\Bigg(& s^3\lambda^4\iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3|u|^2\,dx\,dt \\
& + \iint_Q e^{-2s\alpha_m}|f_1|^2\,dx\,dt + s^2\lambda^2\iint_Q e^{-2s\alpha_m}\xi_m^2|f_2|^2\,dx\,dt \\
& + \ s\lambda\iint_\Sigma e^{-2s\alpha_m^*}\xi_m^*|f_3|^2\,d\sigma\,dt\Bigg)
\end{aligned}
$$

for any $\lambda \geq C$ *and* $s \geq C(T^{2m} + T^{2m-1})$.

This lemma was essentially proved in [8]. In fact, the inequality proved in [8] concerned the same weight functions as in (24), but with $m = 1$. Then one can follow the steps of the proof in [8] (see Theorem 1 in that reference) and adapt the arguments by just taking into account that

$$\partial_t \alpha_m := \alpha_{m,t} \leq CT\xi_m^{(m+1)/m} \quad \text{and} \quad \partial_{tt}\alpha_m := \alpha_{m,tt} \leq CT^2\xi_m^{(m+2)/m},$$

with $C > 0$ independent of s, λ, and T.

2.1. New Carleman estimate for φ. Here we deal with the problem

$$\begin{cases} \varphi_t - \Delta\varphi + a\varphi + B \cdot \nabla\varphi = 0 & \text{in } Q, \\ \varphi = 0 & \text{on } \Sigma, \\ \varphi|_{t=0} = \varphi^0 & \text{in } \Omega. \end{cases} \tag{24}$$

Recall that $a \in \mathbf{R}$ and $B \in \mathbf{R}^N$ are constants.

For this system, we prove the following estimate.

LEMMA 7. *There exists a positive constant* C *depending on* Ω *and* ω_0 *such that*

$$I_w(\Delta\varphi) \leq Cs^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3|\Delta\varphi|^2\,dx\,dt, \tag{25}$$

for any $\lambda \geq C$ *and* $s \geq C(T^{2m} + T^m)$.

Remark 8. Observe that, in particular, we deduce from this inequality the following well-known unique continuation property:

$$\Delta\varphi = 0 \ \text{ in } \omega_0 \times (0,T) \Rightarrow \varphi \equiv 0 \ \text{ in } \Omega \times (0,T). \tag{26}$$

As far as we know, it is a new fact that (26) can be quantified in terms of an inequality like (25). On the other hand, we do not know if (26) holds when a and B are not constant with respect to the space variable.

Proof of Lemma 7. We first look at the equation satisfied by $\Delta\varphi$:

$$(\Delta\varphi)_t - \Delta(\Delta\varphi) + a\Delta\varphi + B \cdot \nabla\Delta\varphi = 0 \quad \text{in } Q.$$

Observe that no boundary conditions are prescribed for $\Delta\varphi$. At this point, we can apply Lemma 6 (with $f_2 \equiv 0$) and deduce the existence of a constant $C = C(\Omega, \omega_0) > 0$ such that

$$\begin{aligned} I_w(\Delta\varphi) \leq C \Bigg(& s^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3|\Delta\varphi|^2\,dx\,dt \\ & + s\lambda \iint_\Sigma e^{-2s\alpha_m^*}\xi_m^* \left|\frac{\partial\Delta\varphi}{\partial n}\right|^2 \Bigg)\,dx\,dt \end{aligned} \tag{27}$$

for any $\lambda \geq C$ and $s \geq C(T^{2m} + T^{2m-1})$.

The next step will be to eliminate the last term in the right-hand side of (27). In order to do this, we introduce the function $\varphi^* := \eta(t)\varphi$, where

$$\eta(t) = s^{(1/2)-(1/m)}\lambda e^{-s\alpha_m^*(t)}(\xi_m^*)^{(1/2)-(1/m)}(t). \tag{28}$$

In view of (24), it fulfills

$$\begin{cases} \varphi_t^* - \Delta\varphi^* + a\varphi^* + B\cdot\nabla\varphi^* = \eta_t\varphi & \text{in } \Omega, \\ \varphi^* = 0 & \text{on } \partial\Omega, \\ \varphi^*|_{t=0} = 0 & \text{in } \Omega. \end{cases} \tag{29}$$

Thanks to (27), we are going to deduce that φ^* is a very regular function. In fact, we have that $\eta_t\varphi \in L^2(0,T;H^2(\Omega)\cap H_0^1(\Omega))$, since $\eta_t\Delta\varphi \in L^2(\Omega)$ and

$$\begin{aligned} \|\eta_t\Delta\varphi\|_{L^2(Q)} &\le CTs^{(3/2)-(1/m)}\lambda\|e^{-s\alpha_m^*}(\xi_m^*)^{3/2}\Delta\varphi\|_{L^2(Q)} \\ &\le Cs^{3/2}\lambda\|e^{-s\alpha_m^*}(\xi_m^*)^{3/2}\Delta\varphi\|_{L^2(Q)}, \end{aligned}$$

for $s \ge CT^m$. The square of this last quantity is bounded by the left-hand side of (27) (recall that $e^{-s\alpha_m^*}$ is the minimum of $e^{-s\alpha_m}$).

Consequently, we have that $\varphi^* \in L^2(0,T;(H^4\cap H_0^1)(\Omega))$ (see, for instance, [16]) and

$$\begin{aligned} \|\varphi^*\|^2_{L^2(0,T;H^4(\Omega))} &= s^{1-2/m}\lambda^2\int_0^T e^{-2s\alpha_m^*}(\xi_m^*)^{1-2/m}\|\varphi\|^2_{H^4(\Omega)}\,dt \\ &\le CI_w(\Delta\varphi). \end{aligned} \tag{30}$$

Taking this into account, by a simple integration by parts we deduce that

$$s^{2-1/m}\lambda^3\int_0^T e^{-2s\alpha_m^*}(\xi_m^*)^{2-1/m}\|\varphi\|^2_{H^3(\Omega)}\,dt \le CI_w(\Delta\varphi). \tag{31}$$

From (30) and (31), we obtain in particular that

$$s^{(3/2)-3/(2m)}\lambda^3\int_0^T e^{-2s\alpha_m^*}(\xi_m^*)^{(3/2)-3/(2m)}\left\|\frac{\partial\Delta\varphi}{\partial n}\right\|^2_{L^2(\Sigma)}\,dt \le CI_w(\Delta\varphi). \tag{32}$$

Since $m > 3$, this justifies that the second term in the right-hand side of (27) is absorbed by the left-hand side. As a conclusion, we obtain the desired inequality (25).

2.2. Carleman estimate for φ and conclusion. Finally, we will deal with the particular coupling of ψ and φ.

First, assuming φ is given, we apply a Carleman estimate to the weak solution ψ of (18) (observe that the right-hand side of the equation satisfied by ψ belongs, for instance, to $L^2(0,T;H^{-1}(\Omega))$), which can be found in [15] (for the explicit dependence with respect to λ and T, see [9]):

$$\begin{aligned} &s^4\lambda^6\iint_Q e^{-3s\alpha_m}\xi_m^4|\psi|^2\,dx\,dt + s^2\lambda^4\iint_Q e^{-3s\alpha_m}\xi_m^2|\nabla\psi|^2\,dx\,dt \\ &\qquad\le C\left(s^4\lambda^6\iint_{\omega_0\times(0,T)} e^{-3s\alpha_m}\xi_m^4|\psi|^2\,dx\,dt\right. \\ &\qquad\qquad\left. + s^3\lambda^4\iint_{\mathcal{O}\times(0,T)} e^{-3s\alpha_m}\xi_m^3|\nabla\varphi|^2\,dx\,dt\right), \end{aligned} \tag{33}$$

for any $\lambda \geq C$ and $s \geq C(T^{2m} + T^{2m-1})$. Observe that we have chosen to apply this result for smaller exponentials, that is to say, for $e^{-3s\alpha_m}$ instead of $e^{-2s\alpha_m}$.

Then we easily see that the last integral in the right-hand side of (33) is bounded by $I_w(\Delta\varphi)$, as long as λ is large enough. In fact, if we denote $\widehat{\alpha}_m(t) = \min_{x\in\overline{\Omega}} \alpha_m(x,t)$ and $\widehat{\xi}_m(t) = \max_{x\in\overline{\Omega}} \xi_m(x,t)$, we have

$$s^3\lambda^4 \iint_{\mathcal{O}\times(0,T)} e^{-3s\alpha_m}\xi_m^3 |\nabla\varphi|^2 \,dx\,dt \leq s^3\lambda^4 \int_0^T e^{-3s\widehat{\alpha}_m}(\widehat{\xi}_m)^3 \|\nabla\varphi\|^2_{L^2(\Omega)}\,dt$$
$$\leq Cs^3\lambda^4 \int_0^T e^{-3s\widehat{\alpha}_m}(\widehat{\xi}_m)^3 \|\Delta\varphi\|^2_{L^2(\Omega)}\,dt \leq Cs^3\lambda^4 \iint_Q e^{-2s\alpha_m}\xi_m^3|\Delta\varphi|^2\,dx\,dt,$$

for $\lambda \geq C$. Here we have used the fact that φ has null trace.

Combining this with (25) and (33), we obtain

$$\begin{aligned} &s^2\lambda^4 \iint_Q e^{-3s\alpha_m}\xi_m^2 (s^2\lambda^2\xi_m^2|\psi|^2 + |\nabla\psi|^2)\,dx\,dt + I_w(\Delta\varphi) \\ &\leq C\left(s^3\lambda^4 \iint_{\omega_0\times(0,T)} \xi_m^3 \left(s\lambda^2 e^{-3s\alpha_m}\xi_m|\psi|^2 + e^{-2s\alpha_m}|\Delta\varphi|^2\right)\,dx\,dt\right), \end{aligned} \tag{34}$$

for any $\lambda \geq C$ and $s \geq C(T^{2m} + T^m)$.

Now, since $\omega_0 \subset \mathcal{O}$, from the equation satisfied by ψ, we find

$$\Delta\varphi = -\psi_t - \Delta\psi + a\psi - B\cdot\nabla\psi \ \text{ in } \omega_0\times(0,T).$$

Then we plug this into the expression of the last integral in (34) and obtain

$$s^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3|\Delta\varphi|^2\,dx\,dt$$
$$= s^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3(\Delta\varphi)(-\psi_t - \Delta\psi + a\psi - B\cdot\nabla\psi)\,dx\,dt.$$

We define a positive function $\theta \in C_c^2(\omega)$ such that $\theta \equiv 1$ in ω_0. Then the task turns to estimating the following integral:

$$s^3\lambda^4 \iint_{\omega\times(0,T)} \theta e^{-2s\alpha_m}\xi_m^3(\Delta\varphi)(-\psi_t - \Delta\psi + a\psi - B\cdot\nabla\psi)\,dx\,dt.$$

After several integration by parts (getting all derivatives out of ψ) with respect to both space and time, we get

$$s^3\lambda^4 \iint_{\omega\times(0,T)} \theta e^{-2s\alpha_m}\xi_m^3(\Delta\varphi)(-\psi_t - \Delta\psi + a\psi - B\cdot\nabla\psi)\,dx\,dt$$
$$= s^3\lambda^4 \iint_{\omega\times(0,T)} \theta(e^{-2s\alpha_m}\xi_m^3)_t\,\Delta\varphi\,\psi\,dx\,dt$$
$$- s^3\lambda^4 \iint_{\omega\times(0,T)} \Delta(\theta e^{-2s\alpha_m}\xi_m^3)\,\Delta\varphi\,\psi\,dx\,dt$$
$$- 2s^3\lambda^4 \iint_{\omega\times(0,T)} \nabla(\theta e^{-2s\alpha_m}\xi_m^3)\cdot\nabla\Delta\varphi\,\psi\,dx\,dt$$
$$+ s^3\lambda^4 \iint_{\omega\times(0,T)} B\cdot\nabla(\theta e^{-2s\alpha_m}\xi_m^3)\,\Delta\varphi\,\psi\,dx\,dt.$$

Here, we have used the equation satisfied by φ and the fact that θ has compact support in ω. Let us do some computations involving the weight functions:

$$(e^{-2s\alpha_m}\xi_m^3)_t \leq CTse^{-2s\alpha_m}(\xi_m)^{4+1/m},$$

for $s \geq CT^{2m}$ and

$$\Delta(e^{-2s\alpha_m}\xi_m^3) \leq Cs^2\lambda^2 e^{-2s\alpha_m}\xi_m^5,$$

for $s \geq CT^{2m}$ and $\lambda \geq C$. With this, we obtain

$$\begin{aligned} &s^3\lambda^4 \iint_{\omega\times(0,T)} \theta e^{-2s\alpha_m}\xi_m^3(\Delta\varphi)(-\psi_t - \Delta\psi + a\psi - B\cdot\nabla\psi)\,dx\,dt \\ &\leq \varepsilon I_w(\Delta\varphi) + C(1+T^2)s^7\lambda^8 \iint_{\omega\times(0,T)} e^{-2s\alpha_m}\xi_m^7|\psi|^2\,dx\,dt, \end{aligned}$$

which, combined with (34), gives the desired inequality (23).

3. Proof of Theorem 3. In this section we will prove Theorem 3. As indicated in the introduction, in order to prove Theorem 4 one can follow the same ideas of the proof of Theorem 3.

For simplicity, in this section we will keep the notation η^0 for the function defined in (21). In the present situation, ω_0 will stand for an open set contained in ω_2, which was also contained in ω (see the paragraph between (9) and (10)).

Throughout this section we will work with the following system (see (10)):

$$\begin{cases} -\varphi_t - \Delta\varphi + c\varphi - E\cdot\nabla\varphi = (P_1^*(t,x;D)\psi)\theta_1 & \text{in } Q, \\ -\psi_t - \Delta\psi + h\psi - K\cdot\nabla\psi = (P_2^*\varphi)\theta_2 & \text{in } Q, \\ \varphi = \psi = 0 & \text{on } \Sigma, \\ \varphi|_{t=T} = \varphi^0, \quad \psi|_{t=T} = \psi^0 & \text{in } \Omega. \end{cases} \tag{35}$$

Recall that, by (11), P_2^* is a second order differential operator in space with constant coefficients.

In order to prove Theorem 3, it suffices to establish the following observability inequality for the solutions of (35).

PROPOSITION 8. *There exists $C(\Omega,\omega,T) > 0$ independent of (φ^0,ψ^0) such that*

$$\|\varphi|_{t=0}\|^2_{L^2(\Omega)} + \|\psi|_{t=0}\|^2_{L^2(\Omega)} \leq C\iint_{\omega\times(0,T)} |\psi|^2\,dx\,dt. \tag{36}$$

As in the previous section, the strategy will consist of proving the corresponding Carleman inequality for system (35). It is presented in the following lemma.

LEMMA 9. *There exists a positive constant $C(\Omega,\omega)$ such that*

$$\begin{aligned} &I_w(P_2^*\varphi) + s^6\lambda^8 \iint_Q e^{-2s\alpha_m}\xi_m^6|\psi|^2\,dx\,dt \\ &\qquad \leq C(1+T^2)s^{10}\lambda^8 \iint_{\omega\times(0,T)} e^{-6s\alpha_m+4s\alpha_m^*}\xi_m^{10}|\psi|^2\,dx\,dt \end{aligned} \tag{37}$$

for any $\lambda \geq C$ and $s \geq C(T^{2m}+T^m)$.

Thanks to (12), the observability inequality (36) readily follows from (37).

As in the proof of Lemma 7, we first deal with the heat equation satisfied by φ and we try to obtain an independent Carleman inequality, viewing $(P_1^*(t,x;D)\psi)\theta_1$ as a right-hand side. More precisely, we consider the heat equation satisfied by $P_2^*\varphi$:

$$(P_2^*\varphi)_t + \Delta(P_2^*\varphi) + cP_2^*\varphi + E\cdot\nabla(P_2^*\varphi) = \Delta((P_1^*(t,x;D)\psi)\theta_1) \ \text{ in } Q.$$

To $P_2^*\varphi$ (as solution of the previous heat equation), we apply Lemma 6 and we obtain

$$\begin{aligned} I_w(P_2^*\varphi) \le \Bigg(& s^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3 |P_2^*\varphi|^2 \,dx\,dt \\ & + s\lambda \iint_\Sigma e^{-2s\alpha_m^*}\xi_m^* \left(\left|\frac{\partial P_2^*\varphi}{\partial n}\right|^2 + \left|\frac{\partial}{\partial n}(P_1^*(t,x;D)\psi)\right|^2 \right) dx\,dt \\ & + s^2\lambda^2 \iint_Q e^{-2s\alpha_m}\xi_m^2 (|\Delta\psi|^2 + |\nabla\psi|^2 + |\psi|^2)\,dx\,dt \Bigg), \end{aligned} \tag{38}$$

for any $\lambda \ge C$ and $s \ge C(T^{2m} + T^{2m-1})$.

Next, we estimate the boundary term in the right-hand side of (38). To this end, we define $\varphi^* = \eta(t)\varphi$, with $\eta(t)$ given by (28). This function fulfills the following system:

$$\begin{cases} \varphi_t^* + \Delta\varphi^* + c\varphi^* + E\cdot\nabla\varphi^* = \eta(t)(P_1^*(t,x;D)\psi)\theta_1 + \eta_t\varphi & \text{in } \Omega, \\ \varphi^* = 0 & \text{on } \partial\Omega, \\ \varphi^*|_{t=T} = 0 & \text{in } \Omega. \end{cases} \tag{39}$$

Assuming that the right-hand side of (39) belongs to $L^2(0,T;H^2(\Omega))$, we have $\varphi^* \in L^2(0,T;(H^4\cap H_0^1)(\Omega)) \cap H^1(0,T;H^2(\Omega))$ (see, for instance, [16]) and

$$\begin{aligned} & \|\varphi^*\|^2_{L^2(0,T;H^4(\Omega))} + \|\varphi_t^*\|^2_{L^2(0,T;H^2(\Omega))} \\ & \qquad \le C(I_w(P_2^*\varphi) + \|\eta(t)(P_1^*(t,x;D)\psi)\theta_1\|_{L^2(0,T;H^2(\Omega))}). \end{aligned} \tag{40}$$

Here we have used (12).

With the same argument as in the previous section, we get

$$\begin{aligned} & s^{(3/2)-3/(2m)}\lambda^3 \int_0^T e^{-2s\alpha_m^*}(\xi_m^*)^{(3/2)-3/(2m)} \left\| \frac{\partial P_2^*\varphi}{\partial n}\right\|^2_{L^2(\Sigma)} dt \\ & \qquad \le C(I_w(P_2^*\varphi) + \|\eta(t)(P_1^*(t,x;D)\psi)\theta_1\|_{L^2(0,T;H^2(\Omega))}). \end{aligned} \tag{41}$$

Again, since $m > 3$, this justifies that the second term in the right-hand side of (38) is absorbed.

As a conclusion, we obtain from (38)

$$\begin{aligned} I_w(P_2^*\varphi) \le C\Bigg(& s^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3 |P_2^*\varphi|^2 \,dx\,dt \\ & + s^{1-2/m}\lambda^2 \int_0^T e^{-2s\alpha_m^*}(\xi_m^*)^{1-2/m} \|(P_1^*(t,x;D)\psi)\theta_1\|^2_{H^2(\Omega)}\,dt \\ & + s^2\lambda^2 \iint_Q e^{-2s\alpha_m}\xi_m^2 (|\Delta\psi|^2 + |\nabla\psi|^2 + |\psi|^2)\,dx\,dt \Bigg), \end{aligned} \tag{42}$$

for any $\lambda \geq C$ and $s \geq C(T^{2m} + T^m)$.

Now we deal with the equation satisfied by ψ. Thus, we apply the classical Carleman inequality for the heat equation with right-hand side in $L^2(Q)$. Let us define

$$I_0(\psi) = s^6\lambda^8 \iint_Q e^{-2s\alpha_m}\xi_m^6|\psi|^2\,dx\,dt + s^4\lambda^6 \iint_Q e^{-2s\alpha_m}\xi_m^4|\nabla\psi|^2\,dx\,dt + s^2\lambda^4 \iint_Q e^{-2s\alpha_m}\xi_m^2|\Delta\psi|^2\,dx\,dt.$$

Then we have

$$I_0(\psi) \leq C\bigg(s^6\lambda^8 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^6|\psi|^2\,dx\,dt + s^3\lambda^4 \iint_Q e^{-2s\alpha_m}\xi_m^3|P_2^*\varphi|^2\,dx\,dt\bigg),$$

for $\lambda \geq C$ and $s \geq C(T^{2m} + T^{2m-1})$. Combining this with (42), we obtain

$$\begin{aligned} I_0(\psi) + I_w(P_2^*\varphi) \leq C\bigg(&s^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3|P_2^*\varphi|^2\,dx\,dt \\ &+ s^{1-2/m}\lambda^2 \int_0^T e^{-2s\alpha_m^*}(\xi_m^*)^{1-2/m}\|\psi\|^2_{H^3(\Omega)}\,dt \\ &+ s^6\lambda^8 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^6|\psi|^2\,dx\,dt\bigg), \end{aligned} \tag{43}$$

for $\lambda \geq C$ and $s \geq C(T^{2m} + T^m)$.

Let us now introduce the function $\widehat{\psi} = \rho_0(t)\psi$, with

$$\rho_0(t) = s^{1/2}\lambda e^{-s\alpha_m^*}(\xi_m^*)^{1/2}.$$

Then $\widehat{\psi}$ satisfies

$$\begin{cases} \widehat{\psi}_t + \Delta\widehat{\psi} + h\widehat{\psi} + K\cdot\nabla\widehat{\psi} = \rho_0(t)(P_2^*\varphi)\theta_2 + \rho_{0,t}\psi & \text{in } \Omega, \\ \widehat{\psi} = 0 & \text{on } \partial\Omega, \\ \widehat{\psi}|_{t=T} = 0 & \text{in } \Omega. \end{cases} \tag{44}$$

Since the right-hand side belongs to $L^2(0,T;H^1(\Omega))$, we deduce that $\widehat{\psi} \in L^2(0,T;H^3(\Omega))$ and

$$\begin{aligned} \|\widehat{\psi}\|^2_{L^2(0,T;H^3(\Omega))} \leq C\bigg(&s\lambda^2 \iint_Q e^{-2s\alpha_m}\xi_m(|\nabla(P_2^*\varphi)|^2 + |P_2^*\varphi|^2)\,dx\,dt \\ &+ T^2s^3\lambda^2 \iint_Q e^{-2s\alpha_m^*}(\xi_m^*)^{3+2/m}|\nabla\psi|^2\,dx\,dt\bigg) \\ &\leq C(I_0(\psi) + I_w(P_2^*\varphi)). \end{aligned} \tag{45}$$

Observe that in inequality (45) we did not use all the information we had about φ, because in the argument of estimating the normal derivative of $P_2^*\varphi$, we obtained

good estimates for $\|\varphi\|^2_{L^2(0,T;H^3(\Omega))}$ with the power $s^{2-1/m}$. We chose not to profit from this for the sake of simplicity.

In particular, from (45) we deduce that the second term in the right-hand side of (43) is absorbed.

For the moment, we have

$$
\begin{aligned}
&I_0(\psi) + I_w(P_2^*\varphi) + \|\widehat{\psi}\|^2_{L^2(0,T;H^3(\Omega))} \\
&\qquad \le C\left(s^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3 |P_2^*\varphi|^2 \,dx\,dt \right. \\
&\qquad\qquad \left. + s^6\lambda^8 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^6 |\psi|^2 \,dx\,dt \right),
\end{aligned} \tag{46}
$$

for $\lambda \ge C$ and $s \ge C(T^{2m} + T^m)$.

We will finally estimate the local term concerning $P_2^*\varphi$. From the equation satisfied by ψ, we have

$$
P_2^*\varphi = \psi_t + \Delta\psi + h\psi + K\cdot\nabla\psi \ \text{ in } \omega_0\times(0,T)
$$

(recall that supp $\theta_2 \subset \overline{\omega}_2$ and $\omega_0 \subset \omega_2$). This gives

$$
\begin{aligned}
&s^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3 |P_2^*\varphi|^2 \,dx\,dt \\
&= s^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3 (P_2^*\varphi)(\psi_t + \Delta\psi + h\psi + K\cdot\nabla\psi)\,dx\,dt.
\end{aligned}
$$

As in the previous section, with the help of a cut-off function $\theta_0 \in C_c^2(\omega)$ with $\theta_0 \equiv 1$ in ω_0, we can integrate by parts with respect to both time and space, and we find

$$
\begin{aligned}
&s^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3 |P_2^*\varphi|^2 \,dx\,dt \le \varepsilon(I_w(\Delta\varphi) + I_0(\psi)) \\
&+ C(1+T^2)s^6\lambda^6 \iint_{\omega\times(0,T)} f(\theta_0) e^{-4s\alpha_m + 2s\alpha_m^*}\xi_m^6 (|\psi|^2 + |\nabla\psi|^2)\,dx\,dt,
\end{aligned}
$$

for some $f(\theta_0) \in C_c^1(\omega)$. Observe that in order to estimate the first term

$$
\iint_{\omega\times(0,T)} e^{-2s\alpha_m}\xi_m^3\theta_0\, P_2^*\varphi\, \psi_t \,dx\,dt,
$$

we have integrated in time and used (40) for the integral

$$
\iint_{\omega\times(0,T)} e^{-2s\alpha_m}\xi_m^3\theta_0\, P_2^*\varphi_t\, \psi \,dx\,dt.
$$

We integrate by parts again and obtain

$$
\begin{aligned}
&s^3\lambda^4 \iint_{\omega_0\times(0,T)} e^{-2s\alpha_m}\xi_m^3 |P_2^*\varphi|^2 \,dx\,dt \\
&\qquad \le \varepsilon\left(I_w(P_2^*\varphi) + I_0(\psi) + s^2\lambda^4 \iint_Q e^{-2s\alpha_m}\xi_m^2 |\Delta\psi|^2 \,dx\,dt \right) \\
&\qquad + C(1+T^2)s^{10}\lambda^8 \iint_{\omega\times(0,T)} e^{-6s\alpha_m + 4s\alpha_m^*}\xi_m^{10} |\psi|^2 \,dx\,dt.
\end{aligned}
$$

From this and (46), we deduce the desired Carleman inequality (37). This finishes the proof of Theorem 3. □

REFERENCES

[1] F. Ammar Khodja, A. Benabdallah, C. Dupaix, and I. Kostin, *Controllability to the trajectories of phase-field models by one control force*, SIAM J. Control Optim., 42 (2003), pp. 1661–1680.

[2] F. Ammar Khodja, A. Benabdallah, C. Dupaix, and I. Kostin, *Null-controllability of some systems of parabolic type by one control force*, ESAIM Control Optim. Calc. Var., 11 (2005), pp. 426–448.

[3] V. Barbu, *Local controllability of the phase field system*, Nonlinear Anal. Ser. A Theory Methods, 50 (2002), pp. 363–372.

[4] O. Bodart and C. Fabre, *Controls insensitizing the norm of the solution of a semilinear heat equation*, J. Math. Anal. Appl., 195 (1995), pp. 658–683.

[5] O. Bodart, M. González-Burgos, and R. Pérez-García, *Existence of insensitizing controls for a semilinear heat equation with a superlinear nonlinearity*, Comm. Partial Differential Equations, 29 (2004), pp. 1017–1050.

[6] L. De Teresa, *Insensitizing controls for a semilinear heat equation*, Comm. Partial Differential Equations, 25 (2000), pp. 39–72.

[7] L. De Teresa and O. Kavian, *Unique Continuation Principle for Systems of Parabolic Equations*, in preparation.

[8] E. Fernández-Cara, M. González-Burgos, S. Guerrero, and J.-P. Puel, *Null controllability of the heat equation with boundary Fourier conditions: The linear case*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 442–465.

[9] E. Fernández-Cara and S. Guerrero, *Global Carleman inequalities for parabolic systems and applications to controllability*, SIAM J. Control Optim., 45 (2006), pp. 1395–1446.

[10] E. Fernández-Cara and S. Guerrero, *Local exact controllability of micropolar fluids*, J. Math. Fluid Mech., to appear.

[11] A. V. Fursikov and O.-Y. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes 34, Seoul National University, Korea, 1996.

[12] M. González-Burgos and R. Pérez-García, *Controllability results for some nonlinear coupled parabolic systems by one control force*, Asymptot. Anal., 46 (2006), pp. 123–162.

[13] S. Guerrero, *Controllability of systems of Stokes equations: Existence of insensitizing controls*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[14] O.-Y. Imanuvilov, *Controllability of parabolic equations*, Sb. Math., 186 (1995), pp. 879–900.

[15] O.-Y. Imanuvilov and M. Yamamoto, *Carleman inequalities for parabolic equations in Sobolev spaces of negative order and exact controllability for semilinear parabolic equations*, Publ. Res. Inst. Math. Sci., 39 (2003), pp. 227–274.

[16] O. A. Ladyzenskaya, V. A. Solonnikov, and N. N. Uraltzeva, *Linear and Quasilinear Equations of Parabolic Type*, Trans. Math. Monographs: Moscow 23, AMS, Providence, RI, 1967.

[17] J.-L. Lions, *Quelques notions dans l'analyse et le contrôle de systèmes à données incomplètes [Some notions in the analysis and control of systems with incomplete data]*, in Proceedings of the XIth Congress on Differential Equations and Applications/First Congress on Applied Mathematics (Spanish), University of Málaga, Málaga, Spain, 1990, pp. 43–54.

[18] J.-L. Lions, *Sentinelles pour les systèmes distribués à données incomplètes [Sentinels for distributed systems with incomplete data],* Recherches en Mathématiques Appliquées, 21, Masson, Paris, 1992.

[19] R. Pérez-García, *Algunos resultados de control para algunos problemas parabólicos acoplados no lineales: Controlabilidad y controles insensibilizantes*, Ph.D. thesis, University of Sevilla, Sevilla, Spain, 2004.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 395–426

EXPLICIT SOLUTION TO AN OPTIMAL SWITCHING PROBLEM IN THE TWO-REGIME CASE*

VATHANA LY VATH† AND HUYÊN PHAM†

Abstract. This paper considers the problem of determining the optimal sequence of stopping times for a diffusion process subject to regime switching decisions. This is motivated in the economics literature by the investment problem under uncertainty for a multi-activity firm involving opening and closing decisions. We use a viscosity solutions approach combined with the smooth-fit property, and explicitly solve the problem in the two-regime case when the state process is of geometric Brownian nature. The results of our analysis take several qualitatively different forms, depending on model parameter values.

Key words. optimal switching, system of variational inequalities, viscosity solutions, smooth-fit principle

AMS subject classifications. 60G40, 49L25, 62L15

DOI. 10.1137/050638783

1. Introduction.

The theory of optimal stopping and its generalization, thoroughly studied in the 1970s, has received a renewed interest with a variety of applications in economics and finance. These applications include asset pricing (American options, swing options), firm investment, and real options. We refer to [4] for a classical and well-documented reference on the subject.

In this paper, we consider the optimal switching problem for a one-dimensional stochastic process X. The diffusion process X may take a finite number of regimes that are switched at stopping time decisions. For example, in the firm's investment problem under uncertainty, a company (oil tanker, electricity station, etc.) manages several production activities operating in different modes or regimes representing a number of different economic outlooks (e.g., state of economic growth, open or closed production activity). The process X is the price of input or output goods of the firm and its dynamics may differ according to the regimes. The firm's project yields a running payoff that depends on the commodity price X and on the regime choice. The transition from one regime to another is realized sequentially at time decisions and incurs certain fixed costs. The problem is to find the switching strategy that maximizes the expected value of profits resulting from the project.

Optimal switching problems were studied by several authors; see [1] or [10]. These control problems lead, via the dynamic programming principle, to a system of variational inequalities. Applications to option pricing, real options, and investment under uncertainty were considered in [2], [5], and [7]. In this last paper, the drift and volatility of the state process depend on an uncontrolled finite-state Markov chain, and the author provides an explicit solution to the optimal stopping problem with applications to Russian options. In [2], an explicit solution is found for a resource extraction problem with two regimes (open or closed field), a linear profit function, and a price process following a geometric Brownian motion. In [5], a similar model is solved with a general profit function in one regime and equal to zero in the other. In

*Received by the editors August 24, 2005; accepted for publication (in revised form) October 24, 2006; published electronically April 17, 2007.
http://www.siam.org/journals/sicon/46-2/63878.html

†Laboratoire de Probabilités et Modèles Aléatoires, CNRS, UMR 7599, Université Paris 7, 75251 Paris, Cedex 05, France (lyvath@math.jussieu.fr, pham@math.jussieu.fr).

both models [2], [5], there is no switching in the diffusion process: changes of regimes affect only the payoff functions. Their method of resolution is to construct a solution to the dynamic programming system by guessing a priori the form of the strategy, and then validate a posteriori the optimality of their candidate by a verification argument. Our model combines regime switchings on both the diffusion process and the general profit functions. We use a viscosity solutions approach for determining the solution to the system of variational inequalities. In particular, we derive directly the smooth-fit property of the value functions and the structure of the switching regions. Explicit solutions are provided in the following cases:

- the drift and volatility terms of the diffusion take two different regime values, and the profit functions are identical of power type;
- there is no switching on the diffusion process, and the two different profit functions satisfy a general condition, including typically power functions.

We also consider the cases for which both switching costs are positive, and for which one of the two is negative. This last case is interesting in applications where a firm chooses between an open or closed activity, and may regain a fraction of its opening costs when it decides to close. The results of our analysis take several qualitatively different forms, depending on model parameter values, essentially the payoff functions and the switching costs.

The paper is organized as follows. We formulate in section 2 the optimal switching problem. In section 3, we state the system of variational inequalities satisfied by the value functions in the viscosity sense. The smooth-fit property for this problem, proved in [9], plays an important role in our subsequent analysis. We also state some useful properties on the switching regions. In section 4, we explicitly solve the problem in the two-regime case when the state process is of geometric Brownian nature.

2. Formulation of the optimal switching problem. We consider a stochastic system that can operate in d modes or regimes. The regimes can be switched at a sequence of stopping times decided by the operator (individual, firm, etc.). The indicator of the regimes is modeled by a cadlag process I_t valued in $\mathbb{I}_d = \{1, \dots, d\}$. The stochastic system X (commodity price, salary, etc.) is valued in $\mathbb{R}_+^* = (0, \infty)$ and satisfies the SDE.

$$dX_t = b_{I_t} X_t dt + \sigma_{I_t} X_t dW_t, \tag{2.1}$$

where W is a standard Brownian motion on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t \geq 0}, P)$ satisfying the usual conditions. $b_i \in \mathbb{R}$ and $\sigma_i > 0$ are the drift and volatility, respectively, of the system X once in regime $I_t = i$ at time t.

A strategy decision for the operator is an impulse control α consisting of a double sequence $\tau_1, \dots, \tau_n, \dots, \kappa_1, \dots, \kappa_n, \dots$, $n \in \mathbb{N}^* = \mathbb{N} \setminus \{0\}$, where τ_n are stopping times, $\tau_n < \tau_{n+1}$ and $\tau_n \to \infty$ a.s., representing the switching regimes time decisions, and κ_n are $\mathcal{F}_{\tau_n}$-measurable valued in $\mathbb{I}_d$ representing the new value of the regime at time $t = \tau_n$. We denote by $\mathcal{A}$ the set of all such impulse controls. Now, for any initial condition $(x, i) \in (0, \infty) \times \mathbb{I}_d$, and any control $\alpha = (\tau_n, \kappa_n)_{n \geq 1} \in \mathcal{A}$, there exists a unique strong solution valued in $(0, \infty) \times \mathbb{I}_d$ to the controlled stochastic system:

$$X_0 = x, \quad I_{0^-} = i, \tag{2.2}$$

$$dX_t = b_{\kappa_n} X_t dt + \sigma_{\kappa_n} X_t dW_t, \quad I_t = \kappa_n, \quad \tau_n \leq t < \tau_{n+1}, \quad n \geq 0. \tag{2.3}$$

Here, we set $\tau_0 = 0$ and $\kappa_0 = i$. We denote by $(X^{x,i}, I^i)$ this solution (as usual, we omit the dependence in α for notational simplicity). We notice that $X^{x,i}$ is a continuous process and I^i is a cadlag process, possibly with a jump at time 0 if $\tau_1 = 0$, and so $I_0 = \kappa_1$.

We are given a running profit function $f : \mathbb{R}_+ \times \mathbb{I}_d \to \mathbb{R}$ and we set $f_i(.) = f(., i)$ for $i \in \mathbb{I}_d$. We assume that for each $i \in \mathbb{I}_d$, the function f_i is nonnegative and is Hölder continuous on $\mathbb{R}_+$: there exists $\gamma_i \in (0, 1]$ such that (s.t.)

$$|f_i(x) - f_i(\hat{x})| \leq C|x - \hat{x}|^{\gamma_i} \quad \forall x, \hat{x} \in \mathbb{R}_+, \tag{2.4}$$

for some positive constant C. Without loss of generality (see Remark 2.1), we may assume that $f_i(0) = 0$. We also assume that for all $i \in \mathbb{I}_d$, the conjugate of f_i is finite on $(0, \infty)$:

$$\tilde{f}_i(y) := \sup_{x \geq 0} \big[f_i(x) - xy \big] \ < \ \infty \quad \forall y > 0. \tag{2.5}$$

The cost for switching from regime i to j is a constant equal to g_{ij}, with the convention $g_{ii} = 0$, and we assume the triangular condition

$$g_{ik} < g_{ij} + g_{jk}, \quad j \neq i, k. \tag{2.6}$$

This last condition means that it is less expensive to switch directly in one step from regime i to k than in two steps via an intermediate regime j. Notice that a switching cost g_{ij} may be negative, and condition (2.6) for $i = k$ prevents arbitrage by switching back and forth, i.e.,

$$g_{ij} + g_{ji} > 0, \quad i \neq j \in \mathbb{I}_d. \tag{2.7}$$

The expected total profit of running the system when the initial state is (x, i) and when using the impulse control $\alpha = (\tau_n, \kappa_n)_{n \geq 1} \in \mathcal{A}$ is

$$J_i(x, \alpha) = E \left[\int_0^\infty e^{-rt} f(X_t^{x,i}, I_t^i) dt - \sum_{n=1}^\infty e^{-r\tau_n} g_{\kappa_{n-1}, \kappa_n} \right].$$

Here $r > 0$ is a positive discount factor, and we use the convention that $e^{-r\tau_n(\omega)} = 0$ when $\tau_n(\omega) = \infty$. We also make the standing assumption

$$r > b := \max_{i \in \mathbb{I}_d} b_i. \tag{2.8}$$

The objective is to maximize this expected total profit over all strategies α. Accordingly, we define the value functions

$$v_i(x) = \sup_{\alpha \in \mathcal{A}} J_i(x, \alpha), \quad x \in \mathbb{R}_+^*, \ i \in \mathbb{I}_d. \tag{2.9}$$

We shall see in the next section that under (2.5) and (2.8), the expectation defining $J_i(x)$ is well defined and the value function v_i is finite.

Remark 2.1. The initial values $f_i(0)$ of the running profit functions received by the firm manager (the controller) before any decisions occur are considered to be included in the switching costs during changing of the regime. This means that

w.l.o.g. we may assume that $f_i(0) = 0$. Indeed, for any profit function f_i, and by setting $\tilde{f}_i = f_i - f_i(0)$, we have for all $x > 0$, $\alpha \in \mathcal{A}$,

$$\begin{aligned}
J_i(x,\alpha) &= E\left[\sum_{n=1}^{\infty}\int_{\tau_{n-1}}^{\tau_n} e^{-rt} f(X_t^{x,i},\kappa_{n-1})dt - \sum_{n=1}^{\infty} e^{-r\tau_n} g_{\kappa_{n-1},\kappa_n}\right] \\
&= E\left[\sum_{n=1}^{\infty}\int_{\tau_{n-1}}^{\tau_n} e^{-rt}\left(\tilde{f}(X_t^{x,i},\kappa_{n-1}) + f_{\kappa_{n-1}}(0)\right)dt \; - \sum_{n=1}^{\infty} e^{-r\tau_n} g_{\kappa_{n-1},\kappa_n}\right] \\
&= E\left[\sum_{n=1}^{\infty}\int_{\tau_{n-1}}^{\tau_n} e^{-rt}\tilde{f}(X_t^{x,i},\kappa_{n-1})dt + \frac{f_{\kappa_0}(0)}{r}\right. \\
&\qquad\qquad \left. - \sum_{n=1}^{\infty} e^{-r\tau_n}\left(g_{\kappa_{n-1},\kappa_n} + \frac{f_{\kappa_n}(0) - f_{\kappa_{n-1}}(0)}{r}\right)\right] \\
&= \frac{f_i(0)}{r} + E\left[\int_0^{\infty} e^{-rt}\tilde{f}(X_t^{x,i},I_t^i)dt - \sum_{n=1}^{\infty} e^{-r\tau_n}\tilde{g}_{\kappa_{n-1},\kappa_n}\right],
\end{aligned}$$

with modified switching costs that take into account the possibly different initial values of the profit functions,

$$\tilde{g}_{ij} = g_{ij} + \frac{f_j(0) - f_i(0)}{r},$$

and that are assumed to satisfy the triangle inequality $\tilde{g}_{ik} < \tilde{g}_{ij} + \tilde{g}_{jk}$, $j \neq i,k$.

3. System of variational inequalities, switching regions, and viscosity solutions. We first state the linear growth property and the boundary condition on the value functions.

LEMMA 3.1. *We have for all $i \in \mathbb{I}_d$,*

$$\max_{j\in\mathbb{I}_d}[-g_{ij}] \; \leq v_i(x) \leq \; \frac{xy}{r-b} \; + \; \max_{j\in\mathbb{I}_d}\frac{\tilde{f}_j(y)}{r} + \max_{j\in\mathbb{I}_d}[-g_{ij}] \; \forall x > 0 \; \forall y > 0. \tag{3.1}$$

In particular, we have $v_i(0^+) = \max_{j\in\mathbb{I}_d}[-g_{ij}]$.

Proof. By considering the particular strategy $\tilde{\alpha} = (\tilde{\tau}_n, \tilde{\kappa}_n)$ of immediately switching from the initial state (x,i) to state (x,j), $j \in \mathbb{I}_d$ (eventually equal to i), at cost g_{ij} and then doing nothing, i.e., $\tilde{\tau}_1 = 0$, $\tilde{\kappa}_1 = j$, $\tilde{\tau}_n = \infty$, $\tilde{\kappa}_n = j$ for all $n \geq 2$, we have

$$J_i(x,\tilde{\alpha}) = E\left[\int_0^{\infty} e^{-rt} f_j(\tilde{X}_t^{x,j})dt - g_{ij}\right],$$

where $\tilde{X}^{x,j}$ denotes the geometric Brownian in regime j starting from x at time 0. Since f_j is nonnegative, and by the arbitrariness of j, we get the lower bound in (3.1).

Given an initial state $(X_0, I_{0^-}) = (x,i)$ and an arbitrary impulse control $\alpha = (\tau_n,\kappa_n)$, we get from the dynamics (2.2)–(2.3) the following explicit expression of $X^{x,i}$:

$$\begin{aligned}
X_t^{x,i} \; &= \; xY_t(i) \\
&:= x\left(\prod_{l=0}^{n-1} e^{b_{\kappa_l}(\tau_{l+1}-\tau_l)} Z^{\kappa_l}_{\tau_l,\tau_{l+1}}\right) e^{b_{\kappa_n}(t-\tau_n)} Z^{\kappa_n}_{\tau_n,t}, \quad \tau_n \leq t < \tau_{n+1}, \quad n \in \mathbb{N},
\end{aligned} \tag{3.2}$$

where

$$Z^j_{s,t} = \exp\left(\sigma_j(W_t - W_s) - \frac{\sigma_j^2}{2}(t-s)\right), \quad 0 \le s \le t, \quad j \in \mathbb{I}_d. \tag{3.3}$$

Here, we used the conventions that $\tau_0 = 0$, $\kappa_0 = i$, and that the product term from l to $n-1$ in (3.2) is equal to 1 when $n = 1$. We then deduce the inequality $X^{x,i}_t \le xe^{bt}M_t$ for all t, where

$$M_t = \left(\prod_{l=0}^{n-1} Z^{\kappa_l}_{\tau_l,\tau_{l+1}}\right) Z^{\kappa_n}_{\tau_n,t}, \quad \tau_n \le t < \tau_{n+1}, \quad n \in \mathbb{N}. \tag{3.4}$$

Now, we notice that (M_t) is a martingale obtained by continuously patching the martingales $(Z^{\kappa_{n-1}}_{\tau_{n-1},t})$ and $(Z^{\kappa_n}_{\tau_n,t})$ at the stopping times τ_n, $n \ge 1$. In particular, we have $E[M_t] = M_0 = 1$ for all t.

We set $\tilde{f}(y) = \max_{j\in\mathbb{I}_d} \tilde{f}_i(y)$, $y > 0$, and we notice by definition of $\tilde{f}_i$ in (2.5) that $f(X^{x,i}_t, I^i_t) \le yX^{x,i}_t + \tilde{f}(y)$ for all t, y. Moreover, we show by induction on N that for all $N \ge 1$, $\tau_1 \le \cdots \le \tau_N$, $\kappa_0 = i$, $\kappa_n \in \mathbb{I}_d$, $n = 1, \dots, N$,

$$-\sum_{n=1}^{N} e^{-r\tau_n} g_{\kappa_{n-1},\kappa_n} \le \max_{j\in\mathbb{I}_d}[-g_{ij}] \quad \text{a.s.}$$

Indeed, the above assertion is obviously true for $N = 1$. Suppose now it holds true at step N. Then, at step $N+1$, we distinguish two cases:

- If $g_{\kappa_N,\kappa_{N+1}} \ge 0$, then we have $-\sum_{n=1}^{N+1} e^{-r\tau_n} g_{\kappa_{n-1},\kappa_n} \le -\sum_{n=1}^{N} e^{-r\tau_n} g_{\kappa_{n-1},\kappa_n}$ and we conclude by the induction hypothesis at step N.
- If $g_{\kappa_N,\kappa_{N+1}} < 0$, then by (2.6), and since $\tau_N \le \tau_{N+1}$, we have $-e^{-r\tau_N} g_{\kappa_{N-1},\kappa_N} - e^{-r\tau_{N+1}} g_{\kappa_N,\kappa_{N+1}} \le e^{-r\tau_N} g_{\kappa_{N-1},\kappa_{N+1}}$, and so $-\sum_{n=1}^{N+1} e^{-r\tau_n} g_{\kappa_{n-1},\kappa_n} \le -\sum_{n=1}^{N} e^{-r\tau_n} g_{\tilde{\kappa}_{n-1},\tilde{\kappa}_n}$, with $\tilde{\kappa}_n = \kappa_n$ for $n = 1, \dots, N-1$, $\tilde{\kappa}_N = \kappa_{N+1}$.

We then conclude by the induction hypothesis at step N.

It follows that

$$\begin{aligned} J_i(x,\alpha) &\le E\left[\int_0^\infty e^{-rt}\left(yxe^{bt}M_t + \tilde{f}(y)\right)dt + \max_{j\in\mathbb{I}_d}[-g_{ij}]\right] \\ &= \int_0^\infty e^{-(r-b)t} yxE[M_t]dt + \int_0^\infty e^{-rt}\tilde{f}(y)dt + \max_{j\in\mathbb{I}_d}[-g_{ij}] \\ &= \frac{xy}{r-b} + \frac{\tilde{f}(y)}{r} + \max_{j\in\mathbb{I}_d}[-g_{ij}]. \end{aligned}$$

From the arbitrariness of α, this shows the upper bound for v_i.

By sending x to zero and then y to infinity into the r.h.s. of (3.1), and recalling that $\tilde{f}_i(\infty) = f_i(0) = 0$ for $i \in \mathbb{I}_d$, we conclude that v_i goes to $\max_{j\in\mathbb{I}_d}[-g_{ij}]$ when x tends to zero. □

We next show the Hölder continuity of the value functions.

LEMMA 3.2. *For all $i \in \mathbb{I}_d$, v_i is Hölder continuous on $(0,\infty)$:*

$$|v_i(x) - v_i(\hat{x})| \le C|x - \hat{x}|^\gamma \quad \forall x, \hat{x} \in (0,\infty) \ \text{ with } \ |x - \hat{x}| \le 1,$$

for some positive constant C, and where $\gamma = \min_{i\in\mathbb{I}_d} \gamma_i$ of condition (2.4).

Proof. By definition (2.9) of v_i and under condition (2.4), we have for all $x, \hat{x} \in (0,\infty)$, with $|x-\hat{x}| \leq 1$,

$$
\begin{aligned}
|v_i(x) - v_i(\hat{x})| &\leq \sup_{\alpha\in\mathcal{A}} |J_i(x,\alpha) - J_i(\hat{x},\alpha)| \\
&\leq \sup_{\alpha\in\mathcal{A}} E\left[\int_0^\infty e^{-rt}\left|f(X_t^{x,i},I_t^i) - f(X_t^{\hat{x},i},I_t^i)\right| dt\right] \\
&\leq C \sup_{\alpha\in\mathcal{A}} E\left[\int_0^\infty e^{-rt}\left|X_t^{x,i} - X_t^{\hat{x},i}\right|^{\gamma_{I_t^i}} dt\right] \\
&= C \sup_{\alpha\in\mathcal{A}} \int_0^\infty E\left[e^{-rt}|x-\hat{x}|^{\gamma_{I_t^i}} |Y_t(i)|^{\gamma_{I_t^i}} dt\right] \\
&\leq C|x-\hat{x}|^\gamma \sup_{\alpha\in\mathcal{A}} \int_0^\infty e^{-(r-b)t} E|M_t|^{\gamma_{I_t^i}} dt \qquad (3.5)
\end{aligned}
$$

by (3.2) and (3.4). For any $\alpha = (\tau_n,\kappa_n)_n \in \mathcal{A}$, by the independence of $(Z^{\kappa_n}_{\tau_n,\tau_{n+1}})_n$ in (3.3), and since

$$
E\left[\left|Z^{\kappa_n}_{\tau_n,\tau_{n+1}}\right|^{\gamma_{\kappa_n}} \middle| \mathcal{F}_{\tau_n}\right] = E\left[\exp\left(\gamma_{\kappa_n}(\gamma_{\kappa_n}-1)\frac{\sigma^2_{\kappa_n}}{2}(\tau_{n+1}-\tau_n)\right)\middle|\mathcal{F}_{\tau_n}\right] \leq 1 \quad \text{a.s.},
$$

we clearly see that $E|M_t|^{\gamma_{I_t^i}} \leq 1$ for all $t \geq 0$. We thus conclude with (3.5). □

The dynamic programming principle combined with the notion of viscosity solutions is known to be a general and powerful tool for characterizing the value function of a stochastic control problem via a PDE representation; see [6]. We recall the definition of viscosity solutions for a PDE in the form

$$
H(x,v,D_xv,D^2_{xx}v) = 0, \quad x \in \mathcal{O}, \qquad (3.6)
$$

where $\mathcal{O}$ is an open subset in $\mathbb{R}^n$ and H is a continuous function and nonincreasing in its last argument (with respect to the order of symmetric matrices).

DEFINITION 3.3. *Let v be a continuous function on $\mathcal{O}$. We say that v is a viscosity solution to* (3.6) *on $\mathcal{O}$ if it is*

(i) *a viscosity supersolution to* (3.6) *on $\mathcal{O}$: For any $\bar{x} \in \mathcal{O}$ and any C^2 function φ in a neighborhood of $\bar{x}$ s.t. $\bar{x}$ is a local minimum of $v-\varphi$, we have*

$$
H(\bar{x}, v(\bar{x}), D_x\varphi(\bar{x}), D^2_{xx}\varphi(\bar{x})) \geq 0;
$$

and

(ii) *a viscosity subsolution to* (3.6) *on $\mathcal{O}$: For any $\bar{x} \in \mathcal{O}$ and any C^2 function φ in a neighborhood of $\bar{x}$ s.t. $\bar{x}$ is a local maximum of $v-\varphi$, we have*

$$
H(\bar{x}, v(\bar{x}), D_x\varphi(\bar{x}), D^2_{xx}\varphi(\bar{x})) \leq 0.
$$

Remark 3.1. 1. By misuse of notation, we shall say that v is a viscosity supersolution (resp., subsolution) to (3.6) by writing

$$
H(x,v,D_xv,D^2_{xx}v) \geq (\text{resp.}, \leq)\ 0, \quad x \in \mathcal{O}, \qquad (3.7)
$$

2. We recall that if v is a smooth C^2 function on $\mathcal{O}$, supersolution (resp., subsolution) in the classical sense to (3.7), then v is a viscosity supersolution (resp., subsolution) to (3.7).

3. There is an equivalent formulation of viscosity solutions, which is useful for proving uniqueness results (see [3]).

(i) A continuous function v on $\mathcal{O}$ is a viscosity supersolution to (3.6) if

$$H(x, v(x), p, M) \geq 0 \quad \forall x \in \mathcal{O},\ \forall (p, M) \in J^{2,-}v(x).$$

(ii) A continuous function v on $\mathcal{O}$ is a viscosity subsolution to (3.6) if

$$H(x, v(x), p, M) \leq 0 \quad \forall x \in \mathcal{O},\ \forall (p, M) \in J^{2,+}v(x).$$

Here $J^{2,+}v(x)$ is the second order superjet defined by

$$J^{2,+}v(x) = \left\{ (p, M) \in \mathbb{R}^n \times S^n : \right.$$

$$\left. \limsup_{\substack{x' \to x \\ x \in \mathcal{O}}} \frac{v(x') - v(x) - p.(x' - x) - \frac{1}{2}(x' - x).M(x' - x)}{|x' - x|^2} \ \leq\ 0 \right\},$$

S^n is the set of symmetric $n \times n$ matrices, and $J^{2,-}v(x) = -J^{2,+}(-v)(x)$.

In what follows, we shall denote by $\mathcal{L}_i$ the second order operator associated with the diffusion X when we are in regime i: for any C^2 function φ on $(0, \infty)$,

$$\mathcal{L}_i \varphi = \frac{1}{2}\sigma_i^2 x^2 \varphi'' + b_i x \varphi'.$$

We then have the following PDE characterization of the value functions v_i by means of viscosity solutions.

THEOREM 3.4. *The value functions v_i, $i \in \mathbb{I}_d$, are the unique viscosity solutions, with linear growth conditions on $(0, \infty)$ and boundary conditions $v_i(0^+) = \max_{j \in \mathbb{I}_d}[-g_{ij}]$, to the system of variational inequalities:*

$$\min \left\{ rv_i - \mathcal{L}_i v_i - f_i\ ,\ v_i - \max_{j \neq i}(v_j - g_{ij}) \right\} = 0, \quad x \in (0, \infty), \quad i \in \mathbb{I}_d. \tag{3.8}$$

This means we have the following properties.

(1) *Viscosity property. For each $i \in \mathbb{I}_d$, v_i is a viscosity solution to*

$$\min \left\{ rv_i - \mathcal{L}_i v_i - f_i\ ,\ v_i - \max_{j \neq i}(v_j - g_{ij}) \right\} = 0, \quad x \in (0, \infty). \tag{3.9}$$

(2) *Uniqueness property. If w_i, $i \in \mathbb{I}_d$, are viscosity solutions, with linear growth conditions on $(0, \infty)$ and boundary conditions $w_i(0^+) = \max_{j \in \mathbb{I}_d}[-g_{ij}]$, to the system of variational inequalities* (3.8)*, then $v_i = w_i$ on $(0, \infty)$.*

Proof. (1) The viscosity property follows from the dynamic programming principle and is proved in [9].

(2) Uniqueness results for switching problems have been proved in [10] in the finite horizon case under different conditions. For sake of completeness, we provide in the appendix a proof of the comparison principle in our infinite horizon context, which implies the uniqueness result. □

Remark 3.2. For fixed $i \in \mathbb{I}_d$, we also have uniqueness of viscosity solutions to (3.9) in the class of continuous functions with linear growth conditions on $(0, \infty)$ and

given boundary conditions on 0. In the next section, we shall use either uniqueness of viscosity solutions to the system (3.8) or for fixed i to (3.9), for the identification of an explicit solution in the two-regime case $d = 2$.

We shall also combine the uniqueness result for the viscosity solutions with the smooth-fit property on the value functions that we state below.

For any regime $i \in \mathbb{I}_d$, we introduce the switching region

$$\mathcal{S}_i = \left\{ x \in (0,\infty) : v_i(x) = \max_{j \neq i} (v_j - g_{ij})(x) \right\}.$$

$\mathcal{S}_i$ is a closed subset of $(0,\infty)$ and corresponds to the region where it is optimal for the operator to change regime. The complement set $\mathcal{C}_i$ of $\mathcal{S}_i$ in $(0,\infty)$ is the so-called continuation region:

$$\mathcal{C}_i = \left\{ x \in (0,\infty) : v_i(x) > \max_{j \neq i} (v_j - g_{ij})(x) \right\},$$

where the operator remains in regime i. In this open domain, the value function v_i is smooth C^2 on $\mathcal{C}_i$ and satisfies, in a classical sense,

$$r v_i(x) - \mathcal{L}_i v_i(x) - f_i(x) = 0, \quad x \in \mathcal{C}_i.$$

As a consequence of the condition (2.6), we have the following elementary partition property of the switching regions (see Lemma 4.2 in [9]):

$$\mathcal{S}_i = \cup_{j \neq i} \mathcal{S}_{ij}, \quad i \in \mathbb{I}_d,$$

where

$$\mathcal{S}_{ij} = \{ x \in \mathcal{C}_j : v_i(x) = (v_j - g_{ij})(x) \}.$$

$\mathcal{S}_{ij}$ represents the region where it is optimal to switch from regime i to regime j and to remain for a moment, i.e., without changing instantaneously from regime j to another regime. The following lemma gives some partial information about the structure of the switching regions.

LEMMA 3.5. *For all $i \neq j$ in $\mathbb{I}_d$, we have*

$$\mathcal{S}_{ij} \subset Q_{ij} := \{ x \in \mathcal{C}_j : (\mathcal{L}_j - \mathcal{L}_i) v_j(x) + (f_j - f_i)(x) - r g_{ij} \geq 0 \}.$$

Proof. Let $x \in \mathcal{S}_{ij}$. By setting $\varphi_j = v_j - g_{ij}$, this means that x is a minimum of $v_i - \varphi_j$ with $v_i(x) = \varphi_j(x)$. Moreover, since x lies in the open set $\mathcal{C}_j$ where v_j is smooth, we have that φ_j is C^2 in a neighborhood of x. By the supersolution viscosity property of v_i to the PDE (3.8), this yields

$$r\varphi_j(x) - \mathcal{L}_i \varphi_j(x) - f_i(x) \geq 0. \tag{3.10}$$

Now recall that for $x \in \mathcal{C}_j$, we have

$$r v_j(x) - \mathcal{L}_j v_j(x) - f_j(x) = 0,$$

so that by substituting into (3.10), we obtain

$$(\mathcal{L}_j - \mathcal{L}_i) v_j(x) + (f_j - f_i)(x) - r g_{ij} \geq 0,$$

which is the required result. □

We quote the smooth-fit property on the value functions, proved in [9].

THEOREM 3.6. *For all $i \in \mathbb{I}_d$, the value function v_i is continuously differentiable on $(0,\infty)$.*

Remark 3.3. In a given regime i, the variational inequality satisfied by the value function v_i is a free-boundary problem as in the optimal stopping problem, which divides the state space into the switching region (stopping region in the pure optimal stopping problem) and the continuation region. The main difficulty with regard to optimal stopping problems for proving the smooth-fit property through the boundaries of the switching regions, is that the switching region for the value function v_i depends also on the other value functions v_j. The method in [9] uses viscosity solutions arguments, and the condition of one-dimensional state space is critical for proving the smooth-fit property. The crucial conditions in this paper require that the diffusion coefficient in any regime of the system X be strictly positive on the interior of the state space, which is the case here since $\sigma_i > 0$ for all $i \in \mathbb{I}_d$, and a triangular condition (2.6) on the switching costs. Under these conditions, on a point x of the switching region $\mathcal{S}_i$ for regime i, there exists some $j \neq i$ s.t. $x \in \mathcal{S}_{ij}$, i.e., $v_i(x) = v_j(x) - g_{ij}$, and the C^1 property of the value functions is written as $v_i'(x) = v_j'(x)$ since g_{ij} is constant.

The next result provides suitable conditions for determining a viscosity solution to the variational inequality type arising in our switching problem.

LEMMA 3.7. *Fix $i \in \mathbb{I}_d$. Let $\mathcal{C}$ be an open set in $(0,\infty)$, $\mathcal{S} = (0,\infty)\setminus\mathcal{C}$ assumed to be the union of a finite number of closed intervals in $(0,\infty)$, and w, h two continuous functions on $(0,\infty)$, with $w = h$ on $\mathcal{S}$ such that*

$$w \text{ is } C^1 \text{ on } \partial\mathcal{S}, \tag{3.11}$$

$$w \geq h \text{ on } \mathcal{C}, \tag{3.12}$$

w is C^2 on $\mathcal{C}$, solution to

$$rw - \mathcal{L}_i w - f_i = 0 \quad \text{on } \mathcal{C}, \tag{3.13}$$

and w is a viscosity supersolution to

$$rw - \mathcal{L}_i w - f_i \geq 0 \quad \text{on } \operatorname{int}(\mathcal{S}). \tag{3.14}$$

Here $\operatorname{int}(\mathcal{S})$ is the interior of $\mathcal{S}$ and $\partial\mathcal{S} = \mathcal{S} \setminus \operatorname{int}(\mathcal{S})$ its boundary. Then, w is a viscosity solution to

$$\min\{rw - \mathcal{L}_i w - f_i, w - h\} = 0 \quad \text{on } (0,\infty). \tag{3.15}$$

Proof. Take some $\bar{x} \in (0,\infty)$ and distinguish the following cases:

• $\bar{x} \in \mathcal{C}$. Since $w = v$ is C^2 on $\mathcal{C}$ and satisfies $rw(\bar{x}) - \mathcal{L}_i w(\bar{x}) - f_i(\bar{x}) = 0$ by (3.13), and recalling $w(\bar{x}) \geq h(\bar{x})$ by (3.12), we obtain the classical solution property, and so a fortiori the viscosity solution property (3.15) of w at $\bar{x}$.

• $\bar{x} \in \mathcal{S}$. Then $w(\bar{x}) = h(\bar{x})$ and the viscosity subsolution property is trivial at $\bar{x}$. It remains to show the viscosity supersolution property at $\bar{x}$. If $\bar{x} \in \operatorname{int}(\mathcal{S})$, this follows directly from (3.14). Suppose now $\bar{x} \in \partial\mathcal{S}$, and to fix the idea, we consider that $\bar{x}$ is on the left-boundary of $\mathcal{S}$ so that from the assumption on the form of $\mathcal{S}$, there exists $\varepsilon > 0$ s.t. $(\bar{x} - \varepsilon, \bar{x}) \subset \mathcal{C}$ on which w is smooth C^2 (the same argument holds true when $\bar{x}$ is on the right-boundary of $\mathcal{S}$). Take some smooth C^2 function φ

s.t. $\bar{x}$ is a local minimum of $w-\varphi$. Since w is C^1 by (3.11), we have $\varphi'(\bar{x}) = w'(\bar{x})$. We may also assume w.l.o.g. (by taking ε small enough) that $(w-\varphi)(\bar{x}) \leq (w-\varphi)(x)$ for $x \in (\bar{x}-\varepsilon, \bar{x})$. Moreover, by Taylor's formula, we have

$$w(\bar{x}-\eta) \;=\; w(\bar{x}) - \eta \int_0^1 w'(\bar{x}-t\eta)dt, \varphi(\bar{x}-\eta) \;=\; \varphi(\bar{x}) - \eta \int_0^1 \varphi'(\bar{x}-t\eta)dt,$$

so that

$$\int_0^1 \varphi'(\bar{x}-t\eta) - w'(\bar{x}-t\eta)\, dt \geq 0 \quad \forall\, 0 < \eta < \varepsilon.$$

Since $\varphi'(\bar{x}) = w'(\bar{x})$, this last inequality is written as

$$\int_0^1 \frac{\varphi'(\bar{x}-t\eta) - \varphi'(\bar{x})}{\eta} - \frac{w'(\bar{x}-t\eta) - w'(\bar{x})}{\eta}\, dt \geq 0 \quad \forall\, 0 < \eta < \varepsilon. \tag{3.16}$$

Now, from (3.13), we have $rw(x) - \mathcal{L}_i w(x) - f_i(x) = 0$ for $x \in (\bar{x}-\varepsilon, \bar{x})$. By sending x towards $\bar{x}$ into this last equality, this shows that $w''(\bar{x}^-) = \lim_{x \nearrow \bar{x}} w''(x)$ exists, and

$$rw(\bar{x}) - b_i \bar{x} w'(\bar{x}) - \frac{1}{2}\sigma_i^2 \bar{x}^2 w''(\bar{x}^-) - f_i(\bar{x}) = 0. \tag{3.17}$$

Moreover, by sending η to zero into (3.16), we obtain

$$\int_0^1 t\big[-\varphi''(\bar{x}) + w''(\bar{x}^-)\big]dt \geq 0,$$

and so $\varphi''(\bar{x}) \leq w''(\bar{x}^-)$. By substituting into (3.17), and recalling that $w'(\bar{x}) = \varphi'(\bar{x})$, we then obtain

$$rw(\bar{x}) - \mathcal{L}_i \varphi(\bar{x}) - f_i(\bar{x}) \geq 0,$$

which is the required supersolution inequality and ends the proof. □

Remark 3.4. Since $w = h$ on $\mathcal{S}$, relation (3.14) means equivalently that h is a viscosity supersolution to

$$rh - \mathcal{L}_i h - f_i \geq 0 \quad \text{on } \operatorname{int}(\mathcal{S}). \tag{3.18}$$

Practically, Lemma 3.7 shall be used as follows in the next section: We consider two C^1 functions v and h on $(0,\infty)$ s.t.

$$\begin{aligned} v(x) \;=\; h(x),\; v'(x) = h'(x), \quad x \;\in \partial\mathcal{S}, \\ v \geq h \quad \text{on } \mathcal{C}, \end{aligned}$$

v is C^2 on $\mathcal{C}$, solution to

$$rv - \mathcal{L}_i v - f_i = 0 \quad \text{on } \mathcal{C},$$

and h is a viscosity supersolution to (3.18). Then, the function w defined on $(0,\infty)$ by

$$w(x) = \begin{cases} v(x), & x \in \mathcal{C}, \\ h(x), & x \in \mathcal{S} \end{cases}$$

satisfies the conditions of Lemma 3.7 and is a viscosity solution to (3.15). This lemma combined with uniqueness viscosity solution results may be viewed as an alternative to the classical verification approach in the identification of the value function. Moreover, with our viscosity solutions approach, we shall see in section 4.2 that Lemma 3.5 and the smooth-fit property of the value functions in Theorem 3.6 provide a direct derivation for the structure of the switching regions, and thus of the solution to our problem.

4. Explicit solution in the two-regime case. In this section, we consider the case where $d = 2$. In this two-regime case, we know from Theorem 3.4 that the value functions v_i, $i = 1, 2$, are the unique continuous viscosity solutions, with linear growth conditions on $(0, \infty)$, and boundary conditions $v_i(0^+) = (-g_{ij})_+ := \max(-g_{ij}, 0)$, $j \neq i$, to the system

$$\min\{rv_1 - \mathcal{L}_1 v_1 - f_1, v_1 - (v_2 - g_{12})\} = 0, \tag{4.1}$$

$$\min\{rv_2 - \mathcal{L}_2 v_2 - f_2, v_2 - (v_1 - g_{21})\} = 0. \tag{4.2}$$

Moreover, the switching regions are

$$\mathcal{S}_i \;=\; \mathcal{S}_{ij} = \{x > 0 : v_i(x) = v_j(x) - g_{ij}\}, \quad i, j = 1, 2,\ i \neq j.$$

We set

$$\underline{x}_i^* \;=\; \inf \mathcal{S}_i \;\in\; [0, \infty] \bar{x}_i^* \;=\; \sup \mathcal{S}_i \;\in\; [0, \infty],$$

with the usual convention that $\inf \emptyset = \infty$.

Let us also introduce some other notation. We consider the second order ODE for $i = 1, 2$:

$$rv - \mathcal{L}_i v - f_i = 0, \tag{4.3}$$

whose general solution (without second member f_i) is given by

$$v(x) = Ax^{m_i^+} + Bx^{m_i^-},$$

for some constants A, B, and where

$$m_i^- = -\frac{b_i}{\sigma_i^2} + \frac{1}{2} - \sqrt{\left(-\frac{b_i}{\sigma_i^2} + \frac{1}{2}\right)^2 + \frac{2r}{\sigma_i^2}} \;<\; 0,$$

$$m_i^+ = -\frac{b_i}{\sigma_i^2} + \frac{1}{2} + \sqrt{\left(-\frac{b_i}{\sigma_i^2} + \frac{1}{2}\right)^2 + \frac{2r}{\sigma_i^2}} \;>\; 1.$$

We also denote

$$\hat{V}_i(x) = E\left[\int_0^\infty e^{-rt} f_i(\hat{X}_t^{x,i}) dt\right],$$

with $\hat{X}^{x,i}$ the solution to the SDE $d\hat{X}_t = b_i \hat{X}_t dt + \sigma_i \hat{X}_t dW_t$, $\hat{X}_0 = x$. Actually, $\hat{V}_i$ is a particular solution to ODE (4.3), with boundary condition $\hat{V}_i(0^+) = f_i(0) = 0$. It corresponds to the reward function associated with the no switching strategy from initial state (x, i), and so $\hat{V}_i \leq v_i$.

Remark 4.1. If $g_{ij} > 0$, then from (2.7), we have $v_i(0^+) = 0 > (-g_{ji})_+ - g_{ij} = v_j(0^+) - g_{ij}$. Therefore, by continuity of the value functions on $(0,\infty)$, we get $\underline{x}_i^* > 0$.

We now give the explicit solution to our problem in the following two situations:

• The diffusion operators are different and the running profit functions are identical.

• The diffusion operators are identical and the running profit functions are different.

We also consider the cases for which both switching costs are positive, and for which one of the two is negative, the other being then positive according to (2.7). This last case is interesting in applications where a firm chooses between an open or closed activity, and may regain a fraction of its opening costs when it decides to close.

4.1. Identical profit functions with different diffusion operators. In this subsection, we suppose that the running functions are identical in the form

$$f_1(x) \;=\; f_2(x) = x^\gamma, \quad 0 < \gamma < 1, \tag{4.4}$$

and the diffusion operators are different. A straightforward calculation shows that under (4.4), we have

$$\hat{V}_i(x) = K_i x^\gamma \quad \text{with} \;\; K_i \;=\; \frac{1}{r - b_i\gamma + \frac{1}{2}\sigma_i^2\gamma(1-\gamma)} \;>\; 0, \;\; i = 1,2.$$

We show that the structure of the switching regions depends actually only on the sign of $K_2 - K_1$, and of the sign of the switching costs g_{12} and g_{21}. More precisely, we have the following explicit result.

THEOREM 4.1. *Let i, $j = 1,2$, $i \neq j$.*

(1) *If $K_i = K_j$, then*

$$v_i(x) = \hat{V}_i(x) + (-g_{ij})_+, \qquad x \in (0,\infty),$$
$$\mathcal{S}_i = \begin{cases} \emptyset & \text{if } \; g_{ij} > 0, \\ (0,\infty) & \text{if } \; g_{ij} \leq 0. \end{cases}$$

It is always optimal to switch from regime i to j if the corresponding switching cost is nonpositive, and never optimal to switch otherwise.

(2) *If $K_j > K_i$, then we have the following situations depending on the switching costs:*

(a) $g_{ij} \leq 0$: *We have $\mathcal{S}_i = (0,\infty)$, $\mathcal{S}_j = \emptyset$, and*

$$v_i \;=\; \hat{V}_j - g_{ij}, v_j \;=\; \hat{V}_j.$$

(b) $g_{ij} > 0$:

• *If $g_{ji} \geq 0$, then $\mathcal{S}_i = [\underline{x}_i^*, \infty)$ with $\underline{x}_i^* \in (0,\infty)$, $\mathcal{S}_j = \emptyset$, and*

$$v_i(x) = \begin{cases} Ax^{m_i^+} + \hat{V}_i(x), & x < \underline{x}_i^*, \\ v_j(x) - g_{ij}, & x \geq \underline{x}_i^*, \end{cases} \tag{4.5}$$

$$v_j(x) = \hat{V}_j(x), \quad x \in (0,\infty), \tag{4.6}$$

where the constants A and $\underline{x}_i^*$ are determined by the continuity and smooth-fit conditions of v_i at $\underline{x}_i^*$, and explicitly given by

$$\underline{x}_i^* = \left(\frac{m_i^+}{m_i^+ - \gamma}\frac{g_{ij}}{K_j - K_i}\right)^{\frac{1}{\gamma}}, \tag{4.7}$$

$$A = (K_j - K_i)\frac{\gamma}{m_i^+}(\underline{x}_i^*)^{\gamma - m_i^+}. \tag{4.8}$$

When we are in regime i, it is optimal to switch to regime j whenever the state process X exceeds the threshold $\underline{x}_i^*$, while when we are in regime j, it is optimal to never switch.

- *If $g_{ji} < 0$, then $\mathcal{S}_i = [\underline{x}_i^*, \infty)$ with $\underline{x}_i^* \in (0, \infty)$, $\mathcal{S}_j = (0, \bar{x}_j^*]$, and*

$$v_i(x) = \begin{cases} Ax^{m_i^+} + \hat{V}_i(x), & x < \underline{x}_i^*, \\ v_j(x) - g_{ij}, & x \geq \underline{x}_i^*, \end{cases} \tag{4.9}$$

$$v_j(x) = \begin{cases} v_i(x) - g_{ji}, & x \leq \bar{x}_j^*, \\ Bx^{m_j^-} + \hat{V}_j(x), & x > \bar{x}_i^*, \end{cases} \tag{4.10}$$

where the constants A, B and $\bar{x}_j^* < \underline{x}_i^*$ are determined by the continuity and smooth-fit conditions of v_i and v_j at $\underline{x}_i^*$ and $\bar{x}_j^*$, and explicitly given by

$$\begin{aligned} \overline{x}_j &= \left[\frac{-m_j^-(g_{ji} + g_{ij} y^{m_i^+})}{(K_i - K_j)(\gamma - m_j^-)(1 - y^{m_i^+ - \gamma})} \right]^{\frac{1}{\gamma}}, \\ \underline{x}_i &= \frac{\overline{x}_j}{y}, \\ B &= \frac{(K_i - K_j)(m_i^+ - \gamma)\underline{x}_i^{\gamma - m_j^-} + m_i^+ g_{ij} \underline{x}_i^{-m_j}}{m_i^+ - m_j^-}, \\ A &= B\underline{x}_i^{m_j^- - m_i^+} - (K_i - K_j)\underline{x}_i^{\gamma - m_i^+} - g_{ij}\underline{x}_i^{-m_i^+}, \end{aligned}$$

with y solution in

$$\left(0, \left(-\frac{g_{ji}}{g_{ij}}\right)^{\frac{1}{m_i^+}}\right)$$

to the equation

$$\begin{aligned} &m_i^+(\gamma - m_j^-)\left(1 - y^{m_i^+ - \gamma}\right)\left(g_{ij} y^{m_j^-} + g_{ji}\right) \\ &\quad + m_j^-(m_i^+ - \gamma)\left(1 - y^{m_j^- - \gamma}\right)\left(g_{ij} y^{m_i^+} + g_{ji}\right) = 0. \end{aligned}$$

When we are in regime i, it is optimal to switch to regime j whenever the state process X exceeds the threshold $\underline{x}_i^*$, while when we are in regime j, it is optimal to switch to regime i for values of the state process X under the threshold $\bar{x}_j^*$.

Economic interpretation. In the particular case where $\sigma_1 = \sigma_2$, we have that $K_2 - K_1 > 0$ means that regime 2 provides a higher expected return b_2 than b_1 of regime 1 for the same volatility coefficient σ_i. Moreover, if the switching cost g_{21} from regime 2 to regime 1 is nonnegative, it is intuitively clear that it is in our best interest to always stay in regime 2, which is formalized by the property that $\mathcal{S}_2 = \emptyset$. However, if one receives some gain compensation to switch from regime 2 to regime 1, i.e., the corresponding cost g_{21} is negative, then it is in our best interest to change regime for small values of the current state. This is formalized by the property that $\mathcal{S}_2 = (0, \bar{x}_2^*]$. On the other hand, in regime 1, our best interest is to switch to regime 2, for all current

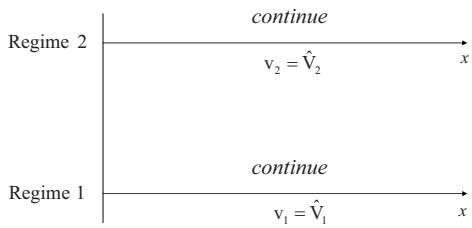


Figure I.1.a: $f_1 = f_2,\ K_1 = K_2,\ g_{12} > 0,\ g_{21} > 0$

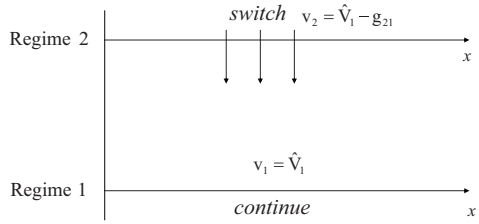


Figure I.1.b: $f_1 = f_2,\ K_1 = K_2,\ g_{12} > 0,\ g_{21} \le 0$

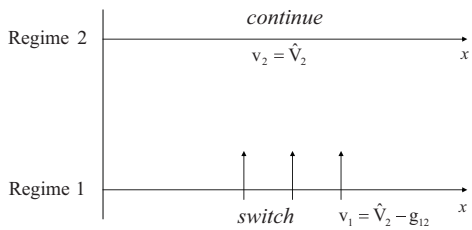


Figure I.2.a: $f_1 = f_2,\ K_2 > K_1,\ g_{12} \le 0$

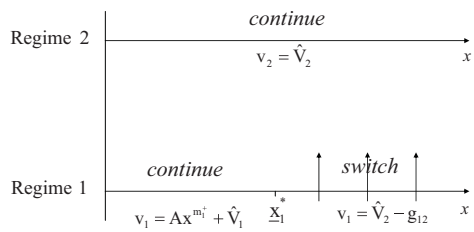


Figure I.2.bi: $f_1 = f_2,\ K_2 > K_1,\ g_{12} > 0,\ g_{21} \ge 0$

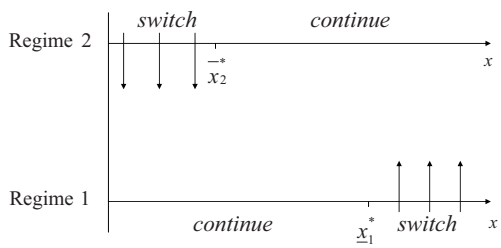


Figure I.2.bii: $f_1 = f_2,\ K_2 > K_1,\ g_{12} > 0,\ g_{21} < 0$

FIG. 1.

values of the state if the corresponding switching cost g_{12} is nonpositive, or from a certain threshold $\underline{x}_1^*$ if the switching cost g_{12} is positive. A similar interpretation holds when $b_1 = b_2$, and $K_2 - K_1 > 0$, i.e., $\sigma_2 < \sigma_1$. Theorem 4.1 extends these results for general coefficients b_i and σ_i, and shows that the critical parameter value determining the form of the optimal strategy is given by the sign of $K_2 - K_1$ and the switching costs. The different optimal strategy structures are depicted in Figure 1.

Proof of Theorem 4.1.

(1) If $K_i = K_j$, then $\hat{V}_i = \hat{V}_j$. We consider the smooth functions $w_i = \hat{V}_i + (-g_{ij})_+$ for $i, j = 1, 2$ and $j \neq i$. Since $\hat{V}_i$ are solutions to (4.3), we see that w_i satisfy

$$rw_i - \mathcal{L}w_i - f_i = r(-g_{ij})_+, \tag{4.11}$$

$$w_i - (w_j - g_{ij}) = g_{ij} + (-g_{ij})_+ - (-g_{ji})_+. \tag{4.12}$$

Notice that the l.h.s. of (4.11) and (4.12) are both nonnegative by (2.7). Moreover,

if $g_{ij} > 0$, then the l.h.s. of (4.11) is zero, and if $g_{ij} \leq 0$, then $g_{ji} > 0$ and the l.h.s. of (4.12) is zero. Therefore, w_i, $i = 1, 2$ is solution to the system

$$\min\{rw_i - \mathcal{L}_i w_i - f_i, w_i - (w_j - g_{ij})\} = 0.$$

Since $\hat{V}_i(0^+) = 0$, we have $w_i(0^+) = (-g_{ij})_+$. Moreover, w_i like $\hat{V}_i$ satisfies a linear growth condition. Therefore, from the uniqueness solutions to the PDE system (4.1)–(4.2), we deduce that $v_i = w_i$. As observed above, if $g_{ij} \leq 0$, then the l.h.s. of (4.12) is zero, and so $\mathcal{S}_i = (0, \infty)$. Finally, if $g_{ij} > 0$, then the l.h.s. of (4.12) is positive, and so $\mathcal{S}_i = \emptyset$.

(2) We now suppose w.l.o.g. that $K_2 > K_1$.

(a) Consider first the case where $g_{12} \leq 0$, and so $g_{21} > 0$. We set $w_1 = \hat{V}_2 - g_{12}$ and $w_2 = \hat{V}_2$. Then, by construction, we have $w_1 = w_2 - g_{12}$ on $(0, \infty)$, and by definition of $\hat{V}_1$ and $\hat{V}_2$:

$$rw_1(x) - \mathcal{L}_1 w_1(x) - f_1(x) = \frac{K_2 - K_1}{K_1} x^\gamma - rg_{12} \ > \ 0 \quad \forall\, x > 0.$$

On the other hand, we also have $rw_2 - \mathcal{L}_2 w_2 - f_2 = 0$ on $(0, \infty)$, and $w_2 \geq w_1 - g_{21}$ since $g_{12} + g_{21} \geq 0$. Hence, w_1 and w_2 are smooth (hence, viscosity) solutions to the system (4.1)–(4.2), with linear growth conditions and boundary conditions $w_1(0^+) = V_1(0^+) - g_{12} = (-g_{12})_+$, $w_2(0^+) = \hat{V}_2(0^+) = 0 = (-g_{21})_+$. By the uniqueness result of Theorem 3.4, we deduce that $v_1 = w_1$, $v_2 = w_2$, and thus $\mathcal{S}_1 = (0, \infty)$, $\mathcal{S}_2 = \emptyset$.

(b) Consider now the case where $g_{12} > 0$. We already know from Remark 4.1 that $\underline{x}_1^* > 0$, and we claim that $\underline{x}_1^* < \infty$. Otherwise, v_1 should be equal to $\hat{V}_1$. Since $v_1 \geq v_2 - g_{12} \geq \hat{V}_2 - g_{12}$, this would imply $(\hat{V}_2 - \hat{V}_1)(x) = (K_2 - K_1)x^\gamma \leq g_{12}$ for all $x > 0$, an obvious contradiction. By definition of $\underline{x}_1^*$, we have $(0, \underline{x}_1^*) \subset \mathcal{C}_1$. We shall prove actually the equality $(0, \underline{x}_1^*) = \mathcal{C}_1$, i.e., $\mathcal{S}_1 = [\underline{x}_1^*, \infty)$. On the other hand, the form of $\mathcal{S}_2$ will depend on the sign of g_{21}.

Case: $g_{21} \geq 0$.

We shall prove that $\mathcal{C}_2 = (0, \infty)$, i.e., $\mathcal{S}_2 = \emptyset$. To this end, let us consider the function

$$w_1(x) = \begin{cases} Ax^{m_1^+} + \hat{V}_1(x), & 0 < x < x_1, \\ \hat{V}_2(x) - g_{12}, & x \geq x_1, \end{cases}$$

where the positive constants A and x_1 satisfy

$$Ax_1^{m_1^+} + \hat{V}_1(x_1) = \hat{V}_2(x_1) - g_{12}, \tag{4.13}$$

$$Am_1^+ x_1^{m_1^+ - 1} + \hat{V}_1'(x_1) = \hat{V}_2'(x_1), \tag{4.14}$$

and are explicitly determined by

$$(K_2 - K_1)x_1^\gamma = \frac{m_1^+}{m_1^+ - \gamma} g_{12}, \tag{4.15}$$

$$A = (K_2 - K_1)\frac{\gamma}{m_1^+} x_1^{\gamma - m_1^+}. \tag{4.16}$$

Notice that by construction, w_1 is C^2 on $(0, x_1) \cup (x_1, \infty)$, and C^1 on x_1.

By using Lemma 3.7, we now show that w_1 is a viscosity solution to

$$\min\left\{rw_1 - \mathcal{L}_1 w_1 - f_1, w_1 - (\hat{V}_2 - g_{12})\right\} = 0 \quad \text{on } (0, \infty). \tag{4.17}$$

We first check that

$$w_1(x) \geq \hat{V}_2(x) - g_{12} \quad \forall\, 0 < x < x_1, \tag{4.18}$$

i.e.,

$$G(x) \;:=\; Ax^{m_1^+} + \hat{V}_1(x) - \hat{V}_2(x) + g_{12} \geq 0 \quad \forall\, 0 < x < x_1.$$

Since $A > 0$, $0 < \gamma < 1 < m_1^+$, $K_2 - K_1 > 0$, a direct derivation shows that the second derivative of G is positive, i.e., G is strictly convex. By (4.14), we have $G'(x_1) = 0$, and so G' is negative, i.e., G is strictly decreasing on $(0, x_1)$. Now, by (4.13), we have $G(x_1) = 0$, and thus G is positive on $(0, x_1)$, which proves (4.18).

By definition of w_1 on $(0, x_1)$, we have in the classical sense

$$rw_1 - \mathcal{L}_1 w_1 - f_1 = 0 \quad \text{on } (0, x_1). \tag{4.19}$$

We now check that

$$rw_1 - \mathcal{L}_1 w_1 - f_1 \geq 0 \quad \text{on } (x_1, \infty) \tag{4.20}$$

holds true in the classical sense, and so a fortiori in the viscosity sense. By definition of w_1 on (x_1, ∞), and K_1, we have for all $x > x_1$,

$$rw_1(x) - \mathcal{L}_1 w_1(x) - f_1(x) = \frac{K_2 - K_1}{K_1} x^\gamma - rg_{12} \quad \forall x > x_1,$$

so that (4.20) is satisfied iff $\frac{K_2 - K_1}{K_1} x_1^\gamma - rg_{12} \geq 0$, or equivalently by (4.15),

$$\frac{m_1^+}{m_1^+ - \gamma} \geq rK_1 \;=\; \frac{r}{r - b_1\gamma + \frac{1}{2}\sigma_1^2\gamma(1-\gamma)}. \tag{4.21}$$

Now, since $\gamma < 1 < m_1^+$, and by definition of m_1^+, we have

$$\frac{1}{2}\sigma_1^2 m_1^+(\gamma - 1) < \frac{1}{2}\sigma_1^2 m_1^+(m_1^+ - 1) \;=\; r - b_1 m_1^+,$$

which proves (4.21) and thus (4.20).

Relations (4.13)–(4.14) and (4.18)–(4.20) mean that the conditions of Lemma 3.7 are satisfied with $\mathcal{C} = (0, x_1)$, $h = \hat{V}_2 - g_{12}$, and we thus get the required assertion (4.17).

On the other hand, we check that

$$\hat{V}_2(x) \geq w_1(x) - g_{21} \quad \forall x > 0, \tag{4.22}$$

which amounts to showing

$$H(x) \;:=\; Ax^{m_1^+} + \hat{V}_1(x) - \hat{V}_2(x) - g_{21} \leq 0 \quad \forall\, 0 < x < x_1.$$

Since $A > 0$, $0 < \gamma < 1 < m_1^+$, $K_2 - K_1 > 0$, a direct derivation shows that the second derivative of H is positive, i.e., H is strictly convex. By (4.14), we have $H'(x_1) = 0$ and so H' is negative, i.e., H is strictly decreasing on $(0, x_1)$. Now, we have $H(0) = -g_{21} \leq 0$ and thus H is negative on $(0, x_1)$, which proves (4.22). Recalling that $\hat{V}_2$ is

a solution to $r\hat{V}_2 - \mathcal{L}_2\hat{V}_2 - f_2 = 0$ on $(0,\infty)$, we deduce that, obviously from (4.22), $\hat{V}_2$ is a classical, hence a viscosity, solution to

$$\min\left\{r\hat{V}_2 - \mathcal{L}_2\hat{V}_2 - f_2, \hat{V}_2 - (w_1 - g_{21})\right\} = 0 \quad \text{on } (0,\infty). \tag{4.23}$$

Since $w_1(0^+) = 0 = (-g_{12})_+$, $\hat{V}_2(0^+) = 0 = (-g_{21})_+$, and w_1, $\hat{V}_2$ satisfy a linear growth condition, we deduce from (4.17), (4.23), and uniqueness to the PDE system (4.1)–(4.2) that

$$v_1 = w_1, v_2 = \hat{V}_2 \quad \text{on } (0,\infty).$$

This proves $\underline{x}_1^* = x_1$, $\mathcal{S}_1 = [x_1,\infty)$, and $\mathcal{S}_2 = \emptyset$.

Case: $g_{21} < 0$.

We shall prove that $\mathcal{S}_2 = (0,\bar{x}_2^*]$. To this end, let us consider the functions

$$w_1(x) = \begin{cases} Ax^{m_1^+} + \hat{V}_1(x), & x < \underline{x}_1, \\ w_2(x) - g_{12}, & x \geq \underline{x}_1, \end{cases}$$

$$w_2(x) = \begin{cases} w_1(x) - g_{21}, & x \leq \bar{x}_2, \\ Bx^{m_2^-} + \hat{V}_2(x), & x > \bar{x}_2, \end{cases}$$

where the positive constants A, B, $\underline{x}_1 > \bar{x}_2$, are the solution to

$$A\underline{x}_1^{m_1^+} + \hat{V}_1(\underline{x}_1) = w_2(\underline{x}_1) - g_{12} = B\underline{x}_1^{m_2^-} + \hat{V}_2(\underline{x}_1) - g_{12}, \tag{4.24}$$

$$Am_1^+\underline{x}_1^{m_1^+-1} + \hat{V}_1'(\underline{x}_1) = w_2'(\underline{x}_1) = Bm_2^-\underline{x}_1^{m_2^--1} + \hat{V}_2'(\underline{x}_1), \tag{4.25}$$

$$A\bar{x}_2^{m_1^+} + \hat{V}_1(\bar{x}_2) - g_{21} = w_1(\bar{x}_2) - g_{21} = B\bar{x}_2^{m_2^-} + \hat{V}_2(\bar{x}_2), \tag{4.26}$$

$$Am_1^+\bar{x}_2^{m_1^+-1} + \hat{V}_1'(\bar{x}_2) = w_1'(\bar{x}_2) = Bm_2^-\bar{x}_2^{m_2^--1} + \hat{V}_2'(\bar{x}_2), \tag{4.27}$$

exist and are explicitly determined after some calculations by

$$\overline{x}_2 = \left[\frac{-m_2^-(g_{21} + g_{12}y^{m_1^+})}{(K_1 - K_2)(\gamma - m_2^-)(1 - y^{m_1^+-\gamma})}\right]^{\frac{1}{\gamma}}, \tag{4.28}$$

$$\underline{x}_1 = \frac{\overline{x}_2}{y}, \tag{4.29}$$

$$B = \frac{(K_1 - K_2)(m_1^+ - \gamma)\underline{x}_1^{\gamma - m_2^-} + m_1^+ g_{12}\underline{x}_1^{-m_2}}{m_1^+ - m_2^-}, \tag{4.30}$$

$$A = B\underline{x}_1^{m_2^- - m_1^+} - (K_1 - K_2)\underline{x}_1^{\gamma - m_1^+} - g_{12}\underline{x}_1^{-m_1^+}, \tag{4.31}$$

with y a solution in

$$\left(0, \left(-\frac{g_{21}}{g_{12}}\right)^{\frac{1}{m_1^+}}\right)$$

to the equation

$$\begin{aligned} &m_1^+(\gamma - m_2^-)\left(1 - y^{m_1^+-\gamma}\right)\left(g_{12}y^{m_2^-} + g_{21}\right) \\ &\quad + m_2^-(m_1^+ - \gamma)\left(1 - y^{m_2^--\gamma}\right)\left(g_{12}y^{m_1^+} + g_{21}\right) = 0. \end{aligned} \tag{4.32}$$

Using (2.7), we have $y < \big(-\frac{g_{21}}{g_{12}}\big)^{\frac{1}{m_1^+}} < 1$. As such, $0 < \bar{x}_2 < \underline{x}_1$. Furthermore, by using (4.29), and (4.32) satisfied by y, we may easily check that A and B are positive constants.

Notice that by construction, w_1 (resp., w_2) is C^2 on $(0, \underline{x}_1) \cup (\underline{x}_1, \infty)$ (resp., $(0, \bar{x}_2) \cup (\bar{x}_2, \infty)$) and C^1 at $\underline{x}_1$ (resp., $\bar{x}_2$).

By using Lemma 3.7, we now show that w_i, $i = 1, 2$, is a viscosity solution to the system

$$\min\{rw_i - \mathcal{L}_i w_i - f_i, w_i - (w_j - g_{ij})\} = 0 \quad \text{on } (0, \infty),\ i, j = 1, 2,\ j \neq i. \tag{4.33}$$

Since the proof is similar for both w_i, $i = 1, 2$, we prove the result only for w_1. We first check that

$$w_1 \geq w_2 - g_{12} \quad \forall\, 0 < x < \underline{x}_1. \tag{4.34}$$

From the definition of w_1 and w_2 and using the fact that $g_{12} + g_{21} > 0$, it is straightforward to see that

$$w_1 \geq w_2 - g_{12} \quad \forall\, 0 < x \leq \overline{x}_2. \tag{4.35}$$

Now, we need to prove that

$$G(x) := Ax^{m_1^+} + \hat{V}_1(x) - Bx^{m_2^-} - \hat{V}_2(x) + g_{12} \geq 0 \quad \forall\, \overline{x}_2 < x < \underline{x}_1. \tag{4.36}$$

We have $G(\bar{x}_2) = g_{12} + g_{21} > 0$ and $G(\underline{x}_1) = 0$. Suppose that there exists some $x_0 \in (\bar{x}_2, \underline{x}_1)$ such that $G(x_0) = 0$. We then deduce that there exists $x_3 \in (\bar{x}_0, \underline{x}_1)$ such that $G'(x_3) = 0$. As such, the equation $G'(x) = 0$ admits at least three solutions in $[\bar{x}_2, \underline{x}_1]$: $\{\bar{x}_2, x_3, \underline{x}_1\}$. However, a straightforward study of the function G shows that G' can take the value zero at most at two points in $(0, \infty)$. This leads to a contradiction, proving therefore (4.36).

By definition of w_1, we have in the classical sense

$$rw_1 - \mathcal{L}_1 w_1 - f = 0 \quad \text{on } (0, \underline{x}_1). \tag{4.37}$$

We now check that

$$rw_1 - \mathcal{L}_1 w_1 - f \geq 0 \quad \text{on } (\underline{x}_1, \infty) \tag{4.38}$$

holds true in the classical sense, and so a fortiori in the viscosity sense. By definition of w_1 on (x_1, ∞), and K_1, we have for all $x > \underline{x}_1$,

$$H(x) := rw_1(x) - \mathcal{L}_1 w_1(x) - f(x) = \frac{K_2 - K_1}{K_1} x^\gamma + m_2^- LBx^{m_2^-} - rg_{12} \quad \forall x > \underline{x}_1, \tag{4.39}$$

where $L = \frac{1}{2}(\sigma_2^2 - \sigma_1^2)(m_2^- - 1) + b_2 - b_1$.

We distinguish two cases:

• First, if $L \geq 0$, the function H would be nondecreasing on $(0, \infty)$ with $\lim_{x \to 0^+} H(x) = -\infty$ and $\lim_{x \to \infty} H(x) = +\infty$. As such, it suffices to show that $H(\underline{x}_1) \geq 0$. From (4.24)–(4.25), we have

$$H(\underline{x}_1) = (K_2 - K_1)\left[\frac{m_1^+ - m_2^-}{K_1} - (m_1^+ - \gamma) m_2^- L\right] - rg_{12} + m_1^+ m_2^- g_{12} L.$$

Using relations (4.21), (4.24), (4.25), (4.29), and the definition of m_1^+ and m_2^-, we then obtain

$$H(\underline{x}_1) = \frac{m_1^+(m_1^+ - m_2^-)}{K_1(m_1^+ - \gamma)} - r \geq \frac{m_1^+}{K_1(m_1^+ - \gamma)} - r \geq 0.$$

- Second, if $L < 0$, it suffices to show that

$$\frac{K_2 - K_1}{K_1} x^\gamma - r g_{12} \geq 0 \quad \forall\, x > \underline{x}_1,$$

which is rather straightforward from (4.21) and (4.29).

Relations (4.34), (4.37), (4.38), and the regularity of w_i, $i = 1, 2$, as constructed, mean that the conditions of Lemma 3.7 are satisfied and we thus get the required assertion (4.33).

Since $w_1(0^+) = 0 = (-g_{12})_+$, $w_2(0^+) = -g_{21} = (-g_{21})_+$, and w_1, $\hat{V}_2$ satisfy a linear growth condition, we deduce from (4.33) and uniqueness to the PDE system (4.1)–(4.2) that

$$v_1 = w_1, v_2 = w_2 \quad \text{on } (0, \infty).$$

This proves $\underline{x}_1^* = \underline{x}_1$, $\mathcal{S}_1 = [x_1, \infty)$ and $\bar{x}_2^* = \bar{x}_2$, $\mathcal{S}_2 = (0, \bar{x}_2]$.

4.2. Identical diffusion operators with different profit functions. In this subsection, we suppose that $\mathcal{L}_1 = \mathcal{L}_2 = \mathcal{L}$, i.e., $b_1 = b_2 = b$, $\sigma_1 = \sigma_2 = \sigma > 0$. We then set $m^+ = m_1^+ = m_2^+$, $m^- = m_1^- = m_2^-$, and $\hat{X}^x = \hat{X}^{x,1} = \hat{X}^{x,2}$. Notice that in this case, the set Q_{ij}, $i, j = 1, 2$, $i \neq j$, introduced in Lemma 3.5, satisfies

$$\begin{aligned} Q_{ij} &= \{x \in \mathcal{C}_j : (f_j - f_i)(x) - r g_{ij} \geq 0\} \\ &\subset \hat{Q}_{ij} := \{x > 0 : (f_j - f_i)(x) - r g_{ij} \geq 0\}. \end{aligned} \tag{4.40}$$

Once we are given the profit functions f_i, f_j, the set $\hat{Q}_{ij}$ can be explicitly computed. Moreover, we prove in the next key lemma that the structure of $\hat{Q}_{ij}$, when it is connected, determines the same structure for the switching region $\mathcal{S}_i$.

LEMMA 4.2. *Let $i, j = 1, 2$, $i \neq j$.*

(1) *Assume that*

$$\sup_{x>0} (\hat{V}_j - \hat{V}_i)(x) > g_{ij}. \tag{4.41}$$

(a) *If there exists $0 < \underline{x}_{ij} < \infty$ such that*

$$\hat{Q}_{ij} = [\underline{x}_{ij}, \infty), \tag{4.42}$$

then $0 < \underline{x}_i^* < \infty$ and

$$\mathcal{S}_i = [\underline{x}_i^*, \infty).$$

(b) *If $g_{ij} \leq 0$ and there exists $0 < \bar{x}_{ij} < \infty$ such that*

$$\hat{Q}_{ij} = (0, \bar{x}_{ij}], \tag{4.43}$$

then $0 < \bar{x}_i^* < \infty$ and

$$\mathcal{S}_i = (0, \bar{x}_i^*].$$

(2) *If there exist* $0 < \underline{x}_{ij} < \bar{x}_{ij} < \infty$ *such that*

$$\hat{Q}_{ij} = [\underline{x}_{ij}, \bar{x}_{ij}], \tag{4.44}$$

then $0 < \underline{x}_i^* < \bar{x}_i^* < \infty$ *and*

$$\mathcal{S}_i = [\underline{x}_i^*, \bar{x}_i^*].$$

(3) *If* $g_{ij} \leq 0$ *and* $\hat{Q}_{ij} = (0, \infty)$*, then* $\mathcal{S}_i = (0, \infty)$ *and* $\mathcal{S}_j = \emptyset$.

Proof. (1a) Consider the case of condition (4.42). Since $\mathcal{S}_i \subset \hat{Q}_{ij}$ by Lemma 3.5, this implies $\underline{x}_i^* := \inf \mathcal{S}_i \geq \underline{x}_{ij} > 0$. We now claim that $\underline{x}_i^* < \infty$. On the contrary, the switching region $\mathcal{S}_i$ would be empty, and so v_i would satisfy on $(0, \infty)$

$$rv_i - \mathcal{L}v_i - f_i = 0 \quad \text{on } (0, \infty).$$

Then, v_i would be in the form

$$v_i(x) = Ax^{m^+} + Bx^{m^-} + \hat{V}_i(x), \quad x > 0.$$

Since $0 \leq v_i(0^+) < \infty$ and v_i is a nonnegative function satisfying a linear growth condition, and using the fact that $m^- < 0$ and $m^+ > 1$, we deduce that v_i should be equal to $\hat{V}_i$. Now, since we have $v_i \geq v_j - g_{ij} \geq \hat{V}_j - g_{ij}$, this would imply

$$\hat{V}_j(x) - \hat{V}_i(x) \leq g_{ij} \quad \forall x > 0.$$

This contradicts condition (4.41), and so $0 < \underline{x}_i^* < \infty$.

By definition of $\underline{x}_i^*$, we already know that $(0, \underline{x}_i^*) \subset \mathcal{C}_i$. We prove actually the equality, i.e., $\mathcal{S}_i = [\underline{x}_i^*, \infty)$ or $v_i(x) = v_j(x) - g_{ij}$ for all $x \geq \underline{x}_i^*$. Consider the function

$$w_i(x) = \begin{cases} v_i(x), & 0 < x < \underline{x}_i^*, \\ v_j(x) - g_{ij}, & x \geq \underline{x}_i^*. \end{cases}$$

We now check that w_i is a viscosity solution of

$$\min\{rw_i - \mathcal{L}w_i - f_i \,,\; w_i - (v_j - g_{ij})\} = 0 \quad \text{on } (0, \infty). \tag{4.45}$$

From Theorem 3.6, the function w_i is C^1 on $(0, \infty)$ and in particular at $\underline{x}_i^*$, where $w_i'(\underline{x}_i^*) = v_i'(\underline{x}_i^*) = v_j'(\underline{x}_i^*)$. We also know that $w_i = v_i$ is C^2 on $(0, \underline{x}_i^*) \subset \mathcal{C}_i$, and satisfies $rw_i - \mathcal{L}w_i - f_i = 0$, $w_i \geq (v_j - g_{ij})$ on $(0, \underline{x}_i^*)$. Hence, from Lemma 3.7, we need only check the viscosity supersolution property of w_i to

$$rw_i - \mathcal{L}w_i - f_i \geq 0 \quad \text{on } (\underline{x}_i^*, \infty). \tag{4.46}$$

For this, take some point $\bar{x} > \underline{x}_i^*$ and some smooth test function φ s.t. $\bar{x}$ is a local minimum of $w_i - \varphi$. Then, $\bar{x}$ is a local minimum of $v_j - (\varphi + g_{ij})$, and by the viscosity solution property of v_j to its Bellman PDE, we have

$$rv_j(\bar{x}) - \mathcal{L}\varphi(\bar{x}) - f_j(\bar{x}) \geq 0.$$

Now, since $\underline{x}_i^* \geq \underline{x}_{ij}$, we have $\bar{x} > \underline{x}_{ij}$ and so by (4.42), $\bar{x} \in \hat{Q}_{ij}$. Hence,

$$(f_j - f_i)(\bar{x}) - rg_{ij} \geq 0.$$

By adding the two previous inequalities, we also obtain the required supersolution inequality:

$$rw_i(\bar{x}) - \mathcal{L}\varphi(\bar{x}) - f_i(\bar{x}) \geq 0,$$

and so (4.45) is proved.

Since $w_i(0^+) = v_i(0^+)$ and w_i satisfies a linear growth condition, and from uniqueness of the viscosity solution to PDE (4.45), we deduce that w_i is equal to v_i. In particular, we have $v_i(x) = v_j(x) - g_{ij}$ for $x \geq \underline{x}_i^*$, which shows that $\mathcal{S}_i = [\underline{x}_i^*, \infty)$.

(1b) The case of condition (4.43) is dealt with by same arguments as above: we first observe that $0 < \bar{x}_i^* := \sup \mathcal{S}_i < \infty$ under (4.41), and then show with Lemma 3.7 that the function

$$w_i(x) = \begin{cases} v_j(x) - g_{ij}, & 0 < x < \bar{x}_i^*, \\ v_i(x), & x \geq \bar{x}_i^*, \end{cases}$$

is a viscosity solution to

$$\min\{rw_i - \mathcal{L}w_i - f_i \ , \ w_i - (v_j - g_{ij})\} = 0 \quad \text{on} \quad (0, \infty).$$

Then, under the condition that $g_{ij} \leq 0$, we see that $g_{ji} > 0$ by (2.7), and so $v_i(0^+) = -g_{ij} = (-g_{ji})_+ - g_{ij} = v_j(0^+) - g_{ij} = w_i(0^+)$. From uniqueness of the viscosity solution to PDE (4.45), we conclude that $v_i = w_i$, and so $\mathcal{S}_i = (0, \bar{x}_i^*]$.

(2) By Lemma 3.5 and (4.40), condition (4.44) implies $0 < \underline{x}_{ij} \leq \underline{x}_i^* \leq \bar{x}_i^* \leq \bar{x}_{ij} < \infty$. We claim that $\underline{x}_i^* < \bar{x}_i^*$. Otherwise, $\mathcal{S}_i = \{\bar{x}_i^*\}$ and v_i would satisfy $rv_i - \mathcal{L}v_i - f_i = 0$ on $(0, \bar{x}_i^*) \cup (\bar{x}_i^*, \infty)$. By continuity and the smooth-fit condition of v_i at $\bar{x}_i^*$, this implies that v_i satisfies actually

$$rv_i - \mathcal{L}v_i - f_i = 0, \quad x \in (0, \infty),$$

and so is in the form

$$v_i(x) = Ax^{m^+} + Bx^{m^-} + \hat{V}_i(x), \quad x \in (0, \infty).$$

Since $0 \leq v_i(0^+) < \infty$ and v_i is a nonnegative function satisfying a linear growth condition, this implies $A = B = 0$. Therefore, v_i is equal to $\hat{V}_i$, which also means that $\mathcal{S}_i = \emptyset$, a contradiction.

We now prove that $\mathcal{S}_i = [\underline{x}_i^*, \bar{x}_i^*]$. Let us consider the function

$$w_i(x) = \begin{cases} v_i(x), & x \in (0, \underline{x}_i^*) \cup (\bar{x}_i^*, \infty), \\ v_j(x) - g_{ij}, & x \in [\underline{x}_i^*, \bar{x}_i^*], \end{cases}$$

which is C^1 on $(0, \infty)$ and in particular on $\underline{x}_i^*$ and $\bar{x}_i^*$ from Theorem 3.6. Hence, by similar arguments as in case (1), using Lemma 3.7, we then show that w_i is a viscosity solution of

$$\min\{rw_i - \mathcal{L}w_i - f_i \ , \ w_i - (v_j - g_{ij})\} = 0. \tag{4.47}$$

Since $w_i(0^+) = v_i(0^+)$ and w_i satisfies a linear growth condition, and from uniqueness of the viscosity solution to PDE (4.47), we deduce that w_i is equal to v_i. In particular, we have $v_i(x) = v_j(x) - g_{ij}$ for $x \in [\underline{x}_i^*, \bar{x}_i^*]$, which shows that $\mathcal{S}_i = [\underline{x}_i^*, \bar{x}_i^*]$.

(3) Suppose that $g_{ij} \leq 0$ and $\hat{Q}_{ij} = (0,\infty)$. We shall prove that $\mathcal{S}_i = (0,\infty)$ and $\mathcal{S}_j = \emptyset$. To this end, we consider the smooth functions $w_i = \hat{V}_j - g_{ij}$ and $w_j = \hat{V}_j$. Then, recalling the ODE satisfied by $\hat{V}_j$, and inequality (2.7), we get

$$rw_j - \mathcal{L}w_j - f_j \;=\; 0, w_j - (w_i - g_{ji}) \;=\; g_{ij} + g_{ji} \;\geq\; 0.$$

Therefore w_j is a smooth (and so a viscosity) solution to

$$\min\big[rw_j - \mathcal{L}w_j - f_j, w_j - (w_i - g_{ji})\big] = 0 \qquad \text{on } \; (0,\infty).$$

On the other hand, by definition of $\hat{Q}_{ij}$, which is assumed equal to $(0,\infty)$, we have

$$\begin{aligned} rw_i(x) - \mathcal{L}w_i(x) - f_i(x) &= r\hat{V}_j(x) - \mathcal{L}\hat{V}_j(x) - f_j(x) + f_j(x) - f_i(x) - rg_{ij} \\ &= f_j(x) - f_i(x) - rg_{ij} \;\geq\; 0 \quad \forall x > 0. \end{aligned}$$

Moreover, by construction we have $w_i = w_j - g_{ij}$. Therefore w_i is a smooth (and so a viscosity) solution to

$$\min\big[rw_i - \mathcal{L}w_i - f_i, w_i - (w_j - g_{ij})\big] = 0 \qquad \text{on } \; (0,\infty).$$

Notice also that $g_{ji} > 0$ by (2.7) and since $g_{ij} \leq 0$. Hence, $w_i(0^+) = -g_{ij} = (-g_{ij})_+ = v_i(0^+)$, $w_j(0^+) = 0 = (-g_{ji})_+ = v_j(0^+)$. From the uniqueness result of Theorem 3.4, we deduce that $v_i = w_i$, $v_j = w_j$, which proves that $\mathcal{S}_i = (0,\infty)$, $\mathcal{S}_j = \emptyset$. □

We shall now provide explicit solutions to the switching problem under general assumptions on the running profit functions, which include several interesting cases for applications:

(**HF**) There exists $\hat{x} \in \mathbb{R}_+$ s.t. the function $F := f_2 - f_1$
is decreasing on $(0,\hat{x})$, increasing on $[\hat{x},\infty)$,
and $F(\infty) := \lim_{x\to\infty} F(x) > 0, \quad g_{12} \;>\; 0.$

Under **(HF)**, there exists some $\bar{x} \in \mathbb{R}_+$ ($\bar{x} > \hat{x}$ if $\hat{x} > 0$ and $\bar{x} = 0$ if $\hat{x} = 0$) from which F is positive: $F(x) > 0$ for $x > \bar{x}$. Economically speaking, condition **(HF)** means that the profit in regime 2 is "better" than the profit in regime 1 from a certain level $\bar{x}$, and the improvement then becomes better and better. Moreover, since profit in regime 2 is better than in regime 1, it is natural to assume that the corresponding switching cost g_{12} from regime 1 to 2 should be positive. However, we shall consider both cases, where g_{21} is positive and nonpositive. Notice that $F(\hat{x}) < 0$ if $\hat{x} > 0$, $F(\hat{x}) = 0$ if $\hat{x} = 0$, and we do not assume necessarily $F(\infty) = \infty$.

Example 4.1. A typical example of different running profit functions satisfying **(HF)** is given by

$$f_i(x) = k_i x^{\gamma_i}, \;\; i = 1,2, \quad \text{with } 0 < \gamma_1 < \gamma_2 < 1, \;\; k_1 \in \mathbb{R}_+, \; k_2 > 0. \tag{4.48}$$

In this case, $\hat{x} = \big(\frac{k_1\gamma_1}{k_2\gamma_2}\big)^{\frac{1}{\gamma_2-\gamma_1}}$, and $\lim_{x\to\infty} F(x) = \infty$.

Another example of profit functions of interest in applications is the case when the profit function in regime 1 is $f_1 = 0$, and the other f_2 is increasing. In this case, assumption **(HF)** is satisfied with $\hat{x} = 0$.

The next proposition states the form of the switching regions in regimes 1 and 2, depending on the parameter values.

PROPOSITION 4.3. *Assume that* **(HF)** *holds.*

(1)
  (i) *If* $rg_{12} \geq F(\infty)$, *then* $\underline{x}_1^* = \infty$, *i.e.,* $\mathcal{S}_1 = \emptyset$.
  (ii) *If* $rg_{12} < F(\infty)$, *then* $\underline{x}_1^* \in (0,\infty)$ *and* $\mathcal{S}_1 = [\underline{x}_1^*, \infty)$.

(2)
  (i) *If* $rg_{21} \geq -F(\hat{x})$, *then* $\mathcal{S}_2 = \emptyset$.
  (ii) *If* $0 < rg_{21} < -F(\hat{x})$, *then* $0 < \underline{x}_2^* < \bar{x}_2^* < \underline{x}_1^*$, *and* $\mathcal{S}_2 = [\underline{x}_2^*, \bar{x}_2^*]$.
  (iii) *If* $g_{21} \leq 0$ *and* $-F(\infty) < rg_{21} < -F(\hat{x})$, *then* $0 = \underline{x}_2^* < \bar{x}_2^* < \underline{x}_1^*$, *and* $\mathcal{S}_2 = (0, \bar{x}_2^*]$.
  (iv) *If* $rg_{21} \leq -F(\infty)$, *then* $\mathcal{S}_2 = (0,\infty)$.

Proof. (1) From Lemma 3.5, we have

$$\hat{Q}_{12} = \{x > 0 : F(x) \geq rg_{12}\}. \tag{4.49}$$

Since $g_{12} > 0$, and $f_i(0) = 0$, we have $F(0) = 0 < rg_{12}$. Under **(HF)**, we then distinguish the two following cases:

(i) If $rg_{12} \geq F(\infty)$, then $\hat{Q}_{12} = \emptyset$, and so by Lemma 3.5 and (4.40), $\mathcal{S}_1 = \emptyset$.

(ii) If $rg_{12} < F(\infty)$, then there exists $\hat{x}_{12} \in (0,\infty)$ such that

$$\hat{Q}_{12} = [\underline{x}_{12}, \infty). \tag{4.50}$$

Moreover, since

$$(\hat{V}_2 - \hat{V}_1)(x) = E\left[\int_0^\infty e^{-rt} F(\hat{X}_t^x) dt\right] \quad \forall x > 0,$$

and F is lower bounded, we obtain by Fatou's lemma:

$$\liminf_{x\to\infty} (\hat{V}_2 - \hat{V}_1)(x) \geq E\left[\int_0^\infty e^{-rt} F(\infty) dt\right] = \frac{F(\infty)}{r} > g_{12}.$$

Hence, conditions (4.41)–(4.42) with $i = 1$, $j = 2$, are satisfied, and we obtain the first assertion by Lemma 4.2(1).

(2) From Lemma 3.5, we have

$$\hat{Q}_{21} = \{x > 0 : -F(x) \geq rg_{21}\}. \tag{4.51}$$

Under **(HF)**, we distinguish the following cases:

(i1) If $rg_{21} > -F(\hat{x})$, then $\hat{Q}_{21} = \emptyset$, and so $\mathcal{S}_2 = \emptyset$.

(i2) If $rg_{21} = -F(\hat{x})$, then either $\hat{x} = 0$ and so $\mathcal{S}_2 = \hat{Q}_{21} = \emptyset$, or $\hat{x} > 0$ and so $\hat{Q}_{21} = \{\hat{x}\}$, $\mathcal{S}_2 \subset \{\hat{x}\}$. In this last case, v_2 satisfies $rv_2 - \mathcal{L}v_2 - f_2 = 0$ on $(0,\hat{x}) \cup (\hat{x},\infty)$. By continuity and the smooth-fit condition of v_2 at $\hat{x}$, this implies that v_2 satisfies actually

$$rv_2 - \mathcal{L}v_2 - f_2 = 0, \quad x \in (0,\infty),$$

and so is in the form

$$v_2(x) = Ax^{m^+} + Bx^{m^-} + \hat{V}_2(x), \quad x \in (0,\infty).$$

Recalling that $0 \leq v_2(0^+) < \infty$ and v_2 is a nonnegative function satisfying a linear growth condition, this implies $A = B = 0$. Therefore, v_2 is equal to $\hat{V}_2$, which also means that $\mathcal{S}_2 = \emptyset$.

If $rg_{21} < -F(\hat{x})$, we need to distinguish three subcases depending on g_{21}:

- If $g_{21} > 0$, then there exist $0 < \underline{x}_{21} < \hat{x} < \bar{x}_{21} < \infty$ such that

$$\hat{Q}_{21} = [\underline{x}_{21}, \bar{x}_{21}]. \tag{4.52}$$

  We then conclude with Lemma 4.2(2) for $i = 2$, $j = 1$.
- If $g_{21} \leq 0$ with $rg_{21} > -F(\infty)$, then there exists $\bar{x}_{21} < \infty$ s.t.

$$\hat{Q}_{21} = (0, \bar{x}_{21}].$$

  Moreover, we clearly have $\sup_{x>0}(\hat{V}_1 - \hat{V}_2)(x) > (\hat{V}_1 - \hat{V}_2)(0) = 0 \geq g_{21}$. Hence, conditions (4.41) and (4.43) with $i = 2$, $j = 1$, are satisfied, and we deduce from Lemma 4.2(1) that $\mathcal{S}_2 = (0, \bar{x}_2^*]$ with $0 < \bar{x}_2^* < \infty$.
- If $rg_{21} \leq -F(\infty)$, then $\hat{Q}_{21} = (0, \infty)$, and we deduce from Lemma 4.2(3) for $i = 2$, $j = 1$, that $\mathcal{S}_2 = (0, \infty)$.

Finally, in the two above subcases when $\mathcal{S}_2 = [\underline{x}_2^*, \bar{x}_2^*]$ or $(0, \bar{x}_2^*]$, we notice that $\bar{x}_2^* < \underline{x}_1^*$ since $\mathcal{S}_2 \subset \mathcal{C}_1 = (0, \infty) \setminus \mathcal{S}_1$, which is equal, from (1), either to $(0, \infty)$ when $\underline{x}_1^* = \infty$ or to $(0, \underline{x}_1^*)$. □

Remark 4.2. In our viscosity solutions approach, the structure of the switching regions is derived from the smooth-fit property of the value functions, uniqueness result for viscosity solutions, and Lemma 3.5. This contrasts with the classical verification approach, where the structure of switching regions should be guessed ad hoc and checked a posteriori by a verification argument.

Economic interpretation. The previous proposition shows that, under **(HF)**, the switching region in regime 1 has two forms depending on the size of its corresponding positive switching cost: If g_{12} is larger than the "maximum net" profit $F(\infty)$ that one can expect by changing regime (case 1(i), which may occur only if $F(\infty) < \infty$), then one has no interest in switching regime, and one always stays in regime 1, i.e., $\mathcal{C}_1 = (0, \infty)$. However, if this switching cost is smaller than $F(\infty)$ (case 1(ii), which always holds true when $F(\infty) = \infty$), then there is some positive threshold from which it is optimal to change regime.

The structure of the switching region in regime 2 exhibits several different forms depending on the sign and size of its corresponding switching cost g_{21} with respect to the values $-F(\infty) < 0$ and $-F(\hat{x}) \geq 0$. If g_{21} is nonnegative larger than $-F(\hat{x})$ (case 2(i)), then one has no interest in switching regime, and one always stays in regime 2, i.e., $\mathcal{C}_2 = (0, \infty)$. If g_{21} is positive, but not too large (case 2(ii)), then there exists some bounded closed interval, which is not a neighborhood of zero, where it is optimal to change regime. Finally, when the switching cost g_{21} is negative, it is optimal to switch to regime 1 at least for small values of the state. Actually, if the negative cost g_{21} is larger than $-F(\infty)$ (case 2(iii), which always holds true for negative cost when $F(\infty) = \infty$), then the switching region is a bounded neighborhood of 0. Moreover, if the cost is negative large enough (case 2(iv), which may occur only if $F(\infty) < \infty$), then it is optimal to change regime for every value of the state.

By combining the different cases for regimes 1 and 2, and observing that case 2(iv) is not compatible with case 1(ii) by (2.7), we then have a priori seven different forms for both switching regions. These forms reduce actually to three when $F(\infty) = \infty$. The various structures of the switching regions are depicted in Figure 2.

Finally, we complete results of Proposition 4.3 by providing the explicit solutions for the value functions and the corresponding boundaries of the switching regions in the seven different cases depending on the model parameter values.

THEOREM 4.4. *Assume that* **(HF)** *holds.*

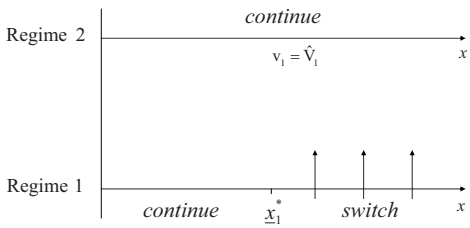


Figure II.1: $rg_{12} < F(\infty)$, $rg_{21} \geq -F(\hat{x})$

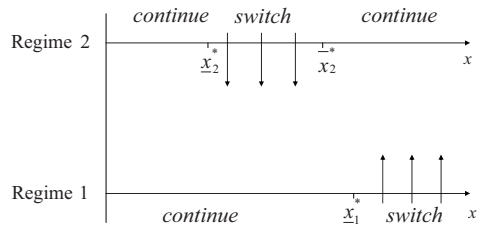


Figure II.2: $rg_{12} < F(\infty)$, $0 < rg_{21} < -F(\hat{x})$

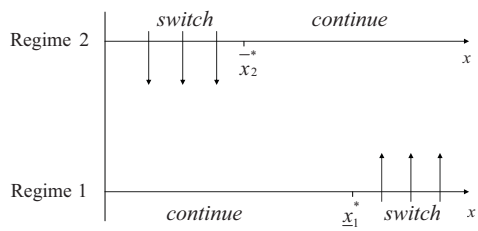


Figure II.3: $rg_{12} < F(\infty)$, $g_{21} \leq 0$, $-F(\infty) < rg_{21} < -F(\hat{x})$

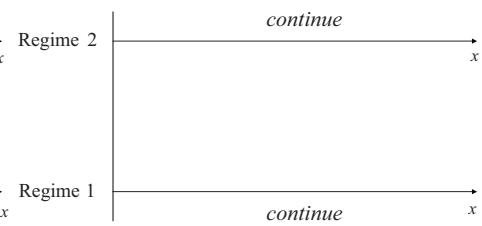


Figure II.4: $rg_{12} \geq F(\infty)$, $rg_{21} > -F(\hat{x})$

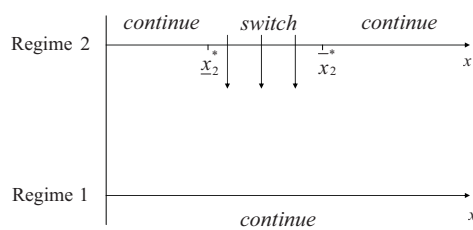


Figure II.5: $rg_{12} \geq F(\infty)$, $0 < rg_{21} < -F(\hat{x})$

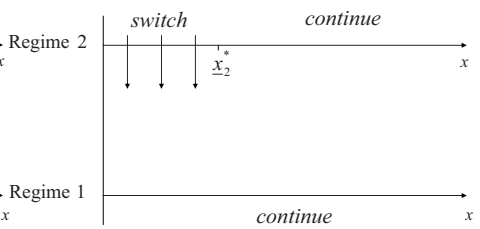


Figure II.6: $rg_{12} \geq F(\infty)$, $g_{21} \leq 0$, $F(\infty) < rg_{21} < -F(\hat{x})$

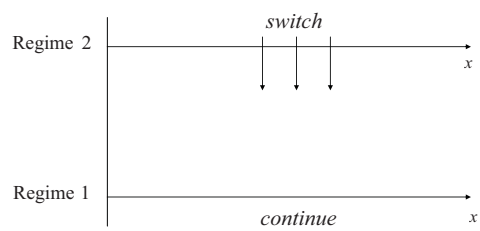


Figure II.7: $rg_{12} \geq F(\infty)$, $g_{21} \leq -F(\infty)$

FIG. 2.

(1) *If* $rg_{12} < F(\infty)$ *and* $rg_{21} \geq -F(\hat{x})$, *then*

$$v_1(x) = \begin{cases} Ax^{m^+} + \hat{V}_1(x), & x < \underline{x}_1^*, \\ v_2(x) - g_{12}, & x \geq \underline{x}_1^*, \end{cases}$$
$$v_2(x) = \hat{V}_2(x),$$

where the constants A and $\underline{x}_1^*$ are determined by the continuity and smooth-fit conditions of v_1 at $\underline{x}_1^*$:

$$A(\underline{x}_1^*)^{m^+} + \hat{V}_1(\underline{x}_1^*) = \hat{V}_2(\underline{x}_1^*) - g_{12},$$
$$Am^+(\underline{x}_1^*)^{m^+-1} + \hat{V}_1'(\underline{x}_1^*) = \hat{V}_2'(\underline{x}_1^*).$$

In regime 1, it is optimal to switch to regime 2 whenever the state process X exceeds the threshold $\underline{x}_1^*$, while when we are in regime 2, it is optimal to never switch.

(2) *If $rg_{12} < F(\infty)$ and $0 < rg_{21} < -F(\hat{x})$, then*

$$v_1(x) = \begin{cases} A_1 x^{m^+} + \hat{V}_1(x), & x < \underline{x}_1^*, \\ v_2(x) - g_{12}, & x \geq \underline{x}_1^*, \end{cases} \tag{4.53}$$

$$v_2(x) = \begin{cases} A_2 x^{m^+} + \hat{V}_2(x), & x < \underline{x}_2^*, \\ v_1(x) - g_{21}, & \underline{x}_2^* \leq x \leq \bar{x}_2^*, \\ B_2 x^{m^-} + \hat{V}_2(x), & x > \bar{x}_2^*, \end{cases} \tag{4.54}$$

where the constants A_1 and $\underline{x}_1^*$ are determined by the continuity and smooth-fit conditions of v_1 at $\underline{x}_1^*$, and the constants A_2, B_2, $\underline{x}_2^*$, $\bar{x}_2^*$ are determined by the continuity and smooth-fit conditions of v_2 at $\underline{x}_2^*$ and $\bar{x}_2^*$:

$$A_1(\underline{x}_1^*)^{m^+} + \hat{V}_1(\underline{x}_1^*) = B_2(\underline{x}_1^*)^{m^-} + \hat{V}_2(\underline{x}_1^*) - g_{12}, \tag{4.55}$$
$$A_1 m^+(\underline{x}_1^*)^{m^+-1} + \hat{V}_1'(\underline{x}_1^*) = B_2 m^-(\underline{x}_1^*)^{m^--1} + \hat{V}_2'(\underline{x}_1^*), \tag{4.56}$$
$$A_2(\underline{x}_2^*)^{m^+} + \hat{V}_2(\underline{x}_2^*) = A_1(\underline{x}_2^*)^{m^+} + \hat{V}_1(\underline{x}_2^*) - g_{21}, \tag{4.57}$$
$$A_2 m^+(\underline{x}_2^*)^{m^+-1} + \hat{V}_2'(\underline{x}_2^*) = A_1 m^+(\underline{x}_2^*)^{m^+-1} + \hat{V}_1'(\underline{x}_2^*), \tag{4.58}$$
$$A_1(\bar{x}_2^*)^{m^+} + \hat{V}_1(\bar{x}_2^*) - g_{21} = B_2(\bar{x}_2^*)^{m^-} + \hat{V}_2(\bar{x}_2^*), \tag{4.59}$$
$$A_1 m^+(\bar{x}_2^*)^{m^+-1} + \hat{V}_1'(\bar{x}_2^*) = B_2 m^-(\bar{x}_2^*)^{m^--1} + \hat{V}_2'(\bar{x}_2^*). \tag{4.60}$$

In regime 1, it is optimal to switch to regime 2 whenever the state process X exceeds the threshold $\underline{x}_1^*$, while when we are in regime 2, it is optimal to switch to regime 1 whenever the state process lies between $\underline{x}_2^*$ and $\bar{x}_2^*$.

(3) *If $rg_{12} < F(\infty)$ and $g_{21} \leq 0$ with $-F(\infty) < rg_{21} < -F(\hat{x})$, then*

$$v_1(x) = \begin{cases} Ax^{m^+} + \hat{V}_1(x), & x < \underline{x}_1^*, \\ v_2(x) - g_{12}, & x \geq \underline{x}_1^*, \end{cases}$$

$$v_2(x) = \begin{cases} v_1(x) - g_{21}, & 0 < x \leq \bar{x}_2^*, \\ Bx^{m^-} + \hat{V}_2(x), & x > \bar{x}_2^*, \end{cases}$$

where the constants A and $\underline{x}_1^*$ are determined by the continuity and smooth-fit conditions of v_1 at $\underline{x}_1^*$, and the constants B and $\bar{x}_2^*$ are determined by the continuity and smooth-fit conditions of v_2 at $\bar{x}_2^*$:

$$A(\underline{x}_1^*)^{m^+} + \hat{V}_1(\underline{x}_1^*) = B(\underline{x}_1^*)^{m^-} + \hat{V}_2(\underline{x}_1^*) - g_{12},$$
$$Am^+(\underline{x}_1^*)^{m^+-1} + \hat{V}_1'(\underline{x}_1^*) = Bm^-(\underline{x}_1^*)^{m^--1} + \hat{V}_2'(\underline{x}_1^*),$$
$$A(\bar{x}_2^*)^{m^+} + \hat{V}_1(\bar{x}_2^*) - g_{21} = B(\bar{x}_2^*)^{m^-} + \hat{V}_2(\bar{x}_2^*),$$
$$Am^+(\bar{x}_2^*)^{m^+-1} + \hat{V}_1'(\bar{x}_2^*) = Bm^-(\bar{x}_2^*)^{m^--1} + \hat{V}_2'(\bar{x}_2^*).$$

(4) *If* $rg_{12} \geq F(\infty)$ *and* $rg_{21} \geq -F(\hat{x})$, *then* $v_1 = \hat{V}_1$, $v_2 = \hat{V}_2$. *It is optimal to never switch in both regimes* 1 *and* 2.

(5) *If* $rg_{12} \geq F(\infty)$ *and* $0 < rg_{21} < -F(\hat{x})$, *then*

$$v_1(x) = \hat{V}_1(x),$$
$$v_2(x) = \begin{cases} Ax^{m^+} + \hat{V}_2(x), & x < \underline{x}_2^*, \\ v_1(x) - g_{21}, & \underline{x}_2^* \leq x \leq \bar{x}_2^*, \\ Bx^{m^-} + \hat{V}_2(x), & x > \bar{x}_2^*, \end{cases}$$

where the constants A, B, $\underline{x}_2^*$, $\bar{x}_2^*$ *are determined by the continuity and smooth-fit conditions of* v_2 *at* $\underline{x}_2^*$ *and* $\bar{x}_2^*$:

$$\begin{aligned} A(\underline{x}_2^*)^{m^+} + \hat{V}_2(\underline{x}_2^*) &= \hat{V}_1(\underline{x}_2^*) - g_{21}, \\ Am^+(\underline{x}_2^*)^{m^+-1} + \hat{V}_2'(\underline{x}_2^*) &= \hat{V}_1'(\underline{x}_2^*), \\ \hat{V}_1(\bar{x}_2^*) - g_{21} &= B(\bar{x}_2^*)^{m^-} + \hat{V}_2(\bar{x}_2^*), \\ \hat{V}_1'(\bar{x}_2^*) &= Bm^-(\bar{x}_2^*)^{m^--1} + \hat{V}_2'(\bar{x}_2^*). \end{aligned}$$

In regime 1, *it is optimal to never switch, while when we are in regime* 2, *it is optimal to switch to regime* 1 *whenever the state process lies between* $\underline{x}_2^*$ *and* $\bar{x}_2^*$.

(6) *If* $rg_{12} \geq F(\infty)$ *and* $g_{21} \leq 0$ *with* $-F(\infty) < rg_{21} < -F(\hat{x})$, *then*

$$v_1(x) = \hat{V}_1(x),$$
$$v_2(x) = \begin{cases} v_1(x) - g_{21}, & 0 < x \leq \bar{x}_2^*, \\ Bx^{m^-} + \hat{V}_2(x), & x > \bar{x}_2^*, \end{cases}$$

where the constants B *and* $\bar{x}_2^*$ *are determined by the continuity and smooth-fit conditions of* v_2 *at* $\bar{x}_2^*$:

$$\begin{aligned} \hat{V}_1(\bar{x}_2^*) - g_{21} &= B(\bar{x}_2^*)^{m^-} + \hat{V}_2(\bar{x}_2^*), \\ \hat{V}_1'(\bar{x}_2^*) &= Bm^-(\bar{x}_2^*)^{m^--1} + \hat{V}_2'(\bar{x}_2^*). \end{aligned}$$

In regime 1, *it is optimal to never switch, while when we are in regime* 2, *it is optimal to switch to regime* 1 *whenever the state process lies below* $\bar{x}_2^*$.

(7) *If* $rg_{12} \geq F(\infty)$ *and* $rg_{21} \leq -F(\infty)$, *then* $v_1 = \hat{V}_1$ *and* $v_2 = v_1 - g_{12}$. *In regime* 1, *it is optimal to never switch, while when we are in regime* 2, *it is always optimal to switch to regime* 1.

Proof. We prove the result only for case (2) since the other cases are dealt with similarly and are even simpler. Case (2) corresponds to the combination of cases 1(ii) and 2(ii) in Proposition 4.3. We then have $\mathcal{S}_1 = [\underline{x}_1^*, \infty)$, which means that $v_1 = v_2 - g_{12}$ on $[\underline{x}_1^*, \infty)$ and v_1 is a solution to $rv_1 - \mathcal{L}v_1 - f_1 = 0$ on $(0, \underline{x}_1^*)$. Since $0 \leq v_1(0^+) < \infty$, v_1 should have the form expressed in (4.53). Moreover, $\mathcal{S}_2 = [\underline{x}_2^*, \bar{x}_2^*]$, which means that $v_2 = v_1 - g_{21}$ on $[\underline{x}_2^*, \bar{x}_2^*]$, and v_2 satisfies $rv_2 - \mathcal{L}v_2 - f_2 = 0$ on $\mathcal{C}_2 = (0, \underline{x}_2^*) \cup (\bar{x}_2^*, \infty)$. Recalling again that $0 \leq v_2(0^+) < \infty$ and v_2 satisfies a linear growth condition, we deduce that v_2 has the form expressed in (4.54). Finally, the constants A_1, $\underline{x}_1^*$, which completely characterize v_1, and the constants A_2, B_2, $\underline{x}_2^*$, $\bar{x}_2^*$, which completely characterize v_2, are determined by the six relations (4.55)–(4.60) resulting from the continuity and smooth-fit conditions of v_1 at $\underline{x}_1^*$ and v_2 at $\underline{x}_2^*$ and $\bar{x}_2^*$, and from recalling that $\bar{x}_2^* < \underline{x}_1^*$. ∎

Remark 4.3. In the classical approach, for instance in case (2), we construct a priori a candidate solution in the form (4.53)–(4.54) and we have to check the existence of a sextuple solution to (4.55)–(4.60), which may be somewhat tedious! Here, by the viscosity solutions approach, and since we already state the smooth-fit C^1 property of the value functions, we know a priori the existence of a sextuple solution to (4.55)–(4.60).

Appendix A. Proof of comparison principle. In this section, we prove a comparison principle for the system of variational inequalities (3.8). The comparison result in [10] for switching problems in a finite horizon does not apply in our context. Inspired by [8], we first produce some suitable perturbation of a viscosity supersolution to deal with the switching obstacle, and then follow the general viscosity solution technique; see, e.g., [3].

THEOREM A.1. *Suppose u_i, $i \in \mathbb{I}_d$, are continuous viscosity subsolutions to the system of variational inequalities* (3.8) *on $(0,\infty)$, and w_i, $i \in \mathbb{I}_d$, are continuous viscosity supersolutions to the system of variational inequalities* (3.8) *on $(0,\infty)$, satisfying the boundary conditions $u_i(0^+) \le w_i(0^+)$, $i \in I_d$, and the linear growth condition*

$$|u_i(x)| + |w_i(x)| \le C_1 + C_2 x \quad \forall x \in (0,\infty),\ i \in \mathbb{I}_d, \tag{A.1}$$

for some positive constants C_1 and C_2. Then,

$$u_i \le w_i \quad \text{on } (0,\infty)\ \ \forall i \in \mathbb{I}_d.$$

Proof. Step 1. Let u_i and w_i, $i \in \mathbb{I}_d$, as in Theorem A.1. We first construct strict supersolutions to the system (3.8) with suitable perturbations of w_i, $i \in \mathbb{I}_d$. We set

$$h(x) = C_1' + C_2' x^p, \quad x > 0,$$

where C_1', $C_2' > 0$, and $p > 1$ are positive constants to be determined later. We then define for all $\lambda \in (0,1)$ the continuous functions on $(0,\infty)$ by

$$w_i^\lambda = (1-\lambda) w_i + \lambda(h + \alpha_i), \quad i \in \mathbb{I}_d,$$

where $\alpha_i = \min_{j \ne i} g_{ji}$. We then see that for all $\lambda \in (0,1)$, $i \in \mathbb{I}_d$,

$$\begin{aligned} w_i^\lambda - \max_{j \ne i}(w_j^\lambda - g_{ij}) &= \lambda\alpha_i + (1-\lambda)w_i - \max_{j \ne i}[(1-\lambda)(w_j - g_{ij}) + \lambda\alpha_j - \lambda g_{ij}] \\ &\ge (1-\lambda)[w_i - \max_{j \ne i}(w_j - g_{ij})] + \lambda\left(\alpha_i + \min_{j \ne i}(g_{ij} - \alpha_j)\right) \\ &\ge \lambda \min_{i \in \mathbb{I}_d}\left(\alpha_i + \min_{j \ne i}(g_{ij} - \alpha_j)\right) \\ &\ge \lambda \underline{\nu}, \end{aligned} \tag{A.2}$$

where $\underline{\nu} := \min_{i \in \mathbb{I}_d}[\alpha_i + \min_{j \ne i}(g_{ij} - \alpha_j)]$ is a constant independent of i. We now check that $\underline{\nu} > 0$, i.e., $\nu_i := \alpha_i + \min_{j \ne i}(g_{ij} - \alpha_j) > 0$, $\forall i \in \mathbb{I}_d$. Indeed, fix $i \in \mathbb{I}_d$, and let $k \in \mathbb{I}_d$ such that $\min_{j \ne i}(g_{ij} - \alpha_j) = g_{ik} - \alpha_k$ and set $\underline{i}$ such that $\alpha_i = \min_{j \ne i} g_{ji} = g_{\underline{i}i}$. We then have

$$\nu_i = g_{\underline{i}i} + g_{ik} - \min_{j \ne k} g_{jk} \ > \ g_{\underline{i}k} - \min_{j \ne k} g_{jk} \ \ge \ 0$$

by (2.6) and thus $\underline{\nu} > 0$.

By definition of the Fenchel–Legendre function in (2.5), and by setting $\tilde{f}(1) = \max_{i\in\mathbb{I}_d} \tilde{f}_i(1)$, we have for all $i \in \mathbb{I}_d$,

$$f_i(x) \le \tilde{f}(1) + x \ \le \tilde{f}(1) + 1 + x^p \quad \forall x > 0.$$

Moreover, recalling that $r > b := \max_i b_i$, we can choose $p > 1$ s.t.

$$\rho \ := \ r - pb - \frac{1}{2}\sigma^2 p(p-1) > 0,$$

where we set $\sigma := \max_i \sigma_i > 0$. By choosing

$$C_1' \ \ge \ \frac{2 + \tilde{f}(1)}{r} - \min_i \alpha_i, C_2' \ \ge \ \frac{1}{\rho},$$

we then have for all $i \in \mathbb{I}_d$,

$$\begin{aligned} rh(x) - \mathcal{L}_i h(x) - f_i(x) &= rC_1' + C_2' x^p \left[r - pb_i - \frac{1}{2}\sigma_i^2 p(p-1) \right] - f_i(x) \\ &\ge rC_1' + \rho C_2' x^p - f_i(x) \\ &\ge 1 \quad \forall x > 0. \end{aligned} \tag{A.3}$$

From (A.2) and (A.3), we then deduce that for all $i \in \mathbb{I}_d$, $\lambda \in (0,1)$, w_i^λ is a supersolution to

$$\min\left\{ rw_i^\lambda - \mathcal{L}_i w_i^\lambda - f_i, w_i^\lambda - \max_{j\neq i}(w_j^\lambda - g_{ij}) \right\} \ge \lambda\delta \quad \text{on } (0,\infty), \tag{A.4}$$

where $\delta = \underline{\nu} \wedge 1 > 0$.

Step 2. In order to prove the comparison principle, it suffices to show that for all $\lambda \in (0,1)$,

$$\max_{j\in\mathbb{I}_d} \sup_{(0,+\infty)} (u_j - w_j^\lambda) \le 0$$

since the required result is obtained by letting λ go to 0. We argue by contradiction and suppose that there exist some $\lambda \in (0,1)$ and $i \in \mathbb{I}_d$ s.t.

$$\theta \ := \ \max_{j\in\mathbb{I}_d} \sup_{(0,+\infty)} (u_j - w_j^\lambda) = \sup_{(0,+\infty)} (u_i - w_i^\lambda) \ > \ 0. \tag{A.5}$$

From the linear growth condition (A.1), and since $p > 1$, we observe that $u_i(x) - w_i^\lambda(x)$ goes to $-\infty$ when x goes to infinity. By choosing also $C_1' \ge \max_i w_i(0^+)$, we then have $u_i(0^+) - w_i^\lambda(0^+) = u_i(0^+) - w_i(0^+) + \lambda(w_i(0^+) - C_1') \le 0$. Hence, by continuity of the functions u_i and w_i^λ, there exists $x_0 \in (0,\infty)$ s.t.

$$\theta = u_i(x_0) - w_i^\lambda(x_0).$$

For any $\varepsilon > 0$, we consider the functions

$$\begin{aligned} \Phi_\varepsilon(x,y) &= u_i(x) - w_i^\lambda(y) - \phi_\varepsilon(x,y), \\ \phi_\varepsilon(x,y) &= \frac{1}{4}|x - x_0|^4 + \frac{1}{2\varepsilon}|x - y|^2 \end{aligned}$$

for all $x, y \in (0,\infty)$. By standard arguments in the comparison principle, the function Φ_ε attains a maximum in $(x_\varepsilon, y_\varepsilon) \in (0,\infty)^2$, which converges (up to a subsequence) to (x_0, x_0) when ε goes to zero. Moreover,

$$\lim_{\varepsilon\to 0} \frac{|x_\varepsilon - y_\varepsilon|^2}{\varepsilon} = 0. \tag{A.6}$$

Applying Theorem 3.2 in [3], we get the existence of M_ε, $N_\varepsilon \in \mathbb{R}$ such that

$$\begin{aligned}(p_\varepsilon, M_\varepsilon) &\in J^{2,+}u_i(x_\varepsilon),\\ (q_\varepsilon, N_\varepsilon) &\in J^{2,-}w_i^\lambda(y_\varepsilon),\end{aligned}$$

where

$$\begin{aligned}p_\varepsilon &= D_x\phi_\varepsilon(x_\varepsilon, y_\varepsilon) = \frac{1}{\varepsilon}(x_\varepsilon - y_\varepsilon) + (x_\varepsilon - x_{_0})^3,\\ q_\varepsilon &= -D_y\phi_\varepsilon(x_\varepsilon, y_\varepsilon) = \frac{1}{\varepsilon}(x_\varepsilon - y_\varepsilon),\end{aligned}$$

and

$$\begin{pmatrix} M_\varepsilon & 0\\ 0 & -N_\varepsilon\end{pmatrix} \;\le\; D^2\phi_\varepsilon(x_\varepsilon, y_\varepsilon) + \varepsilon\left(D^2\phi_\varepsilon(x_\varepsilon, y_\varepsilon)\right)^2 \tag{A.7}$$

with

$$D^2\phi_\varepsilon(x_\varepsilon, y_\varepsilon) = \begin{pmatrix} 3(x_\varepsilon - x_{_0})^2 + \frac{1}{\varepsilon} & -\frac{1}{\varepsilon}\\ -\frac{1}{\varepsilon} & \frac{1}{\varepsilon}\end{pmatrix}.$$

By writing the viscosity subsolution property (3.9) of u_i and the viscosity strict supersolution property (A.4) of w_i^λ, we have the following inequalities:

$$\begin{aligned}\min\Bigg\{ru_i(x_\varepsilon) - \left(\frac{1}{\varepsilon}(x_\varepsilon - y_\varepsilon) + (x_\varepsilon - x_{_0})^3\right) b_i x_\varepsilon - \frac{1}{2}\sigma_i^2 x_\varepsilon^2 M_\varepsilon - f_i(x_\varepsilon),\\ u_i(x_\varepsilon) - \max_{j\neq i}(u_j - g_{ij})(x_\varepsilon)\Bigg\} \;\le\; 0,\end{aligned} \tag{A.8}$$

$$\begin{aligned}\min\Bigg\{rw_i^\lambda(y_\varepsilon) - \frac{1}{\varepsilon}(x_\varepsilon - y_\varepsilon)b_i y_\varepsilon - \frac{1}{2}\sigma_i^2 y_\varepsilon^2 N_\varepsilon - f_i(y_\varepsilon),\\ w_i^\lambda(y_\varepsilon) - \max_{j\neq i}(w_j^\lambda - g_{ij})(y_\varepsilon)\Bigg\} \;\ge\; \lambda\delta.\end{aligned} \tag{A.9}$$

We then distinguish the following two cases:

(1) $u_i(x_\varepsilon) - \max_{j\neq i}(u_j - g_{ij})(x_\varepsilon) \le 0$ in (A.8).

By sending $\varepsilon \to 0$, this implies

$$u_i(x_0) - \max_{j\neq i}(u_j - g_{ij})(x_0) \le 0. \tag{A.10}$$

On the other hand, we have by (A.9):

$$w_i^\lambda(y_\varepsilon) - \max_{j\neq i}(w_j^\lambda - g_{ij})(y_\varepsilon) \ge \lambda\delta,$$

so that by sending ε to zero,

$$w_i^\lambda(x_0) - \max_{j\neq i}(w_j^\lambda - g_{ij})(x_0) \geq \lambda\delta. \tag{A.11}$$

Combining (A.10) and (A.11), we obtain

$$\begin{aligned}\theta \;=\; u_i(x_0) - w_i^\lambda(x_0) &\leq -\lambda\delta + \max_{j\neq i}(u_j - g_{ij})(x_0) - \max_{j\neq i}(w_j^\lambda - g_{ij})(x_0)\\ &\leq -\lambda\delta + \max_{j\neq i}(u_j - w_j^\lambda)(x_0)\\ &\leq -\lambda\delta + \theta,\end{aligned}$$

which is a contradiction.

(2) $ru_i(x_\varepsilon) - \left(\frac{1}{\varepsilon}(x_\varepsilon - y_\varepsilon) + (x_\varepsilon - x_0)^3\right) b_i x_\varepsilon - \frac{1}{2}\sigma_i^2 x_\varepsilon^2 M_\varepsilon - f_i(x_\varepsilon) \leq 0$ in (A.8). Since by (A.9), we also have

$$rw_i^\lambda(y_\varepsilon) - \frac{1}{\varepsilon}(x_\varepsilon - y_\varepsilon)b_i y_\varepsilon - \frac{1}{2}\sigma_i^2 y_\varepsilon^2 N_\varepsilon - f_i(y_\varepsilon) \geq \lambda\delta,$$

this yields, by combining the above two inequalities,

$$\begin{aligned}ru_i(x_\varepsilon) - rw_i^\lambda(y_\varepsilon) - \frac{1}{\varepsilon}b_i(x_\varepsilon - y_\varepsilon)^2 - (x_\varepsilon - x_0)^3 b_i x_\varepsilon\\ +\frac{1}{2}\sigma_i^2 y_\varepsilon^2 N_\varepsilon - \frac{1}{2}\sigma_i^2 x_\varepsilon^2 M_\varepsilon + f_i(y_\varepsilon) - f_i(x_\varepsilon) \leq -\lambda\delta.\end{aligned} \tag{A.12}$$

Now, from (A.7), we have

$$\frac{1}{2}\sigma_i^2 x_\varepsilon^2 M_\varepsilon - \frac{1}{2}\sigma_i^2 y_\varepsilon^2 N_\varepsilon \leq \frac{3}{2\varepsilon}\sigma_i^2(x_\varepsilon - y_\varepsilon)^2 + \frac{3}{2}\sigma_i^2 x_\varepsilon^2(x_\varepsilon - x_0)^2\left(3\varepsilon(x_\varepsilon - x_0)^2 + 2\right),$$

so that by plugging into (A.12), we have

$$\begin{aligned}r\left(u_i(x_\varepsilon) - w_i^\lambda(y_\varepsilon)\right) \leq \frac{1}{\varepsilon}b_i(x_\varepsilon - y_\varepsilon)^2 + (x_\varepsilon - x_0)^3 b_i x_\varepsilon + \frac{3}{2\varepsilon}\sigma_i^2(x_\varepsilon - y_\varepsilon)^2\\ + \frac{3}{2}\sigma_i^2 x_\varepsilon^2(x_\varepsilon - x_0)^2\left(3\varepsilon(x_\varepsilon - x_0)^2 + 2\right) + f_i(y_\varepsilon) - f_i(x_\varepsilon) - \lambda\delta.\end{aligned}$$

By sending ε to zero, and using (A.6), continuity of f_i, we obtain the required contradiction that $r\theta \leq -\lambda\delta < 0$. This ends the proof of Theorem A.1. □

Acknowledgments. We would like to thank Xin Guo for discussions. We are also grateful to the anonymous referees for constructive suggestions and comments that helped to improve the paper.

REFERENCES

[1] A. Bensoussan and J. L. Lions, *Contrôle impulsionnel et inéquations variationnelles*, Dunod, Paris, 1982.

[2] K. A. Brekke and B. Øksendal, *Optimal switching in an economic activity under uncertainty*, SIAM J. Control Optim., 32 (1994), pp. 1021–1036.

[3] M. Crandall, M. H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[4] A. Dixit and R. Pindick, *Investment under uncertainty*, Princeton University Press, Princeton, NJ, 1994.

[5] K. Duckworth and M. Zervos, *A model for investment decisions with switching costs*, Ann. Appl. Probab., 11 (2001), pp. 239–250.
[6] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[7] X. Guo, *An explicit solution to an optimal stopping problem with regime switching*, J. Appl. Probab., 38 (2001), pp. 464–481.
[8] H. Ishii and P. L. Lions, *Viscosity solutions of fully nonlinear second order elliptic partial differential equations*, J. Differential Equations, 83 (1990), pp. 26–78.
[9] H. Pham, *On the smooth-fit property for one-dimensional optimal switching problem*, in Séminaire de Probabilités, Vol. XL, to appear.
[10] S. Tang and J. Yong, *Finite horizon stochastic optimal switching and impulse controls with a viscosity solution approach*, Stoch. Stoch. Rep., 45 (1993), pp. 145–176.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 427–444

THE RELAXED STOCHASTIC MAXIMUM PRINCIPLE IN SINGULAR OPTIMAL CONTROL OF DIFFUSIONS*

SEID BAHLALI†, BOUALEM DJEHICHE‡, AND BRAHIM MEZERDI†

Abstract. This paper studies optimal control of systems driven by stochastic differential equations, where the control variable has two components, the first being absolutely continuous and the second singular. Our main result is a stochastic maximum principle for relaxed controls, where the first part of the control is a measure valued process. To achieve this result, we establish first order optimality necessary conditions for strict controls by using strong perturbation on the absolutely continuous component of the control and a convex perturbation on the singular one. The proof of the main result is based on the strict maximum principle, Ekeland's variational principle, and some stability properties of the trajectories and adjoint processes with respect to the control variable.

Key words. singular control, maximum principle, adjoint process, variational inequality, relaxed control, variational principle

AMS subject classification. 93Exx

DOI. 10.1137/050644744

1. Introduction.

We consider in this paper mixed, relaxed-singular stochastic control problems of systems governed by stochastic differential equations (SDEs), where the control variable has two components, the first being measure valued and the second singular. More precisely the system under consideration evolves according to the SDE

$$\begin{cases} dx_t^q = \int_{A_1} b\left(t, x_t^q, a\right) q_t\left(da\right) dt + \sigma\left(t, x_t^q\right) dB_t + G_t d\xi_t, \\ x^q(0) = x_0, \end{cases}$$

where b, σ, and G are given deterministic functions, x_0 is the initial state, and $B = (B_t)_{t\geq 0}$ is a standard Brownian motion, defined on a probability space $(\Omega, \mathcal{F}, P)$, equipped with a filtration $(\mathcal{F}_t)_{t\geq 0}$ satisfying the usual conditions. The control variable is a suitable process (q, ξ), where $q : \Omega \times [0,T] \longrightarrow P(A_1)$, $\xi : \Omega \times [0,T] \longrightarrow A_2 = ([0,\infty))^m$ are $\mathcal{F} \otimes B[0,T]$ measurable, $(\mathcal{F}_t)$ adapted, and ξ is an increasing process (componentwise), continuous on the left with limits on the right such that $\xi_0 = 0$.

The expected cost to be minimized over the class of admissible controls has the form

$$J(q,\xi) = E\left[g\left(x_T^q\right) + \int_0^T \int_{A_1} h\left(t, x_t^q, a\right) q_t(da) + \int_0^T k_t d\xi_t\right].$$

A control process that solves this problem is called optimal.

Singular control problems have been studied by many authors including Beněs, Shepp, and Witsenhausen [3], Chow, Menaldi, and Robin [6], Karatzas and Shreve

*Received by the editors November 9, 2005; accepted for publication (in revised form) October 22, 2006; published electronically April 27, 2007. This work is supported by MENA Swedish Algerian Research Partnership Program (348-2002-6874).

http://www.siam.org/journals/sicon/46-2/64474.html

†Laboratory of Applied Mathematics, University of Biskra, P.O. Box 145, Biskra (07000), Algeria (sbahlali@yahoo.fr, bmezerdi@yahoo.fr).

‡Division of Mathematical Statistics, Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden (boualem@math.kth.se).

[13], Davis and Norman [7], and Haussmann and Suo [10, 11, 12]. See [10] for a complete list of references on the subject. The approaches used in these papers are mainly based on dynamic programming. It was shown in particular that the value function is a solution of a variational inequality, and the optimal state is a reflected diffusion at the free boundary. Note that in [10], the authors apply the compactification method to show existence of an optimal relaxed singular control.

The second major method for solving control problems is to derive necessary conditions satisfied by some optimal control, known as the stochastic maximum principle. The first version of the stochastic maximum principle that covers singular control problems was obtained by Cadenillas and Haussmann [5] for linear systems. Second order necessary conditions for optimality for nonlinear SDEs with a controlled diffusion matrix were obtained by Bahlali and Mezerdi [2], extending the Peng maximum principle [17] to singular control problems. A first order weak maximum principle has been derived by Bahlali and Chala [1] in which convex perturbations are used for both absolutely continuous and singular components.

Our main goal in this paper is to establish a maximum principle for relaxed-singular controls, where the first part of the control is a measure valued process. This leads to necessary conditions of optimality satisfied by an optimal control, which exists under general conditions on the coefficients (see [10]). To achieve this program, we first prove a first order stochastic maximum principle for strict controls by using spike variation of the absolutely continuous part of the control and a convex perturbation of the singular part. Then by applying Ekeland's variational principle, we are able to prove necessary conditions for near optimality, satisfied by a sequence of strict controls converging in some sense to the relaxed optimal control, by the so-called chattering lemma. The relaxed maximum principle is then derived by using some stability properties of the trajectories and the adjoint processes with respect to the control variable. Our result generalizes the classical relaxed maximum principle proved in Mezerdi and Bahlali [16], to relaxed-singular control problems. However, we note that our maximum principle does not cover the work of Cadenillas and Haussmann [5]. The systems considered in [5] are linear but with random coefficients. In addition, the control variable in [5] is allowed to enter into the diffusion coefficient.

The plan of the rest of the paper is as follows. In section 2, we formulate the control problem and describe the assumptions of the model. In section 3, we derive the maximum principle for strict controls. The last section is devoted to the maximum principle for relaxed controls, which is the main result of this paper.

2. Assumptions. Let T be a fixed strictly positive real number and $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ be a filtered probability space satisfying the usual conditions, on which a d-dimensional Brownian motion $B = (B_t)_t$ is defined. We assume that $(\mathcal{F}_t)_t$ is the natural filtration of (B_t) augmented by P-null sets of $\mathcal{F}$.

Consider the following sets. A_1 is a nonempty subset of R^k and $A_2 = ([0,\infty))^m$.

DEFINITION 2.1. *An admissible strict control is a pair* (u,ξ) *of* $(A_1 \times A_2)$*-valued, measurable,* $\mathcal{F}_t$*-adapted processes, such that*

(i) ξ *is of bounded variation, nondecreasing left-continuous with right limits and* $\xi_0 = 0$.

(ii)

$$E\left[\sup_{t\in[0,T]} |u_t|^2 + |\xi_T|^2\right] < \infty.$$

We denote by $\mathcal{U} = \mathcal{U}_1 \times \mathcal{U}_2$ *the set of admissible strict controls.*

For any $(u,\xi)\in\mathcal{U}$, we consider the SDE

$$\left\{\begin{array}{l} dx_t = b\left(t,x_t,u_t\right)dt + \sigma\left(t,x_t\right)dB_t + G_t d\xi_t, \\ x\left(0\right) = x_0, \end{array}\right. \tag{2.1}$$

where

$$\begin{array}{l} b : [0,T]\times\mathbb{R}^n\times A_1 \longrightarrow \mathbb{R}^n, \\ \sigma : [0,T]\times\mathbb{R}^n \longrightarrow \mathcal{M}_{n\times d}\left(\mathbb{R}\right), \\ G : [0,T] \longrightarrow \mathcal{M}_{n\times m}\left(\mathbb{R}\right). \end{array}$$

The expected cost is given by

$$J\left(u,\xi\right) = E\left[g\left(x_T\right) + \int_0^T h\left(t,x_t,u_t\right)dt + \int_0^T k_t d\xi_t\right], \tag{2.2}$$

where

$$\begin{array}{l} g : \mathbb{R}^n \longrightarrow \mathbb{R}, \\ h : [0,T]\times\mathbb{R}^n\times A_1 \longrightarrow \mathbb{R}, \\ k : [0,T] \longrightarrow \left([0,\infty)\right)^m. \end{array}$$

The strict optimal control problem is to minimize the functional $J\left(.\right)$ over $\mathcal{U}$. A control that solves this problem is called optimal.

The following assumptions will be in force throughout this paper:

$$\begin{array}{c} b,\sigma,g,h \text{ are continuously differentiable with respect to } x. \text{ They and all their} \\ \text{derivatives, } b_x,\sigma_x,g_x,,h_x, \text{are continuous in } (x,u). \\ \text{The derivatives } \sigma_x, g_x \text{ are bounded and } b_x, h_x \text{ are bounded uniformly in } u. \\ b,\sigma \text{ are bounded by } C\left(1+|x|+|u|\right). \\ G \text{ and } k \text{ are continuous and } G \text{ is bounded.} \end{array} \tag{2.3}$$

Under the above hypothesis, (2.1) has a unique strong solution and the cost is well defined from $\mathcal{U}$ into $\mathbb{R}$.

We need the following matrix notation. We denote by $\mathcal{M}_{n\times d}\left(\mathbb{R}\right)$ the space of $n\times d$ real matrices and by $\mathcal{M}^d_{n\times n}\left(\mathbb{R}\right)$ the linear space of vectors $M=\left(M_1,\ldots,M_d\right)$, where $M_i\in\mathcal{M}_{n\times n}\left(\mathbb{R}\right)$.

For any $M,N\in\mathcal{M}^d_{n\times n}\left(\mathbb{R}\right)$, $L,Q\in\mathcal{M}_{n\times d}\left(\mathbb{R}\right)$, $x,y\in\mathbb{R}^n$, and $z\in\mathbb{R}^d$, we use the following notations:

$xy=\sum_{i=1}^n x_i y_i\in\mathbb{R}$ is the product scalar in $\mathbb{R}^n$;

$LQ=\sum_{i=1}^d L_i Q_i\in\mathbb{R}$, where L_i and Q_i are the ith columns of L and Q;

$ML=\sum_{i=1}^d M_i L_i\in\mathbb{R}^n$;

$Mxz=\sum_{i=1}^d \left(M_i x\right) z_i\in\mathbb{R}^n$;

$MN=\sum_{i=1}^d M_i N_i\in\mathcal{M}_{n\times n}\left(\mathbb{R}\right)$;

$$MLN = \sum_{i=1}^{d} M_i L N_i \in \mathcal{M}_{n\times n}(\mathbb{R});$$
$$MLz = \sum_{i=1}^{d} M_i L z_i \in \mathcal{M}_{n\times n}(\mathbb{R}).$$

We denote by L^* the transpose of the matrix L and $M^* = (M_1^*, \dots, M_d^*)$.

3. The maximum principle for strict controls. Facing a control problem, one may ask for necessary conditions satisfied by some optimal control. Throughout this section, let us suppose that $(u,\xi) \in \mathcal{U}$ is an optimal strict control and denote by (x_t) the corresponding optimal trajectory, i.e., the solution of (2.1) controlled by (u,ξ). The strict maximum principle will be proved in two steps. First we define a family of perturbed controls $\left(u^\theta,\xi\right)$, where u^θ is a spike variation of the absolutely continuous part u on a small time interval. The first variational inequality is derived from the fact that

$$J\left(u^\theta,\xi\right) - J(u,\xi) \geq 0.$$

The second step is to introduce another family of perturbed controls $\left(u,\xi^\theta\right)$, where ξ^θ is a convex perturbation of ξ. The second variational inequality is then obtained from the inequality

$$J\left(u,\xi^\theta\right) - J(u,\xi) \geq 0.$$

3.1. The first variational inequality. To obtain the first variational inequality in the stochastic maximum principle, we define the strong perturbation of the absolutely continuous part of the control, sometimes called the spike variation,

$$\left(u_t^\theta,\xi_t\right) = \begin{cases} (v,\xi_t) & \text{if } t \in [\tau,\tau+\theta], \\ (u_t,\xi_t) & \text{otherwise,} \end{cases} \tag{3.1}$$

where $0 \leq \tau < T$ is fixed, $\theta > 0$ is sufficiently small, and v is an arbitrary A_1-valued, F_τ-measurable random variable such that $E[|v|^2] < +\infty$. Note that the singular part is not affected by the perturbation.

If $x_t^{\left(u^\theta,\xi\right)}$ denotes the trajectory associated with $\left(u^\theta,\xi\right)$, then

$$\begin{aligned} x_t^{\left(u^\theta,\xi\right)} &= x_t, \quad t \leq \tau, \\ dx_t^{\left(u^\theta,\xi\right)} &= b\left(t, x_t^{\left(u^\theta,\xi\right)}, v\right)dt + \sigma\left(t, x_t^{\left(u^\theta,\xi\right)}\right)dB_t + G_t d\xi_t, \quad \tau < t < \tau+\theta, \\ dx_t^{\left(u^\theta,\xi\right)} &= b\left(t, x_t^{\left(u^\theta,\xi\right)}, u_t\right)dt + \sigma\left(t, x_t^{\left(u^\theta,\xi\right)}\right)dB_t + G_t d\xi_t, \quad \tau+\theta < t < T. \end{aligned}$$

It is easy to check by standard arguments that

$$E\left(\sup_{t\in[0,T]}\left|x_t^{\left(u^\theta,\xi\right)} - x_t\right|^2\right) \longrightarrow 0 \text{ as } \theta \longrightarrow 0. \tag{3.2}$$

Since (u,ξ) is optimal, then

$$J(u,\xi) \leq J\left(u^\theta,\xi\right) = J(u,\xi) + \theta\left.\frac{dJ\left(u^\theta,\xi\right)}{d\theta}\right|_{\theta=0} + o(\theta)$$

if the indicated derivative exists. Thus a necessary condition for optimality is that

$$\left.\frac{dJ\left(u^{\theta},\xi\right)}{d\theta}\right|_{\theta=0}\geq 0.$$

The rest of this subsection is devoted to the computation of this derivative.

Note that since $b(t,x,u)$ and $h(t,x,u)$ are sufficiently integrable, then the following property holds:

$$\frac{1}{\theta}\int_{t}^{t+\theta}E\left[\left|f(s,x_{s},u_{s})-f(t,x_{t},u_{t})\right|^{2}\right]\longrightarrow 0\ as\ \theta\longrightarrow 0,\ dt-a.e, \tag{3.3}$$

where f stands for b or h.

Choose τ such that (3.3) holds. We define y as the solution of the linear SDE

$$\begin{cases} dy_{t}=b_{x}\left(s,x_{s},u_{s}\right)y_{s}ds+\sigma_{x}\left(s,x_{s}\right)y_{s}dB_{s},\ \tau\leq s\leq T,\\ y_{\tau}=b\left(\tau,x_{\tau},v\right)-b\left(\tau,x_{\tau},u_{\tau}\right)\end{cases} \tag{3.4}$$

and define ς by

$$\begin{cases} d\varsigma_{t}=h_{x}\left(s,x_{s},u_{s}\right)y_{s}ds,\ \tau\leq s\leq T,\\ \varsigma_{\tau}=h\left(\tau,x_{\tau},v\right)-h\left(\tau,x_{\tau},u_{\tau}\right).\end{cases}$$

Lemma 3.1.

$$\lim_{\theta\to 0}E\left[\left|\frac{x_{T}^{\left(u^{\theta},\xi\right)}-x_{T}}{\theta}-y_{T}\right|^{2}\right]=0, \tag{3.5a}$$

$$\lim_{\theta\to 0}E\left[\left|\frac{1}{\theta}\int_{\tau}^{T}\left(h\left(t,x_{t},u_{t}^{\theta}\right)-h\left(t,x_{t},u_{t}\right)\right)dt-\varsigma_{T}\right|^{2}\right]=0. \tag{3.5b}$$

Proof. Since $x_{T}^{\left(u^{\theta},\xi\right)}-x_{T}$ does not depend on the singular part, the proof follows that of [4, Lemma 2.2]. □

Corollary 3.2.

$$\left.\frac{dJ\left(u^{\theta},\xi\right)}{d\theta}\right|_{\theta=0}=E\left[g_{x}\left(x_{T}\right).y_{T}+\varsigma_{T}\right]. \tag{3.6}$$

Proof. See [4, Corollary 2.1]. □

Let us introduce the adjoint process and the first variational inequality from (3.6). We proceed as in Bensoussan [4].

Let $\Phi(t,\tau)$ be the solution of the linear equation

$$\begin{cases} d\Phi(t,\tau)=b_{x}(t,x_{t},u_{t})\Phi(t,\tau)dt+\sigma_{x}(t,x_{t})\Phi(t,\tau)dB_{t},\ t>\tau,\\ \Phi(\tau,\tau)=I_{d}.\end{cases} \tag{3.7}$$

This equation is linear with bounded coefficients. Hence it admits a unique strong solution which is invertible, and its inverse Ψ_{t} is the unique solution of

$$\begin{cases} d\Psi_{t}=\left[\sigma_{x}(t,x_{t})\Psi_{t}\sigma_{x}^{*}(t,x_{t})-\Psi_{t}b_{x}\left(t,x_{t},u_{t}\right)\right]dt-\Psi_{t}\sigma_{x}\left(t,x_{t}\right)dB_{t},\ t>\tau,\\ \Psi(\tau,\tau)=I_{d}.\end{cases} \tag{3.8}$$

Moreover $\Phi(t,\tau)$ satisfies a semigroup property; that is, if $t > s > r$, then $\Phi(t,r) = \Phi(t,s).\Phi(s,r)$, which implies in particular that $\Phi(t,\tau) = \Phi(t)\Psi(\tau)$, where $\Phi(t) = \Phi(t,0)$ and $\Psi(t) = \Psi(t,0)$.

By the uniqueness property, it is easy to check that

$$y(t) = \Phi(t,\tau)\left(b\left(\tau, x_\tau, v\right) - b\left(\tau, x_\tau, u_\tau\right)\right).$$

Then replacing $y(t)$ with its value in (3.4), it holds that

$$\begin{aligned}\left.\frac{dJ\left(u^\theta,\xi\right)}{d\theta}\right|_{\theta=0} &= E\left(g_x\left(x_T\right).\Phi(T,\tau)\left(b\left(\tau,x_\tau,v\right) - b\left(\tau,x_\tau,u_\tau\right)\right)\right)\\ &\quad + E\left(h\left(\tau,x_\tau,v\right) - h\left(\tau,x_\tau,u_\tau\right)\right)\\ &\quad + E\left(\int_\tau^T h_x\left(s,x_s,u_s\right)\Phi(s,\tau)\left(b\left(\tau,x_\tau,v\right) - b\left(\tau,x_\tau,u_\tau\right)\right)ds\right).\end{aligned}$$

Now if we define the adjoint process by

$$p_t = E\left[\Psi_t^*\Phi_T^* g_x(x_T) + \Psi_t^*\int_t^T \Phi_s^* h_x\left(s,x_s,u_s\right)ds/\ \mathcal{F}_t\right], \tag{3.9}$$

it follows that

$$\left.\frac{dJ\left(u^\theta,\xi\right)}{d\theta}\right|_{\theta=0} = E\left[p_t.\left(b\left(\tau,x_\tau,v\right) - b\left(\tau,x_\tau,u_\tau\right)\right) + \left(h\left(\tau,x_\tau,v\right) - h\left(\tau,x_\tau,u_\tau\right)\right)\right].$$

If we define the Hamiltonian H from $[0,T]\times\mathbb{R}^n\times A_1\times\mathbb{R}^n$ into $\mathbb{R}$ by

$$H\left(t,x,v,p\right) = h\left(t,x,v\right) + p.b\left(t,x,v\right),$$

then we get from the optimality of (u,ξ) the first variational inequality

$$0 \le E\left[H\left(\tau,x_\tau,v,p_\tau\right) - H\left(\tau,x_\tau,u_\tau,p_\tau\right)\right], d\tau - a.e. \tag{3.10}$$

3.2. The second variational inequality. To obtain the second variational inequality of the stochastic maximum principle, we introduce the convex perturbation

$$\left(u_t,\xi_t^\theta\right) = \left(u_t,\xi_t + \theta\left(\eta_t - \xi_t\right)\right), \tag{3.11}$$

where $\theta > 0$ and η is an arbitrary element of $\mathcal{U}_2$. Note that the first part of the control is not affected by the perturbation. Since (u,ξ) is an optimal control, we'll derive the second variational inequality from the fact that

$$0 \le J\left(u,\xi^\theta\right) - J\left(u,\xi\right). \tag{3.12}$$

LEMMA 3.3. *Let $x_t^{\left(u,\xi^\theta\right)}$ be the trajectory associated with $\left(u,\xi^\theta\right)$. Then the following estimation holds:*

$$\lim_{\theta\to 0} E\left[\sup_{t\in[0,T]}\left|x_t^{\left(u,\xi^\theta\right)} - x_t\right|^2\right] = 0. \tag{3.13}$$

Proof. From assumption (2.3) and by using the Burkholder–Davis–Gundy inequality for the martingale part, we get

$$E\left[\sup_{t\in[0,T]}\left|x_t^{(u,\xi^\theta)}-x_t\right|^2\right]\le 6KE\int_0^t\sup_{\alpha\in[0,s]}\left|x_\alpha^{(u,\xi^\theta)}-x_\alpha\right|^2 ds+3M\theta^2E\left|\eta_T-\xi_T\right|^2.$$

From Definition 2.1 and using Gronwall's inequality, the result follows immediately by letting θ go to zero. □

LEMMA 3.4. *Under assumption* (2.3)*, the following estimation holds:*

$$\lim_{\theta\to 0}E\left[\left|\frac{x_t^{(u,\xi^\theta)}-x_t}{\theta}-z_t\right|^2\right]=0, \tag{3.14}$$

where z is the solution of the integral equation

$$z_t=\int_0^t b_x\left(s,x_s,u_s\right)z_s ds+\int_0^t\sigma_x\left(s,x_s\right)z_s dB_s+\int_0^t G_s d\left(\eta-\xi\right)_s.$$

Proof. From Definition 2.1 and assumption (2.3), it is easy to verify by Gronwall's inequality that

$$E\left[\sup_{t\in[0,T]}|z_t|^2\right]<\infty. \tag{3.15}$$

Let

$$\gamma_t^\theta=\frac{x_t^{(u,\xi^\theta)}-x_t^{(u,\xi)}}{\theta}-z_t.$$

It is easy to see that

$$\begin{aligned}E\left|\gamma_t^\theta\right|^2\le{}&3\int_0^t E\left|\int_0^1 b_x\left(s,x_s^{(u,\xi^\theta)}+\lambda\left[x_s^{(u,\xi^\theta)}-x_s\right],u_s\right)\gamma_t^\theta d\lambda\right|^2 dt\\&+3\int_0^t E\left|\int_0^1\sigma_x\left(s,x_s^{(u,\xi^\theta)}+\lambda\left[x_s^{(u,\xi^\theta)}-x_s\right]\right)\gamma_t^\theta d\lambda\right|^2 ds+3E\left|\rho_t^\theta\right|^2,\end{aligned}$$

where ρ_t^θ is given by

$$\begin{aligned}\rho_t^\theta={}&\int_0^t\int_0^1 z_s\left[b_x\left(s,x_s^{(u,\xi^\theta)}+\lambda\left[x_s^{(u,\xi^\theta)}-x_s\right],u_s\right)-b_x\left(s,x_s,u_s\right)\right]d\lambda ds\\&+\int_0^t\int_0^1 z_s\left[\sigma_x\left(s,x_s^{(u,\xi^\theta)}+\lambda\left[x_s^{(u,\xi^\theta)}-x_s\right]\right)-\sigma_x\left(s,x_s\right)\right]d\lambda dB_s.\end{aligned}$$

Since b_x,σ_x are bounded, it holds that

$$E\left|\gamma_t^\theta\right|^2\le 6C\int_0^t E\left|\gamma_s^\theta\right|^2 dt+3E\left|\rho_t^\theta\right|^2.$$

By using (3.13) and (3.15), the dominated convergence theorem, we obtain

$$\lim_{\theta\to 0}E\left|\rho_t^\theta\right|^2=0.$$

We conclude by applying Gronwall's lemma and letting θ go to zero. □

LEMMA 3.5. *The following inequality holds:*

$$0 \leq E\left[g_x\left(x_T\right) z_T\right] + E\int_0^T h_x\left(t, x_t, u_t\right) z_t dt + E\int_0^T k_t d\left(\eta - \xi\right)_t. \tag{3.16}$$

Proof. From (3.12), we have

$$\begin{aligned}
0 \leq\ & \frac{1}{\theta} E\left[g\left(x_T^{(u,\xi^\theta)}\right) - g\left(x_T\right)\right] + \frac{1}{\theta} E\int_0^T \left(h\left(t, x_t^{(u,\xi^\theta)}, u_t\right) - h\left(t, x_t, u_t\right)\right) dt \\
& + E\int_0^T k_t d\left(\eta_t - \xi_t\right) \\
=\ & E\int_0^1 \left(\frac{x_T^{(u,\xi^\theta)} - x_t}{\theta}\right) g_x\left(x_t + \lambda\left(x_T^{(u,\xi^\theta)} - x_t\right)\right) d\lambda \\
& + E\int_0^T \int_0^1 \left(\frac{x_T^{(u,\xi^\theta)} - x_t}{\theta}\right) h_x\left(t, x_t + \lambda\left(x_T^{(u,\xi^\theta)} - x_t\right), u_t\right) d\lambda dt \\
& + E\int_0^T k_t d\left(\eta - \xi\right)_t.
\end{aligned}$$

Since the derivatives g_x and h_x are continuous and bounded, by letting θ go to 0, we see that the result follows from (3.13) and (3.14). □

By the same method as in the last section, we are able to derive the second variational inequality from (3.16). If $\Phi(t,s)$ denotes the solution of (3.7), it is easy to check that z_t is given explicitly by

$$z_t = \int_0^t \Phi(t,s).G_s d\left(\eta - \xi\right)_s.$$

Replacing z_t with its value, we obtain the second variational inequality

$$0 \leq E\int_0^T \left(k_t + G_t^* p_t\right) d\left(\eta - \xi\right)_t, \tag{3.17}$$

where p_t is the adjoint process defined in the last subsection by (3.9).

3.3. The adjoint equation and the stochastic maximum principle. Applying Itô's formula to p_t given by (3.9), it is easy to see that p_t satisfies the linear backward SDE

$$\begin{cases} -dp_t = \left[h_x\left(t, x_t, u_t\right) + b_x^*\left(t, x_t, u_t\right) p_t + \sigma_x^*\left(t, x_t\right) K_t\right] dt - K_t dB_t, \\ p_T = g_x(x_T), \end{cases} \tag{3.18}$$

where K_t is given by

$$K_t = \Psi_t^* Q_t - \sigma_x^*\left(t, x_t\right) p_t; \quad K_t \in L^2\left(\left[0,T\right]; \mathbb{R}^{n\times d}\right), \tag{3.19}$$

and Q_t is given by the Itô representation of Brownian martingales

$$\begin{aligned}
\int_0^t Q_s dB_s = & E\left[\Phi_T^* g_x(x_T) + \int_0^T \Phi_t^* h_x(t, x_t, u_t) dt \ / \ \mathcal{F}_t\right] \\
& - E\left[\Phi_T^* g_x(x_T) + \int_0^T \Phi_t^* h_x(t, x_t, u_t) dt\right].
\end{aligned} \tag{3.20}$$

The stochastic maximum principle in its integral form is given by the following theorem.

THEOREM 3.6 (the strict stochastic maximum principle in integral form). *Let (u,ξ) be a strict optimal control minimizing the cost J over $\mathcal{U}$, and let x be the corresponding optimal trajectory. Then there exists a unique pair of adapted processes*

$$(p,K)\in L^2\left([0,T];\mathbb{R}^n\right)\times L^2\left([0,T];\mathbb{R}^{n\times d}\right),$$

which is the solution of the backward SDE (3.18) *such that for all $a\in A_1$ and $\eta\in\mathcal{U}_2$,*

$$H\left(t,x_t,u_t,p_t\right)\leq H\left(t,x_t,a,p_t\right),\ P-a.s,\ dt-a.e, \tag{3.21}$$

$$0\leq E\int_0^T\left(k_t+G_t^*p_t\right)d\left(\eta-\xi\right)_t. \tag{3.22}$$

Proof. From (3.10) we have

$$0\leq E\left[H\left(t,x_t,v,p_t\right)-H\left(t,x_t,u_t,p_t\right)\right],\quad dt-pp,$$

for every bounded A_1-valued, $\mathcal{F}_t$-measurable random variable v such that $E\left|v\right|^2<+\infty$.
Let $a\in A_1$ be a deterministic element and F be an arbitrary element of the σ-algebra $\mathcal{F}_t$, and set

$$w_t=a1_F+u_t1_{\Omega-F}.$$

It is obvious that w is an admissible control. Applying (3.10) with w we get

$$E\left[1_F\left(H\left(t,x_t,a,p_t\right)-H\left(t,x_t,u_t,p_t\right)\right)\right]\geq 0\quad\forall F\in\mathcal{F}_t,$$

which implies that $E\left[\left(H\left(t,x_t,a,p_t\right)-H\left(t,x_t,u_t,p_t\right)\right)/\mathcal{F}_t\right]\geq 0$.

The quantity inside the conditional expectation is $\mathcal{F}_t$-measurable, and thus the result follows immediately.

The second variational inequality (3.22) is proved in subsection 3.2. □

THEOREM 3.7 (the strict stochastic maximum principle). *Let (u,ξ) be an optimal control minimizing the cost J over $\mathcal{U}$, and let x be the corresponding optimal trajectory. Then there exists a unique pair of adapted processes*

$$(p,K)\in L^2\left([0,T];\mathbb{R}^n\right)\times L^2\left([0,T];\mathbb{R}^{n\times d}\right),$$

which is the solution of the backward SDE (3.18), *such that*

$$H\left(t,x_t,u_t,p_t\right)=\min_{a\in A_1}H\left(t,x_t,a,p_t\right),\ dt-a.e,\ P-a.s, \tag{3.23}$$

$$P\left\{\forall t\in[0,T],\ \forall i\ ;\ k_i\left(t\right)+G_i^*\left(t\right)p_t\geq 0\right\}=1, \tag{3.24}$$

$$P\left\{\sum_{i=1}^m\mathbf{1}_{\{k_i(t)+G_i^*(t)p_t\geq 0\}}d\xi_t^i=0\right\}=1. \tag{3.25}$$

Proof. To prove (3.24) and (3.25) we follow [5, Theorem 4.2]. Since (u, ξ) is optimal, the inequality

$$E \int_0^T (k_t + G_t^* p_t)\, d\,(\eta - \xi)_t \geq 0$$

holds for every $\eta \in \mathcal{U}_2$. In particular, let $\eta \in \mathcal{U}_2$ be defined by

$$d\eta_t^i = \begin{cases} 0 & \text{if } k_i(t) + G_i^*(t)\, p_t > 0, \\ d\xi_t^i & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} E \int_0^T & (k_t + G_t^* p_t)\, d\,(\eta - \xi)_t \\ &= E\left[\sum_{i=1}^m \int_0^T (k_i(t) + G_i^*(t)\, p_t)\, 1_{\{k_i(t)+G_i^*(t)p_t>0\}} d\left(-\xi_t^i\right)\right] \\ &\geq 0, \end{aligned}$$

which implies that

$$E\left[\sum_{i=1}^m \int_0^T (k_i(t) + G_i^*(t)\, p_t)\, 1_{\{k_i(t)+G_i^*(t)p_t>0\}} d\xi_t^i\right] = 0,$$

and relation (3.25) follows immediately.

Let us prove (3.24). For each $i \in \{1, 2, \ldots, m\}$, let

$$A_t^i = \{\omega \in \Omega : k_i(t) + G_i^*(t)\, p_t < 0\},$$

$$A^i = \{(t, \omega) \in [0, T] \times \Omega : k_i(t) + G_i^*(t)\, p_t < 0\},$$

and define

$$\eta_t^i = \xi_t^i + \int_0^t 1_{A^i}(s, \omega) ds.$$

It is easy to see that $\eta_t = (\eta_t^1, \eta_t^2, \ldots, \eta_t^m)$ is in $\mathcal{U}_2$. Moreover

$$E \int_0^T (k_t + G_t^* p_t)\, d\,(\eta - \xi)_t = E\left[\sum_{i=1}^m \int_0^T (k_i(t) + G_i^*(t)\, p_t)\, 1_{A^i} dt\right] < 0,$$

which contradicts (3.22), unless for every $i \in \{1, 2, \ldots, m\}$, $dt \otimes P(A^i) = 0$. This proves the desired result since k, G, and p are continuous. □

4. The relaxed model. The strict control problem, as defined in section 3, may fail to have an optimal solution. The reason is that the set $\mathcal{U}$ of strict controls is too narrow and should be embedded into a wider class with a richer topological structure for which the control problem becomes solvable (see [10]). Our main goal in this section is to establish a maximum principle for relaxed-singular controls. This leads to necessary conditions satisfied by an optimal relaxed-singular control, which exists under general assumptions on the coefficients.

Let us begin by a deterministic example which shows that even in simple cases, existence of a strict optimal control is not ensured. The problem is to minimize the cost function

$$J(u) = \int_0^T x^u(t)^2 dt,$$

over the set U_{ad} of open loop controls, i.e., measurable functions $u : [0,T] \longrightarrow \{-1,1\}$. Let $x^u(t)$ denote the solution of

$$dx_t^u = u\, dt, \qquad x(0) = 0.$$

It is clear that $\inf_{u\in\mathcal{U}} J(u) = 0$. Indeed, consider the sequence of controls

$$u_n(t) = (-1)^k \quad \text{if} \quad \frac{k}{n} \le t \le \frac{k+1}{n}, \qquad 0 \le k \le n-1.$$

Clearly $|x^{u_n}(t)| \le \frac{1}{n}$ and $|J(u_n)| \le \frac{T}{n^2}$, which implies that $\inf_{u\in\mathcal{U}} J(u) = 0$. There is, however, no control u such that $J(u) = 0$. If this were the case, then for every t, $x^u(t) = 0$. This in turn would imply that $u_t = 0$, which is impossible. The problem is that the sequence (u_n) has no limit in the space of strict controls. This limit, if it exists, will be the natural candidate for optimality. If we identify $u_n(t)$ with the Dirac measure $\delta_{u_n(t)}(da)$ and set $q_n(dt,du) = \delta_{u_n(t)}(du)dt$, we get a measure on $[0,1]\times A_1$. Then $(q_n(dt,du))_n$ converges weakly to $\frac{1}{2}dt\cdot[\delta_{-1}+\delta_1](da)$.

The idea of relaxed control is to replace the absolutely continuous part of the control (u_t) with a $P(A_1)$-valued process (q_t), where $P(A_1)$ denotes the space of probability measures equipped with the topology of weak convergence.

DEFINITION 4.1. *A relaxed-singular control is a pair (q,η) of processes such that*

(i) *q is a $P(A_1)$–valued process, progressively measurable with respect to $(\mathcal{F}_t)$ and such that for each t, $1_{(0,t]}.q$ is $\mathcal{F}_t$ measurable.*

(ii) $\eta \in \mathcal{U}_2$.

We denote by $\mathcal{R} = \mathcal{R}_1 \times \mathcal{U}_2$ the set of relaxed-singular controls.

For any $(q,\eta) \in \mathcal{R}$, we consider the relaxed SDE

$$\begin{cases} dx_t^q = \int_{A_1} b\,(t, x_t^q, a)\, q_t\,(da)\, dt + \sigma\,(t, x_t^q)\, dB_t + G_t d\eta_t, \\ x^q(0) = x_0. \end{cases} \tag{4.1}$$

The expected cost associated with a relaxed control (q,η) is defined as

$$J(q,\eta) = E\left[g\,(x_T^q) + \int_0^T \int_{A_1} h\,(t, x_t^q, a)\, q_t(da) + \int_0^T k_t d\eta_t \right]. \tag{4.2}$$

The set $\mathcal{U}_1$ is embedded into the set $\mathcal{R}_1$ of $P(A_1)$-valued processes by the mapping

$$\Psi : u \in \mathcal{U}_1 \longmapsto \Psi\,(u)_t\,(da) = \delta_{u(t)}(da) \in \mathcal{R}_1,$$

where δ_u denotes the Dirac measure at a single point u.

Throughout this section we suppose that

$$b \text{ and } h \text{ are bounded,} \tag{4.3}$$

$$A_1 \text{ is compact.} \tag{4.4}$$

Using the compactification method, Haussmann and Suo [10] proved that the relaxed-singular control problem admits an optimal solution. See also [9] for a complete study of relaxed controls for classical control problems.

4.1. Approximation of trajectories. The next lemma, known as the chattering lemma, tells us that any relaxed control is a weak limit of a sequence of strict controls. This lemma was first proved for deterministic measures and then extended to random measures in [9].

LEMMA 4.2 (chattering lemma). *Let (q_t) be a predictable process with values in the space of probability measures on A_1. Then there exists a sequence of predictable processes (u^n) with values in A_1 such that the sequence of random measures $\left(\delta_{u_t^n}(da)\,dt\right)$ converges weakly to $q_t(da)\,dt$, $P-a.s.$*

The next lemma gives the stability of the controlled SDE with respect to the control variable.

LEMMA 4.3. *Let $(q,\eta)\in\mathcal{R}$ be a relaxed control, and let x^q be the corresponding trajectory. Then there exists a sequence $(v^n,\eta)_n\subset\mathcal{U}$ such that*

$$\lim_{n\to\infty} E\left[\sup_{t\in[0,T]}|x_t^n-x_t^q|^2\right]=0, \tag{4.5}$$

$$\lim_{n\to\infty} J(v^n,\eta)=J(q,\eta), \tag{4.6}$$

where x^n denotes the solution of (2.1) associated with (v^n,η).

Proof. (i) Applying the Burkholder–Davis–Gundy inequality for the martingale part, we get

$$\begin{aligned}E\left[\sup_{t\in[0,T]}|x_t^n-x_t^q|^2\right]&\le\alpha_t^n+3\int_0^t E\left|b(s,x_s^n,v_s^n)-b(s,x_s^q,v_s^n)\right|^2ds\\&\quad+3\int_0^t E\left|\sigma(s,x_s^n)-\sigma\left(s,x_s^q\right)\right|^2ds,\end{aligned}$$

where $q_s^n(da)=\delta_{v_s^n}(da)$ and α_t^n is given by

$$\alpha_t^n=4E\left|\int_0^t\int_{A_1}b(s,x_s^q,a)q_s^n(da)ds-\int_0^t\int_{A_1}b(s,x_s^q,a)q_s(da)ds\right|^2.$$

Since the coefficients b and σ are Lipshitz in the state variable x, then

$$E\left(\sup_{t\in[0,T]}|x_t^q-x_t^n|^2\right)\le\alpha_t^n+6M\int_0^t E\left(\sup_{t\in[0,T]}|x_s^n-x_s^q|^2\right)ds,$$

b is bounded and continuous, and then from Lemma 4.2 and using the dominated convergence theorem, we obtain

$$\lim_{n\to\infty}\alpha_t^n=0.$$

The result follows from Gronwall's lemma.

(ii) On the other hand, since g and h are Lipshitz continuous in x, by using the Cauchy–Schwarz inequality, we get

$$\begin{aligned}|J(q^n,\eta)-J(q,\eta)|&\le C\left(E\left|x_T^n-x_T^q\right|^2\right)^{1/2}+C\int_0^T\left(E\left|x_t^n-x_t^q\right|^2\right)^{1/2}ds\\&\quad+\left(E\left|\int_0^T\int_{A_1}h(t,x_t^q,v_t^n)dt-\int_0^T\int_{A_1}h(t,x_t^q,a)q_t\left(da\right)dt\right|^2\right)^{1/2}.\end{aligned}$$

From (4.5), the first and second terms in the right-hand side converge to zero. Since h is continuous and bounded, by using the dominated convergence theorem, we have that the third term in the right-hand side tends to zero. □

Remark. As a consequence, it is easy to see that the strict and relaxed optimal control problems have the same value function.

4.2. Maximum principle for near optimal controls. In this section we establish necessary conditions of near optimality satisfied by a sequence of nearly optimal strict controls. This result is based on Ekeland's variational principle, which is given by the following.

LEMMA 4.4 (Ekeland [8]). *Let (E,d) be a complete metric space and $f : E \longrightarrow \overline{\mathbb{R}}$ be lower semicontinuous and bounded from below. Given $\varepsilon > 0$, suppose $u^\varepsilon \in E$ satisfies $f(u^\varepsilon) \leq \inf(f) + \varepsilon$. Then for any $\lambda > 0$, there exists $v \in E$ such that*

1. $f(v) \leq f(u^\varepsilon)$.
2. $d(u^\varepsilon, v) \leq \lambda$.
3. $f(v) < f(w) + \frac{\varepsilon}{\lambda} d(v,w)$ *for all* $w \neq v$.

To apply Ekeland's variational principle, we have to endow the set $\mathcal{U}$ of strict controls with an appropriate metric. For any $(u,\xi),(v,\eta) \in \mathcal{U}$, we set

$$
\begin{aligned}
d_1(u,v) &= P \otimes dt\left\{(\omega,t) \in \Omega \times [0,T],\ u(t,\omega) \neq v(t,\omega)\right\},\\
d_2(\xi,\eta) &= E\left(\sup_{t\in[0,T]} |\xi_t - \eta_t|^2\right)^{1/2},\\
d[(u,\xi),(v,\eta)] &= d_1(u,v) + d_2(\xi,\eta),
\end{aligned}
$$

where $P \otimes dt$ is the product measure of P with the Lebesgue measure dt.

Let us summarize some of the properties satisfied by d.

LEMMA 4.5.

1. *$(\mathcal{U},d)$ is a complete metric space.*
2. *The cost functional J is continuous from $\mathcal{U}$ into $\mathbb{R}$.*

Proof. 1. It is clear that $(\mathcal{U}_2,d_2)$ is a complete metric space. Moreover, it was shown in [15] that $(\mathcal{U}_1,d_1)$ is a complete metric space. Hence $(\mathcal{U},d)$ is a complete metric space as a product of two complete metric spaces.

Item 2 is proved as in [15]. □

Now let $(\mu,\xi) \in \mathcal{R}$ be an optimal relaxed control and denote by x^μ the trajectory of the system controlled by (μ,ξ). From Lemmas 4.2 and 4.3, there exists a sequence $(u^n)_n$ of strict controls such that

$$
\begin{aligned}
dt\mu_t^n(da) = dt\delta_{u_t^n}(da) &\underset{n\longrightarrow\infty}{\longrightarrow} dt\mu_t(da) \text{ weakly}, \quad P-a.s,\\
E\left[\sup_{t\in[0,T]} |x_t^n - x_t^\mu|^2\right] &\underset{n\longrightarrow\infty}{\longrightarrow} 0,
\end{aligned}
$$

where x_t^n is the solution of (4.1) corresponding to the control μ^n.

According to the optimality of (μ,ξ) and (4.6), there exists a sequence (ε_n) of positive real numbers with $\lim_{n\to\infty} \varepsilon_n = 0$ such that

$$
J(u^n,\xi) = J(\mu^n,\xi) \leq J(\mu,\xi) + \varepsilon_n.
$$

A suitable version of Lemma 4.4 implies that, given any $\varepsilon_n > 0$, there exists

$(u^n, \xi) \in \mathcal{U}$ such that

$$J(u^n, \xi) \leq \inf_{(v,\eta)\in\mathcal{U}} J(v, \eta) + \varepsilon_n,$$

$$J(u^n, \xi) \leq J(v, \eta) + \varepsilon_n d[(u^n, \xi); (v, \eta)] \quad \forall (v, \eta) \in \mathcal{U}. \tag{4.7}$$

Let us define the two perturbations

$$\left(u_t^{n,\theta}, \xi_t\right) = \begin{cases} (v, \xi_t) & \text{if } t \in [\tau, \tau + \theta], \\ (u_t^n, \xi_t) & \text{otherwise,} \end{cases} \tag{4.8}$$

$$\left(u_t^n, \xi_t^\theta\right) = (u_t^n, \xi_t + \theta(\eta_t - \xi_t)). \tag{4.9}$$

From (4.7) we have

$$0 \leq J\left(u_t^{n,\theta}, \xi_t\right) - J(u^n, \xi) + \varepsilon_n d\left[(u^n, \xi); \left(u^{n,\theta}, \xi\right)\right],$$

$$0 \leq J\left(u_t^n, \xi_t^\theta\right) - J(u^n, \xi) + \varepsilon_n d\left[(u^n, \xi); \left(u^n, \xi^\theta\right)\right].$$

From the definition of the metric d, we obtain

$$0 \leq J\left(u_t^{n,\theta}, \xi_t\right) - J(u^n, \xi) + \varepsilon_n d_1\left(u^n, u^{n,\theta}\right),$$

$$0 \leq J\left(u_t^n, \xi_t^\theta\right) - J(u^n, \xi) + \varepsilon_n d_2\left(\xi, \xi^\theta\right).$$

Using the definitions of d_1 and d_2, it holds that

$$0 \leq J\left(u_t^{n,\theta}, \xi_t\right) - J(u^n, \xi) + \varepsilon_n M_1 \theta, \tag{4.10}$$

$$0 \leq J\left(u_t^n, \xi_t^\theta\right) - J(u^n, \xi) + \varepsilon_n M_2 \theta, \tag{4.11}$$

where M_i is a positive constant.

From these above inequalities, we shall establish necessary conditions for near optimality.

THEOREM 4.6 (the near maximum principle in integral form). *For each $\varepsilon_n > 0$, there exists $(u^n, \xi) \in \mathcal{U}$ such that there exists a unique pair of adapted processes*

$$(p^n, K^n) \in L^2([0, T]; \mathbb{R}^n) \times L^2\left([0, T]; \mathbb{R}^{n\times d}\right),$$

which is the solution of the backward SDE

$$\begin{cases} -dp_t^n = [h_x(t, x_t^n, u_t^n) + b_x^*(t, x_t^n, u_t^n) p_t^n + \sigma_x^*(t, x_t^n) K_t^n] dt - K_t^n dB_t, \\ p_T^n = g_x(x_T^n), \end{cases} \tag{4.12}$$

such that for all $(v, \eta) \in \mathcal{U}$,

$$0 \leq E[H(t, x_t^n, v, p_t^n) - H(t, x_t^n, u_t^n, p_t^n)] + M_1 \varepsilon_n, \tag{4.13}$$

$$0 \leq E \int_0^T (k_t + G_t^* p_t^n) d(\eta - \xi)_t + M_2 \varepsilon_n, \tag{4.14}$$

where M_i is a positive constant.

Proof. From inequalities (4.10) and (4.11), respectively, we use the same method as in subsection 3.3 to obtain, respectively, (4.13) and (4.14). ∎

4.3. The relaxed stochastic maximum principle.

THEOREM 4.7 (the relaxed maximum principle in integral form). *Let (μ, ξ) be an optimal relaxed control minimizing the cost J over $\mathcal{R}$, and let x_t^μ be the corresponding optimal trajectory. Then there exists a unique pair of adapted processes*

$$(p^\mu, K^\mu) \in L^2\left([0,T];\mathbb{R}^n\right) \times L^2\left([0,T];\mathbb{R}^{n\times d}\right),$$

which is the solution of the backward SDE

$$\left\{\begin{array}{ll} -dp_t^\mu & = \left[\int_{A_1} h_x\left(t, x_t^\mu, a\right)\mu_t\left(da\right) + \int_{A_1} b_x^*\left(t, x_t^\mu, a\right)\mu_t\left(da\right)p_t^\mu\right]dt \\ & \quad + \sigma_x^*\left(t, x_t^\mu\right)K_t^\mu dt - K_t^\mu dB_t, \\ p_T^\mu & = g_x(x_T^\mu), \end{array}\right. \tag{4.15}$$

such that for all $(v,\eta) \in \mathcal{U}_1 \times \mathcal{U}_2$, we have

$$0 \le E\left[H\left(t, x_t^\mu, v, p_t^\mu\right) - H\left(t, x_t^\mu, \mu_t, p_t^\mu\right)\right], \tag{4.16}$$

$$0 \le E\int_0^T G_t^* p_t^\mu d\left(\eta - \xi\right)_t, \tag{4.17}$$

where $H\left(t, x_t^\mu, \mu_t, p_t^\mu\right) = \int_{A_1} H\left(t, x_t^\mu, a, p_t^\mu\right)\mu_t\left(da\right)$.

To prove Theorem 4.7, we need the following lemma.

LEMMA 4.8. *Let (p^n, K^n) and (p^μ, K^μ), respectively, be the solutions of* (4.12) *and* (4.15). *Then we have*

$$\lim_{n\longrightarrow\infty}\left(E\left[\sup_{t\in[0,T]}\left|p_t^n - p_t^\mu\right|^2\right] + E\int_0^T \left|K_t^n - K_t^\mu\right|^2 ds\right) = 0. \tag{4.18}$$

Proof. For simplicity of notation we set

$$\begin{aligned} b_t^\mu &= \int_{A_1} b_x^*\left(t, x_t^\mu, a\right)\mu_t(da),\ b_t^n = \int_{A_1} b_x^*\left(t, x_t^n, a\right)\mu_t^n(da), \\ \sigma_t^\mu &= \sigma_x^*\left(t, x_t^\mu\right),\ \sigma_t^n = \sigma_x^*\left(t, x_t^n\right), \\ h_t^\mu &= \int_{A_1} h_x\left(t, x_t^\mu, a\right)\mu_t(da),\ h_t^n = \int_{A_1} h_x\left(t, x_t^n, a\right)\mu_t^n(da), \\ \mu_t^n(da) &= \delta_{u_t^n}(da). \end{aligned}$$

Since b_x, σ_x, and h_x are continuous and bounded, by using Lemmas 4.2 and 4.3 and the dominated convergence theorem, we get

$$\lim_{n\longrightarrow\infty} E\left|b_t^n - b_t^\mu\right|^2 = \lim_{n\longrightarrow\infty} E\left|\sigma_t^n - \sigma_t^\mu\right|^2 = \lim_{n\longrightarrow\infty} E\left|h_t^n - h_t^\mu\right|^2 = 0. \tag{4.19}$$

Applying Itô's formula to $(p_t^n - p_t^\mu)^2$, it follows that

$$\begin{aligned} E\left|p_t^n - p_t^\mu\right|^2 + \int_t^T E\left|K_s^n - K_s^\mu\right|^2 ds &= E\left|g_x(x_T^n) - g_x(x_T^\mu)\right|^2 \\ &\quad + 2\int_t^T E\left|\left(p_s^\mu - p_s^n\right)\left(F_s^n - F_s^\mu\right)\right| ds, \end{aligned}$$

where F_t^μ and F_t^n are given by

$$\begin{aligned} F_t^\mu &= b_t^\mu p_t^\mu + \sigma_t^\mu K_t^\mu + h_t^\mu, \\ F_t^n &= b_t^n p_t^n + \sigma_t^n Q_t^n + h_t^n. \end{aligned}$$

Using Young's inequality $|c_1c_2| \leq \frac{\varepsilon}{2}|c_1|^2 + \frac{1}{2\varepsilon}|c_2|^2$, we obtain

$$\begin{aligned} E|p_t^n - p_t^\mu|^2 + \int_t^T E|K_t^n - K_s^\mu|^2\,dt &\leq E|g_x(x_T^n) - g_x(x_T^\mu)|^2 + \frac{1}{\varepsilon}\int_t^T E|p_s^\mu - p_s^n|^2\,ds \\ &\quad + \varepsilon\int_t^T E|F_s^n - F_s^\mu|^2\,ds \\ &\leq \left(\frac{1}{\varepsilon} + 6M\varepsilon\right)\int_t^T E|p_s^\mu - p_s^n|^2\,ds \\ &\quad + 6M\varepsilon\int_t^T E|K_s^n - K_s^\mu|^2\,ds + \varepsilon\xi_t^n, \end{aligned}$$

where ξ_t^n is given by

$$\begin{aligned} \xi_t^n = \frac{1}{\varepsilon}\left(E|g_x(x_T^n) - g_x(x_T^\mu)|^2\right) + 6E\int_t^T |(b_s^n - b_s^\mu)\,p_s^n|^2\,ds \\ + 6E\int_t^T |(\sigma_s^n - \sigma_s^\mu)\,K_s^n|^2\,ds + 3E\int_t^T |h_s^n - h_s^\mu|^2\,ds. \end{aligned} \tag{4.20}$$

Choose $\varepsilon = \frac{1}{12M}$. It follows that

$$E|p_t^n - p_t^\mu|^2 + \frac{1}{2}\int_t^T E|K_s^n - K_s^\mu|^2\,ds \leq C\int_t^T E|p_s^\mu - p_s^n|^2\,ds + C\xi_t^n,$$

with $C = \max\left\{\frac{1}{12M}, 12M + \frac{1}{2}\right\}$.
Hence

$$E|p_t^n - p_t^\mu|^2 \leq C\int_t^T E|p_s^\mu - p_s^n|^2\,ds + C\xi_t^n, \tag{4.21}$$

$$\int_t^T E|K_s^n - K_s^\mu|^2\,ds \leq 2C\int_t^T E|p_s^\mu - p_s^n|^2\,ds + 2C\xi_t^n. \tag{4.22}$$

Let us prove that $\lim_{n\longrightarrow\infty}\xi_t^n = 0$.
Since b_x is bounded, we have

$$|(b_s^n - b_s^\mu)\,p_s^n| \leq 2M|p_s^n|. \tag{4.23}$$

Then by the Cauchy–Schwarz inequality we get

$$\begin{aligned} E\int_t^T |(b_s^n - b_s^\mu)\,p_s^n|\,ds &\leq \int_t^T \left(E|b_s^n - b_s^\mu|^2\right)^{1/2}\left(E|p_s^n|^2\right)^{1/2}ds \\ &\leq c\int_t^T \left(E|b_s^n - b_s^\mu|^2\right)^{1/2}ds. \end{aligned}$$

Now from (4.19), we have

$$\lim_{n\longrightarrow\infty} E\int_t^T |(b_s^n - b_s^\mu)\,p_s^n|\,ds = 0, \tag{4.24}$$

and from (4.23), (4.24), and the dominated convergence theorem, we obtain

$$\lim_{n\longrightarrow\infty} E\int_t^T |(b_s^n - b_s^\mu)\,p_s^n|^2\,ds = 0. \tag{4.25}$$

Using the same arguments, it follows that

$$\lim_{n\longrightarrow\infty} E\int_t^T \left|\left(\sigma_s^n-\sigma_s^\mu\right)K_s^n\right|^2 ds=0. \tag{4.26}$$

Now since g_x is continuous and bounded,

$$\lim_{n\longrightarrow\infty} E\left|g_x(x_T^n)-g_x(x_T^\mu)\right|^2=0. \tag{4.27}$$

Now it is easy to see that by using the above results, $\lim_{n\longrightarrow\infty}\xi_t^n=0$.

The desired result follows from Gronwall's inequality. □

Proof of Theorem 4.6. Let (μ,ξ) be an optimal relaxed control. From Theorem 4.6, there exists a sequence $(u^n,\xi)_n$ in $\mathcal{U}$ such that for all $(v,\eta)\in\mathcal{U}$, the variational equations (4.13) and (4.14) hold. The result follows immediately by letting n go to infinity in (4.13) and (4.14) and using Lemma 4.8. □

THEOREM 4.9 (the relaxed maximum principle). *Let (μ,ξ) be an optimal relaxed control minimizing the functional cost J over $\mathcal{R}$ and let x_t^μ be the trajectory of the system controlled by (μ,ξ). Then there exists a unique pair of adapted processes*

$$\left(p^\mu,K^\mu\right)\in L^2\left([0,T];\mathbb{R}^n\right)\times L^2\left([0,T];\mathbb{R}^{n\times d}\right),$$

which is the solution of the backward SDE (4.15)*, such that*

$$H\left(t,x_t^\mu,\mu_t,p_t^\mu\right)=\min_{a\in A_1} H\left(t,x_t^\mu,a,p_t^\mu\right);\ dt-a.e\ ,\ P-a.s, \tag{4.28}$$

$$P\left\{\forall t\in[0,T]\ ,\ \forall i\ ;\ k_i\left(t\right)+G_i^*\left(t\right).p_t^\mu\geq 0\right\}=1, \tag{4.29}$$

$$P\left\{\sum_{i=1}^m \mathbf{1}_{\{k_i(t)+G_i^*(t).p_t^\mu\geq 0\}}d\xi_t^i=0\right\}=1. \tag{4.30}$$

Proof. From (4.16), we immediately deduce (4.28) by applying the same arguments as in the proof of Theorem 3.6. Using (4.17), we see that assertions (4.29) and (4.30) are proved exactly as in Theorem 3.7. □

Remarks. (1) If $G=k=0$, we recover the relaxed stochastic maximum principle for classical controls (see Mezerdi and Bahlali [16]).

(2) If $\mu_t(da)=\delta_{u(t)}(da)$, we recover the strict maximum principle (Theorem 3.7).

(3) If $\mu_t(da)=\delta_{u(t)}(da)$ and $G=k=0$, we obtain Kushner's stochastic maximum principle [14].

COROLLARY 4.10. *Under the same conditions as in Theorem* 4.9*, we have*

$$H\left(t,x_t^\mu,\mu_t,p_t^\mu\right)=\min_{q\in P(A_1)} H\left(t,x_t^\mu,q,p_t^\mu\right);\ dt-a.e,\ P-a.s, \tag{4.31}$$

$$P\left\{\forall t\in[0,T]\ ,\ \forall i\ ;\ k_i\left(t\right)+G_i^*\left(t\right).p_t^\mu\geq 0\right\}=1,$$

$$P\left\{\sum_{i=1}^m \mathbf{1}_{\{k_i(t)+G_i^*(t).p_t^\mu\geq 0\}}d\xi_t^i=0\right\}=1.$$

Proof. Equation (4.31) is proved as in [16, Corollary 4.2]. □

Acknowledgments. The authors would like to thank the referees for valuable suggestions that improved the manuscript.

REFERENCES

[1] S. Bahlali and A. Chala, *The stochastic maximum principle in optimal control of singular diffusions with nonlinear coefficients*, Random Oper. Stochastic Equations, 13 (2005), pp. 1–10.

[2] S. Bahlali and B. Mezerdi, *A general stochastic maximum principle for singular control problems*, Electron J. of Probab., 10 (2005), pp. 988–1004.

[3] V. E. Beněs, L. A. Shepp, and H. S. Witsenhausen, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.

[4] A. Bensoussan, *Lectures on stochastic control,* in Nonlinear Filtering and Stochastic Control, Lecture Notes in Math. 972, Springer, Berlin, 1982.

[5] A. Cadenillas and U. G. Haussmann, *The stochastic maximum principle for a singular control problem*, Stoch. Stoch. Rep., 49 (1994), pp. 211–237.

[6] P.-L. Chow, J.-L. Menaldi, and M. Robin, *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23 (1985), pp. 858–899.

[7] M. H. A. Davis and A. Norman, *Portfolio selection with transaction costs*, Math. Oper. Res., 15 (1990), pp. 676–713.

[8] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[9] N. El Karoui, N. Huu Nguyen, and M. J. Piqué, *Compactification methods in the control of degenerate diffusions*, Stochastics, 20 (1987), pp. 169–219.

[10] U. G. Haussmann and W. Suo, *Singular optimal stochastic controls* I: *Existence*, SIAM J. Control Optim., 33 (1995), pp. 916–936.

[11] U. G. Haussmann and W. Suo, *Singular optimal stochastic controls* II: *Dynamic programming*, SIAM J. Control Optim., 33 (1995), pp. 937–959.

[12] U. G. Haussmann and W. Suo, *Existence of singular optimal control laws for stochastic differential equations*, Stoch. Stoch. Rep., 48 (1994), pp. 249–272.

[13] I. Karatzas and S. E. Shreve, *Connections between optimal stopping and stochastic control* I: *Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.

[14] H. J. Kushner, *Necessary conditions for continuous parameter stochastic optimization problems,* SIAM J. Control Optim., 10 (1972), pp. 550–565.

[15] B. Mezerdi, *Necessary conditions for optimality for a diffusion with a nonsmooth drift,* Stoch. Stoch. Rep., 24 (1988), pp. 305–326.

[16] B. Mezerdi and S. Bahlali, *Necessary conditions for optimality in relaxed stochastic control problems*, Stoch. Stoch. Rep., 73 (2002), pp. 201–218.

[17] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 445–467

STABILIZING FEEDBACK CONTROLS FOR QUANTUM SYSTEMS*

MAZYAR MIRRAHIMI† AND RAMON VAN HANDEL‡

Abstract. No quantum measurement can give full information on the state of a quantum system; hence any quantum feedback control problem is necessarily one with partial observations and can generally be converted into a completely observed control problem for an appropriate quantum filter as in classical stochastic control theory. Here we study the properties of controlled quantum filtering equations as classical stochastic differential equations. We then develop methods, using a combination of geometric control and classical probabilistic techniques, for global feedback stabilization of a class of quantum filters around a particular eigenstate of the measurement operator.

Key words. quantum feedback control, quantum filtering equations, stochastic stabilization

AMS subject classifications. 81P15, 81V80, 93D15, 93E15

DOI. 10.1137/050644793

1. Introduction.

Though they are both probabilistic theories, probability theory and quantum mechanics have historically developed along very different lines. Nonetheless, the two theories are remarkably close, and indeed a rigorous development of quantum probability [27, 9] contains classical probability theory as a special case. The embedding of classical into quantum probability has a natural interpretation that is central to the idea of a quantum measurement: any set of *commuting* quantum observables can be represented as random variables on some probability space, and, conversely, any set of random variables can be encoded as commuting observables in a quantum model. The quantum probability model then describes the statistics of any set of measurements that we are allowed to make, whereas the sets of random variables obtained from commuting observables describe measurements that can be performed in a single realization of an experiment. As we are not allowed to make noncommuting observations in a single realization, any quantum measurement yields even in principle only partial information about the system.

The situation in quantum feedback control [18, 19] is thus very close to classical stochastic control with partial observations [7]. A typical quantum control scenario, representative of experiments in quantum optics, is shown in Figure 1.1. We wish to control the state of a cloud of atoms; e.g., we could be interested in controlling their collective angular momentum. To observe the atoms, we scatter a laser probe field off the atoms and measure the scattered light using a homodyne detector (a cavity can be used to increase the interaction strength between the light and the atoms). The observation process is fed into a controller which can feed back a control signal to the atoms through some actuator, e.g., a time-varying magnetic field. The entire setup can be described by a Schrödinger equation for the atoms and the probe field, which takes the form of a "quantum stochastic differential equation" in a Markovian limit. The controller, however, has access only to the observations of the probe. The

*Received by the editors November 10, 2005; accepted for publication (in revised form) October 22, 2006; published electronically April 27, 2007. This work was supported by the ARO under grant DAAD19-03-1-0073.

http://www.siam.org/journals/sicon/46-2/64479.html

†Centre Automatique et Systèmes, Ecole des Mines de Paris, 60 bd Saint-Michel, 75272 Paris Cedex 06, France (mazyar.mirrahimi@polytechnique.org).

‡Department of Physics and Control & Dynamical Systems, California Institute of Technology 266-33, Pasadena, CA 91125 (ramon@its.caltech.edu).

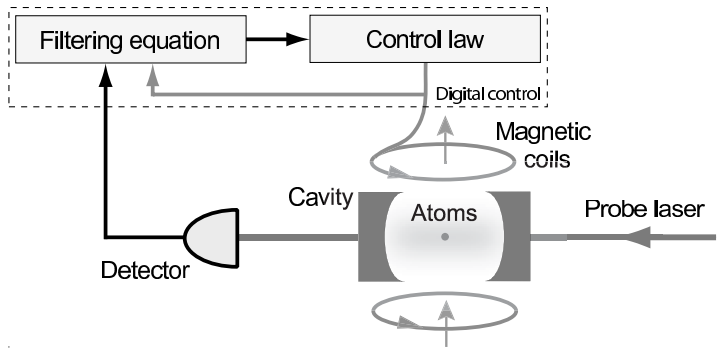


FIG. 1.1. *A typical feedback control scenario in quantum optics. A probe laser scatters off a cloud of atoms in an optical cavity and is ultimately detected. The detected signal is processed by a controller which feeds back to the system through a time-varying magnetic field.*

laser probe itself contributes quantum fluctuations to the observations; hence the observation process can be considered as a noisy observation of an atomic variable.

As in classical stochastic control we can use the properties of the conditional expectation to convert the output feedback control problem into one with complete observations. The conditional expectation $\pi_t(X)$ of an observable X given the observations $\{Y_s : 0 \le s \le t\}$ is the least mean square estimate of X_t (the observable X at time t) given $Y_{s\le t}$. One can obtain a quantum filtering equation [6, 9] that propagates $\pi_t(X)$ or, alternatively, the conditional density matrix ρ_t defined by the relation $\pi_t(X) = \mathrm{Tr}[\rho_t X]$. This is the quantum counterpart of the classical Kushner–Stratonovich equation and plays an equivalent role in quantum stochastic control. In particular, as $\mathbb{E}X_t = \mathbb{E}\pi_t(X)$ we can control the expectations of observables by designing a state feedback control law based on the filter.

Note that as the observation process $Y_{s\le t}$ is measured in a single experimental realization, it is equivalent to a classical stochastic process (i.e., the observables Y_t commute with each other at different times). But as the filter depends only on the observations, it is thus equivalent to a classical stochastic equation; in fact, the filter can be expressed as a classical (Itô) stochastic differential equation for the conditional density matrix ρ_t. Hence ultimately any quantum control problem of this form is reduced to a classical stochastic control problem for the filter.

In this paper we consider a class of quantum control problems of the following form. Rather than specifying a cost function to minimize, as in optimal control theory, we desire to asymptotically prepare a particular quantum state ρ_f in the sense that $\mathbb{E}X_t \to \mathrm{Tr}[\rho_f X]$ as $t \to \infty$ for all X (for a deterministic version, see, e.g., [30]). As $\mathbb{E}X_t = \mathbb{E}\pi_t(X)$, this comes down to finding a feedback control that will ensure the convergence $\rho_t \to \rho_f$ of the conditional density ρ_t. In addition to this convergence, we will show that our controllers also render the filter stochastically stable around the target state, which suggests some degree of robustness to perturbations. (This statement should be interpreted with care, however. See the remark after Proposition 3.4; we will not dwell on this issue.) In section 4 we will discuss the preparation of states in a cloud of atoms where the z-component of the angular momentum has zero variance, whereas in section 5 we will discuss the preparation of correlated states of two spins. Despite their relatively simple description, the creation of such states is not simple. Quantum feedback control may provide a desirable method for reliably preparing such states in practice (though other issues, e.g., the reduction of quan-

tum filters [17] for efficient real-time implementation, must be resolved before such schemes can be realized experimentally; see [15] for a state-of-the-art experimental demonstration of a related quantum control scenario).

Other work. Though we have attempted to indicate the origin of the control problems studied here, a detailed review of the physical and mathematical considerations behind our models is beyond the scope of this paper; nor can we do justice to the full history of the subject, or to results which do not relate directly to the control problems studied in this paper. In these respects we restrict ourselves to providing here a brief historical overview. In the remainder of the paper we will consider the quantum filtering equation as our starting point and concern ourselves exclusively with the associated classical stochastic control problem. For further details on the physical and mathematical basis for our models we refer the reader to the references below.

The theory of quantum nonlinear filtering was developed by Belavkin [6, 5]. The models used in the theory are based on the Hudson–Parthasarathy quantum stochastic calculus [21] and the theory of continuous quantum measurements as in the work of Barchielli and Lupieri [2]; a discrete-time version of the theory that did not require these tools can be found in Belavkin's earlier paper [4]. The potential for feedback control was already realized at that time; [4] develops discrete-time optimal controls for the models considered there, and a continuous-time version was sketched in [5]. We refer the reader to [8] and [14] for recent developments in quantum optimal feedback control, and to [9] for an accessible introduction to quantum probability and filtering. The control problems studied in this paper are not of the optimal control type; they have their origins in [33, 18]. See also [19] for further references.

In the physics literature the theory was independently developed by Carmichael [10] based on earlier work by Davies [11]. The connection to classical filtering theory (as in [6, 5]) was realized only much later; see, e.g., [12]. Wiseman [35] realized that Carmichael's work could be used to describe feedback in the quantum setting, but the controllers used in his work were of a restricted and somewhat unrealistic form: direct (unfiltered) linear feedback of white noise photocurrents with a deterministic gain. We do not consider this type of system here (see [36] for some remarks).

Structure of the paper. In section 2 we first introduce some tools from stochastic stability theory and stochastic analysis that we will use in our proofs. In section 3 we introduce the quantum filtering equation and study issues such as existence and uniqueness of solutions, continuity of the paths, etc. Though some of these issues have been considered in the literature in the absence of control (but in a more general setting; see, e.g., [3] and the references therein), to our knowledge such results are not available in the controlled case. In section 4 we pose the problem of stabilizing an angular momentum eigenstate and prove global stability under a particular control law. It is our expectation that the methods of section 4 are sufficiently flexible to be applied to a wide class of quantum state preparation scenarios. As an example, we use in section 5 the techniques developed in section 4 to stabilize particular entangled states of two spins. Additional results and numerical simulations will appear in [29].

2. Geometric tools for stochastic processes. In this section we briefly review two methods that will allow us to apply geometric control techniques to stochastic systems. The first is a stochastic version of the classical Lyapunov and LaSalle invariance theorems. The second, a support theorem for stochastic differential equations, will allow us to infer properties of stochastic sample paths through the study of a related deterministic system. We refer the reader to the references for proofs of the theorems.

2.1. Lyapunov and LaSalle invariance theorems. The Lyapunov stability theory and LaSalle's invariance theorem are important tools in the analysis of and control design for deterministic systems. Similarly, their stochastic counterparts will play an essential role in what follows. The subject of stochastic stability was studied extensively by Has'minskiĭ [20] and by Kushner [24]. We will cite a small selection of the results that will be needed in the following: a Lyapunov (local) stability theorem for Markov processes and the LaSalle invariance theorem of Kushner [24, 25, 26].

DEFINITION 2.1. *Let x_t^z be a diffusion process on the metric state space X, started at $x_0 = z$, and let $\tilde{z}$ denote an equilibrium position of the diffusion, i.e., $x_t^{\tilde{z}} = \tilde{z}$. Then*

1. *the equilibrium $\tilde{z}$ is said to be* stable in probability *if*

$$\lim_{z\to\tilde{z}} \mathbb{P}\left(\sup_{0\le t<\infty} \|x_t^z - \tilde{z}\| \ge \varepsilon\right) = 0 \qquad \forall \varepsilon > 0, \tag{2.1}$$

2. *the equilibrium $\tilde{z}$ is* globally stable *if it is stable in probability and additionally*

$$\mathbb{P}\left(\lim_{t\to\infty} x_t^z = \tilde{z}\right) = 1 \qquad \forall z \in X. \tag{2.2}$$

In the theorems below we will make the following assumptions.

1. The state space X is a complete separable metric space, and x_t^z is a homogeneous strong Markov process on X with continuous sample paths.
2. $V(\cdot)$ is a nonnegative real-valued continuous function on X.
3. For $\lambda > 0$, let $Q_\lambda = \{x \in X : V(x) < \lambda\}$ and assume Q_λ is nonempty. Let $\tau_\lambda = \inf\{t : x_t^z \notin Q_\lambda\}$ and define the stopped process $\tilde{x}_t^z = x_{t\wedge\tau_\lambda}^z$.
4. $\mathscr{A}_\lambda$ is the weak infinitesimal operator of $\tilde{x}_t$, and V is in the domain of $\mathscr{A}_\lambda$.

The following theorems can be found in Kushner [24, 25, 26].

THEOREM 2.2 (local stability). *Let $\mathscr{A}_\lambda V \le 0$ in Q_λ. Then the following hold:*

1. *$\lim_{t\to\infty} V(\tilde{x}_t^z)$ exists a.s., so $V(x_t^z)$ converges for a.e. path remaining in Q_λ.*
2. *$\mathbb{P}\text{-}\lim_{t\to\infty} \mathscr{A}_\lambda V(\tilde{x}_t^z) = 0$, so $\mathscr{A}_\lambda V(x_t^z) \to 0$ in probability as $t \to \infty$ for almost all paths which never leave Q_λ.*
3. *For $z \in Q_\lambda$ and $\alpha \le \lambda$ we have the uniform estimate*

$$\mathbb{P}\left(\sup_{0\le t<\infty} V(x_t^z) \ge \alpha\right) = \mathbb{P}\left(\sup_{0\le t<\infty} V(\tilde{x}_t^z) \ge \alpha\right) \le \frac{V(z)}{\alpha}. \tag{2.3}$$

4. *If $V(\tilde{z}) = 0$ and $V(x) \ne 0$ for $x \ne \tilde{z}$, then $\tilde{z}$ is stable in probability.*

The following theorem is a stochastic version of the LaSalle invariance theorem. Recall that a diffusion x_t^z is said to be Feller continuous if for fixed t, $\mathbb{E}G(x_t^z)$ is continuous in z for any bounded continuous G.

THEOREM 2.3 (invariance). *Let $\mathscr{A}_\lambda V \le 0$ in Q_λ. Suppose Q_λ has compact closure, $\tilde{x}_t^z$ is Feller continuous, and that $\mathbb{P}(\|\tilde{x}_t^z - z\| > \varepsilon) \to 0$ as $t \to 0$ for any $\varepsilon > 0$, uniformly for $z \in Q_\lambda$. Then $\tilde{x}_t^z$ converges in probability to the largest invariant set contained in $C_\lambda = \{x \in Q_\lambda : \mathscr{A}_\lambda V(x) = 0\}$. Hence x_t^z converges in probability to the largest invariant set contained in C_λ for almost all paths which never leave Q_λ.*

2.2. The support theorem. In the nonlinear control of deterministic systems an important role is played by the application of geometric methods, e.g., Lie algebra techniques, to the vector fields generating the control system. Such methods usually cannot be directly applied to stochastic systems, however, as the processes involved are not (sufficiently) differentiable. The support theorem for stochastic differential equations, in its original form due to Stroock and Varadhan [34], connects events of

probability one for a stochastic differential equation to the solution properties of an associated deterministic system. One can then apply classical techniques to the latter and invoke the support theorem to apply the results to the stochastic system; see, e.g., [22] for the application of Lie algebraic methods to stochastic systems.

We quote the following form of the theorem [23, 22].

THEOREM 2.4. *Let M be a connected, paracompact C^∞-manifold and let X_k, $k = 0, \dots, n$, be C^∞ vector fields on M such that all linear sums of X_k are complete. Let $X_k = \sum_l X_k^l(x)\partial_l$ in local coordinates and consider the Stratonovich equation*

$$dx_t = X_0(x_t)\,dt + \sum_{k=1}^{n} X_k(x_t) \circ dW_t^k, \qquad x_0 = x. \tag{2.4}$$

Consider in addition the associated deterministic control system

$$\frac{d}{dt}x_t^u = X_0(x_t^u) + \sum_{k=1}^{n} X_k(x_t^u)u^k(t), \qquad x_0^u = x, \tag{2.5}$$

with $u^k \in \mathscr{U}$, the set of all piecewise constant functions from $\mathbb{R}_+$ to $\mathbb{R}$. Then

$$\mathscr{S}_x = \overline{\{x_\cdot^u : u \in \mathscr{U}^n\}} \subset \mathscr{W}_x, \tag{2.6}$$

where $\mathscr{W}_x$ is the set of all continuous paths from $\mathbb{R}_+$ to M starting at x, equipped with the topology of uniform convergence on compact sets, and $\mathscr{S}_x$ is the smallest closed subset of $\mathscr{W}_x$ such that $\mathbb{P}(\{\omega \in \Omega : x_\cdot(\omega) \in \mathscr{S}_x\}) = 1$.

3. Solution properties of quantum filters. The purpose of this section is to introduce the dynamical equations for a general quantum system with feedback and to establish their basic solution properties.

We will consider quantum systems with finite dimension $1 < N < \infty$. The state space of such a system is given by the set of density matrices

$$\mathcal{S} = \{\rho \in \mathbb{C}^{N\times N} : \rho = \rho^*,\ \operatorname{Tr}\rho = 1,\ \rho \geq 0\}, \tag{3.1}$$

where ρ^* denotes Hermitian conjugation. In noncommutative probability the space $\mathcal{P}$ is the analogue of the set of probability measures of an N-state random variable. Finite-dimensional quantum systems are ubiquitous in contemporary quantum physics; a system with dimension $N = 2^n$, for example, can represent the collective state of n qubits in the setting of quantum computing, and $N = 2J + 1$ represents a system with fixed angular momentum J. The following lemma describes the structure of $\mathcal{S}$.

LEMMA 3.1. *$\mathcal{S}$ is the convex hull of $\{\rho \in \mathbb{C}^{N\times N} : \rho = vv^*,\ v \in \mathbb{C}^N,\ v^*v = 1\}$.*

Proof. The statement is easily verified by diagonalizing the elements of $\mathcal{P}$. ∎

We now consider continuous measurement of such a system, e.g., by weakly coupling it to an optical probe field and performing a diffusive observation of the field. When the state of the system is conditioned on the observation process, we obtain the following matrix-valued Itô equation for the conditional density, which is a quantum analogue of the Kushner–Stratonovich equation of nonlinear filtering [6, 9, 18]:

$$\begin{aligned} d\rho_t = {} & -i(H_t\rho_t - \rho_t H_t)\,dt + (c\rho_t c^* - \tfrac{1}{2}(c^*c\rho_t + \rho_t c^*c))\,dt \\ & + \sqrt{\eta}\,(c\rho_t + \rho_t c^* - \operatorname{Tr}[(c + c^*)\rho_t]\rho_t)\,dW_t. \end{aligned} \tag{3.2}$$

Here we have introduced the following quantities:

- The Wiener process W_t is the innovation $dW_t = dy_t - \sqrt{\eta}\,\mathrm{Tr}[(c+c^*)\rho_t]dt$. Here y_t, a continuous semimartingale with quadratic variation $\langle y, y\rangle_t = t$, is the observation process obtained from the system.
- $H_t = H_t^*$ is a Hamiltonian matrix which describes the action of external forces on the system. We will consider H_t of the form $H_t = F + u_t G$ with $F = F^*$, $G = G^*$, and the (real) scalar control input u_t.
- u_t is a bounded real càdlàg process that is adapted to $\mathcal{F}_t^y = \sigma(y_s, 0 \le s \le t)$, the filtration generated by the observations up to time t.
- c is a matrix which determines the coupling to the external (readout) field.
- $0 < \eta \le 1$ is the detector efficiency.

Let us begin by studying a different form of (3.2). Consider the linear Itô equation

$$d\tilde\rho_t = -i(H_t\tilde\rho_t - \tilde\rho_t H_t)\,dt + (c\tilde\rho_t c^* - \tfrac{1}{2}(c^*c\tilde\rho_t + \tilde\rho_t c^* c))\,dt + \sqrt{\eta}\,(c\tilde\rho_t + \tilde\rho_t c^*)\,dy_t, \tag{3.3}$$

which is the quantum analogue of the Zakai equation. As it obeys a global (random) Lipschitz condition, this equation has a unique strong solution [32, pp. 249–253].

LEMMA 3.2. *The set of nonnegative nonzero matrices is a.s. invariant for* (3.3).

Proof. We begin by expanding $\tilde\rho_0$ into its eigenstates, i.e., $\tilde\rho_0 = \sum_i \lambda_i v_0^i v_0^{i*}$ with $v_0^i \in \mathbb{C}^N$ being the ith eigenvector and λ_i the ith eigenvalue. As $\tilde\rho_0$ is nonnegative, all the λ_i are nonnegative. Now consider the set of equations

$$d\rho_t^i = -i(H_t\rho_t^i - \rho_t^i H_t)\,dt + (c\rho_t^i c^* - \tfrac{1}{2}(c^*c\rho_t^i + \rho_t^i c^*c))\,dt + (c\rho_t^i + \rho_t^i c^*)\,dW_t' \tag{3.4}$$

with $\rho_0^i = v_0^i v_0^{i*}$. Here we have extended our probability space to admit a Wiener process $\hat W_t$ that is independent of y_t, and $W_t' = \sqrt{\eta}\,y_t + \sqrt{1-\eta}\,\hat W_t$. The process $\tilde\rho_t$ is then equivalent in law to $\mathbb{E}[\rho_t'|\mathcal{F}_t^y]$, where $\rho_t' = \sum_i \lambda_i \rho_t^i$.

Now note that the solution of the set of equations

$$dv_t^i = -iH_t v_t^i\,dt - \tfrac{1}{2}c^*c\,v_t^i\,dt + c\,v_t^i\,dW_t', \qquad v_t^i \in \mathbb{C}^N, \tag{3.5}$$

satisfies $\rho_t^i = v_t^i v_t^{i*}$, as is readily verified by Itô's rule. By [32, pp. 326], we have that $v_t^i = U_t v_0^i$, where the random matrix U_t is a.s. invertible for all t. Hence a.s. $v_t^i \neq 0$ for any finite time unless $v_0^i = 0$. Thus clearly ρ_t' is a.s. a nonnegative nonzero matrix for all t, and the result follows. □

PROPOSITION 3.3. *Equation* (3.2) *has a unique strong solution* $\rho_t = \tilde\rho_t/\mathrm{Tr}\,\tilde\rho_t$ *in* $\mathcal{S}$.

Clearly this must be satisfied if (3.2) is to propagate a density.

Proof. As the set of nonnegative nonzero matrices is invariant for $\tilde\rho_t$, this implies in particular that $\mathrm{Tr}\,\tilde\rho_t > 0$ for all t a.s. Thus the result follows simply from application of Itô's rule to (3.3) and from the fact that if $M = \sum_i \lambda_i v_i$ is a nonnegative nonzero matrix, then $M/\mathrm{Tr}\,M = \sum_i (\lambda_i/\sum_j \lambda_j) v_i \in \mathcal{S}$. □

PROPOSITION 3.4. *The following uniform estimate holds for* (3.2):

$$\mathbb{P}\left(\sup_{0\le\delta\le\Delta} \|\rho_{t+\delta} - \rho_t\| > \varepsilon\right) \le C\Delta(1+\Delta) \qquad \forall \varepsilon > 0, \tag{3.6}$$

where $0 < C < \infty$ *depends only on* ε *and* $\|\cdot\|$ *is the Frobenius norm. Hence the solution of* (3.2) *is stochastically continuous uniformly in* t *and* ρ_0.

Proof. Write $\rho_t = \rho_0 + \Phi_t + \Xi_t$, where

$$\Phi_t = \int_0^t \left[-i(H_s\rho_s - \rho_s H_s) + (c\rho_s c^* - \tfrac{1}{2}(c^*c\rho_s + \rho_s c^*c))\right] ds, \tag{3.7}$$

$$\Xi_t = \int_0^t \sqrt{\eta}\,(c\rho_s + \rho_s c^* - \mathrm{Tr}[(c+c^*)\rho_s]\rho_s)\,dW_s. \tag{3.8}$$

For Ξ_t we have the estimate [1, pp. 81]

$$\mathbb{E}\left(\sup_{0\le\delta\le\Delta}\|\Xi_{t+\delta}-\Xi_t\|^2\right) \le 4\eta\int_t^{t+\Delta}\mathbb{E}\|c\rho_s+\rho_s c^*-\mathrm{Tr}[(c+c^*)\rho_s]\rho_s\|^2\,ds. \tag{3.9}$$

As the integrand is bounded clearly, this expression is bounded by $C_1\Delta$ for some positive constant $C_1<\infty$. For Φ_t we can write

$$\mathbb{E}\left(\sup_{0\le\delta\le\Delta}\|\Phi_{t+\delta}-\Phi_t\|^2\right) \le \mathbb{E}\left[\sup_{0\le\delta\le\Delta}\int_t^{t+\delta}\|G_s\|\,ds\right]^2 = \mathbb{E}\left[\int_t^{t+\Delta}\|G_s\|\,ds\right]^2, \tag{3.10}$$

where G_s denotes the integrand of (3.7). As $\|G_s\|$ is bounded, we can estimate this expression by $C_2\Delta^2$ with $C_2<\infty$. Using $\|A+B\|^2\le 2(\|A\|^2+\|B\|^2)$, we can write

$$\sup_{0\le\delta\le\Delta}\|\rho_{t+\delta}-\rho_t\|^2 \le 2\left(\sup_{0\le\delta\le\Delta}\|\Phi_{t+\delta}-\Phi_t\|^2+\sup_{0\le\delta\le\Delta}\|\Xi_{t+\delta}-\Xi_t\|^2\right). \tag{3.11}$$

Finally, Chebyshev's inequality gives

$$\mathbb{P}\left(\sup_{0\le\delta\le\Delta}\|\rho_{t+\delta}-\rho_t\|>\varepsilon\right) \le \frac{1}{\varepsilon^2}\mathbb{E}\left(\sup_{0\le\delta\le\Delta}\|\rho_{t+\delta}-\rho_t\|^2\right) \le \frac{2C_1\Delta+2C_2\Delta^2}{\varepsilon^2}, \tag{3.12}$$

from which the result follows. □

Remark. The statistics of the observation process y_t should, of course, depend both on the control u_t that is applied to the system and on the initial state ρ_0. We will always assume that the filter initial state ρ_0 matches the state in which the system is initially prepared (i.e., we do not consider "wrongly initialized" filters) and that the same control u_t is applied to the system and to the filter (see Figure 1.1). Quantum filtering theory then guarantees that the innovation W_t is a Wiener process. To simplify our proofs, we make from this point on the following choice: for all initial states and control policies, the corresponding observation processes are defined in such a way that they give rise to the same innovation process W_t.[1]

We now specialize to the following case:

- $u_t = u(\rho_t)$ with $u \in C^1(\mathcal{S},\mathbb{R})$.

In this simple feedback case we can prove several important properties of the solutions. First, however, we must show existence and uniqueness for the filtering equation with feedback: it is not a priori obvious that the feedback $u_t = u(\rho_t)$ results in a well-defined càdlàg control.

[1] This is contrary to the usual stochastic control setup: there the system and observation noises are fixed Wiener processes, and every initial state and control policy gives rise to a different innovation (Wiener) process. However, in the quantum case the system and observation noises do not even commute with the observations process, and thus we cannot use them to fix the innovations. In fact, the observation process y_t that emerges from the quantum probability model is defined only in a "weak" sense as a *-isomorphism between an algebra of observables and a set of random variables on $(\Omega,\mathcal{F},\mathbb{P})$ [9]. Hence we might as well choose the isomorphism for each initial state and control in such a way that all observations $y_t[\rho_0,u_t]$ give rise to the fixed innovations process W_t, regardless of ρ_0,u_t. Note that the only results that depend on the precise choice of $y_t[\rho_0,u_t]$ on $(\Omega,\mathcal{F},\mathbb{P})$ are joint statistics of the filter sample paths for different initial states or controls. However, such results are physically meaningless as the corresponding quantum models generally do not commute.

PROPOSITION 3.5. *Equation* (3.2) *with* $u_t = u(\rho_t)$, $u \in C^1$, *and* $\rho_0 = \rho \in \mathcal{S}$ *has a unique strong solution* $\rho_t \equiv \varphi_t(\rho, u)$ *in* $\mathcal{S}$, *and* u_t *is a continuous bounded control.*

Proof. As $\mathcal{S}$ is compact, we can find an open set $\mathcal{T} \subset \mathbb{C}^{N\times N}$ such that $\mathcal{S}$ is strictly contained in $\mathcal{T}$. Let $C(\rho) : \mathbb{C}^{N\times N} \to [0,1]$ be a smooth function with compact support such that $C(\rho) = 1$ for $\rho \in \mathcal{T}$, and let $U(\rho)$ be a $C^1(\mathbb{C}^{N\times N}, \mathbb{R})$ function such that $U(\rho) = u(\rho)$ for $\rho \in \mathcal{S}$. Then the equation

$$\begin{aligned} d\bar\rho_t = -iC(\bar\rho_t)[F + U(\bar\rho_t)G, \bar\rho_t]\,dt + C(\bar\rho_t)(c\bar\rho_t c^* - \tfrac{1}{2}(c^*c\bar\rho_t + \bar\rho_t c^*c))\,dt \\ + C(\bar\rho_t)\sqrt{\eta}\,(c\bar\rho_t + \bar\rho_t c^* - \mathrm{Tr}[(c+c^*)\bar\rho_t]\bar\rho_t)\,dW_t, \end{aligned}$$

where $[A, B] = AB - BA$, has global Lipschitz coefficients and hence has a unique strong solution in $\mathbb{C}^{N\times N}$ and a.s. continuous adapted sample paths [32]. Moreover, $\bar\rho_t$ must be bounded as $C(\rho)$ has compact support. Hence $U_t = U(\bar\rho_t)$ is an a.s. continuous, bounded adapted process.

Now consider the solution ρ_t of (3.2) with $u_t = U(\bar\rho_t)$ and $\rho_0 = \bar\rho_0 \in \mathcal{S}$. As both ρ_t and $\bar\rho_t$ have a unique solution, the solutions must coincide up to the first exit time from $\mathcal{T}$. But we have already established that ρ_t remains in $\mathcal{S}$ for all $t > 0$, so $\bar\rho_t$ can certainly never exit $\mathcal{T}$. Hence $\bar\rho_t = \rho_t$ for all $t > 0$, and the result follows. ∎

In the following, we will denote by $\varphi_t(\rho, u)$ the solution of (3.2) at time t with the control $u_t = u(\rho_t)$ and initial condition $\rho_0 = \rho \in \mathcal{S}$.

PROPOSITION 3.6. *If* $V(\rho)$ *is continuous, then* $\mathbb{E}V(\varphi_t(\rho, u))$ *is continuous in* ρ; *i.e., the diffusion* (3.2) *is Feller continuous.*

Proof. Let $\{\rho^n \in \mathcal{S}\}$ be a sequence of points converging to $\rho^\infty \in \mathcal{S}$. Let us write $\rho_t^n = \varphi_t(\rho^n, u)$ and $\rho_t^\infty = \varphi_t(\rho^\infty, u)$. First, we will show that

$$\mathbb{E}\|\rho_t^n - \rho_t^\infty\|^2 \to 0 \quad \text{as} \quad n \to \infty, \tag{3.13}$$

where $\|\cdot\|$ is the Frobenius norm ($\|A\|^2 = (A, A)$ with the inner product $(A, B) = \mathrm{Tr}\,(A^*B)$). We will write $\delta_t^n = \rho_t^n - \rho_t^\infty$. Using Itô's rule we obtain

$$\begin{aligned} \mathbb{E}\|\delta_t^n\|^2 = \|\delta_0^n\|^2 &+ \int_0^t \eta\mathbb{E}\mathrm{Tr}\left((c\delta_s^n + \delta_s^n c^* - \mathrm{Tr}[(c+c^*)\rho_s^n]\rho_s^n + \mathrm{Tr}[(c+c^*)\rho_s^\infty]\rho_s^\infty)^2\right) ds \\ &+ \int_0^t 2\,\mathbb{E}\left[\mathrm{Tr}\left((i[\rho_s^n, H(\rho_s^n)] - i[\rho_s^\infty, H(\rho_s^\infty)])\delta_s^n\right) + \mathrm{Tr}\left(c\delta_s^n c^*\delta_s^n - c^*c(\delta_s^n)^2\right)\right] ds, \end{aligned} \tag{3.14}$$

where $[A, B] = AB - BA$. Let us estimate each of these terms. We have

$$\begin{aligned} \mathrm{Tr}\left(c^*c(\delta_t^n)^2\right) &= \|c\delta_t^n\|^2 \le C_1\|\delta_t^n\|^2, \\ \mathrm{Tr}\,(c\delta_t^n c^*\delta_t^n) &= (\delta_t^n c, c\delta_t^n) \le \|\delta_t^n c\|\ \|c\delta_t^n\| \le C_2\|\delta_t^n\|^2, \end{aligned} \tag{3.15}$$

where we have used the Cauchy–Schwarz inequality and the fact that all the operators are bounded. Next we tackle

$$\mathrm{Tr}\left((i[\rho_t^n, H(\rho_t^n)] - i[\rho_t^\infty, H(\rho_t^\infty)])\delta_t^n\right) \le \|i[\rho_t^n, H(\rho_t^n)] - i[\rho_t^\infty, H(\rho_t^\infty)]\|\ \|\delta_t^n\|. \tag{3.16}$$

Now note that $S(\rho) = i[\rho, H(\rho)] = i[\rho, F + u(\rho)G]$ is C^1 in the matrix elements of ρ, and that its derivatives are bounded as $\mathcal{S}$ is compact. Hence $S(\rho)$ is Lipschitz continuous, and we have

$$\|S(\rho_t^n) - S(\rho_t^\infty)\| \le C_3\|\rho_t^n - \rho_t^\infty\| = C_3\|\delta_t^n\|, \tag{3.17}$$

which implies

$$\mathrm{Tr}\left((i[\rho_t^n, H(\rho_t^n)] - i[\rho_t^\infty, H(\rho_t^\infty)])\delta_t^n\right) \le C_3\|\delta_t^n\|^2. \tag{3.18}$$

Finally, we have $\|c\delta_t^n + \delta_t^n c^*\| \le C_4\|\delta_t^n\|$ due to boundedness of multiplication with c, and a Lipschitz argument similar to the one above can be applied to $S'(\rho) = \mathrm{Tr}[(c+c^*)\rho]\rho$, giving

$$\|\mathrm{Tr}[(c+c^*)\rho_t^n]\rho_t^n - \mathrm{Tr}[(c+c^*)\rho_t^\infty]\rho_t^\infty\| \le C_5\|\delta_t^n\|. \tag{3.19}$$

We can now use $\|A+B\|^2 \le \|A\|^2 + 2\|A\|\,\|B\| + \|B\|^2$ to estimate the last term in (3.14) by $C_6\|\delta_t^n\|^2$. Putting all these together, we obtain

$$\mathbb{E}\|\delta_t^n\|^2 \le \|\delta_0^n\|^2 + C\int_0^t \mathbb{E}\|\delta_s^n\|^2 ds, \tag{3.20}$$

and thus by Gronwall's lemma

$$\mathbb{E}\|\delta_t^n\|^2 \le e^{Ct}\|\delta_0^n\|^2 = e^{Ct}\|\rho^n - \rho^\infty\|^2. \tag{3.21}$$

As t is fixed, (3.13) follows.

We have now proved that $\rho_t^n \to \rho_t^\infty$ in mean square as $n \to \infty$, which implies convergence in probability. But then for any continuous V, $V(\rho_t^n) \to V(\rho_t^\infty)$ in probability [16, pp. 60]. As $\mathcal{S}$ is compact, V is bounded, and we have

$$\mathbb{E}V(\rho_t^\infty) = \mathbb{E}[\mathbb{P}\text{-}\lim_{n\to\infty} V(\rho_t^n)] = \lim_{n\to\infty} \mathbb{E}V(\rho_t^n) \tag{3.22}$$

by dominated convergence [16, pp. 72]. But as this holds for any convergent sequence ρ^n, the result follows. ∎

PROPOSITION 3.7. *$\varphi_t(\rho, u)$ is a strong Markov process in $\mathcal{S}$.*

Proof. The proof of the Markov property in [31, pp. 109–110] carries over to our case. But then the strong Markov property follows from Feller continuity [24]. ∎

PROPOSITION 3.8. *Let τ be the first exit time of ρ_t from an open set $Q \subset \mathcal{S}$ and consider the stopped process $\rho_t^Q = \varphi_{t\wedge\tau}(\rho, u)$. Then ρ_t^Q is also a strong Markov process in $\mathcal{S}$. Furthermore, for V s.t. $\mathscr{A}V$ exists and is continuous, where $\mathscr{A}$ is the weak infinitesimal operator associated to $\varphi_t(\rho, u)$, we have $\mathscr{A}_Q V(x) = \mathscr{A}V(x)$ if $x \in Q$ and $\mathscr{A}_Q V(x) = 0$ if $x \ne Q$ for the weak infinitesimal operator $\mathscr{A}_Q$ associated to ρ_t^Q.*

Proof. This follows from [24, pp. 11–12] and Proposition 3.4. ∎

4. Angular momentum systems. In this section we consider a quantum system with fixed angular momentum J ($2J \in \mathbb{N}$), e.g., an atomic ensemble, which is detected through a dispersive optical probe [19]. After conditioning, such systems are described by an equation of the form (3.2), where

- the Hilbert space dimension $N = 2J+1$;
- $c = \beta F_z$, $F = 0$, and $G = \gamma F_y$ with $\beta, \gamma > 0$.

Here F_y and F_z are the (self-adjoint) angular momentum operators defined as follows. Let $\{\psi_k : k = 0, \dots, 2J\}$ be the standard basis in $\mathbb{C}^N$; i.e., ψ_i is the vector with a single nonzero element $\psi_i^i = 1$. Then [28]

$$\begin{aligned} F_y\psi_k &= ic_{k-J}\psi_{k+1} - ic_{J-k}\psi_{k-1}, \\ F_z\psi_k &= (k-J)\psi_k \end{aligned} \tag{4.1}$$

with $c_m = \frac{1}{2}\sqrt{(J-m)(J+m+1)}$. Without loss of generality we will choose $\beta = \gamma = 1$, as we can always rescale time and u_t to obtain any β, γ.

Let us begin by studying the dynamical behavior of the resulting equation,

$$d\rho_t = -iu_t[F_y, \rho_t]\,dt - \tfrac{1}{2}[F_z,[F_z,\rho_t]]\,dt + \sqrt{\eta}\,(F_z\rho_t + \rho_t F_z - 2\,\mathrm{Tr}[F_z\rho_t]\rho_t)\,dW_t \tag{4.2}$$

without feedback $u_t = 0$.

Proposition 4.1 (quantum state reduction). *For any $\rho_0 \in \mathcal{S}$, the solution ρ_t of (4.2) with $u_t = 0$ converges a.s. as $t \to \infty$ to one of $\psi_m\psi_m^*$.*

Proof. We will apply Theorem 2.2 with $Q_\lambda = \mathcal{S}$. Consider the Lyapunov function $v(\rho) = \mathrm{Tr}[F_z^2\rho] - (\mathrm{Tr}[F_z\rho])^2$. One easily calculates $\mathscr{A}v(\rho) = -4\eta\, v(\rho)^2 \le 0$, and hence

$$\mathbb{E}v(\rho_t) = v(\rho_0) - 4\eta\int_0^t \mathbb{E}v(\rho_s)^2\,ds \tag{4.3}$$

by using Itô's rule. Note that $v(\rho) \ge 0$, so

$$4\eta\int_0^t \mathbb{E}v(\rho_s)^2\,ds = v(\rho_0) - \mathbb{E}v(\rho_t) \le v(\rho_0) < \infty. \tag{4.4}$$

Thus we have by monotone convergence

$$\mathbb{E}\int_0^\infty v(\rho_s)^2\,ds < \infty \quad\Longrightarrow\quad \int_0^\infty v(\rho_s)^2\,ds < \infty \quad \text{a.s.} \tag{4.5}$$

By Theorem 2.2 the limit of $v(\rho_t)$ as $t \to \infty$ exists a.s., and hence (4.5) implies that $v(\rho_t) \to 0$ a.s. But the only states ρ that satisfy $v(\rho) = 0$ are $\rho = \psi_m\psi_m^*$. ∎

The main goal of this section is to provide a feedback control law that globally stabilizes (4.2) around the equilibrium solution ($\rho_t \equiv \rho_f, u \equiv 0$), where we select a target state $\rho_f = v_f v_f^*$ from one of $v_f = \psi_m$.

Stabilization of quantum state reduction for low-dimensional angular momentum systems has been studied in [18]. It is shown that the main challenge in such a stabilization problem is due to the geometric symmetry hidden in the state space of the system. Many natural feedback laws fail to stabilize the closed-loop system around the equilibrium point ρ_f because of this symmetry: the ω-limit set contains points other than ρ_f. The approach of [18] uses computer searches to find continuous control laws that break this symmetry and globally stabilize the desired state. Unfortunately, the method is computationally involved and can be applied only to low-dimensional systems. Additionally, it is difficult to prove stability in this way for arbitrary parameter values, as the method is not analytical.

Here we present a different approach which avoids the unwanted limit points by changing the feedback law around them. The approach is entirely analytical and globally stabilizes the desired target state for any dimension N and $0 < \eta \le 1$. The main result of this section can be stated as the following theorem.

Theorem 4.2. *Consider the system (4.2) evolving in the set $\mathcal{S}$. Let $\rho_f = v_f v_f^*$, where v_f is one of ψ_m, and let $\gamma > 0$. Consider the following control law:*

1. *$u_t = -\mathrm{Tr}\,(i[F_y,\rho_t]\rho_f)$ if $\mathrm{Tr}\,(\rho_t\rho_f) \ge \gamma$.*
2. *$u_t = 1$ if $\mathrm{Tr}\,(\rho_t\rho_f) \le \gamma/2$.*
3. *If $\rho_t \in \mathcal{B} = \{\rho : \gamma/2 < \mathrm{Tr}\,(\rho\rho_f) < \gamma\}$, then $u_t = -\mathrm{Tr}\,(i[F_y,\rho_t]\rho_f)$ if ρ_t last entered $\mathcal{B}$ through the boundary $\mathrm{Tr}\,(\rho\rho_f) = \gamma$, and $u_t = 1$ otherwise.*

Then there exists $\gamma > 0$ *s.t.* u_t *globally stabilizes* (4.2) *around* ρ_f *and* $\mathbb{E}\rho_t \to \rho_f$ *as* $t \to \infty$.

Throughout the proofs we use the "natural" distance function

$$V(\rho) = 1 - \mathrm{Tr}\,(\rho\rho_f) : \mathcal{S} \to [0,1]$$

from the state ρ to the target state ρ_f. For future reference, let us define for each $\alpha \in [0,1]$ the level set $\mathcal{S}_\alpha$ to be

$$\mathcal{S}_\alpha = \{\rho \in \mathcal{S} : V(\rho) = \alpha\}.$$

Furthermore, we define the following sets:

$$\begin{aligned}
\mathcal{S}_{>\alpha} &= \{\rho \in \mathcal{S} : \alpha < V(\rho) \le 1\},\\
\mathcal{S}_{\ge\alpha} &= \{\rho \in \mathcal{S} : \alpha \le V(\rho) \le 1\},\\
\mathcal{S}_{<\alpha} &= \{\rho \in \mathcal{S} : 0 \le V(\rho) < \alpha\},\\
\mathcal{S}_{\le\alpha} &= \{\rho \in \mathcal{S} : 0 \le V(\rho) \le \alpha\}.
\end{aligned}$$

The proof of Theorem 4.2 proceeds in four steps.

1. In the first step we show that when the initial state lies in the set $\mathcal{S}_1$, the constant control field $u = 1$ ensures the exit of the trajectories (at least) in expectation from the level set $\mathcal{S}_1$.
2. In the second step we use the result of step 1 to show that there exists a $\gamma > 0$ such that whenever the initial state lies inside the set $\mathcal{S}_{>1-\gamma}$ and the control field is taken to be $u = 1$, the expectation value of the first exit time from this set takes a finite value. Thus if we start the controlled system in the set $\mathcal{S}_{>1-\gamma}$, it will exit this set in finite time with probability one.
3. In the third step we show that whenever the initial state lies inside the set $\mathcal{S}_{\le 1-\gamma}$ and the control is given by the feedback law $u(t) = -\mathrm{Tr}\,(i[F_y, \rho_t]\rho_f)$, the sample paths never exit the set $\mathcal{S}_{<1-\gamma/2}$ with a probability uniformly larger than a strictly positive value. We also show that almost all paths that never leave $\mathcal{S}_{<1-\gamma/2}$ converge to the equilibrium point ρ_f.
4. In the final step, we prove that there is a unique solution ρ_t under the control u_t by piecing together the solutions with fixed controls $u = 1$ and $u = -\mathrm{Tr}\,(i[F_y, \rho_t]\rho_f)$. Combining the results of the second and third steps, we show that the resulting trajectories of the system eventually converge toward the equilibrium state ρ_f with probability one.

Step 1. Let us take a fixed time $T > 0$ and define the nonnegative function

$$\chi(\rho) = \min_{t\in[0,T]} \mathbb{E}V(\varphi_t(\rho, 1)), \qquad \rho \in \mathcal{S}.$$

Recall that $\varphi_t(\rho, 1)$ denotes the solution of (4.2) at time t with the control $u_t = 1$ and initial condition $\rho_0 = \rho$. The goal of the first step is to show the following result.

LEMMA 4.3. $\chi(\rho) < 1$ *for all* $\rho \in \mathcal{S}_1$.

To prove this statement we will first show the following deterministic result.

LEMMA 4.4. *Consider the deterministic differential equation*

$$\frac{d}{dt}v_t = (-iF_y - F_z^2 + CF_z)v_t, \qquad v_0 \in \mathbb{C}^N \setminus \{0\}. \tag{4.6}$$

For sufficiently large $C \gg 1$, v_t *exits the set* $\{v : v^*v_f = 0\}$ *in the interval* $[0,T]$*; i.e., there exists* $t \in [0,T]$ *such that* $v_t^*v_f \neq 0$.

Proof. The matrices F_z and F_y are of the form

$$F_z = \begin{pmatrix} * & & & & 0 \\ & * & & & \\ & & \ddots & & \\ & & & * & \\ 0 & & & & * \end{pmatrix}, \qquad F_y = \begin{pmatrix} 0 & * & & & 0 \\ * & 0 & * & & \\ & \ddots & \ddots & \ddots & \\ & & * & 0 & * \\ 0 & & & * & 0 \end{pmatrix},$$

where F_z has no repeated diagonal entries (F_z has a nondegenerate spectrum) and the starred elements directly above and below the diagonal of F_y are all nonzero.

Now choose a constant κ so that the matrix

$$A = -iF_y - F_z^2 + \kappa F_z$$

admits distinct eigenvalues. This is always possible by choosing sufficiently large κ, as F_z has nondegenerate eigenvalues and the eigenvalues of A depend continuously[2] on κ. For $k \in \{1, \dots, N\}$ define the matrices A_{k-1} and $\tilde{A}_{k+1}$ to be

$$A_{k-1} = [A_{ij}]_{1 \le i,j \le k-1}, \qquad \tilde{A}_{k+1} = [A_{ij}]_{k+1 \le i,j \le N}.$$

The fact that the matrices $[(F_z)_{ij}]_{1 \le i,j \le k-1}$ and $[(F_z)_{ij}]_{k+1 \le i,j \le N}$ have different eigenvalues then implies that for sufficiently large κ the matrices A_{k-1} and $\tilde{A}_{k+1}$ have disjoint spectra as well.

Suppose that the solution of

$$\dot{v} = Av, \qquad v|_{t=0} = v_0,$$

never leaves the set $\{v : v^* v_f = 0\}$ in the interval $t \in [0, T]$. Then in particular

$$\frac{d^n}{dt^n} v^* v_f|_{t=0} = (A^n v_0)^* v_f = 0, \qquad n = 0, 1, \dots .$$

The matrix A is diagonalizable as it has distinct eigenvalues; i.e., $A = PDP^{-1}$ where D is a diagonal matrix. Thus

$$(D^n \tilde{v}_0)^* \tilde{v}_f = 0, \qquad n = 0, 1, \dots, \tag{4.7}$$

where $\tilde{v}_0 = P^{-1} v_0$ and $\tilde{v}_f = P^* v_f$. Equation (4.7) implies that $M \tilde{v}_0 = 0$, where

$$M = \begin{pmatrix} (\tilde{v}_f)_1^* & \dots & (\tilde{v}_f)_N^* \\ (\tilde{v}_f)_1^* D_{11} & \dots & (\tilde{v}_f)_N^* D_{NN} \\ (\tilde{v}_f)_1^* D_{11}^2 & \dots & (\tilde{v}_f)_N^* D_{NN}^2 \\ \vdots & \vdots & \vdots \\ (\tilde{v}_f)_1^* D_{11}^{N-1} & \dots & (\tilde{v}_f)_N^* D_{NN}^{N-1} \end{pmatrix}.$$

The determinant of this Vandermonde matrix is

$$\det M = (\tilde{v}_f)_1^* \cdots (\tilde{v}_f)_N^* \prod_{i>j} (D_{ii} - D_{jj}).$$

[2]Note that the coefficients of the characteristic polynomial of A are continuous functions of κ, and the roots of a polynomial depend continuously on the polynomial coefficients.

As the matrix A has distinct eigenvalues, all the entries $D_{11}, D_{22}, \dots, D_{NN}$ are different. Thus if we can show that all the entries of the vector $\tilde{v}_f$ are nonzero, then the matrix M must be invertible. But then $M\tilde{v}_0 = 0$ implies that $\tilde{v}_0 = 0$, and hence $v_0 = 0$ is the only initial state for which the dynamics does not leave the set $\{v : v^* v_f = 0\}$ in the interval $t \in [0, T]$, proving our assertion.

Let us thus show that in fact all elements of $\tilde{v}_f$ are nonzero. Note that

$$(\tilde{v}_f)_k = (P^* v_f)_k = P^*_{fk},$$

and hence it suffices to show that the eigenvectors of the matrix A have only nonzero elements. Suppose that an eigenvector Ξ of A admits a zero entry, i.e.,

$$A\Xi = \lambda\Xi, \qquad \Xi_k = 0 \text{ for some } k \in \{1, \dots, N\}.$$

Defining $\chi_{k-1} = [\Xi_j]_{j=1,\dots,k-1}$ and $\tilde{\chi}_{k+1} = [\Xi_j]_{j=k+1,\dots,N}$, a straightforward computation shows that due to the structure of the matrix A

$$A_{k-1}\chi_{k-1} = \lambda\chi_{k-1} \quad \text{and} \quad \tilde{A}_{k+1}\tilde{\chi}_{k+1} = \lambda\tilde{\chi}_{k+1}.$$

But by the discussion above, A_{k-1} and $\tilde{A}_{k+1}$ have disjoint spectra, so Ξ can be an eigenvector only if either $\chi_{k-1} = 0$ or $\tilde{\chi}_{k+1} = 0$.

Let us consider the case where $\chi_{k-1} = 0$; the treatment of the second case follows an identical argument. Let $j > k$ be the first nonzero entry of Ξ, i.e.,

$$\Xi_1 = \Xi_2 = \cdots = \Xi_{j-1} = 0 \quad \text{and} \quad \Xi_j \neq 0. \tag{4.8}$$

As $A\Xi = \lambda\Xi$, we have that

$$0 = \lambda\Xi_{j-1} = A_{j-1,j-2}\Xi_{j-2} + A_{j-1,j-1}\Xi_{j-1} + A_{j-1,j}\Xi_j = A_{j-1,j}\Xi_j = -i(F_y)_{j-1,j}\Xi_j.$$

As $(F_y)_{j-1,j} \neq 0$, this relation ensures that $\Xi_j = 0$. But this is in contradiction with (4.8) and thus Ξ cannot admit any zero entry. This completes the proof. ∎

Proof of Lemma 4.3. We begin by restating the problem as in the proof of Lemma 3.2. We can write $\varphi_t(\rho, 1) = \tilde{\rho}_t / \operatorname{Tr} \tilde{\rho}_t$ with $\tilde{\rho}_t = \sum_i \lambda_i \mathbb{E}[v_t^i v_t^{i*} | \mathcal{F}_t^y]$, where λ_i are convex weights and v_t^i are given by the equations

$$dv_t^i = -iF_y v_t^i \, dt - \tfrac{1}{2} F_z^2 v_t^i \, dt + F_z v_t^i \, dW_t', \qquad v_0^i \in \mathbb{C}^N \setminus \{0\}. \tag{4.9}$$

Note that $\mathbb{E}\operatorname{Tr}[\varphi_t(\rho, 1)\rho_f] = 0$ if and only if $\mathbb{E}\operatorname{Tr}[\tilde{\rho}_t \rho_f] = \sum_i \lambda_i \mathbb{E}[v_t^{i*} \rho_f v_t^i] = 0$. But as $v_t^{i*} \rho_f v_t^i \geq 0$, we obtain $\mathbb{E}V(\varphi_t(\rho, 1)) = 1$ if and only if $v_t^{i*} v_f = 0$ a.s. for all i.

To prove the assertion of the lemma, it suffices to show that there exists a $t \in [0, T]$ such that $\mathbb{E}V(\varphi_t(\rho, 1)) < 1$. Thus it is sufficient to prove that

$$\exists t \in [0, T] \quad \text{s.t.} \quad \mathbb{P}(v_t^* v_f \neq 0) > 0, \tag{4.10}$$

where v_t is the solution of an equation of the form (4.9). To this end we will use the support theorem, Theorem 2.4, together with Lemma 4.4.

To apply the support theorem we must first take care of two preliminary issues. First, the support theorem in the form of Theorem 2.4 must be applied to stochastic differential equations with a Wiener process as the driving noise, whereas the noise W_t' of (4.9) is a Wiener process with (bounded) drift:

$$dW_t' = \sqrt{\eta}\, dy_t + \sqrt{1-\eta}\, d\hat{W}_t = 2\eta \operatorname{Tr}[F_z \rho_t] dt + \sqrt{\eta}\, dW_t + \sqrt{1-\eta}\, d\hat{W}_t. \tag{4.11}$$

Using Girsanov's theorem, however, we can find a new measure $\mathbb{Q}$ that is equivalent to $\mathbb{P}$, such that W'_t is a Wiener process under $\mathbb{Q}$ on the interval $[0,T]$. But as the two measures are equivalent,

$$\exists t \in [0,T] \quad \text{s.t.} \quad \mathbb{Q}(v_t^* v_f \neq 0) > 0 \tag{4.12}$$

implies (4.10). Second, the support theorem refers to an equation in the Stratonovich form; however, we can easily find the Stratonovich form

$$dv_t = -iF_y v_t\,dt - F_z^2 v_t\,dt + F_z v_t \circ dW'_t \tag{4.13}$$

which is equivalent to (4.9). It is easily verified that this linear equation satisfies all the requirements of the support theorem.

To proceed, let us suppose that (4.12) does not hold true. Then

$$\mathbb{Q}(v_t^* v_f = 0) = 1 \qquad \forall t \in [0,T]. \tag{4.14}$$

Recall the following sets: $\mathscr{W}_{v_0}$ is the set of continuous paths starting at v_0, and $\mathscr{S}_{v_0}$ is the smallest closed subset of $\mathscr{W}_{v_0}$ such that $\mathbb{Q}(\{\omega \in \Omega : v_\cdot(\omega) \in \mathscr{S}_{v_0}\}) = 1$. Now denote by $\mathscr{T}_{v_0,t}$ the subset of $\mathscr{W}_{v_0}$ such that $v_t^* v_f = 0$, and note that $\mathscr{T}_{v_0,t}$ is closed in the compact uniform topology for any t. Then (4.14) would imply that $\mathscr{S}_{v_0} \subset \mathscr{T}_{v_0,t}$ for all $t \in [0,T]$. But by the support theorem the solutions of (4.6) are elements of $\mathscr{S}_{v_0}$, and by Lemma 4.4 there exist a time $t \in [0,T]$ and a constant C such that the solution of (4.6) is not an element of $\mathscr{T}_{v_0,t}$. Hence we have a contradiction, and the assertion is proved. □

Step 2. We begin by extending the result of Lemma 4.3 to hold uniformly in a neighborhood of the level set $\mathcal{S}_1$.

LEMMA 4.5. *There exists* $\gamma > 0$ *such that* $\chi(\rho) < 1-\gamma$ *for all* $\rho \in \mathcal{S}_{\geq 1-\gamma}$.

Proof. Suppose that for every $\xi > 0$ there exists a matrix $\rho_\xi \in \mathcal{S}_{>1-\xi}$ such that

$$1 - \xi < \chi(\rho_\xi) \leq 1.$$

By extracting a subsequence $\xi_n \searrow 0$ and using the compactness of $\mathcal{S}$, we can assume that $\rho_{\xi_n} \to \rho_\infty \in \mathcal{S}_1$ and that $\chi(\rho_{\xi_n}) \to 1$. But by Lemma 4.3 $\chi(\rho_\infty) = 1 - \epsilon < 1$. Now choose $s \in [0,T]$ such that

$$\mathbb{E}V(\varphi_s(\rho_\infty, 1)) = 1 - \epsilon.$$

Using Feller continuity, Proposition 3.6, we can now write

$$1 = \lim_{n\to\infty} \chi(\rho_{\xi_n}) \leq \lim_{n\to\infty} \mathbb{E}V(\varphi_s(\rho_{\xi_n}, 1)) = \mathbb{E}V(\varphi_s(\rho_\infty, 1)) = 1 - \epsilon < 1,$$

which is a contradiction. Hence there exists $\xi > 0$ such that $\chi(\rho) \leq 1 - \xi$ for all $\rho \in \mathcal{S}_{>1-\xi}$. The result follows by choosing $\gamma = \xi/2$. □

The following lemma is the main result of Step 2.

LEMMA 4.6. *Let* $\tau_\rho(\mathcal{S}_{>1-\gamma})$ *be the first exit time of* $\varphi_t(\rho, 1)$ *from* $\mathcal{S}_{>1-\gamma}$. *Then*

$$\sup_{\rho \in \mathcal{S}_{>1-\gamma}} \mathbb{E}\tau_\rho(\mathcal{S}_{>1-\gamma}) < \infty.$$

Proof. The following result can be found in Dynkin [13, Lemma 4.3, pp. 111]:

$$\mathbb{E}\tau_\rho(\mathcal{S}_{>1-\gamma}) \leq \frac{T}{1 - \sup_{\zeta \in \mathcal{S}} \mathbb{P}\{\tau_\zeta(\mathcal{S}_{>1-\gamma}) > T\}}.$$

We will show that

$$\sup_{\zeta\in\mathcal{S}} \mathbb{P}\{\tau_\zeta(\mathcal{S}_{>1-\gamma}) > T\} < 1. \tag{4.15}$$

This holds trivially for $\zeta \in \mathcal{S}_{\leq 1-\gamma}$, as then $\tau_\zeta(\mathcal{S}_{>1-\gamma}) = 0$. Let us thus suppose that

$$\forall \epsilon > 0 \quad \exists \zeta_\epsilon \in \mathcal{S}_{>1-\gamma} \quad \text{such that} \quad \mathbb{P}\{\tau_{\zeta_\epsilon}(\mathcal{S}_{>1-\gamma}) > T\} > 1-\epsilon.$$

Then for all $s \in [0,T]$, we have that

$$\mathbb{E}V(\varphi_s(\zeta_\epsilon, 1)) > (1-\epsilon) \inf_{\rho\in\mathcal{S}_{>1-\gamma}} V(\rho) = (1-\epsilon)(1-\gamma).$$

By compactness there exist a sequence $\epsilon_n \searrow 0$ and $\zeta_\infty \in \mathcal{S}_{\geq 1-\gamma}$ such that $\zeta_{\epsilon_n} \to \zeta_\infty$ as $n \to \infty$. Thus by Proposition 3.6

$$\mathbb{E}V(\varphi_s(\zeta_\infty, 1)) > 1-\gamma \quad \forall s \in [0,T].$$

But this is in contradiction with the result of Lemma 4.5. Hence there exists an $\epsilon > 0$ such that $\sup_{\zeta\in\mathcal{S}} \mathbb{P}\{\tau_\zeta(\mathcal{S}_{>1-\gamma}) > T\} = 1-\epsilon$, and we obtain

$$\mathbb{E}(\tau_\rho(\mathcal{S}_{>1-\gamma})) \leq \frac{T}{1-(1-\epsilon)} = \frac{T}{\epsilon} < \infty$$

uniformly in ρ. This completes the proof. □

Step 3. In this step we deal with the situation where the initial state lies inside the set $\mathcal{S}_{\leq 1-\gamma}$. We will denote by $u_1(\rho) = -\mathrm{Tr}\,(i[F_y,\rho]\rho_f)$ and by $\varphi_t(\rho,u_1)$ the solution of (4.2) with $\rho_0 = \rho$ and with $u_t = u_1(\rho_t)$. Denote by $\mathscr{A}$ the weak infinitesimal operator of $\varphi_t(\rho,u_1)$. We will apply the stochastic Lyapunov theorems with $Q_\lambda = \mathcal{S}$.

We begin by showing that there is a nonzero probability $p > 0$ that whenever the initial state lies inside $\mathcal{S}_{\leq 1-\gamma}$ the trajectories of the system never exit the set $\mathcal{S}_{<1-\gamma/2}$.

LEMMA 4.7. *For all* $\rho \in \mathcal{S}_{\leq 1-\gamma}$

$$\mathbb{P}\left[\sup_{0\leq t<\infty} V(\varphi_t(\rho,u_1)) \geq 1-\gamma/2\right] \leq 1-p = \frac{1-\gamma}{1-\gamma/2} < 1.$$

Proof. This follows from Theorem 2.2 and $\mathscr{A}V(\rho) = -u_1(\rho)^2 \leq 0$. □

We now restrict ourselves to the paths that never leave $\mathcal{S}_{<1-\gamma/2}$. We will first show that these paths converge toward ρ_f in probability. We then extend this result to prove almost sure convergence.

LEMMA 4.8. *The sample paths of* $\varphi_t(\rho,u_1)$ *that never exit the set* $\mathcal{S}_{<1-\gamma/2}$ *converge in probability to* ρ_f *as* $t \to \infty$.

Proof. Consider the Lyapunov function

$$\mathcal{V}(\rho) = 1 - \mathrm{Tr}\,(\rho\rho_f)^2.$$

It is easily verified that $\mathcal{V}(\rho) \geq 0$ for all $\rho \in \mathcal{S}$ and that $\mathcal{V}(\rho) = 0$ if and only if $\rho = \rho_f$. A straightforward computation gives

$$\mathscr{A}\mathcal{V}(\rho) = -2u_1(\rho)^2\,\mathrm{Tr}\,(\rho\rho_f) - 4\eta\,(\lambda_f - \mathrm{Tr}\,(\rho F_z))^2\,\mathrm{Tr}\,(\rho\rho_f)^2 \leq 0,$$

where λ_f is the eigenvalue of F_z associated to v_f. Now note that all the conditions of Theorem 2.3 are satisfied by virtue of Propositions 3.6 and 3.4. Hence $\varphi_t(\rho,u_1)$

converges in probability to the largest invariant set contained in $\mathcal{C} = \{\rho \in \mathcal{S} : \mathscr{A}\mathcal{V}(\rho) = 0\}$.

In order to satisfy the condition $\mathscr{A}\mathcal{V}(\rho) = 0$, we must have $u_1(\rho)^2 \operatorname{Tr}(\rho\rho_f) = 0$ as well as $(\lambda_f - \operatorname{Tr}(\rho F_z))^2 \operatorname{Tr}(\rho\rho_f)^2 = 0$. The latter implies that

$$\text{either} \quad \operatorname{Tr}(\rho\rho_f) = 0 \qquad \text{or} \quad \operatorname{Tr}(\rho F_z) = \lambda_f.$$

Let us investigate the largest invariant set contained in $\mathcal{C}' = \{\rho \in \mathcal{S} : \operatorname{Tr}(\rho F_z) = \lambda_f\}$. Clearly this invariant set can contain only $\rho \in \mathcal{C}'$ for which $\operatorname{Tr}(\varphi_t(\rho, u_1)F_z)$ is constant. Using Itô's rule we obtain

$$d\operatorname{Tr}(\rho_t F_z) = -iu_1(\rho_t)\operatorname{Tr}([F_y, \rho_t]F_z)\, dt + 2\sqrt{\eta}\,(\operatorname{Tr}(F_z^2\rho_t) - \operatorname{Tr}(F_z\rho_t)^2)\, dW_t.$$

Hence in order for $\operatorname{Tr}(\varphi_t(\rho, u_1)F_z)$ to be constant, we must at least have

$$\operatorname{Tr}(F_z^2\rho) - \operatorname{Tr}(F_z\rho)^2 = 0.$$

But as in the proof of Proposition 4.1, this implies that $\rho = \psi_m\psi_m^*$ for some m, and thus the only possibilities are $V(\rho) = 0$ (for $\rho = v_f v_f^*$) or $V(\rho) = 1$.

From the discussion above it is evident that the largest invariant set contained in $\mathcal{C}$ must be contained inside the set $\{\rho_f\} \cup \mathcal{S}_1$. But then the paths that never exit $\mathcal{S}_{<1-\gamma/2}$ must converge in probability to ρ_f. Thus the assertion is proved. □

LEMMA 4.9. *$\varphi_t(\rho, u_1)$ converges to ρ_f as $t \to \infty$ for almost all paths that never exit the set $\mathcal{S}_{<1-\gamma/2}$.*

Proof. Define the event $P^\rho_{<1-\gamma/2} = \{\omega \in \Omega : \varphi_t(\rho, u_1)$ never exits $\mathcal{S}_{<1-\gamma/2}\}$. Then Lemma 4.8 implies that

$$\lim_{t\to\infty} \mathbb{P}\left(\|\varphi_t(\rho, u_1) - \rho_f\| > \varepsilon \;\middle|\; P^\rho_{<1-\gamma/2}\right) = 0 \qquad \forall \varepsilon > 0.$$

By continuity of V, this also implies

$$\lim_{t\to\infty} \mathbb{P}\left(V(\varphi_t(\rho, u_1)) > \varepsilon \;\middle|\; P^\rho_{<1-\gamma/2}\right) = 0 \qquad \forall \varepsilon > 0.$$

As $V(\rho) \le 1$, we have

$$\begin{aligned}\mathbb{E}\left(V(\varphi_t(\rho, u_1)) \;\middle|\; P^\rho_{<1-\gamma/2}\right) &\le \mathbb{P}\left(V(\varphi_t(\rho, u_1)) > \varepsilon \;\middle|\; P^\rho_{<1-\gamma/2}\right) \\ &\quad + \varepsilon\left[1 - \mathbb{P}\left(V(\varphi_t(\rho, u_1)) > \varepsilon \;\middle|\; P^\rho_{<1-\gamma/2}\right)\right].\end{aligned}$$

Thus

$$\limsup_{t\to\infty} \mathbb{E}\left(V(\varphi_t(\rho, u_1)) \;\middle|\; P^\rho_{<1-\gamma/2}\right) \le \varepsilon \qquad \forall \varepsilon > 0,$$

which implies

$$\lim_{t\to\infty} \mathbb{E}\left(V(\varphi_t(\rho, u_1)) \;\middle|\; P^\rho_{<1-\gamma/2}\right) = 0.$$

But we know by Theorem 2.2 that $V(\varphi_t(\rho, u_1))$ converges a.s. As V is bounded, we obtain by dominated convergence

$$\mathbb{E}\left(\lim_{t\to\infty} V(\varphi_t(\rho, u_1)) \;\middle|\; P^\rho_{<1-\gamma/2}\right) = 0,$$

from which the result follows immediately. □

Step 4. It remains to combine the results of Steps 2 and 3 to prove existence, uniqueness, and global stability of the solution ρ_t. We will denote by u the control law of Theorem 4.2 and by $\varphi_t(\rho, u)$ the associated solution. Note that $\varphi_t(\rho, u)$ is not a Markov process, as the control u depends on the past history of the solution. We will construct $\varphi_t(\rho, u)$ by pasting together the strong Markov processes $\varphi_t(\rho, 1)$ and $\varphi_t(\rho, u_1)$ at the times where the control switches.

LEMMA 4.10. *There is a unique solution $\varphi_t(\rho, u)$ for all $t \in \mathbb{R}_+$. Moreover, for almost every sample path of $\varphi_t(\rho, u)$ there exists a time $T < \infty$ after which the path never exits the set $\mathcal{S}_{<1-\gamma/2}$ and the active control law is u_1.*

Proof. Fix the initial state ρ. We begin by constructing a solution $\varphi_{t\wedge n}(\rho, u)$ up to (at most) an integer time $n \in \mathbb{N}$. To this end, define the predictable stopping time

$$\tau_1^n = \inf\{t \geq 0 : \varphi_t(\rho, 1) \in \mathcal{S}_{\leq 1-\gamma}\} \wedge n.$$

Then we can define $\rho_{\tau_1^n} = \varphi_{\tau_1^n}(\rho, 1)$ and $\varphi_{t\wedge n}(\rho, u) = \varphi_t(\rho, 1)$ for $t < \tau_1^n$. In the following, we will need the two-parameter solution $\varphi_{s,t}(\rho, u')$ of the filtering equation under the simple control u', given the initial state ρ at time s. Define

$$\sigma_1^n = \inf\{t \geq \tau_1^n : \varphi_{\tau_1^n,t}(\rho_{\tau_1^n}, u_1) \in \mathcal{S}_{\geq 1-\gamma/2}\} \wedge n.$$

We can extend our solution by

$$\varphi_{t\wedge n}(\rho, u) = \chi_{t<\tau_1^n}\varphi_t(\rho, 1) + \chi_{\tau_1^n \leq t<\sigma_1^n}\varphi_{\tau_1^n,t}(\rho_{\tau_1^n}, u_1), \qquad t < \sigma_1^n,$$

where χ_A is the indicator function on the set A. To extend the solution further, we continue again with the control law $u = 1$. Recursively, we define an entire sequence of predictable stopping times

$$\sigma_k^n = \inf\{t \geq \tau_k^n : \varphi_{\tau_k^n,t}(\rho_{\tau_k^n}, u_1) \in \mathcal{S}_{\geq 1-\gamma/2}\} \wedge n,$$

$$\tau_k^n = \inf\{t \geq \sigma_{k-1}^n : \varphi_{\sigma_{k-1}^n,t}(\rho_{\sigma_{k-1}^n}, 1) \in \mathcal{S}_{\leq 1-\gamma}\} \wedge n,$$

where

$$\rho_{\sigma_k^n} = \varphi_{\tau_k^n,\sigma_k^n}(\rho_{\tau_k^n}, u_1), \qquad \rho_{\tau_k^n} = \varphi_{\sigma_{k-1}^n,\tau_k^n}(\rho_{\sigma_{k-1}^n}, 1).$$

We can use these times to construct the solution

$$\varphi_{t\wedge n}(\rho, u) = \chi_{t<\tau_1^n}\varphi_t(\rho, 1) + \sum_{k=1}^{\infty}\Big[\chi_{\tau_k^n \leq t<\sigma_k^n}\varphi_{\tau_k^n,t}(\rho_{\tau_k^n}, u_1) + \chi_{\sigma_k^n \leq t<\tau_{k+1}^n}\varphi_{\sigma_k^n,t}(\rho_{\sigma_k^n}, 1)\Big]$$

for all times $t < \Sigma^n = \lim_{k\to\infty}\sigma_k^n \leq n$ (the limit exists, as σ_k is a nondecreasing sequence of stopping times.) Moreover, the solution is a.s. unique, as the segments between each two stopping times are a.s. uniquely defined.

Now note that as anticipated by the notation, it is not difficult to verify that $\varphi_{t\wedge(n+1)}(\rho, u) = \varphi_{t\wedge n}(\rho, u)$ a.s. for $t < \Sigma^n$, and, moreover, $\Sigma^n = \Sigma \wedge n$, $\tau_k^n = \tau_k \wedge n$, $\sigma_k^n = \sigma_k \wedge n$, where $\Sigma = \lim_{t\to\infty}\Sigma^n$, etc. Hence we can let $n \to \infty$ to obtain the unique solution $\varphi_t(\rho, u)$ defined up to the accumulation time Σ, where τ_k, σ_k are the consecutive times at which the control switches. It remains to prove that the solution exists for all time, i.e., that $\Sigma = \infty$ a.s. In particular, this uniquely defines a càdlàg

control u_t, so that by uniqueness $\varphi_t(\rho,u)$ must coincide with the solution of (3.2) with the control u_t. Below we will prove that a.s., only finitely many σ_k are finite. This is sufficient to prove not only existence but also the second statement of the lemma.

To proceed, we use the fact that the strong Markov property holds on each segment between consecutive switching times $\tau_n \le t < \sigma_n$ and $\sigma_n \le t < \tau_{n+1}$. Thus

$$\begin{aligned}\mathbb{P}(\sigma_n < \infty \text{ and } \tau_n < \infty)\\ = \int \chi_{\tau_n<\infty}(\tilde\omega)\,\mathbb{P}(\varphi_t(\rho_{\tau_n}(\tilde\omega),u_1) \text{ exits } \mathcal{S}_{<1-\gamma/2} \text{ in finite time})\,\mathbb{P}(d\tilde\omega),\end{aligned}$$

which implies

$$\begin{aligned}\mathbb{P}(\sigma_n < \infty \,|\, \tau_n < \infty)\\ = \int \mathbb{P}(\varphi_t(\rho_{\tau_n}(\tilde\omega),u_1) \text{ exits } \mathcal{S}_{<1-\gamma/2} \text{ in finite time})\,\mathbb{P}(d\tilde\omega \,|\, \tau_n < \infty).\end{aligned}$$

But $\rho_{\tau_n} \in \mathcal{S}_{\le 1-\gamma}$ on a set Ω_{τ_n} with $\mathbb{P}(\Omega_{\tau_n} \,|\, \tau_n < \infty) = 1$. Hence by Lemma 4.7

$$\mathbb{P}(\sigma_n < \infty \,|\, \tau_n < \infty) \le 1 - p.$$

Through a similar argument, and using Lemma 4.6, we obtain

$$\mathbb{P}(\tau_n < \infty \,|\, \sigma_{n-1} < \infty) = 1.$$

But note that by construction

$$\mathbb{P}(\tau_n < \infty \,|\, \sigma_n < \infty) = \mathbb{P}(\sigma_{n-1} < \infty \,|\, \tau_n < \infty) = 1.$$

Hence we obtain

$$\begin{aligned}\frac{\mathbb{P}(\sigma_n < \infty)}{\mathbb{P}(\sigma_{n-1} < \infty)} &= \frac{\mathbb{P}(\tau_n < \infty \,|\, \sigma_n < \infty)\mathbb{P}(\sigma_n < \infty)}{\mathbb{P}(\tau_n < \infty)}\,\frac{\mathbb{P}(\sigma_{n-1} < \infty \,|\, \tau_n < \infty)\mathbb{P}(\tau_n < \infty)}{\mathbb{P}(\sigma_{n-1} < \infty)}\\ &= \mathbb{P}(\sigma_n < \infty \,|\, \tau_n < \infty)\,\mathbb{P}(\tau_n < \infty \,|\, \sigma_{n-1} < \infty) \le 1 - p.\end{aligned}$$

But $\mathbb{P}(\sigma_1 < \infty) = \mathbb{P}(\sigma_1 < \infty \,|\, \tau_1 < \infty) \le 1 - p$ as $\tau_1 < \infty$ a.s. Hence

$$\mathbb{P}(\sigma_n < \infty) \le (1-p)^n,$$

and thus

$$\sum_{n=1}^{\infty} \mathbb{P}(\sigma_n < \infty) \le \sum_{n=1}^{\infty} (1-p)^n = \frac{1-p}{p} < \infty.$$

By the Borel–Cantelli lemma, we conclude that

$$\mathbb{P}(\sigma_n < \infty \text{ for infinitely many } n) = 0.$$

Hence $\Sigma = \infty$ a.s., and for almost every sample path, there exists an integer $N < \infty$ such that $\sigma_n = \infty$ (and hence also $\tau_{n+1} = \infty$) for all $n \ge N$, and such that $\sigma_n < \infty$ (and hence also $\tau_{n+1} < \infty$) for all $n < N$, which implies the assertion. □

Finally, we can now put together all the ingredients and complete the proof of Theorem 4.2.

Proof of Theorem 4.2. We must check three things: that the target state ρ_f is (locally) stable in probability; that almost all sample paths are attracted to the target state as $t \to \infty$; and that this is also true in expectation. Existence and uniqueness of the solution follows from Lemma 4.10.

(i) To study local stability, we can restrict ourselves to the stopped process

$$\varphi_{t\wedge\tilde{\tau}}(\rho,u) = \varphi_{t\wedge\tilde{\tau}}(\rho,u_1), \quad \tilde{\tau} = \inf\{t : \varphi_t(\rho,u) \notin \mathcal{S}_{<1-\gamma/2}\}.$$

Denote by $\tilde{\mathscr{A}}$ the weak infinitesimal operator of $\varphi_{t\wedge\tilde{\tau}}(\rho,u_1)$, and note that Proposition 3.8 allows us to calculate $\tilde{\mathscr{A}}V$ from (4.2) in the usual way. In particular, we find $\tilde{\mathscr{A}}V(\rho) = -u_1(\rho)^2 \le 0$ for $\rho \in \mathcal{S}_{<1-\gamma/2}$. Hence we can apply Theorem 2.2 with $Q_\lambda = \mathcal{S}_{<1-\gamma/2}$ to conclude stability in probability.

(ii) From Lemmas 4.9 and 4.10, it follows that $\varphi_t(\rho,u) \to \rho_f$ a.s. as $t \to \infty$.

(iii) We have shown that

$$\mathbb{E}\left[\lim_{t\to\infty} V(\varphi_t(\rho,u))\right] = V(\rho_f) = 0.$$

But as V is uniformly bounded, we obtain by dominated convergence

$$V\left(\lim_{t\to\infty} \mathbb{E}\varphi_t(\rho,u)\right) = \lim_{t\to\infty} \mathbb{E}\left[V(\varphi_t(\rho,u))\right] = 0,$$

where we have used that V is linear and continuous. Hence $\mathbb{E}\varphi_t(\rho,u) \to \rho_f$. ∎

5. Two-qubit systems. The methods employed in the previous section can be extended to other quantum feedback control problems. As an example, we treat the case of two qubits in a symmetric dispersive interaction with an optical probe field. Qubits, i.e., two-level quantum systems (having a Hilbert space of dimension two), and in particular correlated (entangled) states of multiple such qubits, play an important role in quantum information processing. Here we investigate the stabilization of two such states in the two-qubit system.

We begin by defining the Pauli matrices

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

and we define the basis $\psi_\uparrow = (1\ 0)^*$ and $\psi_\downarrow = (0\ 1)^*$ in $\mathbb{C}^2$. A system of two qubits lives on the four-dimensional space $\mathbb{C}^2 \otimes \mathbb{C}^2$ with the standard basis $\{\psi_{\uparrow\uparrow} = \psi_\uparrow \otimes \psi_\uparrow$, $\psi_{\uparrow\downarrow} = \psi_\uparrow \otimes \psi_\downarrow$, $\psi_{\downarrow\uparrow} = \psi_\downarrow \otimes \psi_\uparrow$, $\psi_{\downarrow\downarrow} = \psi_\downarrow \otimes \psi_\downarrow\}$. We denote by $\sigma^1_{x,y,z} = \sigma_{x,y,z} \otimes \mathbb{1}$ and $\sigma^2_{x,y,z} = \mathbb{1} \otimes \sigma_{x,y,z}$ the Pauli matrices on the first and second qubit, respectively, and by $F_{x,y,z} = \sigma^1_{x,y,z} + \sigma^2_{x,y,z}$ the (unnormalized) collective angular momentum operators.

The quantum filtering equation for the two-qubit system is given by an equation of the form (3.2):

$$\begin{aligned} d\rho_t = {} & -iu_1(t)[\sigma^1_y, \rho_t]\,dt - iu_2(t)[\sigma^2_y, \rho_t]\,dt \\ & - \tfrac{1}{2}[F_z,[F_z,\rho_t]]\,dt + \sqrt{\eta}\,(F_z\rho_t + \rho_t F_z - 2\,\mathrm{Tr}\,(F_z\rho_t)\,\rho_t)\,dW_t, \end{aligned} \tag{5.1}$$

where u_1 and u_2 are two independent controls acting as local magnetic fields in the y-direction on each of the qubits. The main goal of this section is to stabilize this system around two interesting target states,

$$\rho_s = \frac{1}{2}(\psi_{\uparrow\downarrow} + \psi_{\downarrow\uparrow})(\psi_{\uparrow\downarrow} + \psi_{\downarrow\uparrow})^*, \qquad \rho_a = \frac{1}{2}(\psi_{\uparrow\downarrow} - \psi_{\downarrow\uparrow})(\psi_{\uparrow\downarrow} - \psi_{\downarrow\uparrow})^*.$$

Here ρ_s is a symmetric and ρ_a is an antisymmetric qubit state.

THEOREM 5.1. *Consider the following control law:*

1. $u_1(t) = 1 - \mathrm{Tr}\left(i[\sigma_y^1, \rho_t]\rho_a\right)$, $u_2(t) = 1 - \mathrm{Tr}\left(i[\sigma_y^2, \rho_t]\rho_a\right)$ *if* $\mathrm{Tr}\,(\rho\rho_a) \geq \gamma$;
2. $u_1(t) = 1$, $u_2(t) = 0$ *if* $\mathrm{Tr}\,(\rho\rho_a) \leq \gamma/2$;
3. *if* $\rho_t \in \mathcal{B}_a = \{\rho : \gamma/2 < \mathrm{Tr}\,(\rho\rho_a) < \gamma\}$, *then take* $u_1(t) = 1 - \mathrm{Tr}\left(i[\sigma_y^1, \rho_t]\rho_a\right)$, $u_2(t) = 1 - \mathrm{Tr}\left(i[\sigma_y^2, \rho_t]\rho_a\right)$ *if* ρ_t *last entered the set* $\mathcal{B}_a$ *through the boundary* $\mathrm{Tr}\,(\rho\rho_a) = \gamma$, *and* $u_1(t) = 1$, $u_2(t) = 0$ *otherwise.*

Then there exists $\gamma > 0$ *s.t.* (5.1) *is globally stable around* ρ_a *and* $\mathbb{E}\rho_t \to \rho_a$ *as* $t \to \infty$. *Similarly, consider the following control law:*

1. $u_1(t) = 1 - \mathrm{Tr}\left(i[\sigma_y^1, \rho_t]\rho_s\right)$, $u_2(t) = -1 - \mathrm{Tr}\left(i[\sigma_y^2, \rho_t]\rho_s\right)$ *if* $\mathrm{Tr}\,(\rho\rho_s) \geq \gamma$;
2. $u_1(t) = 1$, $u_2(t) = 0$ *if* $\mathrm{Tr}\,(\rho\rho_s) \leq \gamma/2$;
3. *if* $\rho_t \in \mathcal{B}_s = \{\rho : \gamma/2 < \mathrm{Tr}\,(\rho\rho_s) < \gamma\}$, *then take* $u_1(t) = 1 - \mathrm{Tr}\left(i[\sigma_y^1, \rho_t]\rho_s\right)$, $u_2(t) = -1 - \mathrm{Tr}\left(i[\sigma_y^2, \rho_t]\rho_s\right)$ *if* ρ_t *last entered the set* $\mathcal{B}_s$ *through the boundary* $\mathrm{Tr}\,(\rho\rho_s) = \gamma$, *and* $u_1(t) = 1$, $u_2(t) = 0$ *otherwise.*

This stabilizes the system around the symmetric state ρ_s.

We will prove the result for the antisymmetric case; the proof for the symmetric case may be done exactly in the same manner. We proceed in the same way as in the proof of Theorem 4.2.

Step 1. The proof of Lemma 4.3 carries over directly to the two-qubit case. The proof of Lemma 4.4 also carries over after minor modifications; in particular, in the two-qubit case we can explicitly compute that

$$A = -i\sigma_y^1 - F_z^2 + 2F_z = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & -8 \end{pmatrix}$$

admits the diagonalization $A = PDP^{-1}$ with

$$P = \begin{pmatrix} 1 & 1 & 0 & 0 \\ -i & i & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & .1270 & 7.8730 \end{pmatrix}, \qquad D = \begin{pmatrix} i & 0 & 0 & 0 \\ 0 & -i & 0 & 0 \\ 0 & 0 & -.1270 & 0 \\ 0 & 0 & 0 & -7.8730 \end{pmatrix}.$$

Hence the matrix A has a nondegenerate spectrum, and, moreover,

$$\tilde{v}_a = \tfrac{1}{\sqrt{2}}\, P^*(\psi_{\uparrow\downarrow} - \psi_{\downarrow\uparrow}) = \tfrac{1}{\sqrt{2}}\,(i \;\; -i \;\; -1 \;\; -1)^*$$

has only nonzero entries. The remainder of the proof is identical to that of Lemma 4.3.

Step 2. The proofs of Lemmas 4.5 and 4.6 carry over directly.

Step 3. The proofs of Lemmas 4.7 and 4.9 carry over directly. The following replaces Lemma 4.8. We denote by $U_1(\rho) = 1 - \mathrm{Tr}\left(i[\sigma_y^1, \rho]\rho_a\right)$, by $U_2(\rho) = 1 - \mathrm{Tr}\left(i[\sigma_y^2, \rho]\rho_a\right)$, and by $\varphi_t(\rho, U_1, U_2)$ the associated solution of (5.1).

LEMMA 5.2. *The sample paths of* $\varphi_t(\rho, U_1, U_2)$ *that never exit the set* $\mathcal{S}_{<1-\gamma/2}$ *converge in probability to* ρ_a *as* $t \to \infty$.

Proof. Consider the Lyapunov function

$$\mathcal{V}(\rho) = 1 - \mathrm{Tr}\,(\rho\rho_a)^2\,.$$

It is easily verified that $\mathcal{V}(\rho) \geq 0$ for all $\rho \in \mathcal{S}$ and that $\mathcal{V}(\rho) = 0$ if and only if $\rho = \rho_a$. A straightforward computation gives

$$\mathscr{A}\mathcal{V}(\rho) = -2\left[(U_1(\rho) - 1)^2 + (U_2(\rho) - 1)^2\right]\mathrm{Tr}\,(\rho\rho_a) - 4\eta\,\mathrm{Tr}\,(\rho F_z)^2\;\mathrm{Tr}\,(\rho\rho_a)^2 \leq 0,$$

where $\mathscr{A}$ is the weak infinitesimal operator associated to $\varphi_t(\rho, U_1, U_2)$ (here we have used $[F_y, \rho_a] = 0$ in calculating this expression). Now note that all the conditions of Theorem 2.3 are satisfied by virtue of Propositions 3.6 and 3.4. Hence $\varphi_t(\rho, U_1, U_2)$ converges in probability to the largest invariant set contained in $\mathcal{C} = \{\rho \in \mathcal{S} : \mathscr{A}\mathcal{V}(\rho) = 0\}$.

In order to satisfy the condition $\mathscr{A}\mathcal{V}(\rho) = 0$ we must have at least

$$\text{either} \quad \operatorname{Tr}(\rho\rho_a) = 0 \qquad \text{or} \quad \operatorname{Tr}(\rho F_z) = 0.$$

Let us investigate the largest invariant set contained in $\mathcal{C}' = \{\rho \in \mathcal{S} : \operatorname{Tr}(\rho F_z) = 0\}$. Clearly this invariant set can contain only $\rho \in \mathcal{C}'$ for which $\operatorname{Tr}(\varphi_t(\rho, U_1, U_2)F_z)$ is constant. Using Itô's rule we obtain

$$d\operatorname{Tr}(\rho_t F_z) = -\sum_{j=1}^{2} U_j(\rho_t) \operatorname{Tr}\left(i[\sigma_y^j, \rho_t]F_z\right) dt + 2\sqrt{\eta}\left(\operatorname{Tr}\left(F_z^2\rho_t\right) - \operatorname{Tr}(F_z\rho_t)^2\right) dW_t.$$

Hence in order for $\operatorname{Tr}(\varphi_t(\rho, U_1, U_2)F_z)$ to be constant, we must at least have

$$\operatorname{Tr}\left(F_z^2\rho\right) - \operatorname{Tr}(F_z\rho)^2 = 0,$$

which implies that ρ must be an eigenstate of F_z. The latter can take only one of the following forms: either $\rho = \psi_{\uparrow\uparrow}\psi_{\uparrow\uparrow}^*$ or $\rho = \psi_{\downarrow\downarrow}\psi_{\downarrow\downarrow}^*$, or ρ is any state of the form

$$\rho = \alpha\psi_{\uparrow\downarrow}\psi_{\uparrow\downarrow}^* + \beta\psi_{\uparrow\downarrow}\psi_{\downarrow\uparrow}^* + \beta^*\psi_{\downarrow\uparrow}\psi_{\uparrow\downarrow}^* + (1-\alpha)\psi_{\downarrow\uparrow}\psi_{\downarrow\uparrow}^*. \tag{5.2}$$

Let us investigate in particular the latter case. Note that any density matrix of the form (5.2) satisfies $F_z\rho = \rho F_z = 0$. Suppose that (5.1) with $u_1 = U_1$, $u_2 = U_2$ leaves the set (5.2) invariant; then the solution at time t of

$$\frac{d}{dt}\rho_t = -i[F_y, \rho_t] \tag{5.3}$$

must coincide with $\varphi_t(\rho, U_1, U_2)$ when ρ is of the form (5.2), and in particular (5.3) must leave the set (5.2) invariant (here we have used that $U_1(\rho) = U_2(\rho) = 1$ for ρ of the form (5.2)). We claim that this is only the case if $\rho = \rho_a$, which implies that of all states of the form (5.2) only ρ_a is in fact invariant. To see this, note that by Lemma 3.1 we can write any ρ of the form (5.2) as a convex combination $\sum_i \lambda_i \psi^i\psi^{i*}$ of unit vectors $\psi^i \in \operatorname{span}\{\psi_{\uparrow\downarrow}, \psi_{\downarrow\uparrow}\}$. Thus the solution of (5.3) at time t is given by $\sum_i \lambda_i \psi_t^i\psi_t^{i*}$ with

$$\frac{d}{dt}\psi_t^i = -iF_y\psi_t^i, \qquad \psi_0^i = \psi^i.$$

But $F_y\psi^i \notin \operatorname{span}\{\psi_{\uparrow\downarrow}, \psi_{\downarrow\uparrow}\}$ unless $\psi^i \propto \psi_{\uparrow\downarrow} - \psi_{\downarrow\uparrow}$, which implies the assertion.

From the discussion above it is evident that the largest invariant set contained in $\mathcal{C}$ must be contained inside the set $\{\rho_a\} \cup \mathcal{S}_1$. But then the paths that never exit $\mathcal{S}_{<1-\gamma/2}$ must converge in probability to ρ_a. Thus the lemma is proved. □

Step 4. The remainder of the proof of Theorem 5.1 carries over directly.

Acknowledgments. The authors thank Hideo Mabuchi and Houman Owhadi for helpful discussions.

REFERENCES

[1] L. Arnold, *Stochastic Differential Equations: Theory and Applications*, Wiley, New York, London, Sydney, 1974.
[2] A. Barchielli and G. Lupieri, *Quantum stochastic calculus, operation valued stochastic processes, and continual measurements in quantum mechanics*, J. Math. Phys., 26 (1985), pp. 2222–2230.
[3] A. Barchielli and A. M. Paganoni, *Stochastic differential equations for trace-class operators and quantum continual measurements*, in Stochastic Partial Differential Equations and Applications (Trento, 2002), Lecture Notes Pure Appl. Math. 227, Dekker, New York, 2002, pp. 53–67.
[4] V. P. Belavkin, *Theory of the control of observable quantum systems*, Autom. Remote Control, 44 (1983), pp. 178–188.
[5] V. P. Belavkin, *Nondemolition stochastic calculus in Fock space and nonlinear filtering and control in quantum systems*, in Stochastic Methods in Mathematics and Physics (Karcapz, 1988), R. Guelerak and W. Karwowski, eds., World Scientific, Teaneck, NJ, 1989, pp. 310–324.
[6] V. P. Belavkin, *Quantum stochastic calculus and quantum nonlinear filtering*, J. Multivariate Anal., 42 (1992), pp. 171–201.
[7] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.
[8] L. Bouten and R. van Handel, *On the Separation Principle of Quantum Control*, preprint, 2005; available online at http://arxiv.org/math-ph/0511021.
[9] L. Bouten, R. van Handel, and M. R. James, *An Introduction to Quantum Filtering*, preprint, 2005; available online at http://arxiv.org/math.OC/0601741.
[10] H. J. Carmichael, *An Open Systems Approach to Quantum Optics*, Springer, Berlin, 1993.
[11] E. B. Davies, *Quantum Theory of Open Systems*, Academic Press, London, San Francisco, New York, 1976.
[12] A. C. Doherty, S. Habib, K. Jacobs, H. Mabuchi, and S. M. Tan, *Quantum feedback control and classical control theory*, Phys. Rev. A(3), 62 (2000), p. 012105.
[13] E. B. Dynkin, *Markov Processes*, Vol. I, Springer, Berlin, 1965.
[14] S. C. Edwards and V. P. Belavkin, *Optimal Quantum Filtering and Quantum Feedback Control*, preprint, 2005; available online at http://arxiv.org/quant-ph/0506018.
[15] J. M. Geremia, J. K. Stockton, and H. Mabuchi, *Real-time quantum feedback control of atomic spin-squeezing*, Science, 304 (2004), pp. 270–273.
[16] I. I. Gikhman and A. V. Skorokhod, *Introduction to the Theory of Random Processes*, Dover, Mineola, NY, 1996.
[17] R. Van Handel and H. Mabuchi, *Quantum projection filter for a highly nonlinear model in cavity QED*, J. Opt. B Quantum Semiclass. Opt., 7 (2005), pp. S226–S236.
[18] R. Van Handel, J. K. Stockton, and H. Mabuchi, *Feedback control of quantum state reduction*, IEEE Trans. Automat. Control, 50 (2005), pp. 768–780.
[19] R. Van Handel, J. K. Stockton, and H. Mabuchi, *Modelling and feedback control design for quantum state preparation*, J. Opt. B Quantum Semiclass. Opt., 7 (2005), pp. S179–S197.
[20] R. Z. Has'minskiĭ, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Alphen aan den Rijn, Germantown, MD, 1980.
[21] R. L. Hudson and K. R. Parthasarathy, *Quantum Itô's formula and stochastic evolutions*, Comm. Math. Phys., 93 (1984), pp. 301–323.
[22] H. Kunita, *Supports of diffusion processes and controllability problems*, in Proceedings of the International Symposium on Stochastic Differential Equations, Kyoto, 1976, Wiley, New York, 1978, pp. 163–185.
[23] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, Cambridge, UK, 1990.
[24] H. J. Kushner, *Stochastic Stability and Control*, Academic Press, New York, London, 1967.
[25] H. J. Kushner, *The concept of invariant set for stochastic dynamical systems and applications to stochastic stability*, in Stochastic Optimization and Control, H. F. Karreman, ed., Wiley, New York, 1968, pp. 47–57.
[26] H. J. Kushner, *Stochastic stability*, in Stability of Stochastic Dynamical Systems, Lecture Notes in Math. 294, R. F. Curtain, ed., Springer, Berlin, 1972, pp. 97–123.
[27] H. Maassen, *Quantum probability applied to the damped harmonic oscillator*, in Quantum Probability Communications XII, S. Attal and J. M. Lindsay, eds., World Scientific, River Edge, NJ, 2003, pp. 23–58.
[28] E. Merzbacher, *Quantum Mechanics*, 3rd ed., Wiley, New York, 1998.
[29] M. Mirrahimi, R. Van Handel, A. E. Miller, and H. Mabuchi, 2005, in preparation.
[30] M. Mirrahimi, P. Rouchon, and G. Turinici, *Lyapunov control of bilinear Schrödinger*

equations, Automatica J. IFAC, 41 (2005), pp. 1987–1994.
[31] B. Øksendal, *Stochastic Differential Equations*, 5th ed., Springer, Berlin, 1998.
[32] P. E. Protter, *Stochastic Integration and Differential Equations*, 2nd ed., Springer, Berlin, 2004.
[33] J. K. Stockton, R. Van Handel, and H. Mabuchi, *Deterministic Dicke-state preparation with continuous measurement and control*, Phys. Rev. A(3), 70 (2004), p. 022106.
[34] D. W. Stroock and S. R. Varadhan, *On the support of diffusion processes with applications to the strong maximum principle*, in Proceedings of the Sixth Annual Berkeley Symposium on Mathematical Statistics and Probability, Vol. III, University of California Press, Berkeley, 1972, pp. 333–359.
[35] H. M. Wiseman, *Quantum theory of continuous feedback*, Phys. Rev. A(3), 49 (1994), p. 2133.
[36] H. M. Wiseman, S. Mancini, and J. Wang, *Bayesian feedback versus Markovian feedback in a two-level atom*, Phys. Rev. A(3), 66 (2002), p. 013807.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 468–495

ON THE NEW BARRIER FUNCTION AND SPECIALIZED ALGORITHMS FOR A CLASS OF SEMIDEFINITE PROGRAMS*

CHUNG-YAO KAO† AND ALEXANDRE MEGRETSKI‡

Abstract. Semidefinite programs (SDPs) arising from the Kalman–Yakubovich–Popov (KYP) lemma are frequently encountered in systems robustness analysis, filter design, and other control/signal processing related applications. These programs possess a special structure that can be exploited to construct specialized algorithms that substantially outperform general-purpose SDP solvers. In this paper, a new interior path-following algorithm that utilizes this structure is proposed. The main idea behind the algorithm is a new barrier function for these specially structured SDPs. Convergence of the new algorithm is shown and a measure of the accuracy of suboptimal solutions produced by the algorithm is provided. The algorithm is tested in numerical experiments and the results indicate that the new algorithm is indeed favorable against general-purpose SDP solvers in many circumstances.

Key words. interior point method, barrier function, semidefinite program, KYP lemma, robustness analysis

AMS subject classification. 93D09

DOI. 10.1137/050623796

1. Introduction.

Consider a semidefinite problem (SDP) of the form

$$\begin{aligned} &\inf_{\lambda,P} \ c'\lambda \quad \text{subject to} \\ &\mathcal{F}(P,\lambda) := \begin{bmatrix} PA + A'P & PB \\ B'P & 0 \end{bmatrix} + M_0 + \sum_{i=1}^{n} \lambda_i M_i > 0, \end{aligned} \tag{1.1}$$

where the variables to be optimized are the vector $\lambda \in \mathbf{R}^n$ and the matrix P. The real matrices A and B are of dimensions $m \times m$ and $m \times p$, respectively. The matrix A is assumed to be Hurwitz; i.e., the real part of each eigenvalue of A is strictly negative. The matrices P and M_i, $i = 0, 1, \ldots, n$ are symmetric. The inequality sign in (1.1) denotes matrix inequality; i.e., $X > 0$ means $v'Xv > 0$ for all nonzero v.

The SDP (1.1) is referred to as the KYP–SDP for its tight connection to the Kalman–Yakubovich–Popov lemma [21, 33]. The KYP–SDP frequently appears in control and signal processing related applications. A partial list includes systems robustness analysis [5, 18, 23, 31, 32], filter design [1], and linear control system synthesis [3, 9]. We note that in (1.1), the number of decision variables in P is proportional to the square of the dimension of the matrix A, while the matrices A and B usually correspond to the state space realization of the linear time-invariant (LTI) part of the system to be analyzed or designed. Hence, in practical applications where the LTI systems under consideration have state spaces of large dimension,

*Received by the editors February 3, 2005; accepted for publication (in revised form) October 31, 2006; published electronically April 27, 2007. This research was supported in parts by the NSF (U.S.), the AFOSR (U.S.), the Mittag-Leffler Institute (Sweden), the Göran Gustafson Foundation (Sweden), and the Australian Research Council (Australia).

http://www.siam.org/journals/sicon/46-2/62379.html

†Department of Electrical and Electronic Engineering, University of Melbourne, Parkville 3010, VIC Australia (cykao@ee.unimelb.edu.au).

‡Laboratory of Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 (ameg@mit.edu).

the number of decision variables in P is often order-of-magnitude larger than the number of decision variables in λ. In these cases, the matrix variable P becomes the major computational burden in solving SDP (1.1). It is often observed that when the dimension of A is of several hundred, SDP (1.1) becomes difficult or even impossible to solve using a general-purpose SDP solver. It is not unusual in control system applications to encounter complex dynamical systems having many states. Typical examples of such systems are aircraft autopilots, electrical power grids, large-scale networked control systems, and vibration controllers for flexible structures. Hence, it is of practical importance to develop more efficient and specialized solvers to solve SDPs in the form of (1.1).

To improve the efficiency of the conventional SDP approach, some specialized algorithms have been proposed recently [7, 8, 25, 28, 29, 30]. These algorithms are based on the primal-dual interior point algorithms for solving standard SDPs. Efficiency is achieved by exploring and exploiting the special structure of (1.1). In [25, 29], it is reported that such specialized algorithms are able to solve KYP–SDPs with tens of thousands of decision variables at a reasonable speed, while general-purpose SDP solvers either fail to solve the problem due to lack of sufficient computer memory or become extremely slow, occupying most of the computational capacity of the computer for days.

On the other hand, since computing the matrix variable P constitutes the main computational burden when solving a KYP–SDP, another approach for improving computational efficiency is to avoid this matrix variable. According to the KYP lemma, the SDP (1.1) can be equivalently formulated as the following SDP:

$$\begin{aligned} &\inf_{\lambda} \ c'\lambda \quad \text{subject to} \\ &\begin{bmatrix} (j\omega I - A)^{-1} \\ B \end{bmatrix}^* \left(M_0 + \sum_{i=1}^{n} \lambda_i M_i \right) \begin{bmatrix} (j\omega I - A)^{-1} \\ B \end{bmatrix} > 0 \quad \forall\, \omega \in [0, \infty], \end{aligned} \tag{1.2}$$

where the feasible set is defined by an infinite number of linear matrix inequalities. Furthermore, the KYP lemma also gives rise to a computationally efficient algorithm for checking the feasibility of a given λ. Based on these observations, several cutting plane algorithms were proposed to solve KYP–SDP (1.1) [14, 15, 20]. These algorithms appear to work well, especially when the size of A is large and the number of decision variables in λ is relatively small. However, a common disadvantage of cutting plane and similar methods is that they generally require many iterations to converge to a suboptimal solution of good accuracy, and the number of iterations grows rapidly as the number of decision variables increases. Furthermore, it is commonly known that when attempting to find a very accurate suboptimal solution using such algorithms one often encounters numerical difficulty.

In light of the disadvantage of the cutting plane algorithm, we propose a new interior point algorithm for solving KYP–SDP (1.1). The main idea behind the algorithm is a new barrier function defined on the feasible set of the equivalent SDP (1.2). The barrier function is efficiently computable: the main computation for obtaining the barrier function's first and second derivatives is to solve Lyapunov equations, for which efficient computational routines are widely available. Based on the new barrier function, a path-following algorithm for solving KYP–SDP (1.2) is developed. Results of several numerical experiments indicate that the proposed path-following algorithm is able to solve the KYP–SDP in a much more efficient fashion. Regarding the issue of computational complexity, it can be shown that the proposed barrier

function is self-concordant. However, the self-concordant coefficient appears to be dependent on the problem data, and the exact dependency is yet to be discovered. Hence, Nesterov and Nemirovskii's theorem regarding polynomial-time complexity of the path-following methods [19] is not applicable to the algorithm proposed in this paper, and its worst-case complexity is still to be determined.

The main contribution of this paper is twofold. On the technical side, while the path-following algorithm described in this paper is rather standard and well known, the barrier function on which the algorithm is based has not been considered previously for KYP–SDP (1.1). We show how to evaluate the function and calculate the gradient and the Hessian at any given point in an efficient fashion. On the practical side, the proposed algorithm is implemented, tested in a number of numerical experiments, and compared against SeDuMi [24], a popular general-purpose SDP solver. The results show that the proposed algorithm is indeed favorable in many circumstances.

The paper is organized as follows. In the rest of this section, notations and terminology used throughout the paper are defined. In section 2, the semi-infinite optimization problem equivalent to KYP–SDP (1.1) is formally defined and its dual problem is stated. Section 3 contains the main technical results of the paper, where the new barrier function is described. Furthermore, how to efficiently evaluate the function and compute its gradient and Hessian are shown. In sections 4 and 5 the central path of the semi-infinite optimization problem and the path-following algorithm based on the new barrier function are discussed. Convergence of the algorithm is shown, and a measure of the accuracy of the suboptimal solution produced by the algorithm is provided. In section 6, we explain why the proposed algorithm is more efficient than the general-purpose SDP solvers from the point of view of computational complexity. Numerical experiments are conducted for testing the efficiency of the proposed algorithm. The results and a comparison against the general-purpose SDP solver SeDuMi are presented in section 7. Concluding remarks are drawn in section 8.

Notations and terminology. Given a function $F(\lambda) : \mathbf{R}^n \to \mathbf{R}$, the notations $\nabla F(\lambda)$ and $\nabla^2 F(\lambda)$ are used to denote the *gradient vector* and the *Hessian matrix* of $F(\lambda)$ (the *gradient* and the *Hessian* for short), respectively. The partial derivative of $F(\lambda)$ with respect to the ith component of λ is denoted by $\partial_i F(\lambda)$. The second partial derivative of $F(\lambda)$ with respect to the ith and jth components of λ is denoted by $\partial^2_{ij} F(\lambda)$. If $F(\lambda)$ is at least k times differentiable, then the notation

$$\nabla^k F(\lambda)[h_1, \ldots, h_k]$$

denotes the value of the kth differential of F taken at λ along the collection of directions $h_1, \ldots, h_k$, where $h_i \in \mathbf{R}^n$.

We use I_n to denote the $n \times n$ *identity* matrix. Sometimes the subscript n is dropped when the dimension of I_n is obvious from the context. Given a matrix M, the transposition and the conjugate transposition are denoted by M' and M^*, respectively. We call a matrix M *symmetric* if $M = M'$, and *Hermitian* if $M = M^*$. As mentioned before, the notation $M > 0$ is used to denote positive definiteness. The positive semidefiniteness, negative definiteness, and negative semidefiniteness are denoted using $\geq$, $<$, and $\leq$, respectively. A matrix M is called *Hurwitz* if all its eigenvalues have strictly negative real part. The notation $\mathbf{tr}(M)$ denotes the trace of M. The *Frobenius norm* of a square matrix M is defined as $\|M\|_F := \sqrt{\mathbf{tr}(M'M)}$.

Let $M_1, \ldots, M_n$ be square matrices. Then $M = \mathrm{diag}(M_1, \ldots, M_n)$ (sometimes abbreviated as $M = \mathrm{diag}_i(M_i)$) defines the block diagonal matrix

$$M = \begin{bmatrix} M_1 & & \\ & \ddots & \\ & & M_n \end{bmatrix}.$$

2. The semi-infinite optimization problem and its dual problem. Consider the KYP–SDP (1.1) and its equivalent semi-infinite optimization problem (1.2). To simplify the notation, let $\mathbf{H}_i(\omega, \lambda)$, $i = 0, \ldots, n$, be

$$\mathbf{H}_i(\omega) = \begin{bmatrix} (j\omega I_m - A)^{-1}B \\ I_p \end{bmatrix}^* M_i \begin{bmatrix} (j\omega I_m - A)^{-1}B \\ I_p \end{bmatrix}.$$

For the development in the sequential sections, let M_i be further partitioned into

$$\begin{bmatrix} Q_i & S_i \\ S_i' & R_i \end{bmatrix}, \tag{2.1}$$

where Q_i and R_i are $m \times m$ and $p \times p$ symmetric matrices, respectively. Also let us define $Q(\lambda)$, $F(\lambda)$, and $R(\lambda)$ to be

$$Q_0 + \sum_{i=1}^{n} \lambda_i Q_i, \quad S_0 + \sum_{i=1}^{n} \lambda_i S_i, \quad R_0 + \sum_{i=1}^{n} \lambda_i R_i. \tag{2.2}$$

Finally, define $\mathbf{H}(\omega, \lambda)$ to be $\mathbf{H}_0(\omega) + \sum_{i=1}^{n} \lambda_i \mathbf{H}_i(\omega)$, and the semi-infinite optimization problem (1.2) can be stated as

$$\inf_{\lambda} \; c'\lambda \quad \text{subject to} \quad \mathbf{H}(\omega, \lambda) > 0 \;\; \forall\, \omega \in [0, \infty]. \tag{2.3}$$

The feasible set of problem (2.3) is denoted as Ω, i.e.,

$$\Omega := \{\lambda \mid \mathbf{H}(\omega, \lambda) > 0 \;\forall\, \omega \in [0, \infty]\}. \tag{2.4}$$

Without loss of generality, matrices M_i are assumed to be linearly independent, which ensures that none of the decision variables is redundant. Furthermore, we also assume that Ω is bounded and nonempty.

Note that the constraint in (2.3) and (2.4) should be understood as: there exist an $\varepsilon > 0$ such that $\mathbf{H}(\omega, \lambda) \geq \varepsilon I_p$ for all $\omega \in [0, \infty)$. Hence, by definition, $R(\lambda_F)$ is (strictly) positive definite for any $\lambda_F \in \Omega$.

The semi-infinite optimization problem (2.3) and the KYP–SDP (1.1) are equivalent in the sense that for any pair (λ_F, P_F) that is feasible to (1.1), λ_F is a feasible solution of (2.3). On the other hand, given any feasible solution λ_F of (2.3), there exists a matrix P_F such that (λ_F, P_F) is feasible to (1.1).

The Lagrange dual problem associated with problem (2.3) is given as

$$\begin{aligned} &\sup_{Z(\omega)} \; -\langle \mathbf{H}_0(\omega), Z(\omega) \rangle \quad \text{subject to} \\ &\langle \mathbf{H}_i(\omega), Z(\omega) \rangle = c_i, \;\; i = 1, \ldots, n, \quad \text{and} \quad Z(\omega) \in \mathcal{P}_{NBV}^{m \times m}, \end{aligned} \tag{2.5}$$

where $\langle \mathbf{H}_i(\omega), Z(\omega) \rangle$ is defined by the Stieltjes integral

$$\int_{-\infty}^{\infty} \mathbf{tr}(\mathbf{H}_i(\omega) dZ(\omega)) := 2 \lim_{N \to \infty} \sum_{k=1}^{N} \mathbf{tr}(\mathbf{H}_i(\omega_{k-1})(Z(\omega_k) - Z(\omega_{k-1}))),$$

where $0 = \omega_0 \le \omega_1 \le \cdots \le \omega_N = \infty$ is a partition of $[0, \infty]$ which satisfies

$$\max_{k \in \{1,\ldots,N-1\}} |\omega_k - \omega_{k-1}| \to 0 \quad \text{and} \quad \omega_{N-1} \to \infty \quad \text{as} \quad N \to \infty.$$

$\mathcal{P}_{NBV}^{m\times m}$ is the positive cone defined as

$$\mathcal{P}_{NBV}^{m\times m} := \{Z(\omega) \in \mathcal{S}_{NBV}^{m\times m} \mid Z(\omega_1) \ge Z(\omega_2) \ \forall\ \omega_1 \ge \omega_2 > 0\},$$

where $\mathcal{S}_{NBV}^{m\times m}$ denotes the space of normalized bounded variation functions which map $\mathbf{R} \cup \{\infty\}$ to the space of $m \times m$ complex-valued Hermitian matrices. Readers are referred to [17] for the properties of $\mathcal{S}_{NBV}^{m\times m}$ and $\mathcal{P}_{NBV}^{m\times m}$.

Given any pair of primal-dual feasible solution $(\lambda, Z(\omega))$, we have

$$c'\lambda + \langle \mathbf{H}_0(\omega), Z(\omega)\rangle = \langle \mathbf{H}(\omega, \lambda), Z(\omega)\rangle \ge 0.$$

The last inequality follows $\mathbf{H}(\omega, \lambda) > 0$ for all $\omega \in [0, \infty]$ and $Z(\omega) \in \mathcal{P}_{NBV}^{m\times m}$. Hence, the weak duality theorem holds. The strong duality theorem, which states that the primal and the dual optimization problems have the same optimal objectives, holds under standard regularity assumption. See [10, 11] for details.

3. The barrier function for the semi-infinite optimization problem. Consider the following function from Ω to $\mathbf{R}$:

$$\mathcal{B}(\lambda) = \log \mathcal{G}(\lambda), \tag{3.1}$$

where

$$\mathcal{G}(\lambda) := \frac{1}{\pi} \int_{-\infty}^{\infty} \mathbf{tr}(\mathbf{H}(\omega, \lambda)^{-1}) \frac{d\omega}{1+\omega^2}. \tag{3.2}$$

It is obvious that $\mathcal{B}(\lambda)$ is smooth on Ω. The ith element of the gradient and the (i, j) entry of the Hessian of $\mathcal{B}(\lambda)$ are given as follows:

$$\partial_i \mathcal{B}(\lambda) = \mathcal{G}(\lambda)^{-1} \partial_i \mathcal{G}(\lambda), \tag{3.3}$$

$$\partial_{ij}^2 \mathcal{B}(\lambda) = \mathcal{G}(\lambda)^{-1} \partial_{ij}^2 \mathcal{G}(\lambda) - \mathcal{G}^{-2}(\lambda) \partial_i \mathcal{G}(\lambda) \partial_j \mathcal{G}(\lambda)', \tag{3.4}$$

where

$$\partial_i \mathcal{G}(\lambda) = -\frac{1}{\pi} \int_{-\infty}^{\infty} \mathbf{tr}(\mathbf{H}(\omega, \lambda)^{-1} \mathbf{H}_i(\omega) \mathbf{H}(\omega, \lambda)^{-1}) \frac{d\omega}{1+\omega^2}, \tag{3.5}$$

$$\begin{aligned} \partial_{ij}^2 \mathcal{G}(\lambda) = {} & \frac{1}{\pi} \int_{-\infty}^{\infty} \mathbf{tr}(\mathbf{H}(\omega, \lambda)^{-1} \mathbf{H}_i(\lambda) \mathbf{H}(\omega, \lambda)^{-1} \mathbf{H}_j(\omega) \mathbf{H}(\omega, \lambda)^{-1}) \frac{d\omega}{1+\omega^2} \\ & + \frac{1}{\pi} \int_{-\infty}^{\infty} \mathbf{tr}(\mathbf{H}(\omega, \lambda)^{-1} \mathbf{H}_j(\lambda) \mathbf{H}(\omega, \lambda)^{-1} \mathbf{H}_i(\omega) \mathbf{H}(\omega, \lambda)^{-1}) \frac{d\omega}{1+\omega^2}. \end{aligned} \tag{3.6}$$

The following proposition shows that $\mathcal{B}(\lambda)$ is also a convex function.

PROPOSITION 3.1. *$\mathcal{B}(\lambda)$ is a convex function.*

Proof. We show that the Hessian of $\mathcal{B}(\lambda)$ is strictly positive definite on Ω. It can be easily verified that

$$\nabla^2 \mathcal{B}(\lambda) = \mathcal{G}^{-1}(\lambda) \nabla^2 \mathcal{G}(\lambda) - \nabla \mathcal{G}(\lambda) \mathcal{G}^{-2}(\lambda) \nabla \mathcal{G}(\lambda)'.$$

Since $\mathcal{G}(\lambda) > 0$ for any $\lambda \in \Omega$, therefore, given any $\lambda \in \Omega$, $\nabla^2\mathcal{B}(\lambda) > 0$ if and only if

$$\nabla^2\mathcal{G}(\lambda) - \nabla\mathcal{G}(\lambda)\mathcal{G}^{-1}(\lambda)\nabla\mathcal{G}(\lambda)' > 0.$$

To prove $\nabla^2\mathcal{G}(\lambda) - \nabla\mathcal{G}(\lambda)\mathcal{G}^{-1}(\lambda)\nabla\mathcal{G}(\lambda)' > 0$, it is sufficient to show that

$$\begin{bmatrix} \nabla^2\mathcal{G}(\lambda) & \nabla\mathcal{G}(\lambda) \\ \nabla\mathcal{G}(\lambda)' & \mathcal{G}(\lambda) \end{bmatrix} > 0. \tag{3.7}$$

Matrix inequality (3.7) holds if $\nabla^2\mathcal{G}(\lambda)[h,h] + 2\nabla\mathcal{G}(\lambda)[h] + \mathcal{G}(\lambda) > 0$ for all nonzero h in $\mathbf{R}^n$. Let $\mathcal{A} = \sum_i^n h_i\mathbf{H}_i(\omega)$. By (3.5) and (3.6), we have

$$\begin{aligned}
&\nabla^2\mathcal{G}(\lambda)[h,h] + 2\nabla\mathcal{G}(\lambda)[h] + \mathcal{G}(\lambda) = \sum_{i,j} h_ih_j\partial^2_{ij}\mathcal{G}(\lambda) + 2\sum_i h_i\partial_i\mathcal{G}(\lambda) + \mathcal{G}(\lambda) \\
&= \frac{1}{\pi}\int_{-\infty}^{\infty} \mathbf{tr}(2\mathbf{H}^{-1}\mathcal{A}\mathbf{H}^{-1}\mathcal{A}\mathbf{H}^{-1} - 2\mathbf{H}^{-1}\mathcal{A}\mathbf{H}^{-1} + \mathbf{H}^{-1})\frac{d\omega}{1+\omega^2} \\
&= \frac{1}{\pi}\int_{-\infty}^{\infty} (2\|\mathbf{H}^{-\frac{1}{2}}\mathcal{A}\mathbf{H}^{-1} - 0.5\mathbf{H}^{\frac{1}{2}}\|_F^2 + 0.5\|\mathbf{H}^{-\frac{1}{2}}\|_F^2)\frac{d\omega}{1+\omega^2}.
\end{aligned}$$

Hence, for any $\lambda \in \Omega$, $\nabla^2\mathcal{G}(\lambda)[h,h] + 2\nabla\mathcal{G}(\lambda)[h] + \mathcal{G}(\lambda) > 0$ for all $h \neq 0$. This in turn implies that $\nabla^2\mathcal{B}(\lambda) > 0$ for all $\lambda \in \Omega$, and thus $\mathcal{B}(\lambda)$ is a convex function. ∎

Although the integral over an infinite horizon makes it seemingly difficult to evaluate $\mathcal{B}(\lambda)$ for any given λ, the evaluation can be performed by using a rather efficient computational procedure. The main computation for evaluating $\mathcal{B}(\lambda)$ is to solve one Riccati equation and one Lyapunov equation. Furthermore, it can also be shown that the value of $\mathcal{B}(\lambda)$ approaches infinity as λ approaches the boundary of Ω. This property, together with smoothness and convexity, makes $\mathcal{B}(\lambda)$ a natural barrier for Ω.

Given $\lambda_F \in \Omega$, the following factorization formulas for $\mathbf{H}(\omega, \lambda_F)$ lead to an efficient computational procedure for evaluating $\mathcal{B}(\lambda_F)$.

PROPOSITION 3.2. *Suppose that $A \in \mathbf{R}^{m\times m}$ is a Hurwitz matrix. Given $\lambda_F \in \Omega$, the following factorization formulas hold for $\mathbf{H}(\omega, \lambda_F)$:*

(1) *$\mathbf{H}(\omega, \lambda_F)^{-1}$ can be factorized as $D_H + G_H(j\omega) + G_H(j\omega)^*$, where*

$$G_H(s) = C_H(sI - A_H)^{-1}B_H, \tag{3.8}$$

$$A_H = A - BR_F^{-1}(PB + S_F)', \tag{3.9}$$

$$B_H = BR_F^{-1} - Y(PB + S_F)R_F^{-1}, \tag{3.10}$$

$$C_H = -R_F^{-1}(PB + S_F)', \tag{3.11}$$

$$D_H = R_F^{-1}, \tag{3.12}$$

and P, Y satisfy the following Riccati and Lyapunov equations, respectively:

$$PA + A'P + Q_F - (PB + S_F)R_F^{-1}(PB + S_F)' = 0, \tag{3.13}$$

$$A_HY + YA_H' + BR_F^{-1}B' = 0. \tag{3.14}$$

Matrices Q_F, S_F, and R_F are equal to $Q(\lambda_F)$, $S(\lambda_F)$, and $R(\lambda_F)$, respectively. Furthermore, matrix A_H is also a Hurwitz matrix and the dimension of A_H is the same as that of A.

(2) $\mathbf{H}(\omega,\lambda_F)^{-1}$ *can be factorized as* $\Psi(j\omega)\Psi(j\omega)^*$, *where*

$$\Psi(s) = R_F^{-\frac{1}{2}} + C_H(sI - A_H)^{-1}BR_F^{-\frac{1}{2}}. \tag{3.15}$$

Proof. See Appendix A. ∎

The next proposition offers an algebraic form (as opposed to the integral form) for $\mathcal{G}(\lambda)$. Hence, computation of $\mathcal{G}(\lambda)$ and $\mathcal{B}(\lambda)$ can be performed without numerically integrating $\frac{1}{1+\omega^2}\mathbf{tr}(\mathbf{H}(\omega,\lambda)^{-1})$ over an infinite horizon.

PROPOSITION 3.3. *Let* $\lambda_F \in \Omega$. *Then* $\mathcal{G}(\lambda_F) = \mathbf{tr}(D_H) + 2\mathbf{tr}(C_H(I - A_H)^{-1}B_H)$, *where* A_H, B_H, C_H, D_H *are defined as in* (3.9)–(3.12).

Proof. First, notice that the orders of the trace operator and the integral operator can be exchanged; therefore, by the first factorization formula given in Proposition 3.2, we have

$$\begin{aligned}\mathcal{G}(\lambda_F) &= \mathbf{tr}\left(\frac{1}{\pi}\int_{-\infty}^{\infty}(D_H + G_H(j\omega) + G_H(j\omega)^*)\frac{d\omega}{1+\omega^2}\right)\\ &= \mathbf{tr}(D_H) + \mathbf{tr}\left(\frac{1}{\pi}\int_{-\infty}^{\infty}(G_H(j\omega) + G_H(j\omega)^*)\frac{d\omega}{1+\omega^2}\right).\end{aligned}$$

Now, if we treat $G_H(j\omega)$ as the Fourier transform of the stable causal system $G_H(s)$ and let $g(t)$ be the impulse response of $G_H(s)$, by the Plancherel theorem,

$$\frac{1}{\pi}\int_{-\infty}^{\infty} G_H(j\omega)\frac{1}{1+\omega^2}\,d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} G_H(j\omega)\frac{2}{1+\omega^2}\,d\omega = g(t) * e^{-|t|}\Big|_{t=0},$$

where $*$ denotes the convolution operator. Since $G_H(s)$ is stable and causal, we have

$$g(t) * e^{-|t|}\Big|_{t=0} = \int_0^{\infty} g(\tau)e^{-\tau}d\tau = G_H(1) = C_H(I - A_H)^{-1}B_H.$$

Similarly, we have

$$\frac{1}{\pi}\int_{-\infty}^{\infty} G_H(j\omega)^*\frac{1}{1+\omega^2}d\omega = C_H(I - A_H)^{-1}B_H.$$

This concludes the proof. ∎

Therefore, the computation of $\mathcal{G}(\lambda)$ mainly involves solving Riccati equation (3.13) and Lyapunov equation (3.14). Solving Riccati equations and Lyapunov equations has been well studied. Efficient computational routines for solving such equations are widely available.

The following proposition shows that $\mathcal{B}(\lambda)$ approaches infinity as λ approaches the boundary of Ω.

PROPOSITION 3.4. *The value of the barrier function* $\mathcal{B}(\lambda)$ *is unbounded on the boundary of* Ω.

Proof. Let λ_b belong to the boundary Ω; i.e., $\mathbf{H}(\omega,\lambda_b)$ is only semipositive definite. Suppose that $\mathbf{H}(\omega,\lambda_b)$ is singular at the set $\Gamma = \{\pm\omega_i,\ i = 1,\dots,l\}$. It can then be shown (Chapter 13.4 of [33]) that

(1) matrix $R_b := R(\lambda_b)$ is singular if $\infty \in \Gamma$.

(2) the corresponding A_H has pure imaginary eigenvalues $j\omega_i$, $\omega_i \in \Gamma$, $\omega_i \neq \pm\infty$.

If R_b is singular, unboundedness of $\mathcal{G}(\lambda_b)$ and $\mathcal{B}(\lambda_b)$ immediately follows (3.12). If R_b is not singular, then $\mathbf{H}(\omega,\lambda_b)$ can be factorized as $\Psi(j\omega)\Psi(j\omega)^*$, where $\Psi(s)$ is defined as in (3.15). Note that

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\mathbf{tr}(\mathbf{H}(\omega,\lambda_b)^{-1})\frac{d\omega}{1+\omega^2} = 2\left(\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathbf{tr}(\widetilde{\Psi}(j\omega)\widetilde{\Psi}(j\omega)^*)d\omega\right), \tag{3.16}$$

where $\widetilde{\Psi}(s) = \frac{1}{s+1}\Psi(s)$. The integral in the parentheses gives the $\mathcal{H}_2$-norm of $\widetilde{\Psi}(s)$. Since A_H has eigenvalues on the imaginary axis, $\widetilde{\Psi}(s)$ is not a stable transfer matrix. Therefore, its $\mathcal{H}_2$-norm is unbounded, which in turn implies that $\mathcal{G}(\lambda_b)$ is unbounded, and thus so is $\mathcal{B}(\lambda_b)$. This concludes the proof. ∎

3.1. Gradients and Hessians of $\mathcal{B}(\lambda)$. Proposition 3.3 gives the following equivalent expression for $\mathcal{G}(\lambda)$:

$$\mathcal{G}(\lambda) = \mathbf{tr}(D_H(\lambda) + 2C_H(\lambda)(I - A_H(\lambda))^{-1}B_H(\lambda)), \tag{3.17}$$

where $A_H(\lambda)$, $B_H(\lambda)$, $C_H(\lambda)$, and $D_H(\lambda)$ are defined as in (3.8)–(3.14) with Q_F, S_F, and R_F replaced by $Q(\lambda)$, $F(\lambda)$, and $R(\lambda)$. Differentiating (3.17) with respect to λ_i, we obtain the following expressions for the ith component of the gradient of $\mathcal{G}(\lambda)$:

$$\begin{aligned}\partial_i\mathcal{G}(\lambda) &= \mathbf{tr}(\partial_i D_H + 2(\partial_i C_H)(I - A_H)^{-1}B_H - 2C_H(I - A_H)^{-1}(\partial_i A_H)(I - A_H)^{-1}B_H \\ &\qquad + 2C_H(I - A_H)^{-1}(\partial_i B_H)) \\ &= \mathbf{tr}(\partial_i D_H + 2(I + C_H(I - A_H)^{-1}B)((\partial_i C_H)(I - A_H)^{-1}B_H) \\ &\qquad + 2C_H(I - A_H)^{-1}(\partial_i B_H)),\end{aligned} \tag{3.18}$$

where the second equality is obtained by noting that $\partial_i A_H = B(\partial_i C_H)$. It can be easily verified that the partial derivatives of B_H, C_H, and D_H with respect to λ_1 have the following expressions:

$$\partial_i B_H = (\partial_i Y)C_H' + Y(\partial_i C_H)', \tag{3.19}$$

$$\partial_i C_H = R^{-1}R_i R^{-1}(PB + F)' - R^{-1}((\partial_i P)B + F_i)', \tag{3.20}$$

$$\partial_i D_H = -R^{-1}R_i R^{-1}, \tag{3.21}$$

and the partial derivatives of P and Y with respect to λ_i satisfy the following equations:

$$(\partial_i P)A_H + A_H'(\partial_i P) + (Q_i + F_i C_H + C_H' F_i' + C_H' R_i C_H) = 0, \tag{3.22}$$

$$A_H(\partial_i Y) + (\partial_i Y)A_H' + (B(\partial_i C_H)Y + Y(\partial_i C_H)'B' - BR^{-1}R_i R^{-1}B') = 0. \tag{3.23}$$

For a given point $\lambda \in \Omega$, computation of $\partial_i\mathcal{B}(\lambda)$ can be performed as follows: First, notice that for a fixed λ, (3.22) is a Lyapunov equation with respect to $\partial_i P$. Thus, the value of $\partial_i P$ can be obtained by solving the Lyapunov equation. Then the values of $\partial_i D_H$ and $\partial_i C_H$ can be computed according to expressions (3.21) and (3.20), respectively. As soon as the value of $\partial_i C_H$ is available, one can solve another Lyapunov equation in the form of (3.23) to obtain the value of $\partial_i Y$, and then $\partial_i B_H$ can be computed using expression (3.19). Finally, $\partial_i\mathcal{G}(\lambda)$ can be evaluated using (3.18) and $\partial_i\mathcal{B}(\lambda)$ evaluated using (3.3).

If we further differentiate (3.18) with respect to λ_j, we obtain an expression for $\partial_{ij}\mathcal{G}(\lambda)$:

$$\begin{aligned}\partial_{ij}\mathcal{G}(\lambda) &= \mathbf{tr}(\partial_{ij}^2 D_H) \\ &\qquad + 2\mathbf{tr}(C_H(I - A_H)^{-1}(\partial_{ij}^2 B_H) + (I + C_H(I - A_H)^{-1}B)T_1),\end{aligned} \tag{3.24}$$

where

$$T_1 = (\partial^2_{ij} C_H)(I - A_H)^{-1} B_H + (\partial_i C_H)(I - A_H)^{-1}(\partial_j B_H + B(\partial_j C_H)$$
$$\times \quad (I - A_H)^{-1} B_H) + (\partial_j C_H)(I - A_H)^{-1}(\partial_i B_H) + B(\partial_i C_H) \tag{3.25}$$
$$\times \quad (I - A_H)^{-1} B_H),$$

$$\partial^2_{ij} B_H = (\partial^2_{ij} Y) C'_H + Y(\partial^2_{ij} C_H)' + (\partial_i Y)(\partial_j C_H)' + (\partial_j Y)(\partial_i C_H)', \tag{3.26}$$

$$\partial^2_{ij} C_H = -R^{-1}(R_j(\partial_i C_H) + R_i(\partial_j C_H) + B'(\partial^2_{ij} P)'), \tag{3.27}$$

$$\partial^2_{ij} D_H = R^{-1} R_i R^{-1} R_j R^{-1} + R^{-1} R_j R^{-1} R_i R^{-1}. \tag{3.28}$$

$\partial^2_{ij} P$ and $\partial^2_{ij} Y$ satisfy the following Lyapunov equations:

$$(\partial^2_{ij} P) A_H + A'_H(\partial^2_{ij} P) + (T_2 + T'_2) = 0, \tag{3.29}$$

$$(\partial^2_{ij} Y) A_H + A'_H(\partial^2_{ij} Y) + (T_3 + T'_3) = 0, \tag{3.30}$$

where T_2 and T_3 denote the following expressions:

$$T_2 = ((\partial_i P) B + C_H R_i + F_i)(\partial_j C_H), \tag{3.31}$$

$$T_3 = B((\partial_i C_H)(\partial_j Y) + (\partial_j C_H)(\partial_i Y) + (\partial^2_{ij} C_H) Y) + B R^{-1} R_i R^{-1} R_j R^{-1} B'. \tag{3.32}$$

For a given $\lambda \in \Omega$, the computational procedure of evaluating $\partial^2_{ij} \mathcal{B}(\lambda)$ is similar to that of computing $\partial_i \mathcal{B}(\lambda)$. Assume that the values of the first partial derivatives of B_H, C_H, D_H, P, and Y are available. Then, T_2 can be evaluated, and Lyapunov equation (3.29) can be solved for the value of $\partial^2_{ij} P$. As soon as the value of $\partial^2_{ij} P$ is available, one can evaluate $\partial^2_{ij} C_H$ and $\partial^2_{ij} D_H$ using expressions (3.27) and (3.28). Once the value of $\partial^2_{ij} C_H$ is obtained, Lyapunov equation (3.30) can be solved for the value of $\partial^2_{ij} Y$, which in turn is used to obtain the value of $\partial^2_{ij} B_H$ using (3.26). Finally, $\partial^2_{ij} \mathcal{G}(\lambda)$ can be computed according to expression (3.24) and $\partial^2_{ij} \mathcal{B}(\lambda)$ computed by (3.4).

The computational complexity of evaluating the barrier function and its derivatives can be estimated as follows: The computation mainly involves solving one Riccati equation of the form (3.13), solving $\mathbf{O}(n^2)$ Lyapunov equations of the form (3.14), inverting matrices of sizes $m \times m$ and $p \times p$, and multiplying matrices of sizes $m \times m$, $m \times p$, $p \times m$, and $p \times p$. The computational complexity of solving (3.13) is $\mathbf{O}(m^3)$. Complexity of solving (3.14) is also $\mathbf{O}(m^3)$, or $\mathbf{O}(m^2)$ if diagonalization of the A_H matrix is permissible. The complexity of inverting an $m \times m$ matrix is $\mathbf{O}(m^3)$, as is the complexity of multiplying two $m \times m$ matrices. Therefore, the complexity of the algorithm for evaluating the barrier function and its derivatives is estimated to be $\mathbf{O}(n^2(m^3 + p^3))$.

4. The central path. In this section we describe the central path of the primal optimization problem (2.3) and give some of its properties. Let $t \geq 1$ and define

$$\varphi_t(\lambda) = tc'\lambda + \mathcal{B}(\lambda).$$

For any fixed t, consider the minimization problem

$$\inf \ \varphi_t(\lambda) \quad \text{subject to } \lambda \in \Omega. \tag{4.1}$$

Problem (4.1) is often referred to as *centering*, and the minimizer $\lambda^\star(t)$ is called the center of (4.1). Note that $\varphi_t(\lambda)$ is a convex function, which becomes unbounded at

the boundary of Ω. Under the assumption imposed on $\mathbf{H}(\omega, \lambda)$, problem (4.1) has a unique minimizer, which is strictly inside Ω. The curve $\lambda^\star(t)$, parameterized by $t \geq 1$, is referred to as the *primal central path*.

The minimizer $\lambda^\star(t)$ of (4.1) is characterized by the optimality condition $\nabla\varphi_t(\lambda) = 0$. Using expression (3.5), we have

$$tc_i = \left(\int_{-\infty}^{\infty} \mathbf{tr}(\mathbf{H}(\omega,\lambda^\star)^{-1})\frac{d\omega}{1+\omega^2}\right)^{-1} \int_{-\infty}^{\infty} \mathbf{tr}(\mathbf{H}(\omega,\lambda^\star)^{-1}\mathbf{H}_i(\omega)\mathbf{H}(\omega,\lambda^\star)^{-1})\frac{d\omega}{1+\omega^2}$$

for $i = 1, \ldots, n$. Hence, the matrix function $Z^\star(\omega, t)$, which satisfies

$$\frac{dZ^\star(\omega,t)}{d\omega} = \frac{1}{t}\left(\int_{-\infty}^{\infty} \mathbf{tr}(\mathbf{H}(\omega,\lambda^\star(t))^{-1})\frac{d\omega}{1+\omega^2}\right)^{-1} \frac{1}{1+\omega^2}\mathbf{H}^{-2}(\omega,\lambda^\star(t)), \tag{4.2}$$

is strictly dual feasible. Therefore, the primal-dual duality gap associated with $\lambda^\star(t)$ and $Z^\star(\omega, t)$ can be found as follows:

$$\begin{aligned} c'\lambda^\star(t) + \langle \mathbf{H}_0(\omega), Z^\star(\omega,t)\rangle &= \left\langle \sum_{k=1}^{n} \mathbf{H}_i(\omega)\lambda_i^\star(t), Z^\star(\omega,t)\right\rangle + \langle \mathbf{H}_0(\omega), Z^\star(\omega,t)\rangle \\ &= \langle \mathbf{H}(\omega,\lambda^\star), Z(\omega,t)\rangle = \frac{1}{t}. \end{aligned} \tag{4.3}$$

This implies that the central path converges to the solution of the semi-infinite optimization (2.3) as t approaches infinity.

Newton's method for centering. Consider the minimization problem (4.1). We employ Newton's method with line search to solve this problem.

Newton's method for minimizing $\boldsymbol{\varphi_t(\lambda)}$.
Start at a strictly feasible point λ_0. Select tolerance ϵ and set $n = 0$.
Repeat
(a) Compute the Newton descent direction $\delta\lambda_n = -(\nabla^2\varphi_t(\lambda_n))^{-1}\nabla\varphi_t(\lambda_n)$.
(b) Compute $\rho = (\nabla\varphi_t(\lambda_n)'(\nabla^2\varphi_t(\lambda_n))^{-1}\nabla\varphi_t(\lambda_n))^{\frac{1}{2}}$.
(c) Line minimization: compute $\alpha^\star = \operatorname{argmin}\ \varphi_t(\lambda_n + \alpha\cdot\delta\lambda_n)$.
(d) Update $\lambda_{n+1} := \lambda_n + \alpha^\star\delta\lambda_n$ and $n := n + 1$.
(e) If $\rho < \epsilon$, then stop the loop and return λ_n.
End

The quantity ρ is called the Newton decrement, which is used to measure the closeness to the central path for a strictly feasible λ.

It is well known that Newton's method converges quadratically asymptotically. The global convergence of Newton's method was analyzed by Nesterov and Nemirovskii. In [19] they show that if the function to be minimized has a certain property called *self-concordance*, Newton's method converges in polynomial time. The bound on the number of iterations before the algorithm terminated is also explicitly given.

Let $X \subset \mathbf{R}^n$ be a convex open set. A smooth convex function $F : X \to \mathbf{R}$ is called self-concordant with the parameter value a (or *a-self-concordant* for short) if there exists a constant a such that the following inequality holds for all $x \in X$ and for all $h \in \mathbf{R}^n$:

$$|\nabla^3 F(x)[h,h,h]| \leq 2a^{\frac{-1}{2}}(\nabla^2 F(x)[h,h])^{\frac{3}{2}}.$$

$F(x)$ is called strongly a-self-concordant if $F(x) \to \infty$ as x approaches the boundary of X. Nesterov and Nemirovskii gave a complete characterization of the speed

of convergence of Newton's method applied to minimize a strongly self-concordant function. Readers interested in this result are referred to Chapter 2 of [19]. Since any linear function is trivially self-concordant, $\varphi_t(\lambda)$ is self-concordant if $\mathcal{B}(\lambda)$ is. It can be shown that the barrier function $\mathcal{B}(\lambda)$ is indeed self-concordant.

THEOREM 4.1. *For a given $\mathcal{B}(\lambda)$, there exists a constant a such that*

$$|\nabla^3\mathcal{B}(\lambda)[h,h,h]| \leq a(\nabla^2\mathcal{B}(\lambda)[h,h])^{\frac{3}{2}} \quad \forall\ h \in \mathbf{R}^n \tag{4.4}$$

holds for all $\lambda \in \Omega$.

Proof. See Appendix B. □

The proof of Theorem 4.1 is based on asymptotic analysis. Basically, what is shown is that the ratio

$$\frac{|\nabla^3\mathcal{B}(\lambda)[h,h,h]|}{(\nabla^2\mathcal{B}(\lambda)[h,h])^{\frac{3}{2}}}$$

does not become unbounded as λ approaches the boundary of Ω. However, neither the value nor an upper bound of a is obtained in our proof. Since such information is essential for applying Nesterov and Nemirovskii's results to construct polynomial-time algorithms, the complexity of minimizing $\varphi_t(\lambda)$ is yet to be determined. Note that Theorem 4.1 plays no important role in our algorithm and is not claimed to be a major contribution of this paper.

5. Path-following algorithms. Path-following algorithms can be dated back to Fiacco and McCormick's work in [6], where the path-following algorithm was called the sequential unconstrained minimization method. Worst-case convergence analysis was first attempted by Renegar, who proved the polynomial complexity of a path-following algorithm for linear programming [22]. In the case of general nonlinear convex optimization problems, convergence analysis was studied by Nesterov and Nemirovskii [19]. They proved worst-case polynomial complexity for the case when the barriers used in the path-following algorithms are self-concordant. Interested readers are referred to [19] for an historical overview.

Barrier functions $\mathcal{B}(\lambda)$ are used to construct path-following algorithms for solving the semi-infinite optimization problem (2.3). The algorithms we consider here are standard, which follows the basic principles described below:

Given: $\lambda_0 \in \Omega$.
Initialization: Select $\mu > 1$ and $\epsilon > 0$. Let $t = 1$.
Repeat
(1) Centering: starting from λ_0, find an approximate solution $\lambda^\star(t)$ to the problem

$$\min_{\lambda}\ \varphi_t(\lambda),$$

where $tc'\lambda + \mathcal{B}(\lambda)$, using Newton's method.
(2) Update λ_0: set $\lambda_0 := \lambda^\star(t)$.
(3) Update t: set $t := \mu t$.
Until $(c'(\lambda^\star(t) - \lambda_{opt}) \leq \epsilon)$.

Here λ_{opt} denotes the optimal solution of (2.3). The initial feasible point λ_0 can be found (or determined not to exist) using the so-called big-M method [4]. The path-following algorithm for the big-M method is essentially the same as the algorithm described above. Note that, should the problem be feasible, the big-M method will produce a feasible solution λ_0, which is sufficiently centered. The algorithm terminates

when a certificate that proves $c'(\lambda^\star(t) - \lambda_{opt}) \leq \epsilon$ is obtained. How to obtain such a certificate is discussed in the next section.

The centering step is often referred to as the *inner loop*, and the loop that involves increasing t by a factor μ and finding $\lambda^\star(\mu t)$ from $\lambda^\star(t)$ is called the *outer loop*. While the inner loop is to find feasible solutions that are close to the central path, the outer loop serves to bring $\lambda^\star(t)$ toward the optimal solution λ_{opt}.

The selection of μ involves a certain trade-off. A large μ incurs a large increase on t per outer iteration, and hence less outer steps are required for approximate central solutions to converge to the optimal one. However, a large μ also implies that $\lambda^\star(\mu t)$ is sufficiently far away from $\lambda^\star(t)$, and hence more steps in the inner loop are required for centralization.

There are various strategies for selecting the factor μ. The strategy ultimately dictates the overall complexity of the path-following algorithm. In [19], variants in selecting μ are discussed and the complexity of the corresponding path-following algorithms is given under the assumption that the barrier functions used in the algorithms are self-concordant. Since the complexity of minimizing $\varphi_t(\lambda)$ by Newton's method is yet to be determined, we will not pursue the issue of how to select μ from the theoretical prospect in this paper. Instead, we test different values of μ in numerical experiments, results of which are reported in section 7.

Stopping criterion. The weak and strong duality theorems imply that given any pair of primal and dual feasible solutions $(\lambda,\, Z(\omega))$, the inequality

$$c'\lambda - c'\lambda_{opt} \leq c'\lambda + \langle \mathbf{H}_0(\omega), Z(\omega)\rangle$$

holds, where λ_{opt} is the optimal solution of (2.3). Furthermore, from (4.3) we see that when a centralized feasible solution $\lambda^\star(t)$ is found, a dual solution $Z^\star(\omega, t)$ is immediately given, as in (4.2), which allows us to conclude that $c'\lambda^\star(t) - c'\lambda_{opt} \leq 1/t$.

Practically, a feasible solution on the central path never could be obtained exactly, and an approximation only up to a certain degree of accuracy can be found. Hence, the form (4.2) for obtaining a dual feasible solution is not directly useful in practice, and so seems the bound (4.3). However, when a feasible solution is sufficiently close to the central path, a dual feasible solution can be constructed according to the following proposition.

PROPOSITION 5.1. *Given $\lambda_F \in \Omega$, let $\mathbf{H}_F(\omega)$ and v_N denote $\mathbf{H}(\omega, \lambda_F)$ and the Newton descent direction at λ_F; i.e.,*

$$v_N = -(\nabla^2\varphi_t(\lambda_F))^{-1}\nabla\varphi_t(\lambda_F).$$

Furthermore, let $\kappa = 1 - \mathcal{G}(\lambda_F)^{-1}\nabla\mathcal{G}(\lambda_F)'v_N$ and consider $Z(\omega, t)$, which satisfies

$$\begin{aligned}\frac{dZ(\omega,t)}{d\omega} = \frac{1}{t\pi}\mathcal{G}(\lambda_F)^{-1}\frac{1}{1+\omega^2}\Bigg(&\kappa\mathbf{H}_F(\omega)^{-2} - \mathbf{H}_F(\omega)^{-2}\left(\sum_{j=1}^{n} v_{N,j}\mathbf{H}_j(\omega)\right)\\ &\times\, \mathbf{H}_F(\omega)^{-1} - \mathbf{H}_F(\omega)^{-1}\left(\sum_{j=1}^{n} v_{N,j}\mathbf{H}_j(\omega)\right)\mathbf{H}_F(\omega)^{-2}\Bigg).\end{aligned} \tag{5.1}$$

Then $Z(\omega, t)$ satisfies $\langle \mathbf{H}_i(\omega), Z(\omega, t)\rangle = c_i$. Moreover, if

$$\frac{\kappa}{2}\mathbf{H}_F(\omega) \geq \sum_{i=1}^{n} v_{N,i}\mathbf{H}_i(\omega) \quad \forall\ \omega \in [0, \infty], \tag{5.2}$$

then $Z(\omega,t)$ *is dual feasible and* $c'\lambda_F + \langle \mathbf{H}_0(\omega), Z(\omega,t)\rangle$ *is equal to* $(1+\mathcal{G}(\lambda_F)^{-1}\nabla\mathcal{G}(\lambda_F)'v_N)/t$.

Proof. The following equation can be easily verified:

$$\nabla^2\mathcal{G}(\lambda_F)v_N + \kappa\nabla\mathcal{G}(\lambda_F) = -\mathcal{G}(\lambda_F)tc,$$

which implies

$$\begin{aligned}-\mathcal{G}(\lambda_F)tc_i = {}& \frac{1}{\pi}\int_{-\infty}^{\infty}\mathbf{tr}\left(\mathbf{H}_F(\omega)^{-1}\left(\sum_{j=1}^{n}v_{N,j}\mathbf{H}_j(\omega)\right)\mathbf{H}_F(\omega)^{-1}\mathbf{H}_i(\omega)\mathbf{H}_F(\omega)^{-1}\right)\frac{d\omega}{1+\omega^2}\\ &+\frac{1}{\pi}\int_{-\infty}^{\infty}\mathbf{tr}\left(\mathbf{H}_F(\omega)^{-1}\mathbf{H}_i(\omega)\mathbf{H}_F(\omega)^{-1}\left(\sum_{j=1}^{n}v_{N,j}\mathbf{H}_j(\omega)\right)\mathbf{H}_F(\omega)^{-1}\right)\frac{d\omega}{1+\omega^2}\\ &-\frac{\kappa}{\pi}\int_{-\infty}^{\infty}\mathbf{tr}(\mathbf{H}_F(\omega)^{-1}\mathbf{H}_i(\omega)\mathbf{H}_F(\omega)^{-1})\frac{d\omega}{1+\omega^2}.\end{aligned}$$

Hence,

$$\begin{aligned}c_i = {}& \int_{-\infty}^{\infty}\mathbf{tr}\left(\mathbf{H}_i(\omega)\frac{-1}{t\pi}\mathcal{G}(\lambda_F)^{-1}\mathbf{H}_F(\omega)^{-2}\left(\sum_{j=1}^{n}v_{N,j}\mathbf{H}_j(\omega)\right)\mathbf{H}_F(\omega)^{-1}\right)\frac{d\omega}{1+\omega^2}\\ &+\int_{-\infty}^{\infty}\mathbf{tr}\left(\mathbf{H}_i(\omega)\frac{-1}{t\pi}\mathcal{G}(\lambda_F)^{-1}\mathbf{H}_F(\omega)^{-1}\left(\sum_{j=1}^{n}v_{N,j}\mathbf{H}_j(\omega)\right)\mathbf{H}_F(\omega)^{-2}\right)\frac{d\omega}{1+\omega^2}\\ &+\int_{-\infty}^{\infty}\mathbf{tr}\left(\mathbf{H}_i(\omega)\frac{\kappa}{t\pi}\mathcal{G}(\lambda_F)^{-1}\mathbf{H}_F(\omega)^{-2}\right)\frac{d\omega}{1+\omega^2}.\end{aligned}$$

This shows that $Z(\omega,t)$ as defined in (5.1) satisfies $\langle \mathbf{H}_i(\omega), Z(\omega,t)\rangle = c_i$. Furthermore, if (5.2) holds, then

$$\kappa\mathbf{H}_F(\omega)^2 - \mathbf{H}_F(\omega)\left(\sum_{j=1}^{n}v_{N,j}\mathbf{H}_j(\omega)\right) - \left(\sum_{j=1}^{n}v_{N,j}\mathbf{H}_j(\omega)\right)\mathbf{H}_F(\omega) \geq 0 \quad \forall\ \omega\in[0,\infty],$$

which in turn implies that

$$\frac{dZ(\omega,t)}{d\omega} \geq 0 \quad \forall\ \omega\in[0,\infty]$$

and that $Z(\omega,t)\in\mathcal{P}_{NBV}^{m\times m}$. Hence, $Z(\omega,t)$ is dual feasible.

To see $c'\lambda_F + \langle \mathbf{H}_0(\omega), Z(\omega,t)\rangle = (1+\mathcal{G}(\lambda_F)^{-1}\nabla\mathcal{G}(\lambda_F)'v_N)/t$, note that $c'\lambda_F + \langle \mathbf{H}_0(\omega), Z(\omega,t)\rangle = \langle \mathbf{H}_F(\omega), Z(\omega,t)\rangle$, and therefore

$$\begin{aligned}c'\lambda_F + \langle \mathbf{H}_0(\omega), Z(\omega,t)\rangle = {}& \frac{1}{t\pi}\mathcal{G}(\lambda_F)^{-1}\int_{-\infty}^{\infty}\kappa\cdot\mathbf{tr}(\mathbf{H}_F(\omega)^{-1})\frac{d\omega}{1+\omega^2} - 2\frac{1}{t\pi}\mathcal{G}(\lambda_F)^{-1}\\ &\times\int_{-\infty}^{\infty}\mathbf{tr}\left(\mathbf{H}_F(\omega)^{-1}\left(\sum_{j=1}^{n}v_{N,j}\mathbf{H}_j(\omega)\right)\mathbf{H}_F(\omega)^{-1}\right)\frac{d\omega}{1+\omega^2}\\ ={}& \frac{1}{t}(\kappa + 2\mathcal{G}(\lambda_F)^{-1}\nabla\mathcal{G}(\lambda_F)'v_N)\\ ={}& \frac{1}{t}(1+\mathcal{G}(\lambda_F)^{-1}\nabla\mathcal{G}(\lambda_F)'v_N).\end{aligned}$$

This concludes the proof. □

Note that if λ_F is on the central path, i.e., λ_F minimizes $\varphi_t(\lambda)$, then $v_N \equiv 0$, $\kappa = 1$, and expression (5.1) reduces to expression (4.2) with inequality (5.2) trivially satisfied. Hence, as λ_F is sufficiently close to the central path, a dual feasible solution can be obtained as the form of (5.1), which serves as a certificate for proving $c'\lambda - c'\lambda_{opt} \leq (1 + \mathcal{G}(\lambda_F)^{-1}\nabla\mathcal{G}(\lambda_F)'v_N)/t$.

Although a prescribed measure of the "closeness" of λ_F to the central path can be derived and described in terms of the $\mathcal{H}_\infty$ norms of $\mathbf{H}_i(\omega)$ and $\mathbf{H}_F(\omega)$, the length of λ_F and v_N, and the value of $\mathcal{G}(\lambda_F)$, such a measure may be too conservative for all practical purposes. In practice, it may be more favorable to verify inequality (5.2) numerically as the algorithm proceeds, because checking (5.2) can be done rather efficiently by computing the eigenvalues of a certain Hamiltonian matrix of size $2n \times 2n$. See Chapter 13.4 of [33] for details.

6. Remarks on computational complexity. In this section, we comment on the computational complexities of solving (2.3) using the proposed interior path-following algorithm, and solving (1.1) using general-purpose SDP solvers. Arguments are given to explain why the proposed algorithm may outperform a general-purpose SDP solver, which solves (1.1) without exploiting the special structure of the problem. We recall that the dimension of matrix A is $m \times m$, the dimension of matrix B is $m \times p$, and the number of the decision variables is n.

Problem (1.1) has a well-known, strongly self-concordant barrier $-\log\det(\mathcal{F}(P, \lambda))$. The primal-dual interior-point algorithm for solving (1.1) can be proved to converge in $\mathbf{O}(\sqrt{m+p}\log\frac{m+p}{\varepsilon})$ Newton steps [19, 26]. In each Newton step, without utilizing the special structure of the problem, the computational complexity of evaluating the gradient is estimated to be $\mathbf{O}((m+p)^3(\frac{m^2}{2}+n))$. This involves matrix inversion of an $(m+p)\times(m+p)$ matrix, and $\mathbf{O}(\frac{m^2}{2}+n)$ matrix multiplications, where the multiplicands and multipliers are all $(m+p)\times(m+p)$. The complexity for evaluating the Hessian is $\mathbf{O}((m+p)^2(\frac{m^2}{2}+n)^2)$. The calculation involves $\mathbf{O}((\frac{m^2}{2}+n)^2)$ vector multiplications, where the dimension of the vectors is $(m+p)^2$. Finally, calculating the Newton descent direction requires inversion of an $(\frac{m^2+m}{2}+n)\times(\frac{m^2+m}{2}+n)$ matrix, which has complexity $\mathbf{O}((\frac{m^2}{2}+n)^3)$. Therefore, the overall complexity is $\mathbf{O}(\sqrt{m+p}\log\frac{m+p}{\varepsilon})\cdot(\mathbf{O}((m+p)^3(\frac{m^2}{2}+n))+\mathbf{O}((m+p)^2(\frac{m^2}{2}+n)^2)+\mathbf{O}((\frac{m^2}{2}+n)^3))$.

For the interior path-following algorithm proposed in this paper, the total number of Newton steps required for the algorithms to converge is yet to be determined. However, the complexity of each Newton step of the algorithm can be estimated as follows: In each Newton step, the algorithm involves evaluating the gradient, evaluating the Hessian, calculating the Newton descent direction, and performing a line search to find a new point, which requires $\mathbf{O}(1)$ evaluations of the barrier function. The overall complexity of evaluating the barrier function, its gradient, and Hessian is $\mathbf{O}(n^2(m+p)^3)$ as discussed at the end of section 3.1. The complexity of calculating the Newton descent direction is $\mathbf{O}(n^3)$. Therefore, the estimated complexity of each Newton step is roughly $\mathbf{O}(n^2(m+p)^3)+\mathbf{O}(n^3)$.

Suppose that m and p are of the same order, and $n \ll m$. Then each Newton step of a general-purpose SDP solver solving (1.1) requires $\mathbf{O}(m^6)$ arithmetic operations, while each Newton step of the interior path-following algorithm proposed in section 4 requires only $\mathbf{O}(n^2m^3)$. This is why the algorithms proposed in this paper are more efficient. Furthermore, we expect that when the ratio m/n is large enough, the algorithm proposed in section 4 will perform significantly better than the con-

ventional method. The argument here is based on the assumption that the number of iterations which the proposed path-following algorithm requires to solve (2.3) is roughly the same as the number of iterations which a general-purpose SDP solver requires to solve (1.1). According to the numerical experiments demonstrated in the next section, this assumption seems to hold in practice. As we will see, results of the numerical experiments agree with the arguments presented above.

7. Numerical experiments. Consider the standard block diagram for robustness analysis in Figure 7.1. The nominal system G is LTI and has a state space representation

$$\begin{aligned}\dot{x} &= Ax + B_1 w_1 + B_2 w_2,\\ z_1 &= C_1 x + D_{11} w_1 + D_{22} w_2,\\ z_2 &= C_2 x + D_{21} w_1 + D_{22} w_2,\end{aligned}$$

where A is an $m \times m$ Hurwitz matrix, and w_1, w_2, z_1, z_2 are vector-valued signals. Each of them has n components. The uncertainty Δ corresponds to a diagonal, gain bounded, linear time-varying operator. That is, if z_{2i} and w_{2i} denote the ith components of signals z_2 and w_2, respectively, then $z_{2i} = \delta_i(t) w_{2i}$, where $|\delta_i(t)| \le 1$ for all t. We note that the uncertain system described above captures a large class of practical problems [2].

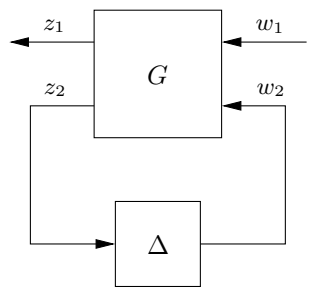


FIG. 7.1. *Standard block diagram for robustness analysis.*

In the experiments, we would like to compute an upper bound of the $\mathbf{L}_2$-gain of the system in Figure 7.1. By the standard integral quadratic constraint (IQC) analysis, an upper bound of the $\mathbf{L}_2$-gain can be found by solving

$$\begin{aligned}&\inf_{\lambda} && \lambda_{n+1}\\ &\text{subject to} && H(\omega,\lambda) > 0 \quad \forall\, \omega \in [0,\infty],\\ &&& \lambda_i > 0, \quad i = 1,\dots,n+1,\end{aligned} \tag{7.1}$$

where

$$H(\omega,\lambda) := \begin{bmatrix} G_{11}(j\omega) & G_{12}(j\omega)\\ G_{21}(j\omega) & G_{22}(j\omega)\\ I & 0\\ 0 & I \end{bmatrix}^* \begin{bmatrix} -I & 0 & 0 & 0\\ 0 & -\Lambda & 0 & 0\\ 0 & 0 & \lambda_{n+1} I & 0\\ 0 & 0 & 0 & \Lambda \end{bmatrix} \begin{bmatrix} G_{11}(j\omega) & G_{12}(j\omega)\\ G_{21}(j\omega) & G_{22}(j\omega)\\ I & 0\\ 0 & I \end{bmatrix},$$

$G_{rs}(j\omega) = C_r(j\omega I - A)^{-1} B_s + D_{rs}$, and $\Lambda = \operatorname{diag}(\lambda_1,\dots,\lambda_n)$. The equivalent KYP–

SDP of problem (7.1) can be expressed as

$$\begin{aligned} &\inf_{P,\lambda} && \lambda_{n+1} \\ &\text{subject to} && \mathcal{F}(P,\lambda) > 0, \\ &&& P = P', \quad \lambda_i > 0, \quad i = 1, \ldots, n+1, \end{aligned} \tag{7.2}$$

where P is a matrix variable and

$$\mathcal{F}(P,\lambda) := \begin{bmatrix} PA + A'P & PB \\ B'P & 0 \end{bmatrix} + \begin{bmatrix} C & D \\ 0 & I \end{bmatrix}' \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix} \begin{bmatrix} C & D \\ 0 & I \end{bmatrix}.$$

Matrices B, C, D, M_1, and M_2 are defined as follows:

$$B = \begin{bmatrix} B_1 & B_2 \end{bmatrix}, C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}, D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}, M_1 = \begin{bmatrix} -I & 0 \\ 0 & -\Lambda \end{bmatrix}, M_2 = \begin{bmatrix} \lambda_{n+1} I & 0 \\ 0 & \Lambda \end{bmatrix}.$$

Comparison with a general-purpose solver. Let $n = 10$ and $m = 10, 20, \ldots, 120$. For each pair of (n, m), 15 problems of the form (7.2) are randomly generated using an algorithm which is virtually identical to that given in Example 3 of [8], except for some minor numerical values. The algorithm is included in Appendix C for interested readers.

These problems are solved using the proposed specialized interior path-following algorithm (SIPA), and the general-purpose SDP solver SeDuMi (version 1.0.5) [24] via the YALMIP interface [16]. The experiments were performed on a 3.2 GHz Pentium IV PC with 2 GB of memory. The solvers were executed on MATLAB platform version 7.0.4.

Figure 7.2 shows the CPU times (in seconds) that SIPA and SeDuMi spent solving the randomly generated L_2-gain minimization problems. Times that SIPA spent are represented by circles, while times that SeDuMi spent are represented by triangles. The solvers stop when either infeasibility is proved or a feasible solution which achieves accuracy no worse than 10^{-4} (proved by a solution of the dual problem) is obtained. We observe that SIPA starts to outperform SeDuMi when the dimension of matrix A (indicated by the number m) reaches 60×60. Significantly better performance of SIPA is observed when m reaches 80 and, when m is equal to 110, SIPA is order-of-magnitude faster than SeDuMi. When m is equal to 120, SeDuMi encounters an out-of-memory problem, while it took only 6 to 12 minutes for SIPA to solve each of the 15 problems. The average CPU times in seconds that the two solvers spent are shown in Table 7.1. Table 7.2 shows the average CPU times per iteration (in seconds) that SIPA and SeDuMi spent solving the randomly generated L_2-gain minimization problems. We can see that, for SeDuMi, the per-iteration CPU time grows much faster, which conforms to the analysis given in section 6.

We note that SIPA is only crudely implemented in MATLAB language. Its custom-made codes are not optimized in any way and its efficiency can be further improved. Hence, the SIPA algorithm potentially could perform faster than it does now.

Comparison with a specialized solver. In this section, comparison is made with a specialized KYP–SDP solver, KYPD [25, 27, 28, 29]. KYPD is a collection of MATLAB functions which implement the dual method described in [25, 28, 29]. The main idea behind KYPD is to reformulate the KYP–SDP into its dual formulation. The dual problem is then solved using any general-purpose SDP solver. KYPD

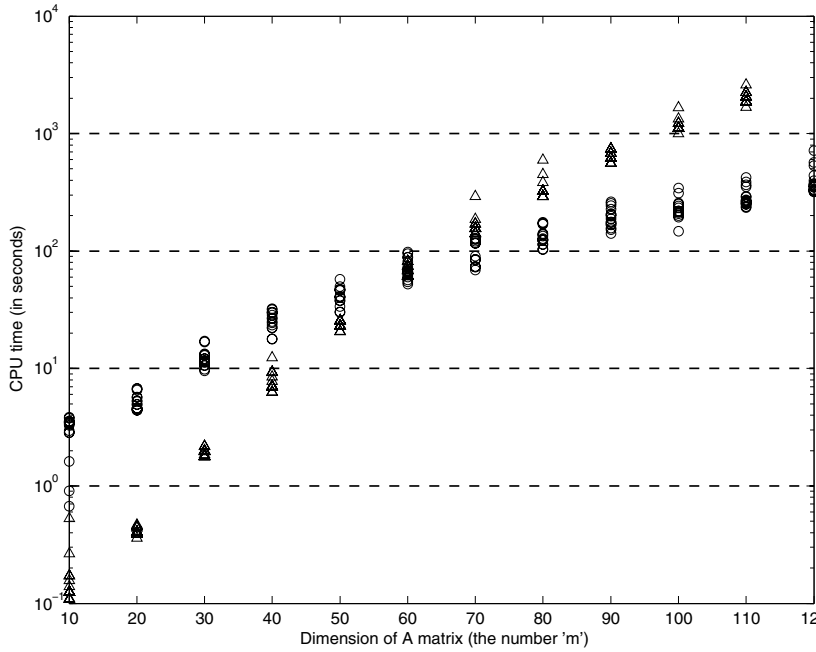


FIG. 7.2. *Comparison of SIPA and SeDuMi. Times that SIPA spent are represented by circles, while times that SeDuMi spent are represented by triangles.*

uses YALMIP to interface with MATLAB. KYPD comes with no SDP solver; the underlying SDP solver we used with KYPD was SeDuMi.

The same sets of problems were solved by KYPD/SeDuMi. The solver is set to terminate when either a feasible solution which achieves accuracy no worse than 10^{-4} is found or infeasibility is proved. Performances of KYPD and SIPA are shown in Figure 7.3. The average CPU times which the two solvers spent are given in Table 7.3, while the average per-iteration CPU times are shown in Table 7.4.

We observed that when m is between 20 and 50, SIPA and KYPD have rather similar performances. Noticeably better performance of SIPA is observed when m reaches 70. When m reaches 90, SIPA is about 2.5 times faster than KYPD on

TABLE 7.1

Comparison of SIPA and SeDuMi. Shown above are the average CPU times (in seconds) that SIPA and SeDuMi spent solving the randomly generated L_2-gain minimization problems.

	$m = 30$	40	50	60	70	80	90	100	110	120
SIPA	12.3	25.9	41.8	71.8	99.1	133	195	232	293	407
SeDuMi	1.9	7.9	23.7	68.8	165.1	341	640	1173	2033	—

TABLE 7.2

The average "per-iteration" CPU times (in seconds) that SIPA and SeDuMi used to solve the randomly generated L_2-gain minimization problems.

	$m = 30$	40	50	60	70	80	90	100	110	120
SIPA	0.26	0.49	0.82	1.28	2.02	2.59	3.42	4.29	5.73	7.46
SeDuMi	0.22	0.79	2.56	6.84	15.7	32.5	61.9	111.3	187.1	—

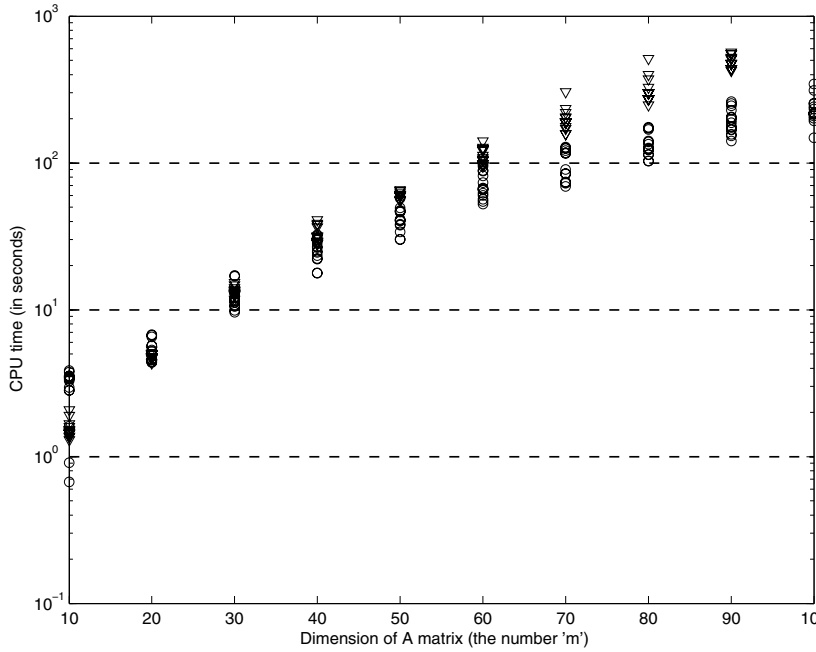


FIG. 7.3. *Comparison of SIPA and KYPD. Times that SIPA spent are represented by circles, while times that KYPD spent are represented by triangles.*

average. Also notice that, according to these tests, the average performance of KYPD is no better than the general-purpose solver SeDuMi when $m \leq 70$. KYPD began to outperform SeDuMi when m reaches 80. Another interesting observation is the memory usage of the three solvers. Under the current implementation, the memory required by SeDuMi and KYPD grows very fast with respect to m, the dimension of the A matrix. Given the same memory space (2 GB), KYPD was able to solve problems with A matrices up to 90×90, while SeDuMi holds up to 110×110. In contrast, SIPA appears to be the most efficient in this regard. The memory space which SIPA requires for all problems is below 400 MB.

Remark. To explain why KYPD broke down at $m = 100$, we note that the value of p (the number of columns of the B matrix) is 20 in all tests. When $m = 100$, the dual

TABLE 7.3

Comparison of SIPA and KYPD. Shown are the average CPU times (in seconds) that SIPA and KYPD spent solving the randomly generated L_2-gain minimization problems.

	$m = 10$	20	30	40	50	60	70	80	90	100
SIPA	2.9	5.3	12.3	25.9	41.8	71.8	99.1	132.5	194.8	231.8
KYPD	1.5	4.8	13.3	32.0	60.1	112.2	197.6	314.2	486.0	—

TABLE 7.4

The average "per-iteration" CPU times (in seconds) that SIPA and KYPD used to solve the randomly generated L_2-gain minimization problems.

	$m = 10$	20	30	40	50	60	70	80	90	100
SIPA	0.07	0.11	0.26	0.49	0.82	1.28	2.02	2.59	3.42	4.29
KYPD	0.18	0.65	1.77	3.89	7.65	13.04	20.93	33.07	52.61	—

formulation of the KYP–SDP involves a 120×120 positive-definite symmetric matrix variable Z which must satisfy certain equality constraints. The KYPD reduces the number of the variables in the dual formulation by calculating a set of bases for Z and by representing Z using those bases. When $m = 100$ and $p = 20$, the number of bases is equal to 2210. It was this huge amount of data—more than two thousand 120×120 matrices—which somehow ate up all the memory resource. As far as we can see, the program broke when YALMIP was processing the data for SeDuMi. We believe that the dual method described in [25, 28, 29] is computationally efficient, especially when p is very small; however, care is required for memory resource management when it comes to implementation.

Remark. By these test results, the authors by no means wish to claim that the proposed SIPA algorithm is superior to KYPD or any other specialized algorithms for KYP–SDP in that regard. The main purpose of this paper, as well as of the selected numerical experiments, is to demonstrate that there is much to gain in computational speed by taking into account the special structure of the KYP–SDP. As much as we agree that a careful comparison among the existing specialized algorithms for KYP–SDP is necessary and important, we believe such a comparison (in terms of computational speed, numerical stability, and required computational resources) must be done by testing the algorithms on a wider range of problems, which is beyond the scope of this paper.

The selection of μ. Recall that in section 5 we refer to the factor by which the weighting coefficient t is increased per outer iteration as μ. Here we present results of numerical experiments which show how the value of μ affects the speed of centralization and overall convergence.

Four randomly generated L_2-gain minimizations were solved using the SIPA algorithm with different values of μ. For different values of μ, the algorithm starts at the same feasible point in each test. In each outer iteration, centralization is considered completed when both of the following conditions are satisfied: (1) a primal feasible solution λ_F is found such that the number $\mathcal{G}(\lambda_F)^{-1}\nabla\mathcal{G}(\lambda_F)'v_N$ is less than 0.05 in absolute value, where v_N is the Newton descent direction at λ_F; (2) the solution λ_F and the corresponding Newton descent direction v_N allow a dual solution to be constructed using formula (5.1).

The results of the tests are shown in Table 7.5. Under each value of μ, the columns "Cen'l" show the average number of Newton steps per outer iteration; namely, these are the average numbers of Newton steps executed for computing a new primal feasible solution, which satisfies the above-mentioned conditions after the weighting coefficient t is increased by a factor of μ. The columns "Tot'l" show the numbers of total Newton steps executed for obtaining a feasible solution of desired accuracy. The observation here is consistent with what is commonly known for the path-following algorithm: the larger the μ is, the smaller the number of outer iterations is required for computing a solution with a prespecified accuracy. On the other hand, the larger the μ is, the larger the number of Newton steps is required for recentering. Hence, the selection of μ involves the trade-off between the number of total outer iterations and the number of Newton steps required for centralization. For the numerical experiments presented here, the most suitable value of μ is apparently between 50 and 100.

8. Concluding remarks. In this paper, a new interior-point method is proposed to efficiently solve the KYP–SDP. The main idea this paper proposes is *not* to solve the original KYP–SDP but to solve an equivalent semi-infinite optimization problem. The main technical contribution of this paper is to give a new bar-

TABLE 7.5
The effect of different values of μ on centralization and overall convergence.

μ	2		10		50		100		300	
	Cen'l	Tot'l	Cen'l	Tot'l	Cen'l	Tot'l	Cen'l	Tot'l	Cen'l	Tot'l
Test 1	1.2	48	2.3	27	3.0	21	3.5	21	4.8	24
Test 2	1.5	58	2.6	31	3.7	26	4.5	27	6.8	34
Test 3	1.4	56	2.7	32	3.7	26	5.0	30	6.2	31
Test 4	1.4	54	2.3	27	3.9	27	4.2	25	6.4	32

rier function for the semi-infinite optimization problem, which allows the standard path-following algorithm to be applied to solve the problem. Numerical experiments show that the proposed path-following method can solve the semi-infinite optimization problem much faster than a general-purpose SDP solver solves the corresponding KYP–SDP. The computational savings are mainly due to the fact that computing a Newton descent direction in the proposed path-following algorithm can be performed much more efficiently than in the algorithms used to solve the corresponding KYP–SDP.

Another natural candidate for the barrier function is the "inverse" barrier; i.e., the function $\mathcal{G}(\lambda)$ itself. We have also investigated this option [12, 13]. The reasons that we consider $\mathcal{G}(\lambda)$ less preferable are

- that function $\mathcal{G}(\lambda)$ is not self-concordant [12], and
- more important, that the primal central path associated with $\mathcal{G}(\lambda)$ does not give a good characterization of convergence.

To be more specific on the second point, we let $\lambda^\star(t)$ be at the primal central path associated with $\mathcal{G}(\lambda)$ and let $Z^\star(\omega, t)$ be the corresponding dual feasible solution which satisfies

$$\frac{dZ^\star(\omega,t)}{d\omega} = \frac{1}{\pi t}\frac{1}{1+\omega^2}\mathbf{H}^{-2}(\omega,\lambda^\star(t)).$$

It can be easily verified that the duality gap $c'\lambda^\star(t) + \langle \mathbf{H}_0(\omega), Z^\star(\omega,t)\rangle$ is equal to $\mathcal{G}(\lambda^\star(t))/t$. Since $\mathcal{G}(\lambda)$ approaches infinity as λ approaches the boundary of the feasible set, the quantity $\mathcal{G}(\lambda^\star(t))/t$ does not give a useful measure of the optimality of $\lambda^\star(t)$. Although this may not imply that $\mathcal{G}(\lambda)$ is inferior to $\log \mathcal{G}(\lambda)$ in practice, we prefer the latter because it offers a better measure of global convergence with respect to t.

Appendix A. Proof of Proposition 3.2. Throughout this section, the notation

$$G := \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$$

is used to denote a rational transfer matrix $G(s) = C(sI - A)^{-1}B + D$. Note that $G^{-1}(s)$ exists if and only if D is invertible. In such cases,

$$G^{-1} = \left[\begin{array}{c|c} A - BD^{-1}C & BD^{-1} \\ \hline -D^{-1}C & D^{-1} \end{array}\right].$$

Furthermore, given two rational transfer matrices of compatible dimensions

$$G_1 := \left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right], \quad G_2 := \left[\begin{array}{c|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right],$$

the following formulas hold:

$$G_1+G_2=\left[\begin{array}{cc|c} A_1 & 0 & B_1 \\ 0 & A_2 & B_2 \\ \hline C_1 & C_2 & D_1+D_2 \end{array}\right], \quad G_1G_2=\left[\begin{array}{cc|c} A_1 & B_1C_2 & B_1D_2 \\ 0 & A_2 & B_2 \\ \hline C_1 & D_1C_2 & D_1D_2 \end{array}\right].$$

Now consider Proposition 3.2. Given $\lambda_F \in \Omega$, $\mathbf{H}(\omega,\lambda_F)$ is a matrix function on ω which has the form

$$\begin{bmatrix} (j\omega I-A)^{-1} \\ B \end{bmatrix}^* \begin{bmatrix} Q_F & S_F \\ S_F' & R_F \end{bmatrix} \begin{bmatrix} (j\omega I-A)^{-1} \\ B \end{bmatrix}.$$

Hence, $\mathbf{H}(\omega,\lambda_F)$ can also be expressed as $\Phi(j\omega)$, where $\Phi(s)$ is the rational transfer matrix

$$\Phi := \left[\begin{array}{cc|c} A & 0 & B \\ Q_F & -A' & S_F \\ \hline S_F' & -B' & R_F \end{array}\right].$$

Let P be a solution of the Riccati equation (3.13), and observe that

$$\begin{bmatrix} I & 0 \\ P & I \end{bmatrix} \begin{bmatrix} A & 0 \\ Q_F & -A' \end{bmatrix} \begin{bmatrix} I & 0 \\ P & I \end{bmatrix}^{-1} = \begin{bmatrix} A & 0 \\ (PB+S_F)R_F^{-1}(PB+S_F)' & -A' \end{bmatrix},$$

$$\begin{bmatrix} I & 0 \\ P & I \end{bmatrix} \begin{bmatrix} B \\ S_F \end{bmatrix} = \begin{bmatrix} B \\ PB+S_F \end{bmatrix}, \quad \begin{bmatrix} S_F' & -B' \end{bmatrix} \begin{bmatrix} I & 0 \\ P & I \end{bmatrix}^{-1} = \begin{bmatrix} (PB+S_F)' & -B' \end{bmatrix}.$$

Therefore, the rational transfer matrix $\Phi(s)$ can also be expressed as

$$\left[\begin{array}{cc|c} A & 0 & B \\ (PB+S_F)R_F^{-1}(PB+S_F)' & -A' & PB+S_F \\ \hline (PB+S_F)' & -B' & R_F \end{array}\right],$$

which in turn can be factorized as $\Phi^{\frac{1}{2}}(-s)'\Phi^{\frac{1}{2}}(s)$, where

$$\Phi^{\frac{1}{2}}(s) := \left[\begin{array}{c|c} A & B \\ \hline (R_F^{-\frac{1}{2}})'(PB+S_F)' & R_F^{\frac{1}{2}} \end{array}\right].$$

Now, it can be readily verified that $\mathbf{H}(\omega,\lambda_F)^{-1}$ is equal to $\Phi(j\omega)^{-1}$, which in turn is equal to $\Psi(j\omega)\Psi(j\omega)^*$, where

$$\Psi(s)=\Phi^{-\frac{1}{2}}(s)=\left[\begin{array}{c|c} A-BR_F^{-1}(PB+S_F)' & BR_F^{-\frac{1}{2}} \\ \hline -R_F^{-1}(PB+S_F)' & R_F^{-\frac{1}{2}} \end{array}\right]=\left[\begin{array}{c|c} A_H & BR_F^{-\frac{1}{2}} \\ \hline C_H & R_F^{-\frac{1}{2}} \end{array}\right].$$

This proves the second factorization. To prove the first factorization, we will show that

$$D_H+G_H(-s)'+G_H(s)=\Psi(s)\Psi(-s)'.$$

Note that

$$\Psi(s)\Psi(-s)'=\left[\begin{array}{cc|c} A_H & -BR_F^{-1}B' & BR_F^{-1} \\ 0 & -A_H' & C_H' \\ \hline C_H & -R_F^{-1}B' & R_F^{-1} \end{array}\right].$$

Let Y be a solution of the Lyapunov equation (3.14). We see

$$\begin{bmatrix} I & Y \\ 0 & I \end{bmatrix} \begin{bmatrix} A_H & -BR_F^{-1}B' \\ 0 & -A_H' \end{bmatrix} \begin{bmatrix} I & Y \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} A_H & 0 \\ 0 & -A_H' \end{bmatrix},$$
$$\begin{bmatrix} I & Y \\ 0 & I \end{bmatrix} \begin{bmatrix} BR_F^{-1} \\ C_H' \end{bmatrix} = \begin{bmatrix} BR_F^{-1} + YC_H' \\ C_H' \end{bmatrix} = \begin{bmatrix} B_H \\ C_H' \end{bmatrix},$$
$$\begin{bmatrix} C_H & -R_F^{-1}B' \end{bmatrix} \begin{bmatrix} I & Y \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} C_H & -C_HY - R_F^{-1}B' \end{bmatrix} = \begin{bmatrix} C_H & -B_H' \end{bmatrix}.$$

Therefore

$$\Psi(s)\Psi(-s)' = \left[\begin{array}{cc|c} A_H & 0 & B_H \\ 0 & -A_H' & C_H' \\ \hline C_H & -B_H' & R_F^{-1} \end{array}\right] = D_H + G_H(s) + G_H(s)'.$$

Finally, existence of a solution P for (3.13) such that $A_H := A - BR_F^{-1}(PB + S_F)'$ is Hurwitz is guaranteed as long as $\lambda_F \in \Omega$. Proof of this is established by results in Chapters 13.2–13.4 of [33]. Existence of a solution Y for (3.14) is guaranteed because A_H has no eigenvalue on the imaginary axis.

Appendix B. Proof of Theorem 4.1. The following two lemmas will be used in the proof.

LEMMA B.1. *Let ω_0 be a real number and $\beta \geq 2$ be a positive even integer. Let*

$$G(t,\omega) = \frac{r_0(t) + r_1(t)(\omega-\omega_0)^\beta + \cdots + r_{n-1}(t)(\omega-\omega_0)^{\beta(n-1)}}{(p(t) + (\omega-\omega_0)^\beta q(t))^n}, \tag{B.1}$$

where $p(t)$, $q(t)$, $r_i(t)$, $i = 0, \ldots, n-1$, are polynomial functions in t. Consider the integral

$$F(t) := \int_{\omega_0-\epsilon}^{\omega_0+\epsilon} G(t,\omega)d\omega, \tag{B.2}$$

where ϵ is a positive number. $F(t)$ can be expressed as

$$F(t) = \sum_{i=0}^{n-1} c_i(t;\epsilon) \cdot r_i(t)p(t)^{i+\frac{1}{\beta}-n} q(t)^{-(i+\frac{1}{\beta})}, \tag{B.3}$$

where each $c_i(t;\epsilon)$ is positive and bounded for all t and ϵ.

Proof. Fix t and let

$$\omega = \omega_0 + \left(\frac{p(t)}{q(t)}\right)^{\frac{1}{\beta}} \widetilde{\omega}.$$

We have

$$\begin{aligned} \int_{\omega_0-\epsilon}^{\omega_0+\epsilon} \frac{r_i(t)(\omega-\omega_0)^{\beta\cdot i}}{(p(t)+(\omega-\omega_0)^\beta q(t))^n}\, d\omega &= \int_{-\eta}^{\eta} \frac{r_i(t)p(t)^{i+\frac{1}{\beta}} q(t)^{-(i+\frac{1}{\beta})} \widetilde{\omega}^{\beta i}}{p(t)^n(1+\widetilde{\omega}^\beta)^n}\, d\widetilde{\omega} \\ &= r_i(t)p(t)^{i+\frac{1}{\beta}-n} q(t)^{-(i+\frac{1}{\beta})} \int_{-\eta}^{\eta} \frac{\widetilde{\omega}^{\beta i}}{(1+\widetilde{\omega}^\beta)^n}\, d\widetilde{\omega}, \end{aligned}$$

where $\eta = \epsilon \left(\frac{p(t)}{q(t)}\right)^{-\frac{1}{\beta}}$. Now we define

$$c_i(t;\epsilon) := \int_{-\eta}^{\eta} \frac{\widetilde{\omega}^{\beta i}}{(1+\widetilde{\omega}^{\beta})^n}\, d\widetilde{\omega}, \tag{B.4}$$

and we reach the expression of $F(t)$. Finally, to see that $c_i(t;\epsilon)$ is bounded, note that the following inequalities hold for all $i = 0, \ldots, n-1$:

$$\int_{-\eta}^{\eta} \frac{\widetilde{\omega}^{\beta i}}{(1+\widetilde{\omega}^{\beta})^n}\, d\widetilde{\omega} \le \int_{-\infty}^{\infty} \frac{\widetilde{\omega}^{\beta i}}{(1+\widetilde{\omega}^{\beta})^n}\, d\widetilde{\omega} \le \int_{-\infty}^{\infty} \frac{1}{(1+\widetilde{\omega}^{\beta})^{n-i}}\, d\widetilde{\omega} = \text{constant}.$$

This concludes the proof. ∎

LEMMA B.2. *Let $\alpha \ge 1$ be an integer and $\beta \ge 2$ be an even integer. Let*

$$G(t,\omega) = \frac{r(t)}{t^{\alpha}p(t) + (\omega-\omega_0)^{\beta}q(t)}, \quad F(t) = \int_{\omega_0-\epsilon}^{\omega_0+\epsilon} G(t,\omega)\, d\omega,$$

where ϵ is a given positive number. Functions $p(t)$, $q(t)$, $r(t)$ are polynomials in t which satisfy $r(0) \neq 0$, $p(0) \neq 0$, and $q(0) \neq 0$. Then

$$F(t) = t^{-n}s_0(t), \quad \dot{F}(t) = t^{-n-1}s_1(t), \tag{B.5}$$

$$\ddot{F}(t) = t^{-n-2}s_2(t), \quad \dddot{F}(t) = t^{-n-3}s_3(t), \tag{B.6}$$

where $n = \alpha - \frac{\alpha}{\beta}$, and each $s_i(t)$ satisfies $s_i(0) \neq 0$ and $s_i(0) < \infty$.

Proof. Notice that $G(t)$ is in the form of (B.1). The corresponding $r_0(t)$, $p(t)$, $q(t)$, and n are $r(t)$, $t^{\alpha}p(t)$, $q(t)$, and 1, respectively. Therefore, by Lemma B.1, we have

$$F(t) = c_0(t;\epsilon)r(t)(t^{\alpha}p(t))^{\frac{1}{\beta}-1}q(t)^{-\frac{1}{\beta}} = t^{-\alpha+\frac{\alpha}{\beta}}c(t)r(t)p(t)^{\frac{1}{\beta}-1}q(t)^{-\frac{1}{\beta}},$$

which is exactly of the form described in (B.5) with $s_0(t) = c_0(t;\epsilon)r(t)p(t)^{\frac{1}{\beta}-1}q(t)^{-\frac{1}{\beta}}$. Note that $c_0(t;\epsilon)$ is of the form (B.4). Since $p(0) \neq 0$, $c_0(0;\epsilon)$ is strictly positive and bounded. This, together with $q(0) \neq 0$ and $r(0) \neq 0$, implies that $s_0(0) \neq 0$ and is bounded.

Now, consider the derivatives of $F(t)$. We have

$$\dot{F}(t) = \int_{\omega_0-\epsilon}^{\omega_0+\epsilon} \dot{G}(t,\omega)\, d\omega, \quad \ddot{F}(t) = \int_{\omega_0-\epsilon}^{\omega_0+\epsilon} \ddot{G}(t,\omega)\, d\omega, \quad \dddot{F}(t) = \int_{\omega_0-\epsilon}^{\omega_0+\epsilon} \dddot{G}(t,\omega)\, d\omega,$$

where

$$\dot{G}(t,\omega) = \frac{d_{10}(t) + d_{11}(t)(\omega-\omega_0)^{\beta}}{(t^{\alpha}p(t) + (\omega-\omega_0)^{\beta}q(t))^2},$$

$$\ddot{G}(t,\omega) = \frac{d_{20}(t) + d_{21}(t)(\omega-\omega_0)^{\beta} + d_{22}(t)(\omega-\omega_0)^{2\beta}}{(t^{\alpha}p(t) + (\omega-\omega_0)^{\beta}q(t))^3},$$

$$\dddot{G}(t,\omega) = \frac{d_{30}(t) + d_{31}(t)(\omega-\omega_0)^{\beta} + d_{32}(t)(\omega-\omega_0)^{2\beta} + d_{33}(t)(\omega-\omega_0)^{3\beta}}{(t^{\alpha}p(t) + (\omega-\omega_0)^{\beta}q(t))^4},$$

and $d_{ij}(t)$ are polynomial functions in t. We note that

$$d_{10}(t) = \dot{r}t^{\alpha}p - r(\alpha t^{\alpha-1}p + t^{\alpha}\dot{p}) = -\alpha t^{\alpha-1}rp + \mathbf{P}(t^{\alpha}),$$

$$d_{20}(t) = \dot{d}_{10}t^{\alpha}p - 2d_{10}(\alpha t^{\alpha-1}p + t^{\alpha}\dot{p}) = \alpha(\alpha+1)t^{2\alpha-2}rp^2 + \mathbf{P}(t^{2\alpha-1}),$$

$$d_{30}(t) = \dot{d}_{20}t^{\alpha}p - 3d_{20}(\alpha t^{\alpha-1}p + t^{\alpha}\dot{p}) = -\alpha(\alpha+1)(\alpha+2)t^{3\alpha-3}rp^3 + \mathbf{P}(t^{3\alpha-2}).$$

Here we use $\mathbf{P}(t^k)$ to denote any polynomial in t, where each term has at least power k. Furthermore,

$$d_{11}(t) = \dot{r}q - r\dot{q} = \mathbf{P}(t^{\gamma_1}),$$

$$d_{21}(t) = t^\alpha \dot{d}_{11}p + \dot{d}_{10}q - 2(d_{11}(\alpha t^{\alpha-1}p + t^\alpha \dot{p}) + d_{10}\dot{q}) = \begin{cases} \mathbf{P}(t^{\alpha-2}) & \text{if } \alpha \geq 2, \\ \mathbf{P}(t^0) & \text{if } \alpha = 1, \end{cases}$$

$$d_{22}(t) = \dot{d}_{11}q - 2d_{11}\dot{q} = \mathbf{P}(t^{\gamma_2}),$$

$$d_{31}(t) = t^\alpha \dot{d}_{21}p + \dot{d}_{20}q - 3(d_{21}(\alpha t^{\alpha-1}p + t^\alpha \dot{p}) + d_{20}\dot{q}) = \begin{cases} \mathbf{P}(t^{2\alpha-3}) & \text{if } \alpha \geq 2, \\ \mathbf{P}(t^0) & \text{if } \alpha = 1, \end{cases}$$

$$d_{32}(t) = t^\alpha \dot{d}_{22}p + \dot{d}_{21}q - 3(d_{22}(\alpha t^{\alpha-1}p + t^\alpha \dot{p}) + d_{21}\dot{q}) = \begin{cases} \mathbf{P}(t^{\alpha-3}) & \text{if } \alpha \geq 3, \\ \mathbf{P}(t^0) & \text{if } \alpha = 2, 1, \end{cases}$$

$$d_{33}(t) = \dot{d}_{22}q - 3d_{22}\dot{q} = \mathbf{P}(t^{\gamma_3}),$$

where γ_i are some nonnegative integers. By Lemma B.1,

$$\dot{F}(t) = c_0(t;\epsilon)d_{10}(t)t^{-2\alpha+\frac{\alpha}{\beta}}p(t)^{\frac{1}{\beta}-2}q(t)^{-\frac{1}{\beta}} + c_1(t;\epsilon)d_{11}(t)t^{-\alpha+\frac{\alpha}{\beta}}p(t)^{\frac{1}{\beta}-1}q(t)^{-(\frac{1}{\beta}+1)}.$$

Substituting $d_{10}(t)$ and $d_{11}(t)$ in the above expression, we obtain

$$\begin{aligned} \dot{F}(t) &= t^{-2\alpha+\frac{\alpha}{\beta}}(-\alpha t^{\alpha-1}r(t)p(t) + \mathbf{P}(t^\alpha))c_0(t;\epsilon)p(t)^{\frac{1}{\beta}-2}q(t)^{-\frac{1}{\beta}} \\ &\quad + t^{-\alpha+\frac{\alpha}{\beta}}\mathbf{P}(t^{\gamma_1})c_1(t;\epsilon)p(t)^{\frac{1}{\beta}-1}q(t)^{-(\frac{1}{\beta}+1)} \\ &= t^{-n-1}(-\alpha c_0(t;\epsilon)r(t)p(t)^{\frac{1}{\beta}-1}q(t)^{-\frac{1}{\beta}} + \mathbf{P}(t)c_0(t;\epsilon)p(t)^{\frac{1}{\beta}-2}q(t)^{-\frac{1}{\beta}} \\ &\qquad + \mathbf{P}(t^{\gamma_1+1})c_1(t;\epsilon)p(t)^{\frac{1}{\beta}-1}q(t)^{-(\frac{1}{\beta}+1)}) \\ &= t^{-n-1}s_1(t). \end{aligned}$$

Note that $s_1(0) = -\alpha c_0(0;\epsilon)r(0)^2p(0)^{\frac{1}{\beta}-1}q(0)^{-\frac{1}{\beta}} = -\alpha s_0(0)$. Hence, $s_1(0) \neq 0$ and is bounded above. This concludes that $\dot{F}(t)$ is of the form described in (B.5). A similar derivation leads to the forms described in (B.6) for $\ddot{F}(t)$ and $\dddot{F}(t)$, where $s_2(t)$ and $s_3(t)$ satisfy $s_2(0) \neq 0$ and $s_3(0) \neq 0$, and where both are bounded above. ∎

Proof of Theorem 4.1. We are now ready to prove Theorem 4.1. Recall that

$$\mathcal{B}(\lambda) = \log(\mathcal{G}(\lambda)) \quad \text{and} \quad \mathcal{G}(\lambda) = \left(\frac{1}{\pi}\int_{-\infty}^{\infty} \mathbf{tr}(\mathbf{H}(\omega,\lambda)^{-1})\frac{d\omega}{1+\omega^2}\right).$$

Given any $\lambda \in \Omega$ and any $h \in \mathbf{R}^n$, let T be the bounded open interval $\{t : \lambda + th \in \Omega\}$. Now, define $F(t) : T \to \mathbf{R} := \mathcal{G}(\lambda + th)$ and $E(t) : T \to \mathbf{R} := \mathcal{B}(\lambda + th) = \log(F(t))$. Let $\gamma(t) := \dddot{E}(t)^2/\ddot{E}(t)^3$. The idea is to show that the supremum of $\gamma(t)$ over T is bounded from above. Since $\gamma(t)$ is a continuous function, this property would imply that $\gamma(t)$ is finite as t approaches the boundary of T, which in turn implies that (4.4) holds.

Note that one may express $\mathbf{tr}(\mathbf{H}(\omega,\lambda+th)^{-1})/(\pi(1+\omega^2))$ as $r(\omega,t)/s(\omega,t)$, where $r(\omega,t)$ and $s(\omega,t)$ are polynomials in ω and t. Without loss of generality, let us assume that 0 is a boundary point of T and $\mathbf{H}(\omega,\lambda+th)$ is singular at $\omega = \omega_1,\ldots,\omega_n$ at $t = 0$, where each $\omega_i < \infty$. In the case when $\mathbf{H}(\omega,\lambda+th)$ has singularity at infinity, $\mathbf{tr}(\mathbf{H}(\omega,\lambda+th)^{-1})/(\pi(1+\omega^2))$ can be expressed as $r(\omega,t)/s(\omega,t) + 1/(t^\nu \cdot \pi(1+\omega^2))$, and the analysis is completely analogous to what is presented below.

Under these assumptions, we have the following expression for $s(\omega,t)$:

$$s(\omega,t) = t^{\alpha}p(\omega,t) + q(\omega,t)\Pi_{i=1}^{n}(\omega-\omega_i)^{\beta_i},$$

where α is an integer greater than or equal to 1 and each β_i is an even integer greater than or equal to 2. Furthermore, $r(\omega,t)$, $p(\omega,t)$, and $q(\omega,t)$ satisfy $r(\omega_i,0)\neq 0$, $p(\omega_i,0)\neq 0$, and $q(\omega_i,0)\neq 0$ for $i=1,\ldots,n$. Let

$$G_i(\omega,t) = \frac{r_i(t)}{t^{\alpha}p_i(t) + q_i(t)(\omega-\omega_i)^{\beta_i}},$$

where $r_i(t) := r(\omega_i,t)$, $p_i(t) := p(\omega_i,t)$, and $q_i(t) := q(\omega_i,t)$. Let ϵ be a small positive number and $\Sigma = \bigcup_{k=1}^{n}[\omega_k-\epsilon,\ \omega_k+\epsilon]$. Note that when ϵ is sufficiently small, $r(\omega,t)/s(\omega,t) \approx G_i(\omega,t)$ for $\omega\in[\omega_i-\epsilon,\omega_i+\epsilon]$. Therefore, the following hold:

$$F(t) = \sum_{i=1}^{n}\int_{\omega_i-\epsilon}^{\omega_i+\epsilon}\frac{r(\omega,t)}{s(\omega,t)}d\omega + \int_{[-\infty,\ \infty]\setminus\Sigma}\frac{r(\omega,t)}{s(\omega,t)}d\omega \tag{B.7}$$

$$\approx \sum_{i=1}^{n}\int_{\omega_i-\epsilon}^{\omega_i+\epsilon} G_i(\omega,t)d\omega + M(t), \tag{B.8}$$

where $M(t)\leq\mathcal{M}<\infty$ for all $t\in T$. To see this, we note that $s(\omega,t)$ is bounded away from 0 for all $t\in T$ and for all $\omega\in[-\infty,\ \infty]\setminus\Sigma$; therefore, the second integral in (B.7) is bounded above for all $t\in T$.

Now consider the kth derivative of $F(t)$. We have

$$\begin{aligned}\frac{d^k}{dt^k}F(t) &= \sum_{i=1}^{n}\int_{\omega_i-\epsilon}^{\omega_i+\epsilon}\frac{d^k}{dt^k}\left(\frac{r(\omega,t)}{s(\omega,t)}\right)d\omega + \int_{[-\infty,\ \infty]\setminus\Sigma}\frac{d^k}{dt^k}\left(\frac{r(\omega,t)}{s(\omega,t)}\right)d\omega\\ &= \sum_{i=1}^{n}\int_{\omega_i-\epsilon}^{\omega_i+\epsilon}\frac{\bar{r}_k(\omega,t)}{s(\omega,t)^k}d\omega + \int_{[-\infty,\ \infty]\setminus\Sigma}\frac{\bar{r}_k(\omega,t)}{s(\omega,t)^k}d\omega,\end{aligned}$$

where $\bar{r}_k(\omega,t)$ is a polynomial in ω and t. Similarly, since $s(\omega,t)^k$ is bounded away from 0 for all $t\in T$ and for all $\omega\in[-\infty,\ \infty]\setminus\Sigma$, and for any $\bar{t}\in T$,

$$\frac{d^k}{dt^k}\left(\frac{r(\omega,\bar{t})}{s(\omega,\bar{t})}\right) \approx \frac{d^k}{dt^k}G_i(\omega,\bar{t}) \quad \text{for } \omega\in[\omega_i-\epsilon,\omega_i+\epsilon],$$

we conclude that

$$\frac{d^k}{dt^k}F(t) \approx \sum_{i=1}^{n}\int_{\omega_i-\epsilon}^{\omega_i+\epsilon}\frac{d^k}{dt^k}G_i(\omega,t)d\omega + M_k(t), \tag{B.9}$$

where $M_k(t)\leq\mathcal{M}_k<\infty$ for all $t\in T$. Now, let

$$F_i(t) = \int_{\omega_i-\epsilon}^{\omega_i+\epsilon} G_i(\omega,t)d\omega, \quad i=1,\ldots,n.$$

By virtue of Lemma B.2, we have $F_i(t) = t^{-m_i}s_{i0}(t)$, $\dot{F}_i(t) = t^{-m_i-1}s_{i1}(t)$, $\ddot{F}_i(t) = t^{-m_i-2}s_{i2}(t)$, $\dddot{F}_i(t) = t^{-m_i-3}s_{i3}(t)$, where $m_i = \alpha - \frac{\alpha}{\beta_k}$, and each $s_{ij}(t)$, $j=0,\ldots,3$,

satisfies that $s_{ij}(0) \neq 0$ and is bounded above. Without loss of generality, let us assume $m_1 \geq m_2 \geq \cdots \geq m_n$. Then (B.8) and (B.9) imply that

$$F(t) \approx t^{-m_1} s_0(t), \quad \dot{F}(t) \approx t^{-m_1-1} s_1(t), \tag{B.10}$$

$$\ddot{F}(t) \approx t^{-m_1-2} s_2(t), \quad \dddot{F}(t) \approx t^{-m_1-3} s_3(t). \tag{B.11}$$

Again, in (B.10) and (B.11), each $s_i(t)$, $i = 0, \ldots, 3$, satisfies that $s_i(0) \neq 0$ and is bounded above. Now, consider $E(t) = \log F(t)$. It can be readily verified that

$$\gamma(t) := \frac{\dddot{E}(t)^2}{\ddot{E}(t)^3} = \frac{(F(t)^2 \dddot{F}(t) - 3F(t)\dot{F}(t)\ddot{F}(t) + 2\dot{F}(t)^3)^2}{(F(t)\ddot{F}(t) - \dot{F}(t)^2)^3}. \tag{B.12}$$

Substituting (B.10) and (B.11) in (B.12), we obtain

$$\gamma(t) \approx \frac{t^{-6m_1-6}(s_0(t)^2 s_3(t) - 3s_0(t)s_1(t)s_2(t) + 2s_1(t)^3)^2}{t^{-6m_1-6}(s_0(t)s_2(t) - s_1(t)^2)^3}.$$

Therefore, as $t \to 0$, we have

$$\gamma(0) \approx \frac{(s_0(0)^2 s_3(0) - 3s_0(0)s_1(0)s_2(0) + 2s_1(0)^3)^2}{(s_0(0)s_2(0) - s_1(0)^2)^3},$$

which is a finite number. This shows that the supremum of $\dddot{E}(t)^2/\ddot{E}(t)^3$ over T is bounded, which in turn implies that $\mathcal{B}(\lambda)$ satisfies (4.4).

Appendix C. The algorithm for generating the random systems in section 7. The random systems in section 7 were generated using an algorithm which is virtually identical to that given in Example 3 of [8]. The algorithm works as follows. Let ν_{ij}, $i = 1, \ldots, 10$, $j = 1, \ldots, l$, be drawn from a uniform distribution on $[-\nu_l^0, -\nu_l^0 + 1]$, where ν_l^0, $l = 1, \ldots, 12$, have the following numerical values, respectively: 4.8, 3.0, 2.5, 2.1, 2.0, 1.95, 1.85, 1.85, 1.84, 1.83, 1.82, 1.81. Define $\bar{A}_i = \operatorname{diag}_j(\nu_{ij}) + A_0$, where A_0 is an $l \times l$ matrix of zeros, except for the first superdiagonal, which has all elements equal to 1. Then let $\bar{A} = \operatorname{diag}_i(\bar{A}_i)$. For $\bar{B}$, we first define $\bar{B}_i \in \mathbf{R}^{l \times 1}$ to be $\begin{bmatrix} 0 & 0 & \cdots & 1 \end{bmatrix}'$ and let $\bar{B}_t = \operatorname{diag}_i(\bar{B}_i)$. Finally, define $\bar{B} = \begin{bmatrix} B_t & 2B_t \end{bmatrix}$. For $\bar{C}$, the construction is similar: first we introduce $\bar{C}_i = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \in \mathbf{R}^{1 \times l}$ and let $\bar{C}_t = \operatorname{diag}_i(\bar{C}_i)$. Finally, let $\bar{C} = [\bar{C}_t \quad 2\bar{C}_t]$. The dimensions of $\bar{A}$, $\bar{B}$, and $\bar{C}$ are $m \times m$, $m \times p$, and $p \times m$, where $m = 10l$ and $p = 20$. By changing l, we vary the dimensions of the systems from 10 to 120. In order to couple all inputs with all outputs, random unitary transformations are introduced. For this, we produce two matrices $\Upsilon_m \in \mathbf{R}^{m \times m}$ and $\Upsilon_p \in \mathbf{R}^{p \times p}$, whose entries are drawn from a uniform distribution on $[0, 1]$. Let the singular value decompositions of Υ_m and Υ_p be $U_m \Sigma_m V_m'$ and $U_p \Sigma_p V_p'$, respectively. Finally, we define $A = U_m' \bar{A} U_m$, $B = U_m' \bar{B} V_p$, and $C = U_p' \bar{C} U_m$. The random systems in section 7 are realized by the pair of matrices $(A, B, C, 0_{p \times p})$.

Acknowledgments. The first author would like to thank Ulf Jönsson of the Division of Optimization and Systems Theory, KTH, and Hisaya Fujioka of the Department of Applied Analysis and Complex Dynamical Systems, Kyoto University, for their valuable comments on this work.

REFERENCES

[1] B. Alkire and L. Vandenberghe, *Convex optimization problems involving finite autocorrelation sequences,* Math. Program. Ser. A, 93 (2002), pp. 331–359.

[2] G. J. Balas, J. C. Doyle, K. Glover, A. Packard, and R. Smith, *μ-Analysis and Synthesis Toolbox*, The Math Works Inc, Natick, MA, 1993.

[3] S. Boyd, L. EL Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[5] J. Doyle, A. Packard, and K. Zhou, *Review of LFTs, LMIs, and μ*, in Proceedings of the 30th IEEE Conference on Decision and Control, Brighton, England, 1991, pp. 1227–1232.

[6] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968.

[7] A. Hansson and L. Vandenberghe, *Efficient solution of linear matrix inequalities for integral quadratic constraints*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sidney, Australia, 2000, pp. 5033–5034.

[8] A. Hansson and L. Vandenberghe, *A primial-dual potential reduction method for integral quadratic constraints*, in Proceedings of the 2001 American Control Conference, Arlington, VA, 2001, pp. 3013–3018.

[9] H. Hinfi, B. Hassibi, and S. Boyd, *Multiobjective H_2/H_∞ optimal control via finite-dimensional Q-parametrization and linear matrix inequalities*, in Proceedings of the 1998 American Control Conference, Philadelphia, PA, 1998, pp. 3244–3249.

[10] U. Jönsson and A. Rantzer, *Duality bonds in robustness analysis*, Automatica, 33 (1997), pp 1835–1844.

[11] U. Jönsson, *Duality in multiplier-based robustness analysis*, IEEE Trans. Automat. Control, 44 (1999), pp. 2246–2256.

[12] C.-Y. Kao, *Efficient Computational Algorithms for Robustness Analysis*, Ph.D. thesis, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 2002.

[13] C.-Y. Kao and A. Megretski, *Specialized fast algorithms for IQC optimization problems*, in Proceedings of the 2001 American Control Conference, Arlington, VA, 2001, pp. 3019–3024.

[14] C.-Y. Kao, A. Megretski, and U. Jönsson, *An algorithm for solving special frequency dependent LMIs*, in Proceedings of the 2000 American Control Conference, Chicago, IL, 2000, pp. 307–311.

[15] C.-Y. Kao, A. Megretski, and U. Jönsson, *Specialized fast algorithms for IQC feasibility and optimization problems*, Automatica, 40 (2004), pp. 239–252.

[16] J. Löfberg, *YALMIP: A toolbox for modeling and optimization in MATLAB*, in Proceedings of the 2004 IEEE CACSD, Taipei, Taiwan, 2004, pp. 284–289.

[17] D. G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[18] A. Megretski and A. Rantzer, *System analysis via integral quadratic constraints*, IEEE Trans. Automat. Control, 42 (1997), pp. 819–830.

[19] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.

[20] P. A. Parrilo, *On the numerical solution of LMIs derived from the KYP lemma*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 2334–2338.

[21] A. Rantzer, *On the Kalman–Yakubovich–Popov lemma*, Systems Control Lett., 28 (1996), pp. 7–10.

[22] J. Renegar, *A polynomial-time algorithm, based on Newton's method, for linear programming*, Math. Programming, 40 (1988), pp. 59–93.

[23] M. G. Safonov and M. Athans, *A multi-loop generalization of the circle criterion for stability margin analysis*, IEEE Trans. Automat. Control, 26 (1981), pp. 415–422.

[24] J. F. Sturm, *Using SeDuMi, a MATLAB Toolbox for Optimization over Symmetric Cones*, citeseer.ist.psu.edu/sturm99using.html (1999).

[25] L. Vandenberghe, V. R. Balakrishnan, R. Wallin, and A. Hansson, *On the implementation of primal-dual interior-point methods for semidefinite programming problems derived from the KYP lemma*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 4658–4663.

[26] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.

[27] R. Wallin, *User's Guide to KYPD Solver*, www.control.isy.liu.se/research/authors/reports/2517/kypd.html (2003).

[28] R. Wallin, *Optimization Algorithms for System Analysis and Identification*, Ph.D. thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, 2004.

[29] R. Wallin and A. Hansson, *KYPD: A solver for semi-definite programs derived from the Kalman-Yakubovich-Popov lemma*, in Proceedings of the 2004 IEEE Conference on Computer Aided Control Systems Design, Taipei, Taiwan, 2004, pp. 1–6.

[30] R. Wallin, A. Hansson, and L. Vandenberghe, *Comparison of two structure-exploiting optimization algorithms for integral quadratic constraints*, in Proceedings of the 4th IFAC Symposium on Robust Control Design, Milan, Italy, 2003.

[31] J. C. Willems, *Dissipative dynamical system, Part* I: *General theory*, Arch. Ration. Mech. Anal., 45 (1972), pp. 321–353.

[32] V. A. Yakubovich, *Frequency conditions for the absolute stability of control systems with several nonlinear or linear nonstationary blocks*, Avtomat. Telemekh., 6 (1967), pp. 5–30.

[33] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice–Hall, Englewood Cliffs, NJ, 1995.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 496–540

THE FAST FOURIER TRANSFORM*

ULRICH OBERST†

Abstract. Fast Fourier transforms (FFTs) are fast algorithms, i.e., of low complexity, for the computation of the discrete Fourier transform (DFT) on a finite abelian group. They are among the most important algorithms in applied and engineering mathematics and in computer science, in particular for one- and multidimensional systems theory and signal processing. We give a relatively short survey of the FFT for arbitrary finite abelian groups, cyclic or not, with complete and partially novel proofs, the main distinction being explicit induction formulas for the FFT in all cases which generalize the original FFT-algorithm due to Cooley and Tukey and, much earlier, to Gauß. We believe that our approach has didactic advantages over the usual ones. We also present the application of the FFT to fast convolution algorithms, and the so-called number theoretic transforms over finite coefficient rings. We do not treat those algorithms which decrease the multiplicative complexity at the expense of many more rational linear combinations, which in this context are considered costless, nor do we discuss the DFT for nonabelian finite groups.

Key words. fast Fourier transform, discrete Fourier transform, fast convolution

AMS subject classification. 65T50

DOI. 10.1137/060658242

1. Introduction.

Fast Fourier transforms (FFTs) are fast algorithms, i.e., of low complexity, for the computation of the *discrete Fourier transform* (DFT) on a finite abelian group which, in turn, is a special case of the Fourier transform on a locally compact abelian group. The FFTs are among the most important algorithms in applied and engineering mathematics and in computer science, in particular for *one- and multidimensional systems theory and signal processing*, as evidenced by references [4], [11], [14], [17], [21], [24], [25], [31], [32], [37]. Various textbooks on the FFT are mentioned at the end of this introduction.

The present article gives a relatively *short* survey of the FFT for arbitrary finite abelian groups, cyclic or not, with complete and partially novel proofs which in our opinion have didactic advantages over the usual ones. The main distinction consists in *explicit induction formulas* for the FFT, proven and announced in 1988 [27], [28], for all possible cases which generalize the FFT-algorithm on the group $\mathbb{Z}/\mathbb{Z}2^r$ due to Cooley and Tukey [16] and, much earlier, to Gauß. We also treat the applications of the FFT to fast convolution algorithms. We do not discuss the algorithms with fewer *essential multiplications* at the expense of many more rational linear combinations, i.e., those with low *multiplicative complexity*, for instance, those of Winograd [40]. Nor do we treat the FFT for *noncommutative* finite groups [5], [12].

An algorithm is called *fast* if it has *low complexity*, where the complexity is the number of elementary computation steps necessary to execute it. In this paper and in most computer processors such a step is of the form $ax + y$ with numbers a, x, y; i.e., it consists of one multiplication together with one addition.

The following *motivational* remarks taken from [6] and [22] on the Fourier theory for general locally compact abelian groups or *harmonic analysis* will not be used in

*Received by the editors April 26, 2006; accepted for publication (in revised form) October 31, 2006; published electronically April 27, 2007.
http://www.siam.org/journals/sicon/46-2/65824.html
†Institut für Mathematik der Universität Innsbruck, Technikerstraße 25, A-6020 Innsbruck, Austria (Ulrich.Oberst@uibk.ac.at).

any way in the rest of this article. For the group $G = \mathbb{R}^r$ the *Fourier transform* of a function $a \in \mathrm{L}^1(\mathbb{R}^r)$ is the bounded, continuous function

$$\widehat{a}(y) := \int_{\mathbb{R}^r} a(x) \exp(-2\pi i x \bullet y) dx,\ y \in \mathbb{R}^r,\ \text{where } x \bullet y := x_1 y_1 + \cdots + x_r y_r$$

is the standard scalar product. Under suitable assumptions, for instance, if $\widehat{a}$ is absolutely integrable, too [20, p. 164], the *Fourier inversion formula*

$$a(x) = \int_{\mathbb{R}^r} \widehat{a}(y) \exp(+2\pi i x \bullet y) dy$$

holds almost everywhere. For fixed y the map $x \mapsto \langle x, y\rangle := \exp(-2\pi i x \bullet y)$ is a character on $\mathbb{R}^r$, i.e., a continuous group homomorphism from $\mathbb{R}^r$ into the circle group $\mathrm{S}^1 := \{z \in \mathbb{C};\ |\ z\ | = 1\}$. Let $\mathrm{Gr}_{\mathrm{cont}}(\mathbb{R}^r, \mathrm{S}^1)$ denote the multiplicative group of all characters with the multiplication of functions. Then, more precisely, the continuous, symmetric, bimultiplicative form $\langle -, -\rangle$ is nondegenerate, i.e., induces the (topological) isomorphism

$$\mathbb{R}^r \cong \mathrm{Gr}_{\mathrm{cont}}(\mathbb{R}^r, \mathrm{S}^1),\ y \mapsto \langle -, y\rangle,$$

and the Fourier inversion has the form

$$\widehat{a}(y) := \int_{\mathbb{R}^r} a(x)\langle x, y\rangle dx,$$
$$a(x) := \int_{\mathbb{R}^r} \widehat{a}(y)\langle -x, y\rangle dy,\ \langle -x, y\rangle = \langle x, y\rangle^{-1} = \overline{\langle x, y\rangle}.$$

In general, the character group $\widehat{G} := \mathrm{Gr}_{\mathrm{cont}}(G, \mathrm{S}^1)$ of a locally compact abelian group G is not isomorphic to G, for instance, $\widehat{\mathbb{Z}^r} \cong (\mathrm{S}^1)^r$, but the form $\langle -, -\rangle : G \times \widehat{G} \to \mathrm{S}^1,\ \langle g, \widehat{g}\rangle := \widehat{g}(g)$, is nondegenerate in the sense that the map $G \to \mathrm{Gr}_{\mathrm{cont}}(\widehat{G}, \mathrm{S}^1),\ g \mapsto \langle g, -\rangle$, is a (topological) isomorphism and the Fourier inversion has the form

$$\widehat{a}(\widehat{g}) := \int_G a(g)\langle g, \widehat{g}\rangle dg,\ a \in \mathrm{L}^1(G),$$
$$a(g) := \int_{\widehat{G}} \widehat{a}(\widehat{g})\langle -g, \widehat{g}\rangle d\widehat{g},\ \langle -g, \widehat{g}\rangle = \langle g, \widehat{g}\rangle^{-1} = \overline{\langle g, \widehat{g}\rangle},$$

where dg, respectively, $d\widehat{g}$, are the suitably normalized *Haar measures* on G, respectively, $\widehat{G}$.

We specialize the preceding considerations to the simple case of a finite abelian group G of exponent $d > 0$, i.e., satisfying $dG = 0$. In various ways one can choose a group $\widehat{G} \cong G$, for instance, $\widehat{G} = G$, and a biadditive form

$$\bullet : G \times \widehat{G} \to \mathbb{Z}/\mathbb{Z}d \text{ such that}$$
$$\widehat{G} \cong \mathrm{Hom}(G, \mathbb{Z}/\mathbb{Z}d),\ \widehat{g} \mapsto (-) \bullet \widehat{g},\ \text{and } G \cong \mathrm{Hom}(\widehat{G}, \mathbb{Z}/\mathbb{Z}d),\ g \mapsto g \bullet (-),$$

are isomorphisms, the latter signifying that the form $\bullet$ is *nondegenerate*. In the engineering literature the groups G and $\widehat{G}$ are called the *time*, respectively, the *frequency* domain, in the standard one-dimensional case of time signals. We choose a primitive dth root of one in $\mathbb{C}$, for instance, $\zeta := \exp(-\frac{2\pi i}{d})$; hence

$$\mathbb{Z}/\mathbb{Z}d \cong \mu := \langle\zeta\rangle = \{1, \zeta, \cdots, \zeta^{d-1}\} \subseteq \mathrm{S}^1, \overline{k} \mapsto \zeta^{\overline{k}} := \zeta^k.$$

The nondegenerate form $\bullet$ thus induces the nondegenerate bimultiplicative form

$$\langle -, -\rangle : G \times \widehat{G} \to \mu,\ \langle g, \widehat{g}\rangle := \zeta^{g \bullet \widehat{g}},\ \text{such that}$$
$$\widehat{G} \cong \mathrm{Gr}(G, \mu),\ \widehat{g} \mapsto \langle -, \widehat{g}\rangle,\ \text{and } G \cong \mathrm{Gr}(\widehat{G}, \mu),\ g \mapsto \langle g, -\rangle.$$

Here $\mathrm{Gr}(G,\mu)$ denotes the *multiplicative* abelian group of homomorphisms from the *additive* abelian group G into the *multiplicative* abelian group μ. The canonical group isomorphisms

$$\widehat{G} \cong \mathrm{Hom}(G, \mathbb{Z}/\mathbb{Z}d) \cong \mathrm{Gr}(G,\mu) = \mathrm{Gr}(G, \mathrm{S}^1)$$

hold. In this article we use the chosen group $\widehat{G}$ instead of the isomorphic *character group* $\mathrm{Gr}(G,\mu)$ for the development of the theory. The standard choices for the one-dimensional DFT are

$$d > 0,\ G := \widehat{G} = \mathbb{Z}/\mathbb{Z}d,\ \overline{k} \bullet \overline{l} = \overline{kl},\ \langle \overline{k}, \overline{l} \rangle = \exp\left(-2\pi i \tfrac{kl}{d}\right).$$

It is a well-known and simple, but for this paper essential, observation that the contravariant *duality functor* $G \mapsto \widehat{G} \cong \mathrm{Gr}(G,\mu)$ is exact on finite abelian groups of exponent d. The Haar integral on $\mathbb{C}^G$ which is unique up to a multiplicative positive constant is the map $\mathbb{C}^G \to \mathbb{C},\ a \mapsto \sum_{g\in G} a(g)$. Therefore we define two DFTs

$$\mathrm{Four}_G : \mathbb{C}^G \to \mathbb{C}^{\widehat{G}},\ a \mapsto \widehat{a},\ \widehat{a}(\widehat{g}) := \textstyle\sum_{g\in G} a(g)\langle g, \widehat{g}\rangle,\ \text{and}$$
$$\mathrm{Four}_{\widehat{G}} : \mathbb{C}^{\widehat{G}} \to \mathbb{C}^G,\ b \mapsto \widehat{b},\ \widehat{b}(g) := \textstyle\sum_{\widehat{g}\in\widehat{G}} b(\widehat{g})\langle g, \widehat{g}\rangle.$$

The map $\mathrm{Four}_{\widehat{G}}$ is sometimes called the *inverse discrete Fourier transform* (IDFT). The Fourier inversion formula has the form

$$N^{-1}\widehat{\widehat{a}}(-g) = a(g),\ \text{where } a \in \mathbb{C}^G,\ N := \mathrm{ord}(G).$$

The form $\langle -, - \rangle$ and the Fourier transform can also be defined if $\mathbb{C}$ is replaced by an arbitrary commutative ring K and if ζ is a primitive dth root of one in K, and we will do this in these notes. However, the Fourier inversion holds under additional assumptions on ζ only [26], [15], [18]. Interesting cases concern finite factor rings $K = \mathbb{Z}/\mathbb{Z}M$ of $\mathbb{Z}$, where the corresponding DFT is also called a *number theoretic transform* (NTT), or rings of algebraic integers. In our opinion the change of the coefficient ring does not justify a change of the terminology, so we will always talk of the DFT.

Any filtration or increasing sequence of subgroups $0 = G_0 \subseteq G_1 \subseteq \cdots \subseteq G_r = G$ of G gives rise to an FFT-algorithm for the computation of Four_G. That nontrivial subgroups H of G and their factor groups G/H are significant for the construction of an FFT for Four_G has been one of the basic observations in this field since [16], and the book [5], for instance, stresses this point of view. For groups of prime order there are no FFTs *in this sense*, and different algorithms have been designed, the first one by Rader [33]. Our description of the recursive FFT-algorithms gives simple explicit recursion formulas and makes essential use of the exactness of the duality functor. For the important case of cyclic groups similar formulas are contained in [8, pp. 188–191].

The central and novel sections of this survey paper are those on the FFT. The sections on duality theory, the DFT, and the complexity of linear maps contain necessary preliminaries and are simple adaptions from the literature. The two short sections on fast convolution algorithms derived from the FFT and on NTTs are included for completeness' sake and are also simple variants of the literature [26].

Since the FFT is so important in engineering applications there are very many papers and books on this subject, too numerous to be available to and be read and known by the author. Therefore the list of references at the end of this survey paper

contains only books and papers which are actually mentioned in the text, and omission of an article is no comment whatsoever on its historical or practical significance. Standard textbooks on the FFT are those of Brigham [8], Nussbaumer [26], and Beth [5] (in German), newer books are those of Clausen and Baum [12], Chu and George [13], and Garg [18]. Besides the signal processing and systems textbooks quoted above, the book [8] and especially that of Briggs and Henson [7] give surveys of the many mathematical and technical applications of the DFT and thus of the FFT from an engineering point of view, for instance, to the computation of Fourier integrals and coefficients, to trigonometric interpolation, and to digital filtering.

We discuss briefly the literature on the construction of FFT and convolution algorithms which minimize the *multiplicative complexity* according to Winograd and which are otherwise not treated in the present paper. The seminal papers in this direction are those of Winograd, Auslander, and Tolimieri and their coworkers [39], [40], [2], [1], [35]. In [29], [38], and the book [30], which unfortunately has not yet appeared, we constructed the *optimal* fast Fourier and Hartley, respectively, Gelfand, transforms on arbitrary finite abelian groups, respectively, finite-dimensional, commutative, semisimple $\mathbb{Q}$-algebras, i.e., algorithms for these transformations of *minimal multiplicative complexity*, and computed the exact value of the latter with the help of [3]. The recent paper [36] emphasizes the renewed interest in such algorithms.

The present paper presupposes the algebraic knowledge of a mathematics student at the end of the second university year. Some results are recalled under the heading *Reminder*.

2. Duality.

Reminder 1 (see [23, p. 46]). Let $G = (G, +)$ be a finite abelian group, written additively. Then there are numbers $d_1 > 0, \cdots, d_r > 0$ and an isomorphism

$$G \cong \mathbb{Z}/\mathbb{Z}d_1 \times \cdots \times \mathbb{Z}/\mathbb{Z}d_r. \tag{1}$$

The least common multiple

$$\exp(G) := \operatorname{lcm}(d_1, \cdots, d_r) \text{ with } \mathbb{Z}\exp(G) = \{k \in \mathbb{Z};\ kG = 0\} \tag{2}$$

is called *the exponent* of G. If, in addition, d_ϱ divides $d_{\varrho+1}$ for all $\varrho = 1, \cdots, r-1$, then the d_ϱ are unique and are called the *invariant factors* of G and $\exp(G) = d_r$. If d is a multiple of $\exp(G)$ or, in other words, if $dG = 0$, we say that G is a *group of exponent* d.

If G and H are additively written abelian groups, the group of all additive or $\mathbb{Z}$-linear homomorphisms from G to H is denoted by $\operatorname{Hom}(G, H) = \operatorname{Hom}_{\mathbb{Z}}(G, H)$ as usual.

If $r > 0$ and K is a field, the map

$$\bullet : K^r \times K^r \to K,\ x \bullet y := x_1y_1 + \cdots + x_ry_r \text{ for } x = (x_1, \cdots, x_r),$$

is a nondegenerate symmetric bilinear form; i.e., the induced map

$$K^r \to \operatorname{Hom}_K(K^r, K),\ y \mapsto (-) \bullet y = y \bullet (-),$$

is a K-isomorphism.

The following symmetric bilinear form is the analogue of the preceding one for finite abelian groups.

THEOREM 2 (nondegenerate bilinear form). *Let*

$$G = \mathbb{Z}/\mathbb{Z}d_1 \times \cdots \times \mathbb{Z}/\mathbb{Z}d_r \ni g = (\overline{g_1}, \cdots, \overline{g_r}),\ g_\varrho \in \mathbb{Z},$$

be the finite abelian group of exponent $d > 0$, i.e., $dG = 0$. Then the map

$$\bullet : G \times G \to \mathbb{Z}/\mathbb{Z}d,\ g \bullet h := \overline{\textstyle\sum_{\varrho=1}^{r} g_\varrho h_\varrho \frac{d}{d_\varrho}}, \tag{3}$$

is well defined and is a nondegenerate, symmetric $\mathbb{Z}$-bilinear form; i.e., the following hold.

(1) *The definition is independent of the representatives g_ϱ, h_ϱ.*

(2) *$g \bullet h = h \bullet g$, $g \bullet (h + h') = g \bullet h + g \bullet h'$ for all g, h, h' in G.*

(3) *$G \cong \mathrm{Hom}(G, \mathbb{Z}/\mathbb{Z}d)$, $h \mapsto (-) \bullet h$.*

Proof. (1) *The map is well defined*: Let $g = (\overline{g_1}, \cdots, \overline{g_r}) = (\overline{g'_1}, \cdots, \overline{g'_r})$; hence $g'_\varrho = g_\varrho + k_\varrho d_\varrho$, $k_\varrho \in \mathbb{Z}$, for $\varrho = 1, \ldots, r$. But then

$$\textstyle\sum_{\varrho=1}^{r} g'_\varrho h_\varrho \frac{d}{d_\varrho} = \sum_{\varrho=1}^{r} g_\varrho h_\varrho \frac{d}{d_\varrho} + \sum_{\varrho=1}^{r} g_\varrho h_\varrho k_\varrho d \in \sum_{\varrho=1}^{r} g_\varrho h_\varrho \frac{d}{d_\varrho} + \mathbb{Z}d, \text{ and hence}$$
$$\overline{\textstyle\sum_{\varrho=1}^{r} g'_\varrho h_\varrho \frac{d}{d_\varrho}} = \overline{\textstyle\sum_{\varrho=1}^{r} g_\varrho h_\varrho \frac{d}{d_\varrho}} = g \bullet h.$$

The independence of the representatives h_ϱ is shown in the same fashion.

(2) The symmetry and bilinearity follow trivially from the definition.

(3) It remains to show that $G \to \mathrm{Hom}(G, \mathbb{Z}/\mathbb{Z}d),\ h \bullet (-) = (-) \bullet h$, is an isomorphism.

(i) *Monomorphism*: Assume that $(-) \bullet h = 0$. For $\varrho = 1, \ldots, r$ let $\delta_\varrho := (0, \cdots, 0, \underset{\varrho}{\overline{1}}, 0, \cdots, 0)$ denote the analogue of the standard basis such that $(\overline{g_1}, \cdots, \overline{g_r}) = \sum_{\varrho=1}^{r} g_\varrho \delta_\varrho$ for all $g \in G$. Then

$$0 = \delta_\varrho \bullet h = \overline{h_\varrho \tfrac{d}{d_\varrho}} \in \mathbb{Z}/\mathbb{Z}d;\ \text{hence for } \varrho = 1, \ldots, r$$
$$d \mid h_\varrho \tfrac{d}{d_\varrho} \text{ or } d_\varrho \mid h_\varrho \text{ and } \overline{h_\varrho} = 0 \text{ in } \mathbb{Z}/\mathbb{Z}d_\varrho,\ \text{i.e., } h = 0.$$

(ii) *Epimorphism*: Let $\varphi : G \to \mathbb{Z}/\mathbb{Z}d$ be any homomorphism. The equation

$$d_\varrho \delta_\varrho = 0 \text{ implies } d_\varrho \varphi(\delta_\varrho) = 0 \text{ in } \mathbb{Z}/\mathbb{Z}d;\ \text{hence } \varphi(\delta_\varrho) = \overline{h_\varrho \tfrac{d}{d_\varrho}} = \delta_\varrho \bullet h,\ h_\varrho \in \mathbb{Z},$$
$$\text{and for } g \in G :\ \varphi(g) = \varphi(\textstyle\sum_{\varrho=1}^{r} g_\varrho \delta_\varrho) = \sum_{\varrho=1}^{r} g_\varrho \varphi(\delta_\varrho)$$
$$= \textstyle\sum_{\varrho=1}^{r} g_\varrho \delta_\varrho \bullet h = (\sum_{\varrho=1}^{r} g_\varrho \delta_\varrho) \bullet h = g \bullet h \text{ and } \varphi = (-) \bullet h. \qquad \square$$

COROLLARY 3. *With the data of the preceding theorem, let G_1 and G_2 be two groups which are isomorphic to G and let $\varphi_i : G_i \cong G$, $i = 1, 2$, be two isomorphisms. Then*

$$\bullet : G_1 \times G_2 \to \mathbb{Z}/\mathbb{Z}d,\ g_1 \bullet g_2 := \varphi_1(g_1) \bullet \varphi_2(g_2), \tag{4}$$

is a nondegenerate bilinear form; i.e., the maps

$$G_1 \to \mathrm{Hom}(G_2, \mathbb{Z}/\mathbb{Z}d),\ g_1 \mapsto g_1 \bullet (-), \quad \text{and} \quad G_2 \to \mathrm{Hom}(G_1, \mathbb{Z}/\mathbb{Z}d),\ g_2 \mapsto (-) \bullet g_2,$$

are isomorphisms.

The proof is obvious. The corollary implies that the following assumptions can be realized in various ways.

Assumption 4. Let $d > 0$. In what follows we consider finite abelian groups G with $dG = 0$. For each such G we choose a group $\widehat{G}$ and a nondegenerate bilinear form $\bullet : G \times \widehat{G} \to \mathbb{Z}/\mathbb{Z}d$, hence the *canonical* isomorphisms

(5)
$$\text{can} : G \cong \text{Hom}(\widehat{G}, \mathbb{Z}/\mathbb{Z}d),\ g \mapsto g \bullet (-),\ \text{and}\ \text{can} : \widehat{G} \cong \text{Hom}(G, \mathbb{Z}/\mathbb{Z}d),\ \widehat{g} \mapsto (-) \bullet \widehat{g}.$$

For the groups $G = \mathbb{Z}/\mathbb{Z}d_1 \times \cdots \times \mathbb{Z}/\mathbb{Z}d_r$ the canonical choices are $\widehat{G} = G$ and the symmetric form of (3). In the context of the FFT the groups G (resp., $\widehat{G}$) are often called the *time domain* (resp., the *frequency domain*), and therefore it is advantageous to make a notational distinction between G and $\widehat{G}$ even if $G = \widehat{G}$.

If G is any finite abelian group, the theory applies for $d = \exp(G)$.

Reminder 5 (see [23, pp. 76,77]). $\text{Hom}(G, H)$ is an *additive functor* in its two variables G and H. In particular, a homomorphism $\varphi : G_1 \to G_2$ of abelian groups induces the homomorphism

$$\text{Hom}(\varphi, \mathbb{Z}/\mathbb{Z}d) : \text{Hom}(G_2, \mathbb{Z}/\mathbb{Z}d) \to \text{Hom}(G_1, \mathbb{Z}/\mathbb{Z}d),\ \chi_2 \mapsto \chi_2\varphi,$$

in the reverse direction. This assignment satisfies the relations

$$\text{Hom}(\text{id}_G, \mathbb{Z}/\mathbb{Z}d) = \text{id}_{\text{Hom}(G,\mathbb{Z}/\mathbb{Z}d)},$$
$$\text{Hom}(\varphi_1, \mathbb{Z}/\mathbb{Z}d)\,\text{Hom}(\varphi_2, \mathbb{Z}/\mathbb{Z}d) = \text{Hom}(\varphi_2\varphi_1, \mathbb{Z}/\mathbb{Z}d) \text{ for } G_1 \xrightarrow{\varphi_1} G_2 \xrightarrow{\varphi_2} G_3,$$
$$\text{Hom}(\varphi^{-1}, \mathbb{Z}/\mathbb{Z}d) = \text{Hom}(\varphi, \mathbb{Z}/\mathbb{Z}d)^{-1} \text{ if } \varphi : G_1 \cong G_2.$$

Corollary 6. *For each finite abelian group G of exponent $d > 0$ there is a* noncanonical *isomorphism $G \cong \widehat{G}$.*

Proof. Choose an isomorphism $\varphi : H = \mathbb{Z}/\mathbb{Z}d_1 \times \cdots \times \mathbb{Z}/\mathbb{Z}d_r \to G$ and on H the bilinear form from (3) which induces the isomorphism $H \cong \text{Hom}(H, \mathbb{Z}/\mathbb{Z}d)$. Then

$$\widehat{G} \cong \text{Hom}(G, \mathbb{Z}/\mathbb{Z}d) \overset{\text{Hom}(\varphi,\mathbb{Z}/\mathbb{Z}d)}{\cong} \text{Hom}(H, \mathbb{Z}/\mathbb{Z}d) \cong H \cong G. \qquad \square$$

Remark 7. If K is a field, V a finite-dimensional K-vector space, and $V^\star := \text{Hom}_K(V, K)$ its dual space, the canonical Gelfand map

$$\text{Gelf} : V \to V^{\star\star},\ v \mapsto \text{Gelf}(v),\ \text{Gelf}(v)(v^\star) := v^\star(v),$$

is a K-isomorphism. The following result is the analogue for finite abelian groups.

Theorem 8. *There is the unique* canonical Gelfand *isomorphism*

$$\text{Gelf}_G : G \cong \widehat{\widehat{G}} \text{ with } g \bullet \widehat{g} = \widehat{g} \bullet \text{Gelf}_G(g) \text{ for all } g \in G,\ \widehat{g} \in \widehat{G}. \tag{6}$$

Proof.

$$G \cong \text{Hom}(\widehat{G}, \mathbb{Z}/\mathbb{Z}d) \cong \widehat{\widehat{G}},\ g \to g \bullet (-) = (-) \bullet \text{Gelf}_G(g) \leftarrow \text{Gelf}_G(g). \qquad \square$$

Lemma and Definition 9. 1. *For each homomorphism $\varphi : G_1 \to G_2$ there is a unique homomorphism*

$$\varphi^\star : \widehat{G_2} \to \widehat{G_1} \text{ such that } \varphi(g_1) \bullet \widehat{g_2} = g_1 \bullet \varphi^\star(\widehat{g_2}) \text{ for all } g_1 \in G_1,\ \widehat{g_2} \in \widehat{G_2}. \tag{7}$$

The map $\varphi^\star$ is called the adjoint *of φ.*

2. *The relations $\text{id}_G^\star = \text{id}_{\widehat{G}}$ and $\varphi_1^\star\varphi_2^\star = (\varphi_2\varphi_1)^\star$ for $G_1 \xrightarrow{\varphi_1} G_2 \xrightarrow{\varphi_2} G_3$ hold.*

Hence the assignment $G \mapsto \widehat{G},\ \varphi \mapsto \varphi^\star$, *is a contravariant functor on finite abelian groups of exponent* $d > 0$ *and is called the* duality functor *in this article. Observe that* $\widehat{G} \cong \mathrm{Hom}(G, \mathbb{Z}/\mathbb{Z}d)$ *can be chosen in various ways.*

Proof. 1. There is a unique homomorphism $\varphi^\star$ such that the following diagram with vertical isomorphisms is commutative:

$$\begin{array}{ccc} \widehat{G_2} & \xrightarrow{\varphi^\star} & \widehat{G_1} \\ \downarrow \mathrm{can}_2 & & \downarrow \mathrm{can}_1 \\ \mathrm{Hom}(G_2, \mathbb{Z}/\mathbb{Z}d) & \xrightarrow{\mathrm{Hom}(\varphi, \mathbb{Z}/\mathbb{Z}d)} & \mathrm{Hom}(G_1, \mathbb{Z}/\mathbb{Z}d) \\ \widehat{g_2} & \mapsto & \varphi^\star(\widehat{g_2}) \\ \downarrow & & \downarrow \\ \chi_2 := (-) \bullet \widehat{g_2} & \mapsto & \chi_2\varphi = \varphi(-) \bullet \widehat{g_2} = (-) \bullet \varphi^\star(\widehat{g_2}) \end{array} \tag{8}$$

viz., $\varphi^\star := \mathrm{can}_1^{-1} \circ \mathrm{Hom}(\varphi, \mathbb{Z}/\mathbb{Z}d) \circ \mathrm{can}_2$. The commutativity signifies that

$$\varphi(g_1) \bullet \widehat{g_2} = g_1 \bullet \varphi^\star(\widehat{g_2}) \text{ for all } g_1 \in G_1,\ \widehat{g_2} \in \widehat{G_2}.$$

2. The relations follow from the commutative diagram (8) and Reminder 5. □

LEMMA 10. *The Gelfand map is a natural transformation; i.e., for* $\varphi : G_1 \to G_2$ *the following diagram is commutative:*

$$\begin{array}{ccc} G_1 & \xrightarrow{\varphi} & G_2 \\ \downarrow \mathrm{Gelf}_1 & & \downarrow \mathrm{Gelf}_2 \\ \widehat{\widehat{G_1}} & \xrightarrow{\varphi^{\star\star}} & \widehat{\widehat{G_2}} \end{array} . \tag{9}$$

Proof. For all $g_1 \in G_1$ and $\widehat{g_2} \in \widehat{G_2}$ we have

$$\begin{gathered} \widehat{g_2} \bullet \mathrm{Gelf}_2(\varphi(g_1)) = \varphi(g_1) \bullet \widehat{g_2} = g_1 \bullet \varphi^\star(\widehat{g_2}) \\ = \varphi^\star(\widehat{g_2}) \bullet \mathrm{Gelf}_1(g_1) = \widehat{g_2} \bullet \varphi^{\star\star}(\mathrm{Gelf}_1(g_1)); \text{ hence} \\ \mathrm{Gelf}_2(\varphi(g_1)) = \varphi^{\star\star}(\mathrm{Gelf}_1(g_1)). \end{gathered}$$ □

Reminder 11 (exactness [25, pp. 16, 77]). 1. Consider a sequence of abelian groups and homomorphisms

$$G_1 \xrightarrow{\varphi_1} G_2 \xrightarrow{\varphi_2} G_3. \tag{10}$$

The sequence is called a *complex* if $\varphi_2\varphi_1 = 0$ or $\mathrm{im}(\varphi_1) \subseteq \ker(\varphi_2)$.

2. The sequence (10) is called *exact* if $\mathrm{im}(\varphi_1) = \ker(\varphi_2)$.

3. A possibly infinite sequence

$$G_* :\ \cdots \to G_{i+1} \xrightarrow{d_{i+1}} G_i \xrightarrow{d_i} G_{i-1} \to \cdots, i \in \mathbb{Z}, \tag{11}$$

is called a *complex* (resp., *exact*) if and only if all three member subsequences have this property, i.e. $B_i := \mathrm{im}(d_{i+1}) \subseteq Z_i := \ker(d_i)$ (resp., $B_i = Z_i$) for all i. The groups $H_i(G_*) := Z_i/B_i$ are called the *homology groups* of the complex and are all zero if and only if G_* is exact.

4. For a sequence

$$0 \longrightarrow G_1 \xrightarrow{\varphi_1} G_2 \xrightarrow{\varphi_2} G_3 \tag{12}$$

the following properties are equivalent.

(a) The sequence is exact.
(b) $\ker(\varphi_1) = 0$, i.e., φ_1 is a monomorphism, and $\operatorname{im}(\varphi_1) = \ker(\varphi_2)$.
(c) The map φ_1 induces an isomorphism $\varphi_{1,\mathrm{ind}} : G_1 \cong \ker(\varphi_2)$.

5. For a sequence

$$G_1 \xrightarrow{\varphi_1} G_2 \xrightarrow{\varphi_2} G_3 \longrightarrow 0 \tag{13}$$

and the *cokernel* $\operatorname{cok}(\varphi_1) := G_2/\operatorname{im}(\varphi_1)$ the following properties are equivalent.
(a) The sequence is exact.
(b) $\operatorname{im}(\varphi_2) = G_3$, i.e., φ_2 is an epimorphism, and $\operatorname{im}(\varphi_1) = \ker(\varphi_2)$.
(c) The map φ_2 induces the isomorphism $\varphi_{2,\mathrm{ind}} : \operatorname{cok}(\varphi_1) \cong G_3$, $\overline{g_2} \mapsto \varphi_2(g_2)$.

6. The Hom-functor is left exact. Moreover, the sequence (13) is exact if and only if for all abelian groups X the derived sequence

$$\operatorname{Hom}(G_1, X) \xleftarrow{\operatorname{Hom}(\varphi_1, X)} \operatorname{Hom}(G_2, X) \xleftarrow{\operatorname{Hom}(\varphi_2, X)} \operatorname{Hom}(G_3, X) \longleftarrow 0 \tag{14}$$

is exact.

The next *duality theorem* states that the duality functor $G \mapsto \widehat{G}$ preserves and reflects exactness.

THEOREM 12 (duality theorem). *A sequence*

$$G_1 \xrightarrow{\varphi_1} G_2 \xrightarrow{\varphi_2} G_3 \tag{15}$$

of finite abelian groups G of exponent d ($dG = 0$) is exact if and only if its dual sequence

$$\widehat{G_1} \xleftarrow{\varphi_1^\star} \widehat{G_2} \xleftarrow{\varphi_2^\star} \widehat{G_3} \tag{16}$$

has this property.

Proof. $\Rightarrow$: 1. Assume first that the sequence

$$G_1 \xrightarrow{\varphi_1} G_2 \xrightarrow{\varphi_2} G_3 \longrightarrow 0 \tag{17}$$

is exact, i.e., φ_2 is surjective. Lemma 9 implies the commutative diagram

$$\begin{array}{ccccccc}
\widehat{G_1} & \xleftarrow{\varphi_1^\star} & \widehat{G_2} & \xleftarrow{\varphi_2^\star} & \widehat{G_3} & \leftarrow & 0 \\
\downarrow \mathrm{can}_1 & & \downarrow \mathrm{can}_2 & & \downarrow \mathrm{can}_3 & & \\
\operatorname{Hom}(G_1, \mathbb{Z}/\mathbb{Z}d) & \xleftarrow{\operatorname{Hom}(\varphi_1, \mathbb{Z}/\mathbb{Z}d)} & \operatorname{Hom}(G_2, \mathbb{Z}/\mathbb{Z}d) & \xleftarrow{\operatorname{Hom}(\varphi_2, \mathbb{Z}/\mathbb{Z}d)} & \operatorname{Hom}(G_3, \mathbb{Z}/\mathbb{Z}d) & \leftarrow & 0
\end{array}$$

with vertical isomorphisms whose lower row is exact according to part 6 of Reminder 11. The commutativity then implies that also the upper row is exact.

2. We prove that $\varphi^\star$ is an epimorphism if $\varphi : G_1 \to G_2$ is a monomorphism.

The sequence

$$0 \leftarrow C := \operatorname{cok}(\varphi^\star) \xleftarrow{\mathrm{can}} \widehat{G_1} \xleftarrow{\varphi^\star} \widehat{G_2}$$

is exact. Part 1 of this proof and Lemma 10 imply the commutative diagram

$$\begin{array}{ccccccc}
 & & & & G_1 & \xrightarrow{\varphi} & G_2 \\
 & & & & \downarrow \mathrm{Gelf}_1 & & \downarrow \mathrm{Gelf}_2 \\
0 & \to & \widehat{C} & \xrightarrow{\mathrm{can}^\star} & \widehat{\widehat{G_1}} & \xrightarrow{\varphi^{\star\star}} & \widehat{\widehat{G_2}}
\end{array}$$

with exact row and vertical isomorphisms. Since φ is a monomorphism, so is $\varphi^{\star\star}$, and hence $\widehat{C} = 0$. Since C and $\widehat{C}$ are isomorphic, we obtain $C = \operatorname{cok}(\varphi^\star) = 0$ or that $\varphi^\star$ is surjective.

3. The exact sequence (15) gives rise to the commutative diagram

$$\begin{array}{ccccccc} G_1 & \xrightarrow{\varphi_1} & G_2 & \xrightarrow{\text{can}} & C := \operatorname{cok}(\varphi_1) & \longrightarrow & 0 \\ & & \downarrow \varphi_2 & & \downarrow \psi & & \\ & & G_3 & = & G_3 & & \end{array},$$

where $\psi(\overline{g_2}) = \varphi_2(g_2)$. Since $C = G_2/\operatorname{im}(\varphi_1) = G_2/\ker(\varphi_2)$, the homomorphism theorem implies that ψ is a monomorphism. Dual to the preceding one is the commutative diagram

$$\begin{array}{ccccccc} \widehat{G_1} & \xleftarrow{\varphi_1^\star} & \widehat{G_2} & \xleftarrow{\text{can}^\star} & \widehat{C} & \longleftarrow & 0 \\ & & \uparrow \varphi_2^\star & & \uparrow \psi^\star & & \\ & & \widehat{G_3} & = & \widehat{G_3} & & \end{array}.$$

Its first row is exact, and $\psi^\star$ is an epimorphism according to parts 1 and 2 of the proof. Since $\varphi_2^\star = \text{can}^\star\, \psi^\star$, we conclude that $\operatorname{im}(\varphi_2^\star) = \operatorname{im}(\text{can}^\star) = \ker(\varphi_1^\star)$ and thus the exactness of (16).

$\Leftarrow$: Assume that (16) is exact. There results the diagram

$$\begin{array}{ccccc} G_1 & \xrightarrow{\varphi_1} & G_2 & \xrightarrow{\varphi_2} & G_3 \\ \downarrow \text{Gelf}_1 & & \downarrow \text{Gelf}_2 & & \downarrow \text{Gelf}_3 \\ \widehat{\widehat{G_1}} & \xrightarrow{\varphi_1^{\star\star}} & \widehat{\widehat{G_2}} & \xrightarrow{\varphi_2^{\star\star}} & \widehat{\widehat{G_3}} \end{array}.$$

The exactness of (16) and the proof "$\Rightarrow$" imply the exactness of its lower row, and Lemma 10 implies its commutativity. Since the Gelfand maps are isomorphisms, the wanted exactness of the upper row follows. □

3. The discrete Fourier transform. In this section we define and investigate the DFT for K-valued functions on a *finite* abelian group where K denotes a suitable coefficient field or even ring.

Assumption 13. The assumptions of section 2 remain in force, in particular $d > 0$. We consider finite additively written abelian groups G of exponent d ($dG = 0$) and the nondegenerate bilinear forms $\bullet : G \times \widehat{G} \to \mathbb{Z}/\mathbb{Z}d$. Let K be a commutative *coefficient* ring. Then K^G is the K-module of functions from G to K with its argumentwise addition and scalar multiplication. The standard case for the FFT will be the coefficient field $\mathbb{C}$ of complex numbers. However, since we are also going to discuss the so-called arithmetic transforms with a finite coefficient ring or field, we consider the more general situation from Assumption 13. Let $\mathrm{U}(K)$ denote the group of units or invertible elements of K. For the definition of the DFT on K^G we also need an analogue of the circle group $S^1 = \{z \in \mathbb{C};\ |\ z\ | = 1\} \subset \mathrm{U}(\mathbb{C})$ in the standard case of complex Fourier transforms. Therefore we make the following additional assumption for the ring K.

Assumption 14. Let $\zeta \in \mathrm{U}(K)$ be a *primitive dth root of one* in K, i.e.,

$$\zeta^d = 1,\ \mu := \langle\zeta\rangle = \{1, \zeta, \cdots, \zeta^{d-1}\} \subseteq \mathrm{U}(K),\ \operatorname{ord}(\zeta) = \operatorname{ord}(\mu) = d.$$

Examples 15.

(1) Let

$$K := \mathbb{C},\ \zeta := \exp\left(-\tfrac{2\pi i}{d}\right).\ \text{Then}\ \mu := \langle\zeta\rangle = \{\eta \in \mathbb{C};\ \eta^d = 1\}$$

is the group of *all* dth roots of one in $\mathbb{C}$ and consists of the vertices of the regular d-gon. These data are those of the standard complex DFT.

(2) Let $d := 2,\ K := \mathbb{R},\ \zeta := -1$. These data are used for the discrete *Walsh–Fourier* transform.

(3) Let $K := \mathbb{C} \times \mathbb{C},\ \zeta := (\zeta_1, \zeta_2) := (\exp(-\frac{2\pi i}{d}), 1)$. This is a primitive dth root of one, but it does not generate the finite group of all dth roots of one which consists of the elements (ζ_1^m, ζ_1^n).

(4) Let K be a finite field of characteristic p and dimension $[K : \mathbb{Z}/\mathbb{Z}p] = n$, hence with $q := p^n$ elements. The group $\mathrm{U}(K) = K \setminus \{0\}$ is cyclic and hence generated by a primitive root of order $d := q-1$. For instance, $\mathrm{U}(\mathbb{Z}/\mathbb{Z}7) = \langle\overline{3}\rangle$, whereas $\mathrm{ord}(\overline{2}) = 3$.

If G_1 and G_2 are arbitrary abelian groups and one of them is multiplicatively written, we denote the group of all homomorphisms from G_1 to G_2 by $\mathrm{Gr}(G_1, G_2)$ instead of $\mathrm{Hom}(G_1, G_2)$.

LEMMA 16. *Consider the situation of Example* 15(1) *and a finite abelian group* G *with* $dG = 0$. *Then* $\mathrm{Gr}(G, \mu) = \mathrm{Gr}(G, S^1)$ *is the group of all* complex characters *on* G.

Proof. Let $\chi : G \to S^1$ be any character, i.e., homomorphism. The relations $dg = 0$ for $g \in G$ imply $\chi(g)^d = 1$ and hence $\chi(g) \in \mu$ since μ is the group of *all* roots of 1. □

This result suggests that we consider the group $\mathrm{Gr}(G, \mu)$ as a suitable analogue of the character group for general coefficient rings, and we will do this; i.e, we call this group the *character group* of G. Notice that, in general, this group depends on the choice of ζ in contrast to the complex case.

COROLLARY AND DEFINITION 17. *The maps*

$$\begin{gathered}\mathbb{Z}/\mathbb{Z}d \cong \mu = \langle\zeta\rangle,\ \overline{k} \mapsto \zeta^{\overline{k}} := \zeta^k,\ \textit{hence also}\\ \mathrm{Hom}(G, \mathbb{Z}/\mathbb{Z}d) \cong \mathrm{Gr}(G, \mu),\ \varphi \mapsto \chi,\ \chi(g) = \zeta^{\varphi(g)},\end{gathered} \tag{18}$$

are isomorphisms. For each group G *(finite abelian,* $dG = 0$*) the nondegenerate bilinear form* $\bullet : G \times \widehat{G} \to \mathbb{Z}/\mathbb{Z}d$ *induces the nondegenerate bimultiplicative form*

$$\langle -, - \rangle : G \times \widehat{G} \to \mu = \langle\zeta\rangle,\ \langle g, \widehat{g}\rangle := \zeta^{g \bullet \widehat{g}};\ \textit{i.e.,} \tag{19}$$

(1) *for all* $g_1, g_2 \in G$ *and* $\widehat{g_1}, \widehat{g_2} \in \widehat{G}$

$$\langle g_1, \widehat{g_1} + \widehat{g_2}\rangle = \langle g, \widehat{g_1}\rangle\langle g, \widehat{g_2}\rangle,\ \langle g_1 + g_2, \widehat{g}\rangle = \langle g_1, \widehat{g}\rangle\langle g_2, \widehat{g}\rangle,$$

(2)

$$G \cong \mathrm{Gr}(\widehat{G}, \mu),\ g \mapsto \langle g, -\rangle,\ \widehat{G} \cong \mathrm{Gr}(G, \mu),\ \widehat{g} \mapsto \langle -, \widehat{g}\rangle.$$

The proof of this corollary is obvious since it consists in just replacing the additive group $\mathbb{Z}/\mathbb{Z}d$ by the multiplicative group $\mu = \langle\zeta\rangle$.

Reminder 18. The K-module K^G of all functions $a = (a(g))_{g \in G} : G \to K$ has the standard basis $\delta_h := (\delta_{h,g})_{g \in G},\ h \in G$, and the basis representation is

$$a = (a(g))_{g \in G} = \textstyle\sum_{g \in G} a(g)\delta_g.$$

We also consider the function module $K^{\widehat{G}}$ with the corresponding structure.

LEMMA AND DEFINITION 19 (DFT). *The data are as introduced above. The map*

$$\mathrm{Four}_G : K^G \to K^{\widehat{G}},\ a \mapsto \widehat{a},\ \widehat{a}(\widehat{g}) := \textstyle\sum_{g\in G} a(g)\langle g, \widehat{g}\rangle,$$

is K-linear and is called the discrete Fourier transform (DFT). *The function $\widehat{a} \in K^{\widehat{G}}$ is also called the Fourier transform of a. The analogous map*

$$\mathrm{Four}_{\widehat{G}} : K^{\widehat{G}} \to K^G,\ b \mapsto \widehat{b},\ \widehat{b}(g) := \textstyle\sum_{\widehat{g}\in\widehat{G}} b(\widehat{g})\langle g, \widehat{g}\rangle,$$

is called the Fourier transform *on $K^{\widehat{G}}$ or* inverse discrete Fourier transform (IDFT). *Notice that* $\mathrm{Four}_{\widehat{G}}$ *maps into K^G and not into $K^{\widehat{\widehat{G}}}$.*

The Fourier transform depends on the choice of the nondegenerate form $\bullet$ and of the primitive dth root ζ.

Examples 20. (1) Let $d := n > 0$, $K := \mathbb{C}$, $\zeta := \exp(-\frac{2\pi i}{n})$, and $G := \mathbb{Z}_n := \mathbb{Z}/\mathbb{Z}n = \widehat{G}$ with $\overline{k} \bullet \overline{l} := \overline{kl} \in \mathbb{Z}_n$ and hence $\langle \widehat{k}, \widehat{l}\rangle = \zeta^{kl} = \exp(-2\pi i \frac{kl}{n})$.

We identify

$$G = \mathbb{Z}_n = \{\overline{0}, \cdots, \overline{n-1}\} = \{0, \cdots, n-1\},$$
$$\mathbb{C}^G = \mathbb{C}^{\widehat{G}} = \mathbb{C}^{\mathbb{Z}_n} = \mathbb{C}^n \ni a = (a(\overline{k}))_{\overline{k}\in\mathbb{Z}_n}$$
$$= (a(\overline{0}), \cdots, a(\overline{n-1}) = (a(0), \cdots, a(n-1)), \text{ and hence}$$
$$\mathrm{Four}_G = \mathrm{Four}_{\widehat{G}} : \mathbb{C}^n \to \mathbb{C}^n.$$

The Fourier transform of $a = (a(0), \cdots, a(n-1))$ is

$$\widehat{a} = (\widehat{a}(0), \cdots, \widehat{a}(n-1)),$$
$$\widehat{a}(l) = \textstyle\sum_{\overline{k}\in\mathbb{Z}_n} a(k)\langle \overline{k}, \overline{l}\rangle = \sum_{k=0}^{n-1} a(k)\zeta^{kl} = \sum_{k=0}^{n-1} a(k)\exp\left(-2\pi i \frac{kl}{n}\right).$$

(2) Let $d := 2$, $K := \mathbb{R}$, $\zeta := -1$, and let $G = \mathbb{Z}_2^r \ni g = (g_1, \cdots, g_r)$ be the finite-dimensional $\mathbb{Z}_2$-vector space which is the typical finite group of exponent 2. We choose

$$\widehat{G} := G,\ g \bullet h := g_1h_1 + \cdots + g_rh_r, \text{ and hence } \langle g, h\rangle = (-1)^{g\bullet h}.$$

The *Fourier transform* $\widehat{a}$ of $a \in \mathbb{R}^G$ is given by $\widehat{a}(h) = \sum_{g\in G} a(g)(-1)^{g\bullet h}$. One also talks about the *Walsh–Fourier transform* in this case.

LEMMA 21. *For each $g \in G$ the Fourier transform of*

$$\delta_g \in K^G \text{ is } \widehat{\delta_g} = \langle g, -\rangle \in \mathrm{Gr}(\widehat{G}, \mu) \subset K^{\widehat{G}}.$$

Proof. $\widehat{\delta_g}(\widehat{g}) = \sum_{h\in G} \delta_g(h)\langle h, \widehat{g}\rangle = \langle g, \widehat{g}\rangle$. □

The K-module K^G admits two structures as commutative K-algebras which are both significant for the DFT.

LEMMA AND DEFINITION 22. *With the argumentwise multiplication*

$$(a_1a_2)(g) := a_1(g)a_2(g),\ a_1, a_2 \in K^G,\ g \in G, \tag{20}$$

the K-module K^G is a commutative K-algebra whose identity 1_{K^G} is the constant function with value 1. *The standard basis consists of complete orthogonal idempotents, i.e.,*

$$\textstyle\sum_{g\in G} \delta_g = 1,\ \delta_g\delta_h = \delta_{g,h}\delta_g.$$

The proof is obvious.

LEMMA AND DEFINITION 23. *With the* convolution *multiplication*

$$(a_1 * a_2)(g) := \sum_{g_1+g_2=g} a_1(g_1)a_2(g_2) = \sum_{h\in G} a_1(g-h)a_2(h) = \sum_{h\in G} a_1(h)a_2(g-h) \tag{21}$$

the K-module K^G is a commutative K-algebra with the identity δ_0. One writes $K[G] := (K^G, *)$ and calls this algebra the group algebra *of G with coefficients in K. The map*

$$\delta : G \to U(K[G]),\ g \mapsto \delta_g, \tag{22}$$

is a group monomorphism, i.e., injective with

$$\delta_0 = 1,\ \delta_{g_1+g_2} = \delta_{g_1} * \delta_{g_2},\ \text{hence } \delta_g^{-1} = \delta_{-g}.$$

The proof is analogous to that for the polynomial algebra $K[X] := K[\mathbb{N}]$ and is omitted.

The map $\delta : G \to U(K[G])$ has the following *universal property*. For two K-algebras A and B let $\mathrm{Al}_K(A, B)$ denote the set of K-algebra homomorphisms from A to B.

THEOREM 24 (universal property). *For each K-algebra B the map*

$$\mathrm{Al}_K(K[G], B) \to \mathrm{Gr}(G, \mathrm{U}(B)),\ \varphi \mapsto \chi := \varphi \circ \delta, \tag{23}$$

is bijective. The inverse map is given by

$$\chi \mapsto \varphi,\ \varphi(a) = \sum_{g\in G} a(g)\chi(g),\ a \in K^G.$$

Proof. The map is injective since $\chi := \varphi \circ \delta,\ \chi(g) = \varphi(\delta_g)$, implies

$$\varphi(a) = \varphi\left(\sum_{g\in G} a(g)\delta_g\right) = \sum_{g\in G} a(g)\varphi(\delta_g) = \sum_{g\in G} a(g)\chi(g). \tag{24}$$

Let, conversely, χ be given and define φ via the K-linear map (24), in particular,

$$\varphi(\delta_g) = \chi(g) \text{ and } \varphi(1_{K[G]}) = \varphi(\delta_0) = \chi(0) = 1_B.$$

Then

$$\varphi(\delta_{g_1} * \delta_{g_2}) = \varphi(\delta_{g_1+g_2}) = \chi(g_1+g_2) = \chi(g_1)\chi(g_2) = \varphi(\delta_{g_1})\varphi(\delta_{g_2}).$$

Therefore φ is multiplicative on the standard basis and therefore a K-algebra homomorphism by bilinear extension. □

COROLLARY 25. *For $B := K$ there results the bijection*

$$\mathrm{Al}_K(K[G], K) \cong \mathrm{Gr}(G, \mathrm{U}(K)),\ \varphi \mapsto \varphi \circ \delta.$$

In particular, for every $\widehat{g} \in \widehat{G}$, the group homomorphism $\langle -, \widehat{g}\rangle : G \to \mu \subseteq \mathrm{U}(K)$ induces the K-algebra homomorphism

$$K[G] = (K^G, *) \to K,\ a \mapsto \sum_{g\in G} a(g)\langle g, \widehat{g}\rangle = \widehat{a}(\widehat{g}).$$

THEOREM 26 (convolution theorem). *The K-linear Fourier transform* $\mathrm{Four}_G : K[G] \to K^{\widehat{G}}$ *is an algebra homomorphism, i.e.,*

$$\widehat{\delta_0} = \langle 0, - \rangle = 1_{K^{\widehat{G}}},\ \widehat{a_1 * a_2}(\widehat{g}) = \widehat{a_1}(\widehat{g})\widehat{a_2}(\widehat{g}).$$

Proof. Since $K^{\widehat{G}}$ is supplied with the argumentwise multiplication, the theorem is a direct consequence of Corollary 25. □

COROLLARY AND DEFINITION 27 (antipode). *The group automorphism $g \mapsto -g$ of G induces the algebra automorphism*

$$\mathrm{S}_G : K^G \cong K^G,\ \delta_g \mapsto \delta_{-g},\ \mathrm{S}_G(a)(g) = a(-g),$$

with respect to both multiplications on K^G. This map is called the antipode *on K^G and is an involution, i.e., $\mathrm{S}_G^2 = \mathrm{id}_{K^G}$ or $\mathrm{S}_G^{-1} = \mathrm{S}_G$. We likewise define $\mathrm{S}_{\widehat{G}}$ on $K^{\widehat{G}}$.*

Proof. For the convolution multiplication this follows from the universal property of $K[G]$, and for the argumentwise multiplication directly from the definition. □

LEMMA 28. *The antipode commutes with the Fourier transform, i.e., the diagram*

$$\begin{array}{ccc} K^G & \xrightarrow{\mathrm{Four}_G} & K^{\widehat{G}} \\ \downarrow\ \mathrm{S}_G & & \downarrow\ \mathrm{S}_{\widehat{G}} \\ K^G & \xrightarrow{\mathrm{Four}_G} & K^{\widehat{G}} \end{array} \quad \textit{is commmutative or } \mathrm{Four}_G\,\mathrm{S}_G = \mathrm{S}_{\widehat{G}}\,\mathrm{Four}_G\,.$$

Proof. The following sequence of equations is valid:

$$\mathrm{Four}_G\,\mathrm{S}_G(\delta_g) = \mathrm{Four}_G(\delta_{-g}) = \langle -g, - \rangle = \mathrm{S}_{\widehat{G}}(\langle g, - \rangle) = \mathrm{S}_{\widehat{G}}\,\mathrm{Four}_G(\delta_g). \qquad \square$$

For the proof of the *Fourier inversion theorem* we need an additional assumption on the root ζ.

Assumption 29 (see [15, Satz 2.8]). For the data of Assumption 14 we assume in what follows that d is invertible in K and that for each divisor $m > 1$ of d and the root $\eta := \zeta^{\frac{d}{m}}$ of order $\mathrm{ord}(\eta) = m$ the relation

$$1 + \eta + \cdots + \eta^{m-1} = 0$$

holds. In Theorem 80 we will give several equivalent conditions for this assumption as in [15, Satz 2.8].

Recall that all considered groups G are finite abelian of exponent d ($dG = 0$). Let $N := \mathrm{ord}(G)$ denote the order of G.

COROLLARY 30. *The preceding Assumption* 29 *is satisfied if K is a field.*

Proof. The second property follows from the relation

$$0 = \eta^m - 1 = (\eta - 1)(\eta^{m-1} + \cdots + \eta + 1)$$

since $\mathrm{ord}(\eta) = m \neq 1$; hence $\eta \neq 1$. Assume that the characteristic p of K divides d or $d = 0$ in K. Then p is prime and

$$d = pk \ \Rightarrow\ 0 = \zeta^d - 1 = (\zeta^k)^p - 1^p = (\zeta^k - 1)^p \ \Rightarrow\ \zeta^k = 1 \ \Rightarrow\ \mathrm{ord}(\zeta) \leq k < d,$$

a contradiction to $\mathrm{ord}(\zeta) = d$. □

COROLLARY 31. *Under Assumption* 29 *the order $N := \mathrm{ord}(G)$ of G is also invertible in K.*

Proof. If $G \cong \mathbb{Z}/\mathbb{Z}d_1 \times \cdots \times \mathbb{Z}/\mathbb{Z}d_r$ with $d_\varrho \mid d$, then $N = d_1 * \cdots * d_r$ divides d^r and therefore is invertible like d. ∎

LEMMA 32. *Under Assumption* 29 *any character* $\chi \in Gr(G, \mu)$ *of the group* G *satisfies the relation*

$$\textstyle\sum_{g\in G} \chi(g) = N\delta_{1,\chi} = \begin{cases} N & \text{if } \chi = 1, \\ 0 & \text{if } \chi \neq 1. \end{cases}$$

Here $1 : G \to \mu,\ g \mapsto 1$, *denotes the trivial character which is the neutral element of the character group.*

Proof. The assertion is obvious for $\chi = 1$. Assume therefore that $\chi \neq 1$ and that the image $\mathrm{im}(\chi)$ has the order $m \neq 1$. Then $\mathrm{im}(\chi)$ is the unique subgroup of order m of the cyclic group $\mu = \langle\zeta\rangle$ and is generated by $\eta := \zeta^{\frac{d}{m}}$; hence $\mathrm{im}(\chi) = \langle\eta\rangle = \{1, \eta, \cdots, \eta^{m-1}\}$. Let $\eta = \chi(g)$. The isomorphism

$$G/\ker(\chi) \cong \mathrm{im}(\chi) = \langle\eta\rangle, \overline{ig} \mapsto \chi(ig) = \chi(g)^i = \eta^i,$$

implies that every element $h \in G$ has a unique representation

$$h = ig + k, 0 \leq i \leq m-1, k \in \ker(\chi).$$

We infer

$$\begin{aligned} \textstyle\sum_{h\in G} \chi(h) &= \textstyle\sum \{\chi(ig+k); 0 \leq i \leq m-1, k \in \ker(\chi)\} \\ &= \textstyle\sum_{i,k} \chi(g)^i = \sum_k \left(\sum_{i=0}^{m-1} \eta^i\right) = 0, \end{aligned}$$

where $\sum_{i=0}^{m-1} \eta^i = 0$ according to Assumption 29. ∎

THEOREM 33. *The following equations hold for* $a \in K^G,\ b \in K^{\widehat{G}},\ g \in G,\ \widehat{g} \in \widehat{G}$:

$$\begin{aligned} Na(0) = \textstyle\sum_{\widehat{g}\in\widehat{G}} \widehat{a}(\widehat{g}),\ Nb(0) = \textstyle\sum_{g\in G} \widehat{b}(g), \\ N\delta_{0,g} = \textstyle\sum_{\widehat{g}\in\widehat{G}} \langle g, \widehat{g}\rangle,\ N\delta_{0,\widehat{g}} = \textstyle\sum_{g\in G} \langle g, \widehat{g}\rangle. \end{aligned}$$

Proof. Since $\widehat{\delta_g} = \langle g, -\rangle$ and $\widehat{\delta_{\widehat{g}}} = \langle -, \widehat{g}\rangle$, only the first equation has to be shown. With $\chi := \langle g, -\rangle : \widehat{G} \to \mu$ Lemma 32 implies $\sum_{\widehat{g}\in\widehat{G}} \langle g, \widehat{g}\rangle = N\delta_{0,g}$; hence

$$\begin{aligned} \textstyle\sum_{\widehat{g}\in\widehat{G}} \widehat{a}(\widehat{g}) = \textstyle\sum_{\widehat{g}\in\widehat{G}, g\in G} a(g)\langle g, \widehat{g}\rangle \\ = \textstyle\sum_g a(g) \sum_{\widehat{g}} \langle g, \widehat{g}\rangle = \sum_g a(g) N\delta_{0,g} = Na(0). \end{aligned}$$ ∎

THEOREM 34 (Fourier inversion theorem). *Under Assumption* 29 *the Fourier transform* Four_G *is an isomorphism and*

$$\begin{aligned} \mathrm{Four}_{\widehat{G}} \circ \mathrm{Four}_G = N\,\mathrm{S}_G \ \textit{ or } \ \mathrm{Four}_G^{-1} = N^{-1}\,\mathrm{S}_G\,\mathrm{Four}_{\widehat{G}} = N^{-1}\,\mathrm{Four}_{\widehat{G}}\,\mathrm{S}_{\widehat{G}} \ \textit{ or} \\ \widehat{\widehat{a}}(g) = Na(-g) \ \textit{or}\ \widehat{\widehat{a}} = N\,\mathrm{S}_G(a). \end{aligned}$$

Proof. All assertions follow from the last equation which has to be shown for the standard basis vectors only, from the invertibility of N and of the antipode and the same properties for $\mathrm{Four}_{\widehat{G}}$ instead of Four_G. But for $g, h \in G$

$$\begin{aligned} \widehat{\widehat{\delta_h}}(g) = \widehat{\langle h, -\rangle}(g) = \textstyle\sum_{\widehat{g}\in\widehat{G}} \langle h, \widehat{g}\rangle\langle g, \widehat{g}\rangle = \sum_{\widehat{g}\in\widehat{G}} \langle g+h, \widehat{g}\rangle \\ = N\delta_{0,g+h} = N\delta_{-h}(g) = N\,\mathrm{S}_G(\delta_h)(g);\ \text{hence } \widehat{\widehat{\delta_h}} = N\,\mathrm{S}_G(\delta_h). \end{aligned}$$ ∎

Example 35. In the situation of Example 20(1), with $d = N = n$ the Fourier inversion has the form

$$a \leftrightarrow \widehat{a},\ \widehat{a}(l) = \textstyle\sum_{k=0}^{n-1} a(k) \exp\left(-2\pi i \frac{kl}{n}\right),\ a(k) = n^{-1} \sum_{l=0}^{n-1} \widehat{a}(l) \exp\left(+2\pi i \frac{kl}{n}\right).$$

THEOREM 36 (product theorem). *The map* $N^{-1} \operatorname{Four}_G : K^G \to K[\widehat{G}]$ *is an algebra isomorphism; i.e.,*

$$N^{-1}\widehat{a_1 a_2} = N^{-1}\widehat{a_1} * N^{-1}\widehat{a_2} \text{ or } \widehat{a_1 a_2} = N^{-1}\widehat{a_1} * \widehat{a_2} \text{ and } N^{-1}\widehat{1} = \delta_0 \text{ or } \widehat{1} = N\delta_0.$$

Proof. The Fourier inversion theorem (Theorem 34) and Lemma 28 imply that $N^{-1} \operatorname{S}_G \operatorname{Four}_{\widehat{G}} : K^{\widehat{G}} \to K[G]$ and $\operatorname{S}_G : K[G] \to K[G]$ are algebra isomorphisms. The same follows for $N^{-1} \operatorname{Four}_{\widehat{G}}$ and then also for $N^{-1} \operatorname{Four}_G$. □

The action of the group G on itself by translation induces an action on K^G by K-algebra automorphisms, viz.,

$$\circ : G \times K^G, (g, a) \mapsto g \circ a := \delta_g * a, \quad (g \circ a)(h) = a(h - g). \tag{25}$$

Similarly $\widehat{G}$ acts on $K^{\widehat{G}}$.

THEOREM 37 (shift theorem). *For* $a \in K^G$, $g \in G$, *and* $\widehat{g} \in \widehat{G}$ *the following relations hold:*

$$\operatorname{Four}_G(g \circ a) = \langle g, - \rangle \widehat{a}, \quad \operatorname{Four}_G(\langle -, \widehat{g} \rangle a) = (-\widehat{g}) \circ \widehat{a}. \tag{26}$$

Proof. The first equation follows from the convolution theorem since

$$\widehat{g \circ a} = \widehat{\delta_g * a} = \widehat{\delta}_g \widehat{a} = \langle g, - \rangle \widehat{a},$$

and the second from

$$\begin{aligned} &\operatorname{Four}_G(\langle -, \widehat{g}_1 \rangle a)(\widehat{g}_2) = \textstyle\sum_{g \in G} \langle g, \widehat{g}_1 \rangle a(g) \langle g, \widehat{g}_2 \rangle \\ &= \textstyle\sum_{g \in G} a(g) \langle g, \widehat{g}_1 + \widehat{g}_2 \rangle = \widehat{a}(\widehat{g}_1 + \widehat{g}_2) = ((-\widehat{g}_1) \circ \widehat{a})(\widehat{g}_2). \end{aligned}$$ □

COROLLARY AND DEFINITION 38 (correlation). *The* correlation function $a \circ b \in K^G$ *of two functions* $a, b \in K^G$ *is defined as*

$$\begin{aligned} a \circ b &:= (\operatorname{S}_G a) * b, \text{ i.e. }, \\ (a \circ b)(h) &:= \textstyle\sum_{g \in G} (\operatorname{S}_G a)(g) b(h - g) \\ &= \textstyle\sum_{g \in G} a(-g) b(h - g) = \sum_{g \in G} a(g) b(h + g). \end{aligned}$$

Then

$$b \circ a = \operatorname{S}_G(a \circ b) \text{ and } \operatorname{Four}_G(a \circ b) = (\operatorname{S}_{\widehat{G}} \widehat{a}) \widehat{b}.$$

Proof. Since S_G is an involution and an algebra homomorphism, we infer

$$S_G(a \circ b) = S_G^2 a * S_G b = S_G b * a = b \circ a.$$

The second equation follows from the convolution theorem and from $\operatorname{S}_{\widehat{G}} \operatorname{Four}_G = \operatorname{Four}_G \operatorname{S}_G$. □

For the coefficient field $K := \mathbb{C}$ the preceding considerations can be slightly changed. For a function $a \in \mathbb{C}^G$ we define the complex conjugate function $\overline{a} \in$

$\mathbb{C}^G$ as $\overline{a}(g) := \overline{a(g)}$ and likewise for $a \in \mathbb{C}^{\widehat{G}}$. On $\mathbb{C}^G$ and likewise on $\mathbb{C}^{\widehat{G}}$ we define the standard hermitian inner product

(27)
$$(a_1, a_2) := \sum_{g\in G} \overline{a_1(g)}a_2(g) = \sum_{g\in G}(S\overline{a_1})(-g)a_2(g) = (S\overline{a_1} * a_2)(0) = (\overline{a_1} \circ a_2)(0),$$

where S denotes either S_G or $S_{\widehat{G}}$.

LEMMA 39. *$\widehat{\overline{a}} = \overline{S\widehat{a}}$, and hence $S\widehat{\overline{a}} = \overline{\widehat{a}}$.*

Proof.

$$\widehat{\overline{a}}(\widehat{g}) = \sum_g \overline{a(g)}\langle g, \widehat{g}\rangle = \overline{\sum_g a(g)\langle g, -\widehat{g}\rangle} = \overline{S\widehat{a}(\widehat{g})}. \qquad \square$$

THEOREM 40 (Plancherel). *For $a_1, a_2 \in \mathbb{C}^G$: $N(a_1, a_2) = (\widehat{a_1}, \widehat{a_2})$.*

Proof. Using (27), Theorem 33, Corollary 38, and finally the preceding lemma, we get

$$N(a_1, a_2) = N(\overline{a_1} \circ a_2)(0) = \sum_{\widehat{g}\in\widehat{G}} \widehat{\overline{a_1} \circ a_2}(\widehat{g})$$
$$= \sum_{\widehat{g}\in\widehat{G}}((S\widehat{\overline{a_1}})\widehat{a_2})(\widehat{g}) = \sum_{\widehat{g}\in\widehat{G}}(\widehat{\overline{a_1}}\widehat{a_2})(\widehat{g}) = (\widehat{a_1}, \widehat{a_2}). \qquad \square$$

COROLLARY 41 (orthogonality relations). *For two characters $a_i := \langle -, \widehat{g_i}\rangle$, $\widehat{g_1}$, $\widehat{g_2} \in \widehat{G}$, on G one obtains the orthogonality relation*

$$(a_1, a_2) = N(\delta_{\widehat{g_1}}, \delta_{\widehat{g_2}}) = N\delta_{\widehat{g_1},\widehat{g_2}}.$$

Hence the characters $\langle -, \widehat{g}\rangle$ are an orthogonal basis of $\mathbb{C}^G$.

Proof. This follows from the preceding theorem and $\widehat{\delta_{\widehat{g_i}}} = \langle -, \widehat{g_i}\rangle$ with the roles of G and $\widehat{G}$ interchanged. $\square$

4. Linear complexity. The FFT is a *fast* algorithm for the computation of the DFT. An algorithm is called fast if it has low complexity. In this section we define the *linear complexity* [9, Chap. 13] of matrices and in particular of the DFT to make this terminology precise. See [34] or the book [9] for a comprehensive treatment of *algebraic complexity theory*.

Let K again denote a commutative ring and I, J finite sets, for instance, G and $\widehat{G}$ in the preceding section. We consider the free column module $K^J := K^{J\times 1}$ with the column vectors $\xi = (\xi_j)_{j\in J}$, the free row module $K^{1\times J}$ with the row vectors $x = (x_j)_{j\in J}$ and the standard basis δ_j, $j \in J$, and the free module $K^{I\times J}$ of $I \times J$ matrices with coefficients in K. We identify

(28)
$$K^{I\times J} = \mathrm{Hom}_K(K^J, K^I),\ A = (\xi \mapsto A\xi),\ \text{in particular,}$$
$$K^{1\times J} = \mathrm{Hom}_K(K^J, K),\ x = (\xi \mapsto x\xi = \textstyle\sum_{j\in J} x_j\xi_j).$$

The following considerations will be applied mainly to the Fourier transform $\mathrm{Four}_G \in K^{\widehat{G}\times G} = \mathrm{hom}_K(K^G, K^{\widehat{G}})$. In the complexity theoretic arguments below we will mostly assume $A \in K^{m\times n}$.

Motivation 42. For $A \in K^{I\times J}$ the complexity or cost of an algorithm for the computation of $A\xi$ for arbitrary ξ will be the number of necessary *elementary computation steps* whose cost is defined to be 1. Such a step could be an addition or a multiplication, but we will use steps of the form $(x, y) \mapsto ax + y$ of one multiplication and one addition for numbers a, x, y in K as realized in many standard computer processors. If, more generally, $a \in K$ is a constant and $v, w \in K^{1\times J}$, $\xi \in K^J$ are vectors,

then $(av+w)\xi = a(v\xi)+(w\xi)$; i.e., the result is obtained from the numbers $v\xi$ and $w\xi$ with one elementary computation step. This motivates the following definitions of an algorithm and its complexity.

DEFINITION 43. *Let I, J be finite sets and let $A \in K^{I\times J}$. A* sequential algorithm *of* complexity *or* cost *$M \geq 0$ for A or, in more detail, for the computation of $A\xi$ for all $\xi \in K^J$ is a sequence $v_1, \cdots, v_M$ of row vectors in $K^{1\times J}$ with the following properties.*

(1) *Each row A_{i-}, $i \in I$, belongs to $V := \{\delta_j;\ j \in J\} \cup \{0\} \cup \{v_1, \cdots, v_M\}$.*
(2) *For each $k = 1, \ldots, M$ the vector v_k is given in the form $v_k = av + w$, where $a \in K$ and $v, w \in \{\delta_j;\ j \in J\} \cup \{0\} \cup \{v_1, \cdots, v_{k-1}\}$.*

The data a, v, w depend on k, but do not get an index for notational simplicity. The algorithm to compute $A\xi$ for arbitrary ξ computes the list of M values $v_1\xi, \cdots, v_M\xi$ with M elementary computation steps $v_k\xi = a(v\xi)+(w\xi)$ for values $v\xi$ and $w\xi$ computed earlier, and the $(A\xi)_i = A_{i-}\xi$, $i \in I$, are among these by condition (1) of Definition 43. In contrast, the computation of $0\xi = 0$ and $\delta_j\xi = \xi_j$ is costless. This signifies that the access time to the components of ξ on a real computer is neglected.

LEMMA 44. *For the computation of $A \in K^{m\times n}$ there is an algorithm of complexity $\leq mn$.*

Proof. The algorithm is the standard one for the matrix-vector product and is given by the sequence of vectors

$$\begin{array}{ccccccc}
v_{1,1} := A_{11}\delta_1 & \cdots & v_{1,j} = A_{1j}\delta_j + v_{1,j-1} & \cdots & v_{1,n} = A_{1-} = A_{1n}\delta_n + v_{1,n-1} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
v_{i,1} := A_{i1}\delta_1 & \cdots & v_{i,j} = A_{ij}\delta_j + v_{i,j-1} & \cdots & v_{i,n} = A_{i-} = A_{in}\delta_n + v_{i,n-1} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
v_{m,1} := A_{m1}\delta_1 & \cdots & v_{m,j} = A_{mj}\delta_j + v_{m,j-1} & \cdots & v_{m,n} = A_{m-} = A_{mn}\delta_n + v_{m,n-1}.
\end{array}$$ ∎

If in the preceding proof $A_{ij} = 0$ and hence $v_{i,j} = v_{i,j-1}$, one of these vectors can be omitted and hence the following corollary holds.

COROLLARY 45. *If N is the number of nonzero components of a matrix $A \in K^{m\times n}$, then there is an algorithm for A of complexity N.*

DEFINITION AND COROLLARY 46. *The* linear complexity *$\mu(A)$ of a matrix $A \in K^{I\times J}$ is the minimal complexity of an algorithm for A. Then*

(1) *$\mu(A) \leq N$, where N is the number of nonzero components of A;*
(2) *$\mu(A) = 0$ if and only if each row of A is either zero or a standard basis vector;*
(3) *$\mu(1, a_2, \cdots, a_n) \leq n-1$ for $a_2, \cdots, a_n \in K$.*

More generally, if ${}_KW$ and ${}_KV$ are free K-modules of finite dimension $n := [W : K]$ (resp., $m := [V : K]$), if $\underline{w} = (w_1, \cdots, w_n)$ (resp., $\underline{v} = (v_1, \cdots, v_m)$) are fixed chosen bases of these modules, and if $f : W \to V$, $f(\underline{w}) = \underline{v}A$, is a linear map with the matrix A with respect to the chosen bases*, then we define the complexity*

$$\mu(f) := \mu_{\underline{w},\underline{v}}(f) := \mu(A)$$

as that of the matrix A. Of course, basis transformations of V do not have complexity zero in general.

Proof. Concerning the last item the $1 \times n$ matrix $A := (1, a_2, \cdots, a_n)$ admits the algorithm $v_2 := \delta_1 + a_2\delta_2, \cdots, v_n := A$ since the computation of $1\xi_1 = \xi_1$ is of complexity zero. ∎

COROLLARY 47. *If Assumption 29 holds and G is a group of order N, the complexity of the Fourier transform $\mathrm{Four}_G = (\langle g, \widehat{g}\rangle)_{\widehat{g}\in\widehat{G}, g\in G} \in K^{\widehat{G}\times G}$ is at most $N(N-1)$.*

Proof. This follows like item 3 of Corollary 46 since for the column index $g := 0$ and any row index $\widehat{g}$ the entry of Four_G is $\langle 0, \widehat{g}\rangle = 1$. □

Definition and Corollary 48. *If $\alpha : I \to J$ is any map between finite index sets, the map*

$$K^\alpha : K^J \to K^I,\ \xi = (\xi_j)_{j\in J} \mapsto \xi \circ \alpha = (\xi_{\alpha(i)})_{i\in I}$$

is called an index transformation *and is of complexity zero.*

Proof. The computation of $K^\alpha(\xi)_i = \xi_{\alpha(i)}$ just reads off one component of ξ, and these operations are costless. □

The following theorem is decisive for the computation of an upper bound of the FFT.

Theorem 49. *If $A \in K^{m\times n}$ and $B \in K^{n\times p}$, then $\mu(AB) \le \mu(A) + \mu(B)$.*

Proof. Let $v_1, \cdots, v_M$ (resp., $w_1, \cdots, w_N$) be algorithms for A (resp., B) of minimal complexity $M := \mu(A)$ and $N := \mu(B)$. We are going to show that $w_1, \cdots, w_N, v_1B, \cdots, v_MB$ is an algorithm for AB; hence $\mu(AB) \le M + N = \mu(A) + \mu(B)$. Let

$$\begin{aligned} V_A &:= \{\delta_i;\ i = 1, \ldots, n\} \cup \{0\} \cup \{v_1, \cdots, v_M\},\\ V_B &:= \{\delta_j;\ j = 1, \ldots, p\} \cup \{0\} \cup \{w_1, \cdots, w_N\},\\ V_{AB} &:= \{\delta_j;\ j = 1, \ldots, p\} \cup \{0\} \cup \{w_1, \cdots, w_N, v_1B, \cdots, v_MB\}. \end{aligned}$$

By definition $V_B \subseteq V_{AB}$. We have to show that V_{AB} satisfies properties (1) and (2) from Definition 43.

(1) We use $A_{i-} \in V_A,\ B_{j-} \in V_B$ and show that $(AB)_{i-} = A_{i-}B \in V_{AB}$.

Case 1. $A_{i-} = 0 \ \Rightarrow\ A_{i-}B = 0 \in V_{AB}$.

Case 2. $A_{i-} = \delta_k \ \Rightarrow\ A_{i-}B = \delta_k B = B_{k-} \in V_B \subseteq V_{AB}$.

Case 3. $A_{i-} = v_k \ \Rightarrow\ A_{i-}B = v_kB \in V_{AB}$.

(2) We have to show that each vector x in $\{w_1, \cdots, v_MB\}$ is obtained from vectors in V_{AB} preceding x by an elementary computation step. For the vectors $w_l \in V_B$ this is obvious. Therefore consider a vector $x = v_kB \in V_{AB}$, where $v_k = au_1 + u_2$ with $a \in K$ and vectors $u_1, u_2 \in V_A$ preceding v_k. Then $x = v_kB = a(u_1B) + (u_2B)$, and we have to show that u_1B and u_2B precede x in V_{AB}.

Case 1. $u_j = 0 \ \Rightarrow\ u_jB = 0 \in V_{AB}$ preceding x.

Case 2. $u_j =$ standard basis vector $\ \Rightarrow\ u_jB =$ row of $B \ \Rightarrow\ u_jB \in V_B \subseteq V_{AB}$ preceding $x = v_kB$.

Case 3. $u_j = v_l,\ l < k \ \Rightarrow\ u_jB = v_lB \in V_{AB}$ preceding $x = v_kB$.

Hence V_{AB} has the properties of an algorithm. □

Corollary 50. *The complexity of block matrices satisfies*

$$\mu\left(\begin{pmatrix} A & 0\\ 0 & B\end{pmatrix}\right) \le \mu(A) + \mu(B),\ A, B \in K^{\bullet\times\bullet} \text{ of arbitrary size.}$$

Proof. Theorem 49, $\left(\begin{smallmatrix} A & 0 \\ 0 & B \end{smallmatrix}\right) = \left(\begin{smallmatrix} A & 0 \\ 0 & \mathrm{id} \end{smallmatrix}\right)\left(\begin{smallmatrix} \mathrm{id} & 0 \\ 0 & B \end{smallmatrix}\right)$, and the trivial relation $\mu\left(\begin{smallmatrix} A & 0 \\ 0 & \mathrm{id} \end{smallmatrix}\right) = \mu(A)$ yield the result. □

Remark 51 (multiplicative complexity). Let $(X - x_1) * \cdots * (X - x_n) \in \mathbb{Q}[X] \subset \mathbb{C}[X]$ be a rational polynomial with n distinct rational roots $x_i,\ i = 1, \ldots, n$. Then Lagrange interpolation or the Chinese remainder theorem implies the canonical $\mathbb{C}$-isomorphism

$$\varphi : \mathbb{C}[X]_{<n} \cong \mathbb{C}^n,\ f \mapsto (f(x_1), \cdots, f(x_n)),$$

where $\mathbb{C}[X]_{\langle n}$ is the space of polynomials of degree less than n. The domain (resp., the codomain) of φ has the basis $1, \cdots, X^{n-1}$ (resp., the standard basis $\delta_1, \cdots, \delta_n$). For fixed j and $f = a_{n-1}X^{n-1} + \cdots + a_0 \in \mathbb{C}[X]$ euclidean division furnishes

$$f = g(X - x_j) + f(x_j),\ g := b_{n-2}X^{n-2} + \cdots + b_0 \text{ with}$$
$$b_{n-2} = a_{n-1},\ b_{i-1} = a_i + x_j b_i,\ i = n-2, \ldots, 1,\ f(x_j) = a_0 + x_j b_0.$$

This shows that $f(x_j)$ can be computed from f with $n-1$ elementary computation steps and hence $\mu(\varphi) \leq n(n-1)$. Observe, however, that the necessary multiplications have the rational factor x_j. In the multiplicative complexity theory due to Winograd [40] which is, for instance, also used in [26] or [18], these rational multiplications—at least if the x_j are small integers—and rational linear combinations are considered costless, and therefore the complexity of φ is considered to be zero. This is not justified for those computers where the elementary computation step consists of one multiplication and one addition. The same cautionary remarks apply to almost all fast algorithms which use the Chinese remainder theorem and which are not discussed in this paper.

5. The fast Fourier transform (FFT). This is the central section of this article. Assumptions 14 and 29 are in force; in particular all groups are finite abelian of exponent $d > 0$.

Reminder 52. If $\varphi : G \to H$ is a group epimorphism of additive groups, a map $\sigma : H \to G$ is called a *section* of φ if $\varphi\sigma = \mathrm{id}_H$. Then σ is injective, and the elements $\sigma(h),\ h \in H$, are a system of representatives of $G/\ker(\varphi)$; i.e., the map

$$H \times \ker(\varphi) \to G,\ (h, k) \mapsto \sigma(h) + k, \tag{29}$$

is bijective. The map (29) is an isomorphism, and especially $G = \sigma(H) \oplus \ker(\varphi)$ if and only if σ is a monomorphism, but, in general, these properties do *not* hold.

We construct the FFT algorithm by means of a given *filtration* or sequence of subgroups

$$G_0 = 0 \subseteq G_1 \subseteq \cdots \subseteq G_r = G. \tag{30}$$

A filtration (30) gives rise to the commutative exact diagrams (with exact rows and

columns) for $i = 1, \dots, r$:

$$
\begin{array}{ccccccccc}
 & & & & & & 0 & & \\
 & & & & & & \downarrow & & \\
 & & 0 & & & & G_i/G_{i-1} & & \\
 & & \downarrow & & & & \downarrow{\scriptstyle \gamma_i:=\text{inj}} & & \\
0 & \longrightarrow & G_{i-1} & \xrightarrow{\alpha_{i-1}:=\text{inj}} & G & \xrightarrow{\lambda_{i-1}:=\text{can}} & G/G_{i-1} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \beta_i:=\text{inj}} & & \| & & \downarrow{\scriptstyle \nu_i:=\text{can}} & & \\
0 & \longrightarrow & G_i & \xrightarrow{\alpha_i:=\text{inj}} & G & \xrightarrow{\lambda_i:=\text{can}} & G/G_i & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \mu_i:=\text{can}} & & & & \downarrow & & \\
 & & G_i/G_{i-1} & & & & 0 & & \\
 & & \downarrow & & & & & & \\
 & & 0 & & & & & &
\end{array}
\tag{31}
$$

where inj (resp., can) are the canonical injections (resp., surjections). Moreover, λ_0 and α_r are isomorphisms, and the compatibility relations $\lambda_{i-1}\alpha_i = \gamma_i\mu_i$ hold. For more flexibility we now make the following, slightly more general assumption.

Assumption 53. Assume that Assumptions 14 and 29 are satisfied and that commutative exact diagrams (32) are given for $i = 1, \dots, r$:

$$
\begin{array}{ccccccccc}
 & & & & & & 0 & & \\
 & & & & & & \downarrow & & \\
 & & 0 & & & & K_i & & \\
 & & \downarrow & & & & \downarrow{\scriptstyle \gamma_i} & & \\
0 & \longrightarrow & G_{i-1} & \xrightarrow{\alpha_{i-1}} & G & \xrightarrow{\lambda_{i-1}} & H_{i-1} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \beta_i} & & \| & & \downarrow{\scriptstyle \nu_i} & & \\
0 & \longrightarrow & G_i & \xrightarrow{\alpha_i} & G & \xrightarrow{\lambda_i} & H_i & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \mu_i} & & & & \downarrow & & \\
 & & K_i & & & & 0 & & \\
 & & \downarrow & & & & & & \\
 & & 0 & & & & & &
\end{array}
\tag{32}
$$

such that the following additional properties hold:

(1) G_0 and H_r are zero or λ_0 and α_r are isomorphisms.

(2) The compatibility relations $\lambda_{i-1}\alpha_i = \gamma_i\mu_i$, $i = 1, \dots, r$, hold.

(3) Sections $\sigma_i : K_i \to G_i$, $i = 1, \dots, r$, with $\mu_i\sigma_i = \mathrm{id}_{K_i}$ and $\sigma_i(0) = 0$ are chosen arbitrarily.

These diagrams, in turn, induce the filtration $0 \subseteq \alpha_1(G_1) \subseteq \cdots \subseteq \alpha_r(G_r) = G$ and the isomorphisms

$$G/\alpha_i(G_i) \cong H_i, \bar{g} \mapsto \lambda_i(g),$$
$$G_i/\beta_i(G_{i-1}) \cong \alpha_i(G_i)/\alpha_{i-1}(G_{i-1}) \cong K_i, \overline{g_i} \mapsto \overline{\alpha_i(g_i)} \mapsto \mu_i(g_i).$$

In the situation of the preceding assumption we define

$$N := \operatorname{ord}(G) \text{ and } e_\varrho := \operatorname{ord}(K_\varrho); \text{ hence } N = e_1 * \cdots * e_r. \tag{33}$$

Recall that every group admits a Jordan–Hölder series, i.e., a filtration (30) or diagrams (31) or (32) with the property that the factors $K_\varrho \cong G_\varrho/G_{\varrho-1}$ are simple or have prime order e_ϱ, and that these prime numbers are uniquely determined by G.

Application of the duality functor $G \mapsto \widehat{G}$ to the preceding diagram yields further commutative exact diagrams.

COROLLARY 54. *Under Assumption* 53 *the following diagrams are commutative and exact:*

$$\begin{array}{ccccccccc}
 & & & & 0 & & & & \\
 & & & & \downarrow & & & & \\
 & & 0 & & \widehat{K_j} & & & & \\
 & & \downarrow & & \downarrow{\scriptstyle \mu_j^\star} & & & & \\
0 & \longrightarrow & \widehat{H_j} & \xrightarrow{\lambda_j^\star} & \widehat{G} & \xrightarrow{\alpha_j^\star} & \widehat{G_j} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \nu_j^\star} & & \| & & \downarrow{\scriptstyle \beta_j^\star} & & \\
0 & \longrightarrow & \widehat{H_{j-1}} & \xrightarrow{\lambda_{j-1}^\star} & \widehat{G} & \xrightarrow{\alpha_{j-1}^\star} & \widehat{G_{j-1}} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \gamma_j^\star} & & & & \downarrow & & \\
 & & \widehat{K_j} & & & & 0 & & \\
 & & \downarrow & & & & & & \\
 & & 0 & & & & & &
\end{array} \tag{34}$$

for $j = r, r-1, \ldots, 1$. *Furthermore, they have the additional properties that*

1. $\lambda_0^\star$ *and* $\alpha_r^\star$ *are isomorphisms,*
2. $\alpha_j^\star \lambda_{j-1}^\star = \mu_j^\star \gamma_j^\star$, *and*
3. *sections* $\widehat{\sigma}_j : \widehat{K_j} \to \widehat{H_{j-1}}$ *with* $\gamma_j^\star \widehat{\sigma}_j = \mathrm{id}_{\widehat{K_j}}$ *and* $\widehat{\sigma}_j(0) = 0$ *are chosen arbitrarily.*

Thus, up to the reverse numbering, the diagrams from (34) satisfy the same properties as the diagrams (32) of Assumption 53, and the same arguments apply to both of them.

LEMMA 55. *Under Assumption* 53 *the map*

$$\operatorname{ind} : \textstyle\prod_{i=1}^r K_i \to G,\ k = (k_i)_{i=1,\ldots,r} \mapsto \operatorname{ind}(k) := \sum_{i=1}^r \alpha_i \sigma_i(k_i),$$

is bijective; i.e., every $g \in G$ *admits a unique representation* $g = \sum_{i=1}^r \alpha_i \sigma_i(k_i)$ *with* $k_i \in K_i$.

Proof. By induction on $i = 0, \dots, r$ we show that $g = \alpha_i(g_i) \in \alpha_i(G_i), g_i \in G_i$, admits a unique representation

$$g = \sum_{j=0}^{i} \alpha_j \sigma_j(k_j), k_j \in K_j.$$

The assertion is trivial for $i = 0$ and $\alpha_0(G_0) = 0$. For $i > 0$ the exact sequence

$$0 \to G_{i-1} \xrightarrow{\beta_i} G_i \underset{\sigma_i}{\overset{\mu_i}{\rightleftarrows}} K_i \to 0$$

with the section σ_i of μ_i and (29) imply the unique representation

$$g_i = \beta_i(g_{i-1}) + \sigma_i(k_i), g_{i-1} \in G_{i-1}, k_i \in K_i,$$

and

$$g = \alpha_i(g_i) = \alpha_i\beta_i(g_{i-1}) + \alpha_i\sigma_i(k_i) = \alpha_{i-1}(g_{i-1}) + \alpha_i\sigma_i(k_i).$$

By induction there are unique $k_j \in K_j,\ j = 0, \dots, i-1$, with

$$\alpha_{i-1}(g_{i-1}) = \sum_{j=0}^{i-1} \alpha_j\sigma_j(k_j);\ \text{hence } g = \alpha_i(g_i) = \sum_{j=0}^{i} \alpha_j\sigma_j(k_j). \qquad \square$$

Application of Lemma 55 to diagram (34) yields the corollary.

COROLLARY 56. *Under Assumption* 53 *every* $\widehat{g} \in \widehat{G}$ *has a unique representation*

$$\widehat{g} = \sum_{j=1}^{r} \lambda_{j-1}^{\star}\widehat{\sigma}_j(\widehat{k}_j), \widehat{k}_j \in \widehat{K_j},$$

or, in other terms, the map

$$\widehat{\mathrm{ind}} : \textstyle\prod_{j=1}^{r} \widehat{K_j} \to \widehat{G}, \widehat{k} = (\widehat{k}_j)_{j=1,\dots,r} \mapsto \widehat{\mathrm{ind}}(\widehat{k}) := \sum_{j=1}^{r} \lambda_{j-1}^{\star}\widehat{\sigma}_j(\widehat{k}_j),$$

is bijective.

COROLLARY AND DEFINITION 57 (index transformations). *The maps*

$$\mathrm{Ind} := K^{\mathrm{ind}} : K^G \to K^{\prod_{i=1}^{r} K_i}, a \mapsto a_0 := a \circ \mathrm{ind},$$
$$a_0(k_1, \cdots, k_r) = a\left(\textstyle\sum_{i=1}^{r} \alpha_i\sigma_i(k_i)\right), k_i \in K_i,$$

and

$$\widehat{\mathrm{Ind}} := K^{\widehat{\mathrm{ind}}} : K^{\widehat{G}} \to K^{\prod_{j=1}^{r} \widehat{K_j}}, b \mapsto b_r := b \circ \widehat{\mathrm{ind}},$$
$$b_r(\widehat{k}_1, \cdots, \widehat{k}_r) = b\left(\textstyle\sum_{j=1}^{r} \lambda_{j-1}^{\star}\widehat{\sigma}_j(\widehat{k}_j)\right)$$

are K*-isomorphisms and* index transformations *according to Definition and Corollary* 48, *and hence are of complexity zero.*

The following easy considerations are central for the fast computation of the Fourier transform $\widehat{a}$ of a function $a \in K^G$ given by $\widehat{a}(\widehat{g}) = \sum_{g \in G} a(g)\langle g, \widehat{g}\rangle$. According to Lemmas 55 and 56 we write g and $\widehat{g}$ as

$$g = \mathrm{ind}(k) = \textstyle\sum_{i=1}^{r} \alpha_i\sigma_i(k_i), k = (k_i)_{i=1,\dots,r} \in \prod_{i=1}^{r} K_i,$$
$$\widehat{g} = \widehat{\mathrm{ind}}(\widehat{k}) = \textstyle\sum_{j=1}^{r} \lambda_{j-1}^{\star}\widehat{\sigma}_j(\widehat{k}_j), \widehat{k} = (\widehat{k}_j)_{j=1,\dots,r} \in \prod_{j=1}^{r} \widehat{K_j},$$

and compute the bimultiplicative form $\langle g, \widehat{g}\rangle$ as

$$\begin{aligned}\langle g, \widehat{g}\rangle &= \Big\langle \textstyle\sum_{i=1}^r \alpha_i\sigma_i(k_i), \sum_{j=1}^r \lambda_{j-1}^\star \widehat{\sigma}_j(\widehat{k}_j)\Big\rangle = \prod_{i,j=1}^r \mathrm{fact}_{ij}(k,\widehat{k}), \text{ where}\\ \mathrm{fact}_{ij}(k,\widehat{k}) &:= \langle \alpha_i\sigma_i(k_i), \lambda_{j-1}^\star\widehat{\sigma}_j(\widehat{k}_j)\rangle = \langle \lambda_{j-1}\alpha_i\sigma_i(k_i), \widehat{\sigma}_j(\widehat{k}_j)\rangle.\end{aligned} \tag{35}$$

For $j > i$ the commutativity of the diagram (32) furnishes

$$\begin{aligned}\lambda_{j-1}\circ\alpha_i = \nu_{j-1}\circ\cdots\circ\nu_{i+1}\circ\lambda_i\circ\alpha_i = 0 \text{ since}\\ \lambda_i\circ\alpha_i = 0;\ \text{hence}\ \mathrm{fact}_{i,j}(k,\widehat{k}) = \langle 0, \widehat{\sigma}_j(\widehat{k}_j)\rangle = 1.\end{aligned}$$

For $j = i$ we use the compatibility condition (2) from Assumption 53 and infer

$$\begin{aligned}\mathrm{fact}_{ii}(k,\widehat{k}) &= \langle \lambda_{i-1}\alpha_i\sigma_i(k_i), \widehat{\sigma}_i(\widehat{k}_i)\rangle = \langle \gamma_i\mu_i\sigma_i(k_i), \widehat{\sigma}_i(\widehat{k}_i)\rangle\\ &= \langle \mu_i\sigma_i(k_i), \gamma_i^\star\widehat{\sigma}_i(\widehat{k}_i)\rangle = \langle k_i, \widehat{k}_i\rangle.\end{aligned}$$

Thus

$$\mathrm{fact}_{ij}(k,\widehat{k}) = \langle \alpha_i\sigma_i(k_i), \lambda_{j-1}^\star\widehat{\sigma}_j(\widehat{k}_j)\rangle = \begin{cases}\langle k_i, \widehat{k}_i\rangle & \text{if } i = j,\\ 1 & \text{if } i < j.\end{cases} \tag{36}$$

From (35) and (36) we infer

$$\begin{aligned}\langle g, \widehat{g}\rangle &= \textstyle\prod_{j\le i} \mathrm{fact}_{ij}(k,\widehat{k}) = \prod_{i=1}^r\prod_{j=1}^i \mathrm{fact}_{ij}(k,\widehat{k})\\ &= \textstyle\prod_{i=1}^r \varphi_i(k_i;\widehat{k}_1,\cdots,\widehat{k}_i), \text{ where } \varphi_i : K_i\times\widehat{K_1}\times\cdots\times\widehat{K_i}\to K,\\ \varphi_i(k_i;\widehat{k}_1,\cdots,\widehat{k}_i) &:= \textstyle\prod_{j=1}^i \mathrm{fact}_{ij}(k,\widehat{k}) = \Big\langle \alpha_i\sigma_i(k_i), \sum_{j=1}^i \lambda_{j-1}^\star\widehat{\sigma}_j(\widehat{k}_j)\Big\rangle.\end{aligned} \tag{37}$$

The decisive property of the functions φ_i is that they depend on the first i components $\widehat{k}_1,\cdots,\widehat{k}_i$ of $\widehat{k}$ only. In the same fashion (35) and (36) give rise to the representation

$$\begin{aligned}\langle g, \widehat{g}\rangle &= \textstyle\prod_{j=1}^r\prod_{i=j}^r \mathrm{fact}_{ij}(k,\widehat{k}) = \prod_{j=1}^r \widehat{\varphi}_j(k_j,\cdots,k_r;\widehat{k}_j) \qquad \text{with}\\ &\widehat{\varphi}_j : K_j\times\cdots\times K_r\times\widehat{K_j}\to K,\\ \widehat{\varphi}_j(k_j,\cdots,k_r;\widehat{k}_j) &:= \textstyle\prod_{i=j}^r \mathrm{fact}_{ij}(k,\widehat{k}) = \Big\langle \sum_{i=j}^r \alpha_i\sigma_i(k_i), \lambda_{j-1}^\star\widehat{\sigma}_j(\widehat{k}_j)\Big\rangle.\end{aligned} \tag{38}$$

For fixed $\widehat{g} = \widehat{\mathrm{ind}}(\widehat{k}) \in \widehat{G}$ Lemma 55 and (37) imply

$$\begin{aligned}\widehat{a}(\widehat{g}) &= \textstyle\sum_{g\in G} a(g)\langle g, \widehat{g}\rangle = \sum_{k\in\prod_{i=1}^r K_i} a(\mathrm{ind}(k))\langle \mathrm{ind}(k), \widehat{\mathrm{ind}}(\widehat{k})\rangle\\ &= \textstyle\sum_{k_1\in K_1,\cdots,k_r\in K_r} a_0(k_1,\cdots,k_r)\prod_{i=1}^r \varphi_i(k_i;\widehat{k}_1,\cdots,\widehat{k}_i),\end{aligned} \tag{39}$$

where $a_0 = a\circ\mathrm{ind} = \mathrm{Ind}(a)$ according to Corollary 57. This formula for $\widehat{a}(\widehat{g})$ suggests that we define, for $\varrho = 1,\dots,r$, intermediate functions

$$\begin{aligned}a_\varrho &: \widehat{K_1}\times\cdots\times\widehat{K_\varrho}\times K_{\varrho+1}\times\cdots\times K_r\to K,\\ &a_\varrho(\widehat{k}_1,\cdots,\widehat{k}_\varrho,k_{\varrho+1},\cdots,k_r)\\ &:= \textstyle\sum_{k_1\in K_1,\cdots,k_\varrho\in K_\varrho} a_0(k_1,\cdots,k_r)\prod_{i=1}^\varrho \varphi_i(k_i;\widehat{k}_1,\cdots,\widehat{k}_i).\end{aligned} \tag{40}$$

By definition and according to (39), we have

$$a_0(k_1,\cdots,k_r) = a(\mathrm{ind}(k)) \text{ and } a_r(\widehat{k}_1,\cdots,\widehat{k}_r) = \widehat{a}(\widehat{\mathrm{ind}}(\widehat{k})).$$

The next theorem is the most important result of this paper. Its main idea, viz., the recursive computation of the DFT, is due to Cooley and Tukey [16] who developed the algorithm for the group $G = \mathbb{Z}/\mathbb{Z}2^r$ on the basis of the filtration

$$G_0 := 0 \subset G_1 := \mathbb{Z}2^{r-1}/\mathbb{Z}2^r \subset \cdots \subset G_i := \mathbb{Z}2^{r-i}/\mathbb{Z}2^r \subset \cdots \subset G_r = G.$$

Later it turned out that the same idea had been used before, in particular, by Gauss. See the introduction of [8] for a short historical survey. The "decimation in time" and "decimation in frequency" terminology used below comes from the application of the DFT to the computation of one-dimensional Fourier integrals or series where $\mathbb{R}$ or $\mathbb{Z}$ are interpreted as time or frequency models, and has been adapted from [8, pp. 188–191].

THEOREM 58 (Cooley–Tukey FFT or decimation in time). *The following recursive algorithm computes the Fourier transform $\widehat{a} \in K^{\widehat{G}}$ of a function $a \in K^G$. By induction on $\varrho = 0, \ldots, r$ define functions*

$$\begin{gathered}
a_\varrho : \widehat{K_1} \times \cdots \times \widehat{K_\varrho} \times K_{\varrho+1} \times \cdots \times K_r \to K \qquad \textit{by} \\
a_0(k_1, \cdots, k_r) := a\left(\textstyle\sum_{i=1}^r \alpha_i \sigma_i(k_i)\right) \ \textit{and for } 1 \le \varrho \le r \\
a_\varrho(\widehat{k}_1, \cdots, \widehat{k}_\varrho, k_{\varrho+1}, \cdots, k_r) \\
:= \textstyle\sum_{k_\varrho \in K_\varrho} a_{\varrho-1}(\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, k_\varrho, k_{\varrho+1}, \cdots, k_r)\varphi_\varrho(k_\varrho; \widehat{k}_1, \cdots, \widehat{k}_\varrho), \qquad \textit{where} \\
\varphi_\varrho(k_\varrho; \widehat{k}_1, \cdots, \widehat{k}_\varrho) = \left\langle \alpha_\varrho \sigma_\varrho(k_\varrho), \textstyle\sum_{j=1}^\varrho \lambda_{j-1}^\star \widehat{\sigma}_j(\widehat{k}_j) \right\rangle .
\end{gathered}$$

Then

$$\widehat{a}(\widehat{g}) = a_r(\widehat{k}_1, \cdots, \widehat{k}_r) \textit{ for } \widehat{g} = \widehat{\mathrm{ind}}(\widehat{k}) = \sum_{j=1}^r \lambda_{j-1}^\star \widehat{\sigma}_j(\widehat{k}_j) \in \widehat{G}, \widehat{k}_j \in \widehat{K_j}.$$

Proof. It remains to show that the functions a_ϱ defined in (40) satisfy these recursive relations. But for $\varrho > 0$

$$\begin{gathered}
a_\varrho(\widehat{k}_1, \cdots, \widehat{k}_\varrho, k_{\varrho+1}, \cdots, k_r) \\
= \textstyle\sum_{k_1, \cdots, k_\varrho} a_0(k_1, \cdots, k_r) \prod_{i=1}^\varrho \varphi_i(k_i; \widehat{k}_1, \cdots, \widehat{k}_i) \\
= \textstyle\sum_{k_\varrho \in K_\varrho} \varphi_\varrho(k_\varrho; \widehat{k}_1, \cdots, \widehat{k}_\varrho) \sum_{k_1, \cdots, k_{\varrho-1}} a_0(k_1, \cdots, k_r) \prod_{i=1}^{\varrho-1} \varphi_i(k_i; \widehat{k}_1, \cdots, \widehat{k}_i) \\
= \textstyle\sum_{k_\varrho \in K_\varrho} \varphi_\varrho(k_\varrho; \widehat{k}_1, \cdots, \widehat{k}_\varrho) a_{\varrho-1}(\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, k_\varrho, \cdots, k_r). \qquad \square
\end{gathered}$$

The induction formula of the preceding theorem can be given another traditional form. From the definition of φ_ϱ in (37) and from (36) we infer

$$\begin{aligned}
\varphi_\varrho(k_\varrho; \widehat{k}_1, \cdots, \widehat{k}_\varrho) &= \left\langle \alpha_\varrho \sigma_\varrho(k_\varrho), \textstyle\sum_{j=1}^\varrho \lambda_{j-1}^\star \widehat{\sigma}_j(\widehat{k}_j) \right\rangle \\
&= \langle k_\varrho, \widehat{k}_\varrho \rangle \left\langle \alpha_\varrho \sigma_\varrho(k_\varrho), \textstyle\sum_{j=1}^{\varrho-1} \lambda_{j-1}^\star \widehat{\sigma}_j(\widehat{k}_j) \right\rangle
\end{aligned}$$

or

$$\begin{gathered}
\varphi_\varrho(k_\varrho; \widehat{k}_1, \cdots, \widehat{k}_\varrho) = \langle k_\varrho, \widehat{k}_\varrho \rangle \tau_\varrho(k_\varrho; \widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}) \text{ with} \\
\tau_\varrho(k_\varrho; \widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}) := \left\langle \alpha_\varrho \sigma_\varrho(k_\varrho), \textstyle\sum_{j=1}^{\varrho-1} \lambda_{j-1}^\star \widehat{\sigma}_j(\widehat{k}_j) \right\rangle .
\end{gathered} \tag{41}$$

The elements τ_ϱ are roots of unity and hence nonzero and are usually called the *twiddle factors* [8, p. 191], [5, p. 121]. With their help we define, for $\varrho = 1, \dots, r$, the isomorphisms

$$
\begin{aligned}
T_\varrho &: K^{\widehat{K_1}\times\cdots\times\widehat{K_{\varrho-1}}\times K_\varrho\times\cdots\times K_r} \to K^{\widehat{K_1}\times\cdots\times\widehat{K_\varrho}\times K_{\varrho+1}\times\cdots\times K_r},\\
&T_\varrho(c)(\widehat{k}_1, \cdots, \widehat{k}_\varrho, k_{\varrho+1}, \cdots, k_r)\\
&:= \textstyle\sum_{k_\varrho\in K_\varrho} c(\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, k_\varrho, \cdots, k_r)\varphi_\varrho(k_\varrho; \widehat{k}_1, \cdots, \widehat{k}_\varrho)\\
&= \mathrm{Four}_{K_\varrho}[c(\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, -, k_{\varrho+1}, \cdots, k_r)\tau_\varrho(-; \widehat{k}_1, \cdots, \widehat{k}_{\varrho-1})](\widehat{k}_\varrho).
\end{aligned}
\tag{42}
$$

Here the argument of $\mathrm{Four}_{K_\varrho}$ is a function in K^{K_ϱ} which depends on the parameters $\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, k_{\varrho+1}, \cdots, k_r$. The map T_ϱ is an isomorphism since the multiplication with τ_ϱ and the Fourier transform $\mathrm{Four}_{K_\varrho}$ are bijective.

THEOREM 59. *In the situation of Theorem* 58 *the induction formula computing* a_ϱ *from* $a_{\varrho-1}$ *can be expressed as* $a_\varrho = T_\varrho(a_{\varrho-1})$, $\varrho = 1, \dots, r$; *hence*

$$\mathrm{Four}_G = \widehat{\mathrm{Ind}}^{-1} \circ T_r \circ \cdots \circ T_1 \circ \mathrm{Ind} : K^G \to K^{\widehat{G}}.$$

With $e_\varrho := \mathrm{ord}(K_\varrho)$ *and* $N := e_1 * \cdots * e_r = \mathrm{ord}(G)$ *the complexity satisfies*

$$\mu(\mathrm{Four}_G) \le N(e_1 + \cdots + e_r - r).$$

Proof. The first assertion is obvious. Concerning the complexity recall the condition $\sigma_\varrho(0) = 0$ from Assumption 53 and

$$\varphi_\varrho(k_\varrho; \widehat{k}_1, \cdots, \widehat{k}_\varrho) = \left\langle \alpha_\varrho\sigma_\varrho(k_\varrho), \textstyle\sum_{j=1}^{\varrho} \lambda_{j-1}^\star \widehat{\sigma}_j(\widehat{k}_j) \right\rangle; \text{ hence } \varphi_\varrho(0; \widehat{k}_1, \cdots, \widehat{k}_\varrho) = 1.$$

From this and (42) we infer $\mu(T_\varrho) \le N(e_\varrho - 1)$ as in the proof of part 3 of Corollary 46. Since index transformations are costless according to Definition and Corollary 48, we conclude by means of Theorem 49 that

$$\mu(\mathrm{Four}_G) \le \textstyle\sum_\varrho \mu(T_\varrho) \le N(e_1 - 1 + \cdots + e_r - 1) = N(e_1 + \cdots + e_r - r). \qquad \square$$

Remark 60 (butterfly diagrams). In the literature special cases of the induction formula of Theorem 58 are often represented by means of a directed graph or so-called butterfly diagram. Such a graph can be introduced in general; it is, however, useless for the actual execution of the fast algorithm. Its graphical representation in the plane is also of no practical significance and, moreover, is complicated except in the simplest cases such as $G = \mathbb{Z}/\mathbb{Z}8$ where it is usually shown. Indeed, consider the graph $\Gamma := (V, E)$ with vertex (resp., edge) sets V (resp., E), where

$$V := \biguplus_{\varrho=0}^r V_\varrho, V_\varrho := \widehat{K_1} \times \cdots \times \widehat{K_\varrho} \times K_{\varrho+1} \times \cdots \times K_r,\ E \subset V \times V,$$

with edges from $(\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, k_\varrho, \cdots, k_r)$ to $(\widehat{k}_1, \cdots, \widehat{k}_\varrho, k_{\varrho+1}, \cdots, k_r)$ or from $V_{\varrho-1}$ to V_ϱ only. For $w = (\widehat{k}_1, \cdots, \widehat{k}_\varrho, k_{\varrho+1}, \cdots, k_r) \in V_\varrho$, $\varrho \ge 1$, there results the bijection

$$
\begin{aligned}
K_\varrho &\cong \{(v, w); (v, w) \text{ is an edge of } \Gamma \text{ with endpoint } w\},\\
k_\varrho &\mapsto (v, w), v := (\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, k_\varrho, \cdots, k_r) \in V_{\varrho-1}.
\end{aligned}
$$

With the abbreviation $a(v) := a_{\varrho-1}(\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, k_\varrho, \cdots, k_r)$ the recursion formula of Theorem 58 has the form

$$a(w) = \sum a(v)\varphi_\varrho(k_\varrho; \widehat{k}_1, \cdots, \widehat{k}_\varrho),$$

where $v = (\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, k_\varrho, \cdots, k_r)$ runs over all sources of edges with sink w.

The next theorem on the "decimation in frequency" FFT computes $\mathrm{Four}_{\widehat{G}} : K^{\widehat{G}} \to K^G$ and is proved in the same fashion as Theorem 58 on the basis of (38) instead of (37). For the choice $\widehat{G} = G$ it yields a second fast algorithm for the computation of Four_G (compare [8, p. 192]).

THEOREM 61 (Sande–Tukey FFT or decimation in frequency). *Using data from Assumption* 53 *and Corollary* 54, *the following algorithm computes the Fourier transform* $\widehat{b} \in K^G$ *of a function* $b \in K^{\widehat{G}}$. *By recursion from* $\varrho = r$ *to* 0 *define functions*

$$b_\varrho : \widehat{K_1} \times \cdots \times \widehat{K_\varrho} \times K_{\varrho+1} \times \cdots \times K_r \to K, \varrho = r, \ldots, 0,$$
$$b_r(\widehat{k}_1, \cdots, \widehat{k}_r) := b\left(\textstyle\sum_{j=1}^r \lambda_{j-1}^\star \widehat{\sigma}_j(\widehat{k}_j)\right), \widehat{k}_j \in \widehat{K}_j, \text{ and for } r \geq \varrho > 0$$
$$b_{\varrho-1}(\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, k_\varrho, \cdots, k_r)$$
$$:= \textstyle\sum_{\widehat{k}_\varrho \in \widehat{K}_\varrho} b_\varrho(\widehat{k}_1, \cdots, \widehat{k}_\varrho, k_{\varrho+1}, \cdots, k_r)\widehat{\varphi}_\varrho(k_\varrho, \cdots, k_r; \widehat{k}_\varrho).$$

Then

$$\widehat{b}(g) = \textstyle\sum_{\widehat{g} \in \widehat{G}} b(\widehat{g})\langle g, \widehat{g}\rangle = b_0(k_1, \cdots, k_r) \text{ for } g = \sum_{i=1}^r \alpha_i\sigma_i(k_i) \in G, k_i \in K_i.$$

Proof. According to (38) we have

$$\widehat{b}(g) = \textstyle\sum_{\widehat{k} \in \prod_{j=1}^r \widehat{K_j}} b(\widehat{\mathrm{ind}}(\widehat{k}))\langle \mathrm{ind}\, k, \widehat{\mathrm{ind}}(\widehat{k})\rangle$$
$$= \textstyle\sum_{\widehat{k}_1 \in \widehat{K_1}, \cdots, \widehat{k}_r \in \widehat{K_r}} b_r(\widehat{k}_1, \cdots, \widehat{k}_r) \prod_{j=1}^r \widehat{\varphi}_j(k_j, \cdots, k_r; \widehat{k}_j).$$

In analogy to (40) we define functions $b_\varrho, r \geq \varrho \geq 0$, by

$$b_r(\widehat{k}_1, \cdots, \widehat{k}_r) := b(\widehat{\mathrm{ind}}(\widehat{k})) \text{ and for } \varrho < r \text{ by}$$
$$b_\varrho(\widehat{k}_1, \cdots, \widehat{k}_\varrho, k_{\varrho+1}, \cdots, k_r)$$
$$= \textstyle\sum_{\widehat{k}_{\varrho+1} \in \widehat{K}_{\varrho+1}, \cdots, \widehat{k}_r \in \widehat{K}_r} b_r(\widehat{k}_1, \cdots, \widehat{k}_r) \prod_{j=\varrho+1}^r \widehat{\varphi}_j(k_j, \cdots, k_r; \widehat{k}_j),$$

and show that they satisfy the recursive relations from which the theorem follows directly. But, indeed,

$$b_{\varrho-1}(\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, k_\varrho, \cdots, k_r)$$
$$= \textstyle\sum_{\widehat{k}_\varrho \in \widehat{K}_\varrho} \widehat{\varphi}_\varrho(k_\varrho, \cdots, k_r; \widehat{k}_\varrho) \sum_{\widehat{k}_{\varrho+1} \in \widehat{K}_{\varrho+1}, \cdots, \widehat{k}_r \in \widehat{K}_r} b_r(\widehat{k}_1, \cdots, \widehat{k}_r)$$
$$* \textstyle\prod_{j=\varrho+1}^r \widehat{\varphi}_j(k_j, \cdots, k_r; \widehat{k}_j)$$
$$= \textstyle\sum_{\widehat{k}_\varrho \in \widehat{K}_\varrho} \widehat{\varphi}_\varrho(k_\varrho, \cdots, k_r; \widehat{k}_\varrho) b_\varrho(\widehat{k}_1, \cdots, \widehat{k}_\varrho, k_{\varrho+1}, \cdots, k_r). \qquad \square$$

In analogy to (41) and (42) we also obtain, for $\varrho = r, \ldots, 1$,

$$\widehat{\varphi}_\varrho(k_\varrho, \cdots, k_r; \widehat{k}_\varrho) = \langle k_\varrho, \widehat{k}_\varrho\rangle \widehat{\tau}_\varrho(k_{\varrho+1}, \cdots, k_r; \widehat{k}_\varrho),$$
$$\widehat{\tau}_\varrho(k_{\varrho+1}, \cdots, k_r; \widehat{k}_\varrho) := \left\langle \textstyle\sum_{i=\varrho+1}^r \alpha_i\sigma_i(k_i), \lambda_{\varrho-1}^\star \widehat{\sigma}_\varrho(\widehat{k}_\varrho)\right\rangle, \tag{43}$$

and define the isomorphism

$$\begin{gathered}\widehat{T}_\varrho : K^{\widehat{K_1}\times\cdots\times\widehat{K_\varrho}\times K_{\varrho+1}\times\cdots\times K_r} \cong K^{\widehat{K_1}\times\cdots\times\widehat{K_{\varrho-1}}\times K_\varrho\times\cdots\times K_r},\\ \widehat{T}_\varrho(c) := \mathrm{Four}_{\widehat{K_\varrho}}[c(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},-,k_{\varrho+1},\cdots,k_r)\widehat{\tau}_\varrho(k_{\varrho+1},\cdots,k_r;-)].\end{gathered} \tag{44}$$

THEOREM 62. *In the situation of Theorem* 61 *and with the isomorphisms* $\widehat{T}_\varrho$ *from* (44) *and* $\mathrm{Ind}, \widehat{\mathrm{Ind}}$ *from Corollary* 57*, we have*

$$\begin{gathered}\mathrm{Four}_{\widehat{G}} = \mathrm{Ind}^{-1}\circ\widehat{T}_1\circ\cdots\circ\widehat{T}_r\circ\widehat{\mathrm{Ind}} \text{ and}\\ \mu(\mathrm{Four}_{\widehat{G}}) \le N(e_1+\cdots+e_r-r),\end{gathered}$$

where $e_\varrho := \mathrm{ord}(K_\varrho)$ *and* $N := e_1 * \cdots * e_r = \mathrm{ord}(G)$.

Theorems 59 and 62 signify that the FFT-algorithms in Theorems 58 and 61 are *fast*, i.e., of relatively *low complexity* $N(e_1+\cdots+e_r-r)$ instead of the $N(N-1)$ of the direct computation of Four_G. Recall that the algorithms and their complexity depend on the diagrams from Assumption 53.

The best FFT-algorithms according to the preceding theorems are obtained when the diagrams from Assumption 53 are constructed by means of a Jordan–Hölder series of G (see the explanation after (33)). Then $N = e_1 * \cdots * e_r$ is the prime factor decomposition of G.

For the next theorem we introduce a logarithm type arithmetic function Λ. Let $\mathbb{N} := \{0,1,\cdots\}$ denote the additive monoid of natural numbers and $\mathbb{N}_{>0} := \{1,2,\cdots\}$ the multiplicative monoid of positive numbers. Every $N \in \mathbb{N}_{>0}$ admits the unique prime factor decomposition

$$N = \textstyle\prod_{p\in\mathcal{P}} p^{\mathrm{ord}_p(N)},\ \mathrm{ord}_p(N) = 0 \text{ for almost all } p,$$

where $\mathcal{P} = \{2,3,5,\cdots\}$ is the set of prime numbers. The standard isomorphism

$$\mathbb{N}_{>0} \cong \mathbb{N}^{(\mathcal{P})} := \{\nu \in \mathbb{N}^{\mathcal{P}};\ \nu(p) = 0 \text{ for almost all } p \in \mathcal{P}\},\ N \mapsto (\mathrm{ord}_p(N))_{p\in\mathcal{P}},$$

follows and induces the composed epimorphism

$$\Lambda : \mathbb{N}_{>0} \cong \mathbb{N}^{(\mathcal{P})} \to \mathbb{N},\ N \mapsto (\mathrm{ord}_p(N))_{p\in\mathcal{P}} \mapsto \Lambda(N) := \textstyle\sum_{p\in\mathcal{P}}(p-1)\,\mathrm{ord}_p(N); \tag{45}$$

hence $\Lambda(1) = 0,$ and $\Lambda(M*N) = \Lambda(M) + \Lambda(N)$. The obvious inequality

$$\begin{gathered}1+(p-1)m < (1+p-1)^m = p^m \text{ for } m \ge 2 \text{ implies } \Lambda(N) \le N-1 \text{ and}\\ \Lambda(N) = N-1 \ \Leftrightarrow\ N = 1 \text{ or } N \text{ is prime.}\end{gathered}$$

THEOREM 63. *Let* G *be an abelian group of exponent* d *and order* N*. Then*

$$\mu(\mathrm{Four}_G) \le N\Lambda(N) \le N(N-1).$$

The equality $N(N-1) = N\Lambda(N)$ *holds if and only if* G *is simple or zero.*

Proof. Choose a Jordan–Hölder series of G, the corresponding diagrams (31) as those in (32), and the FFT-algorithms derived from these diagrams. Then the numbers e_ϱ are exactly the prime factors of $N = e_1 * \cdots * e_r$ and

$$\begin{gathered}\Lambda(e_\varrho) = e_\varrho - 1,\ \Lambda(N) = \Lambda(e_1 * \cdots * e_r) = \textstyle\sum_\varrho \Lambda(e_\varrho) = \sum_\varrho (e_\varrho - 1);\ \text{hence}\\ \mu(\mathrm{Four}_G) \le N(e_1 - 1 + \cdots + e_r - 1) = N\Lambda(N). \qquad \square\end{gathered}$$

Examples 64. Let G be a group of order N.

(1) The first standard case was that of Cooley and Tukey [16]:

$$N = 2^r,\ \Lambda(N) = (2-1) * r = r,\ \mu(\mathrm{Four}_G) \leq 2^r * r = N \log_2(N).$$

(2)

$$N = 675 = 3^3 * 5^2,\ \Lambda(N) = (3-1) * 3 + (5-1) * 2 = 14 \text{ and}$$
$$N\Lambda(N) = 675 * 14 = 9450 < N(N-1) = 454950.$$

(3) $G := \mathbb{Z}/\mathbb{Z}2^{10} \times \mathbb{Z}/\mathbb{Z}2^{10},\ N = 2^{20}$. This group can be considered as a lattice with approximately one million points and may, for instance, be used for digital image processing. The direct computation of Four_G has complexity $N(N-1) \sim 2^{40}$, whereas that of the FFT is $20 * 2^{20} = 1,25 * 2^{24}$. The improvement of the complexity is dramatic.

6. The FFT in the standard cases. Assumption 29 remains in force. In this section we derive the standard special cases of the FFT and start with that of a cyclic group $G = \mathbb{Z}/\mathbb{Z}n$ of exponent $d > 0$, i.e., with $n \mid d$. As usual in the engineering literature we often identify

$$\mathbb{Z}/\mathbb{Z}n = \{0, 1, \cdots, n-1\},\ k = \overline{k},\ 0 \leq k \leq n-1, \tag{46}$$

and emphasize that the necessary care has to be taken in context with this identification. For $G = \mathbb{Z}/\mathbb{Z}n$ we choose

$$\widehat{G} := G = \mathbb{Z}/\mathbb{Z}n, \langle \overline{k}, \overline{l} \rangle := \zeta^{kld/n},\ \overline{k}, \overline{l} \in \mathbb{Z}/\mathbb{Z}n, \tag{47}$$

according to Theorem 2. A factorization $n = n_1 n_2$ of n gives rise to the exact sequence with a natural section $\sigma : \mathbb{Z}/\mathbb{Z}n_2 \to \mathbb{Z}/\mathbb{Z}n$:

(48)

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathbb{Z}/\mathbb{Z}n_1 & \xrightarrow{\text{inj}} & \mathbb{Z}/\mathbb{Z}n & \xrightarrow{\text{can}} & \mathbb{Z}/\mathbb{Z}n_2 & \longrightarrow & 0 \\ & & \| & & \| & & \| & & \\ & & \{0, \cdots, n_1 - 1\} & & \{0, \cdots, n-1\} & & \{0, \cdots, n_2 - 1\}, & & \end{array}$$

where $\mathrm{inj}(\overline{k_1}) := \overline{k_1 n_2},\ \mathrm{can}(\overline{k}) := \overline{k},\ \sigma(\overline{k_2}) := \overline{k_2}$ if $0 \leq k_2 \leq n_2 - 1$. For $\overline{k} \in \mathbb{Z}/\mathbb{Z}n$ and $\overline{k_2} \in \mathbb{Z}/\mathbb{Z}n_2$ the equations

$$\langle \mathrm{can}(\overline{k}), \overline{k_2} \rangle_{\mathbb{Z}/\mathbb{Z}n_2} = \zeta^{kk_2 d/n_2} = \zeta^{k(k_2 n_1)d/n} = \langle \overline{k}, \mathrm{inj}(\overline{k_2}) \rangle_{\mathbb{Z}/\mathbb{Z}n}$$

prove $\mathrm{can}^\star(\overline{k_2}) = \mathrm{inj}(\overline{k_2})$; hence

$$\begin{gathered} (\mathrm{can} : \mathbb{Z}/\mathbb{Z}n \to \mathbb{Z}/\mathbb{Z}n_2)^\star = \mathrm{inj} : \mathbb{Z}/\mathbb{Z}n_2 \to \mathbb{Z}/\mathbb{Z}n \text{ and} \\ (\mathrm{inj} : \mathbb{Z}/\mathbb{Z}n_1 \to \mathbb{Z}/\mathbb{Z}n)^\star = \mathrm{can} : \mathbb{Z}/\mathbb{Z}n \to \mathbb{Z}/\mathbb{Z}n_1, \end{gathered} \tag{49}$$

the second equality following from the first by means of $\mathrm{inj}^\star = (\mathrm{can}^\star)^\star = \mathrm{can}$. Now assume that a factorization of d is given from which we derive the following data:

$$\begin{gathered} d = e_1 * \cdots * e_r,\ d_1(i) := e_1 * \cdots * e_i,\ i = 0, \ldots, r; \text{ hence} \\ d_1(0) = 1,\ d_1(i) = d_1(i-1) * e_i, \\ d_2(j) := d/d_1(j) = e_{j+1} * \cdots * e_r,\ j = r, \ldots, 0, \\ d_2(r) = 1,\ d_2(j-1) = d_2(j) * e_j. \end{gathered} \tag{50}$$

From (50) and by means of (48) we construct commutative exact diagrams of the type (53):

$$
\begin{array}{ccccccccc}
 & & & & 0 & & & & \\
 & & & & \downarrow & & & & \\
 & & 0 & & \mathbb{Z}/\mathbb{Z}e_i & & & & \\
 & & \downarrow & & \downarrow{\scriptstyle \gamma_i=\mathrm{inj}} & & & & \\
0 & \longrightarrow & \mathbb{Z}/\mathbb{Z}d_1(i-1) & \xrightarrow{\alpha_{i-1}=\mathrm{inj}} & \mathbb{Z}/\mathbb{Z}d & \xrightarrow{\lambda_{i-1}=\mathrm{can}} & \mathbb{Z}/\mathbb{Z}d_2(i-1) & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \beta_i=\mathrm{inj}} & & \| & & \downarrow{\scriptstyle \nu_i=\mathrm{can}} & & \\
0 & \longrightarrow & \mathbb{Z}/\mathbb{Z}d_1(i) & \xrightarrow{\alpha_i=\mathrm{inj}} & \mathbb{Z}/\mathbb{Z}d & \xrightarrow{\lambda_i=\mathrm{can}} & \mathbb{Z}/\mathbb{Z}d_2(i) & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \mu_i=\mathrm{can}} & & & & \downarrow & & \\
 & & \mathbb{Z}/\mathbb{Z}e_i & & & & 0 & & \\
 & & \downarrow & & & & & & \\
 & & 0 & & & & & &
\end{array}
\tag{51}
$$

with the natural sections σ from (48) in $\mathbb{Z}/\mathbb{Z}d_1(i) \underset{\sigma_i=\sigma}{\overset{\mu_i=\mathrm{can}}{\rightleftarrows}} \mathbb{Z}/\mathbb{Z}e_i$. Application of the exact duality functor $H \mapsto \widehat{H} = H$ to the cyclic groups of the preceding diagram and the identities (49) yield the dual commutative exact diagrams of the type (34):

$$
\begin{array}{ccccccccc}
 & & & & 0 & & & & \\
 & & & & \downarrow & & & & \\
 & & 0 & & \mathbb{Z}/\mathbb{Z}e_j & & & & \\
 & & \downarrow & & \downarrow{\scriptstyle \mu_j^\star=\mathrm{inj}} & & & & \\
0 & \longrightarrow & \mathbb{Z}/\mathbb{Z}d_2(j) & \xrightarrow{\lambda_j^\star=\mathrm{inj}} & \mathbb{Z}/\mathbb{Z}d & \xrightarrow{\alpha_j^\star=\mathrm{can}} & \mathbb{Z}/\mathbb{Z}d_1(j) & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \nu_j^\star=\mathrm{inj}} & & \| & & \downarrow{\scriptstyle \beta_j^\star=\mathrm{can}} & & \\
0 & \longrightarrow & \mathbb{Z}/\mathbb{Z}d_2(j-1) & \xrightarrow{\lambda_{j-1}^\star=\mathrm{inj}} & \mathbb{Z}/\mathbb{Z}d & \xrightarrow{\alpha_{j-1}^\star=\mathrm{can}} & \mathbb{Z}/\mathbb{Z}d_1(j-1) & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \gamma_j^\star=\mathrm{can}} & & & & \downarrow & & \\
 & & \mathbb{Z}/\mathbb{Z}e_j & & & & 0 & & \\
 & & \downarrow & & & & & & \\
 & & 0 & & & & & &
\end{array}
\tag{52}
$$

with the natural sections $\widehat{\sigma}$ from (48) in $\mathbb{Z}/\mathbb{Z}d_2(j-1) \underset{\widehat{\sigma}_j=\sigma}{\overset{\gamma_j^\star=\mathrm{can}}{\rightleftarrows}} \mathbb{Z}/\mathbb{Z}e_j$. According to Lemma 55 the diagram (51) gives rise to the index bijection

$$\begin{gathered}\operatorname{ind} : \prod_{i=1}^{r} \mathbb{Z}/\mathbb{Z}e_i = \prod_{i=1}^{r} \{0, \cdots, e_i - 1\} \cong \mathbb{Z}/\mathbb{Z}d = \{0, \cdots, d-1\}, \\ \operatorname{ind}(k_1, \cdots, k_r) = \sum_{i=1}^{r} \alpha_i \sigma_i(k_i) = \sum_{i=1}^{r} k_i d/d_1(i) \\ = \sum_{i=1}^{r} k_i d_2(i) = \sum_{i=1}^{r} k_i * e_{i+1} * \cdots * e_r. \end{gathered} \tag{53}$$

Likewise, the diagram (52) induces the bijection

$$\begin{gathered}\widehat{\operatorname{ind}} : \prod_{j=1}^{r} \mathbb{Z}/\mathbb{Z}e_j = \prod_{j=1}^{r} \{0, \cdots, e_j - 1\} \cong \mathbb{Z}/\mathbb{Z}d = \{0, \cdots, d-1\}, \\ \widehat{\operatorname{ind}}(\widehat{k}_1, \cdots, \widehat{k}_r) = \sum_{j=1}^{r} \lambda_{j-1}^{\star} \widehat{\sigma}_j(\widehat{k}_j) = \sum_{j=1}^{r} \widehat{k}_j d/d_2(j-1) \\ = \sum_{j=1}^{r} \widehat{k}_j * e_1 * \cdots * e_{j-1}. \end{gathered} \tag{54}$$

Corollary 65. *The unique representation*

$$n = \sum_{i=1}^{r} k_i * e_{i+1} * \cdots * e_r,\ 0 \le n < d,\ 0 \le k_i < e_i,\ i = 1, \ldots, r,$$

according to (53) *is obtained by recursion with respect to* i *as*

$$n_r := n,\ n_i := n_{i-1} * e_i + k_i,\ 0 \le k_i < e_i,\ i = r, \ldots, 1.$$

Likewise, the unique representation

$$n = \sum_{j=1}^{r} \widehat{k}_j * e_1 * \cdots * e_{j-1},\ 0 \le n < d,\ 0 \le \widehat{k}_j < e_j,\ j = 1, \ldots, r,$$

according to (54) *is obtained by induction with respect to* j *as*

$$\widehat{n}_0 := n,\ \widehat{n}_{j-1} = \widehat{n}_j * e_j + \widehat{k}_j,\ 0 \le \widehat{k}_j < e_j,\ j = 1, \ldots, r.$$

Proof. The proof is the same as that of the q-adic representation of a natural number for $q > 1$. For instance,

$$\begin{gathered} d > n =: n_r := n_{r-1} * e_r + k_r,\ 0 \le k_r < e_r,\ n_{r-1} \le \tfrac{n}{e_r} < \tfrac{d}{e_r} = e_1 * \cdots * e_{r-1}, \\ d > n =: \widehat{n}_0 = \widehat{n}_1 * e_1 + \widehat{k}_1,\ 0 \le \widehat{k}_1 < e_1,\ \widehat{n}_1 < e_2 * \cdots * e_r. \end{gathered}$$

□

For vectors $(k_i; \widehat{k}_1, \cdots, \widehat{k}_i)$ with components $k_i,\ \widehat{k}_i \in \mathbb{Z}/\mathbb{Z}e_i = \{0, \cdots, e_i - 1\}$ the function φ_i according to (37) is defined by

$$\begin{gathered}\varphi_i(k_i; \widehat{k}_1, \cdots, \widehat{k}_i) = \left\langle \alpha_i \sigma_i(k_i), \sum_{j=1}^{i} \lambda_{j-1}^{\star} \widehat{\sigma}_j(\widehat{k}_j) \right\rangle \\ = \left\langle k_i d/d_1(i), \sum_{j=1}^{i} \widehat{k}_j d/d_2(j-1) \right\rangle = \zeta^{\varepsilon_i(k_i; \widehat{k}_1, \cdots, \widehat{k}_i)}, \text{ where} \\ \varepsilon_i(k_i; \widehat{k}_1, \cdots, \widehat{k}_i) := k_i * e_{i+1} * \cdots * e_r * \sum_{j=1}^{i} \widehat{k}_j * e_1 * \cdots * e_{j-1},\ 0 \le k_i,\ \widehat{k}_i < e_i. \end{gathered} \tag{55}$$

Similarly the function $\widehat{\varphi}_j$ from (38) has the form

$$\begin{gathered}\widehat{\varphi}_j(k_j, \cdots, k_r; \widehat{k}_j) = \zeta^{\widehat{\varepsilon}_j(k_j, \cdots, k_r; \widehat{k}_j)},\ j = 1, \cdots, r, \text{ where} \\ \widehat{\varepsilon}_j(k_j, \cdots, k_r; \widehat{k}_j) := \widehat{k}_j * e_1 * \cdots * e_{j-1} * \sum_{i=j}^{r} k_i * e_{i+1} * \cdots * e_r,\ 0 \le k_i,\ \widehat{k}_i < e_i. \end{gathered} \tag{56}$$

Theorem 58 applied to the preceding situation now implies the following theorem.

Theorem 66 (FFT for cyclic groups [8, pp. 188–191]). *Consider a number* $d > 0$ *with a factorization* $d = e_1 * \cdots * e_r$*, the cyclic group* $G := \mathbb{Z}/\mathbb{Z}d$*, and the associated DFT*

$$\begin{gathered}\operatorname{Four}_{\mathbb{Z}/\mathbb{Z}d} : K^{\mathbb{Z}/\mathbb{Z}d} = K^{\{0, \cdots, d-1\}} = K^d \to K^d,\ a \mapsto \widehat{a}, \\ \widehat{a}(l) := \sum_{k=0}^{d-1} a(k) \zeta^{kl},\ 0 \le l < d. \end{gathered}$$

1. *The following "decimation in time" algorithm computes* $\widehat{a}$ *from* a *with complexity* $d(e_1+\cdots+e_r-r)$. *Inductively define functions*

$$a_\varrho : \prod_{i=1}^{r}\{0,\cdots,e_i-1\} \to K \text{ for } \varrho = 0,\dots,r \text{ by}$$
$$a_0(k_1,\cdots,k_r) := a\left(\textstyle\sum_{i=1}^{r} k_i * e_{i+1} * \cdots * e_r\right),$$
$$a_\varrho(\widehat{k}_1,\cdots,\widehat{k}_\varrho,k_{\varrho+1},\cdots,k_r) := \textstyle\sum_{k_\varrho=0}^{e_\varrho-1} a_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},k_\varrho,\cdots,k_r)\zeta^{\varepsilon_\varrho(k_\varrho;\widehat{k}_1,\cdots,\widehat{k}_\varrho)}$$

with ε_ϱ *from* (55). *Then*

$$\widehat{a}(l) = a_r(\widehat{k}_1,\cdots,\widehat{k}_r) \text{ for } l = \textstyle\sum_{j=1}^{r}\widehat{k}_j * e_1 * \cdots * e_{j-1},\ 0 \le l < d, 0 \le \widehat{k}_j < e_j.$$

2. *The following "decimation in frequency" algorithm also computes* $\widehat{a}$ *with complexity* $d(e_1+\cdots+e_r-r)$. *Recursively define functions*

$$b_\varrho : \prod_{\varrho=1}^{r}\{0,\cdots,e_\varrho-1\} \to K \text{ for } \varrho = r,\dots,0 \text{ by}$$
$$b_r(\widehat{k}_1,\cdots,\widehat{k}_r) := a\left(\textstyle\sum_{j=1}^{r}\widehat{k}_j * e_1 * \cdots * e_{j-1}\right),$$
$$b_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},k_\varrho,\cdots,k_r) = \textstyle\sum_{\widehat{k}_\varrho=0}^{e_\varrho-1} b_\varrho(\widehat{k}_1,\cdots,\widehat{k}_\varrho,k_{\varrho+1},\cdots,k_r)\zeta^{\widehat{\varepsilon}_\varrho(k_\varrho,\cdots,k_r;\widehat{k}_\varrho)}$$

with $\widehat{\varepsilon}_\varrho$ *from* (56). *Then*

$$\widehat{a}(k) = b_0(k_1,\cdots,k_r) \text{ for } k = \textstyle\sum_{i=1}^{r} k_i * e_{i+1} * \cdots * e_r, 0 \le k < d, 0 \le k_i < e_i.$$

Example 67. In the situation of the preceding theorem we choose

$$K := \mathbb{C},\ d = 6 = 2*3,\ G := \mathbb{Z}/\mathbb{Z}6 = \{0,\cdots,5\},$$
$$\zeta := \exp(2\pi i/6) = 1/2 + i\sqrt{3}/2,\ \zeta^6 = 1,$$
$$\widehat{a}(l) = \textstyle\sum_{k=0}^{5} a(k)\zeta^{kl},\ 0 \le k,l \le 5.$$

The FFT-algorithm of the preceding theorem computes $\widehat{a}$ from $a = (a(0),\cdots,a(5))$ with $6*(2+3-2) = 18$ elementary computation steps. The root ζ satisfies the cyclotomic equation $\phi_6(\zeta) = \zeta^2-\zeta+1 = 0$ or $\zeta^2 = \zeta-1$, hence the group table

k	0	1	2	3	4	5
ζ^k	1	ζ	$\zeta-1$	-1	$-\zeta$	$-\zeta+1$

The index functions are

$$\mathrm{ind}(k_1,k_2) = k_1 * e_2 + k_2 = 3k_1 + k_2 \text{ and}$$
$$\widehat{\mathrm{ind}}(\widehat{k}_1,\widehat{k}_2) = \widehat{k}_1 + \widehat{k}_2 * e_1 = \widehat{k}_1 + 2\widehat{k}_2, \quad 0 \le k_1,\widehat{k}_1 \le 1,\ 0 \le k_2,\widehat{k}_2 \le 2.$$

The values of ind and $\widehat{\mathrm{ind}}$ are given in the following table:

(k_1,k_2)	$(0,0)$	$(0,1)$	$(0,2)$	$(1,0)$	$(1,1)$	$(1,2)$
$\mathrm{ind}(k_1,k_2)$	0	1	2	3	4	5
$\widehat{\mathrm{ind}}(k_1,k_2)$	0	2	4	1	3	5

The value table of $a_0 := a \circ \mathrm{ind}$ is

(k_1,k_2)	$(0,0)$	$(0,1)$	$(0,2)$	$(1,0)$	$(1,1)$	$(1,2)$
$a_0(k_1,k_2)$	$a(0)$	$a(1)$	$a(2)$	$a(3)$	$a(4)$	$a(5)$

For the computation of a_1 we need the exponent ε_1, where

$$\varepsilon_1(k_1;\widehat{k}_1) = k_1 * \widehat{k}_1 * e_2 = 3k_1\widehat{k}_1,\ \varepsilon_1(0;\widehat{k}_1) = 0,\ \varepsilon_1(1;\widehat{k}_1) = 3\widehat{k}_1,$$
$$a_1(\widehat{k}_1,k_2) = a_0(0,k_2) + a_0(1,k_2)\zeta^{3\widehat{k}_1}.$$

In detail we get

$$\begin{array}{lllllllll}
a_1(0,0) & = & a_0(0,0) & + & a_0(1,0) & = & a(0) & + & a(3),\\
a_1(0,1) & = & a_0(0,1) & + & a_0(1,1) & = & a(1) & + & a(4),\\
a_1(0,2) & = & a_0(0,2) & + & a_0(1,2) & = & a(2) & + & a(5),\\
a_1(1,0) & = & a_0(0,0) & + & a_0(1,0)\zeta^3 & = & a(0) & - & a(3),\\
a_1(1,1) & = & a_0(0,1) & - & a_0(1,1)\zeta^3 & = & a(1) & - & a(4),\\
a_1(1,2) & = & a(0,2) & - & a_0(1,2)\zeta^3 & = & a(2) & - & a(5).
\end{array}$$

For the computation of a_2 we need ε_2, where

$$\varepsilon_2(k_2;\widehat{k}_1;\widehat{k}_2) = k_2(\widehat{k}_1 + 2\widehat{k}_2),$$
$$a_2(\widehat{k}_1,\widehat{k}_2) = a_1(\widehat{k}_1,0) + a_1(\widehat{k}_1,1)\zeta^{\widehat{k}_1+2\widehat{k}_2} + a_1(\widehat{k}_1,2)\zeta^{2(\widehat{k}_1+2\widehat{k}_2)}.$$

In detail, we obtain

$$\begin{aligned}
\widehat{a}(0) = a_2(0,0) &= a_1(0,0) + a_1(0,1) + a_1(0,2)\\
&= a(0) + a(3) + a(1) + a(4) + a(2) + a(5) = \textstyle\sum_{i=0}^{5} a(i)\zeta^{i*0},\\
\widehat{a}(2) = a_2(0,1) &= a_1(0,0) + a_1(0,1)\zeta^2 + a_1(0,2)\zeta^4\\
&= a(0) + a(3) + (a(1) + a(4))\zeta^2 + (a(2) + a(5))(-\zeta)\\
&= a(0) + a(1)\zeta^2 + a(2)(-\zeta) + a(3) + a(4)\zeta^2 + a(5)(-\zeta)\\
&= \textstyle\sum_{i=0}^{5} a(i)\zeta^{i*2},\\
\widehat{a}(4) = a_2(0,2) &= a_1(0,0) + a_1(0,1)\zeta^4 + a_1(0,2)\zeta^8\\
&= (a(0) + a(3)) + (a(1) + a(4))(-\zeta) + (a(2) + a(5))\zeta^2\\
&= a(0) + a(1)(-\zeta) + a(2)\zeta^2 + a(3) + a(4)(-\zeta) + a(5)\zeta^2\\
&= \textstyle\sum_{i=0}^{5} a(i)\zeta^{i*4},\\
\widehat{a}(1) = a_2(1,0) &= a_1(1,0) + a_1(1,1)\zeta^1 + a_1(1,2)\zeta^2\\
&= (a(0) - a(3)) + (a(1) - a(4))\zeta + (a(2) - a(5))\zeta^2\\
&= a(0) + a(1)\zeta + a(2)\zeta^2 + a(3)(-1) + a(4)(-\zeta) + a(5)(-\zeta^2)\\
&= \textstyle\sum_{i=0}^{5} a(i)\zeta^{i*1},\\
\widehat{a}(3) = a_2(1,1) &= a_1(1,0) + a_1(1,1)\zeta^3 + a_1(1,2)\zeta^6\\
&= (a(0) - a(3)) + (a(1) - a(4))(-1) + (a(2) - a(5))\\
&= a(0) + a(1)(-1) + a(2) + a(3)(-1) + a(4) + a(5)(-1)\\
&= \textstyle\sum_{i=0}^{5} a(i)\zeta^{i*3},\\
\widehat{a}(5) = a_2(1,2) &= a_1(1,0) + a_1(1,1)\zeta^5 + a_1(1,2)\zeta^{10}\\
&= (a(0) - a(3)) + (a(1) - a(4))(-\zeta^2) + (a(2) - a(5))(-\zeta)\\
&= a(0) + a(1)(-\zeta^2) + a(2)(-\zeta) + a(3)(-1) + a(4)\zeta^2 + a(5)\zeta\\
&= \textstyle\sum_{i=0}^{5} a(i)\zeta^{i*5}.
\end{aligned}$$

In the following corollary we assume that d in Theorem 66 is a power of a number q; i.e.,

$$d = q^r,\ q > 1,\ r > 1,\ e_1 = \cdots = e_r = q,\ d_1(i) = q^i,\ d_2(i) = q^{r-i}. \tag{57}$$

The associated index functions according to (53) and (54) are

$$\begin{gathered}\operatorname{ind}(k_1,\cdots,k_r)=\textstyle\sum_{i=1}^r k_i q^{r-i}=\sum_{j=1}^r k_{r+1-j}q^{j-1},\ 0\le k_i<q,\\ \widehat{\operatorname{ind}}(\widehat{k}_1,\cdots,\widehat{k}_r)=\textstyle\sum_{j=1}^r \widehat{k}_j q^{j-1},\ 0\le \widehat{k}_j<q,\end{gathered}\tag{58}$$

and they give the q-adic representation of a natural number. The map

$$\begin{gathered}\widehat{\operatorname{ind}}^{-1}\circ\operatorname{ind}=\operatorname{ind}^{-1}\circ\widehat{\operatorname{ind}}:\{0,\cdots,q-1\}^r\to\{0,\cdots,q-1\}^r,\\ (k_1,\cdots,k_r)\mapsto(k_r,\cdots,k_1),\end{gathered}\tag{59}$$

is usually called the *bit reversal map* for an obvious reason. The functions ε_i and $\widehat{\varepsilon}_j$ from (55) and (56) are

(60)
$$\varepsilon_i(k_i;\widehat{k}_1,\cdots,\widehat{k}_i)=\textstyle\sum_{j=1}^i k_i\widehat{k}_j q^{j-1+r-i},\quad \widehat{\varepsilon}_j(k_j,\cdots,k_r;\widehat{k}_j)=\sum_{i=j}^r k_i\widehat{k}_j q^{j-1+r-i}.$$

Corollary 68. *Consider natural numbers $q>1$, $r>1$, and $d:=q^r$, the cyclic group $G:=\mathbb{Z}/\mathbb{Z}q^r$, and the DFT*

$$\operatorname{Four}_{\mathbb{Z}/\mathbb{Z}q^r}K^G=K^{q^r}\to K^{q^r},\ a\mapsto\widehat{a},\ \widehat{a}(l):=\textstyle\sum_{k=0}^{q^r-1}a(k)\zeta^{kl},\ 0\le l<q^r.$$

1. *The following "decimation in time" algorithm computes $\widehat{a}$ from a with complexity $q^r(q-1)r$. Inductively define functions*

$$\begin{gathered}a_\varrho:\{0,\cdots,q-1\}^r\to K\ \text{for}\ \varrho=0,\dots,r\ \text{by}\\ a_0(k_1,\cdots,k_r):=a\left(\textstyle\sum_{i=1}^r k_i q^{r-i}\right),\\ a_\varrho(\widehat{k}_1,\cdots,\widehat{k}_\varrho,k_{\varrho+1},\cdots,k_r):=\textstyle\sum_{k_\varrho=0}^{q-1}a_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},k_\varrho,\cdots,k_r)\zeta^{\varepsilon_\varrho(k_\varrho;\widehat{k}_1,\cdots,\widehat{k}_\varrho)}\end{gathered}$$

with ε_ϱ from (60). *Then*

$$\widehat{a}(l)=a_r(\widehat{k}_1,\cdots,\widehat{k}_r)\ \text{for}\ l=\textstyle\sum_{j=1}^r\widehat{k}_j q^{j-1},\ 0\le l<q^r,\ 0\le\widehat{k}_j<q.$$

2. *The following "decimation in frequency" algorithm also computes $\widehat{a}$ with complexity $q^r(q-1)r$. Recursively define functions*

$$\begin{gathered}b_\varrho:\{0,\cdots,q-1\}^r\to K\ \text{for}\ \varrho=r,\dots,0\ \text{by}\\ b_r(\widehat{k}_1,\cdots,\widehat{k}_r):=a\left(\textstyle\sum_{j=1}^r\widehat{k}_j q^{j-1}\right),\\ b_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},k_\varrho,\cdots,k_r)=\textstyle\sum_{\widehat{k}_\varrho=0}^{q-1}b_\varrho(\widehat{k}_1,\cdots,\widehat{k}_\varrho,k_{\varrho+1},\cdots,k_r)\zeta^{\widehat{\varepsilon}_\varrho(k_\varrho,\cdots,k_r;\widehat{k}_\varrho)}\end{gathered}$$

with $\widehat{\varepsilon}_\varrho$ from (60). *Then*

$$\widehat{a}(k)=b_0(k_1,\cdots,k_r)\ \text{for}\ k=\textstyle\sum_{i=1}^r k_i q^{r-i},\ 0\le k<q^r,\ 0\le k_i<q.$$

Corollary 69 (see [16]). *In the situation of Corollary* 68 *assume that $q=2$ and $G=\mathbb{Z}/\mathbb{Z}2^r$. The FFT-algorithms reduce to the following algorithms. The functions $a,\widehat{a}$ and a_ϱ,b_ϱ belong to K^{2^r} (resp., $K^{\{0,1\}^r}$).*

1. *The following "decimation in time" algorithm computes $\widehat{a}$ from a with complexity $r*2^r$. Inductively define functions*

$$\begin{gathered}a_\varrho:\{0,1\}^r\to K\ \text{for}\ \varrho=0,\dots,r\ \text{by}\\ a_0(k_1,\cdots,k_r):=a\left(\textstyle\sum_{i=1}^r k_i 2^{r-i}\right),\\ a_\varrho(\widehat{k}_1,\cdots,\widehat{k}_\varrho,k_{\varrho+1},\cdots,k_r)\\ :=a_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},0,k_{\varrho+1},\cdots,k_r)+a_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},1,k_{\varrho+1},\cdots,k_r)\zeta^{\varepsilon_\varrho(1;\widehat{k}_1,\cdots,\widehat{k}_\varrho)}\end{gathered}$$

with $\varepsilon_\varrho(1;\widehat{k}_1,\cdots,\widehat{k}_\varrho) := \sum_{j=1}^{\varrho}\widehat{k}_j 2^{j-1+r-\varrho}$. *Then*

$$\widehat{a}(l) = a_r(\widehat{k}_1,\cdots,\widehat{k}_r) \text{ for } l = \textstyle\sum_{j=1}^{r}\widehat{k}_j 2^{j-1},\ 0 \le l < 2^r, 0 \le \widehat{k}_j \le 1.$$

2. *The following "decimation in frequency" algorithm also computes* $\widehat{a}$ *with complexity* $r * 2^r$. *Recursively define functions*

$$b_\varrho : \{0,1\}^r \to K \text{ for } \varrho = r,\dots,0 \text{ by}$$
$$b_r(\widehat{k}_1,\cdots,\widehat{k}_r) := a\left(\textstyle\sum_{j=1}^{r}\widehat{k}_j 2^{j-1}\right),$$
$$b_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},k_\varrho,\cdots,k_r)$$
$$= b_\varrho(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},0,k_{\varrho+1},\cdots,k_r) + b_\varrho(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},1,k_{\varrho+1},\cdots,k_r)\zeta^{\widehat{\varepsilon}_\varrho(k_\varrho,\cdots,k_r;1)}$$

with $\widehat{\varepsilon}_\varrho(k_\varrho,\cdots,k_r;1) := \sum_{i=\varrho}^{r} k_i 2^{\varrho-1+r-i}$. *Then*

$$\widehat{a}(k) = b_0(k_1,\cdots,k_r) \text{ for } k = \sum_{i=1}^{r} k_i 2^{r-i},\ 0 \le k < 2^r,\ 0 \le k_i \le 1.$$

Observe that the computation of $a_\varrho(\widehat{k}_1,\cdots,\widehat{k}_\varrho,k_{\varrho+1},\cdots,k_r)$ *(resp., of* $b_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},k_\varrho,\cdots,k_r)$*) from* $a_{\varrho-1}$ *(resp.,* b_ϱ*) requires just one elementary computation step* $\alpha+\lambda\beta$.

For the next application of Theorem 58 we assume that a direct decomposition of the group G, i.e., an isomorphism

$$\varphi : \textstyle\prod_{i=1}^{r} K_i \cong G, \tag{61}$$

is given. For every subset I of $\{1,\cdots,r\}$ we define

$$G(I) := \textstyle\prod_{i\in I} K_i, \text{ especially } G_i := G(\{1,\cdots,i\}), H_i := G(\{i+1,\cdots,r\}). \tag{62}$$

For $J \subseteq I$ there results the exact sequence

$$0 \to G(J) \xrightarrow{\text{inj}} G(I) \xrightarrow{\text{proj}} G(I\setminus J) \to 0,$$
$$\text{inj}((l_j)_{j\in J}) := (k_i)_{i\in I}, \text{ where } k_i := \begin{cases} l_i & \text{if } i \in J,\\ 0 & \text{if } i \in I\setminus J,\end{cases} \tag{63}$$
$$\text{proj}((k_i)_{i\in I}) := (k_i)_{i\in I\setminus J},$$

where, moreover, inj : $G(I\setminus J) \to G(I)$ is a *homomorphic* section of the canonical projection proj. The groups $\widehat{K_i}$ and the forms $\langle -,-\rangle_{K_i}$ being given arbitrarily, we now choose

$$\widehat{G(I)} := \textstyle\prod_{i\in I}\widehat{K_i}, \quad \langle (k_i)_{i\in I}, (\widehat{k}_i)_{i\in I}\rangle := \prod_{i\in I}\langle k_i,\widehat{k}_i\rangle_{K_i}. \tag{64}$$

It is then easily seen that

$$(\text{inj} : G(J) \to G(I))^\star = \text{proj} : \widehat{G(I)} \to \widehat{G(J)},$$
$$(\text{proj} : G(I) \to G(J))^\star = \text{inj} : \widehat{G(J)} \to \widehat{G(I)}. \tag{65}$$

The isomorphism φ from (61) and the exact sequences (63) and (65) now imply the exact sequences

$$
\begin{aligned}
0 \to G_i &\xrightarrow{\varphi\circ\mathrm{inj}} G \xrightarrow{\mathrm{proj}\circ\varphi^{-1}} H_i \to 0,\\
0 \to \widehat{H_i} &\xrightarrow{(\varphi^\star)^{-1}\circ\mathrm{inj}} \widehat{G} \xrightarrow{\mathrm{proj}\circ\varphi^\star} \widehat{G_i} \to 0.
\end{aligned} \tag{66}
$$

Finally we use these data to construct the diagrams (32) and (34) in the form

$$
\begin{array}{ccccccccc}
 & & & & & & 0 & & \\
 & & & & & & \downarrow & & \\
 & & 0 & & & & K_i & & \\
 & & \downarrow & & & & \downarrow{\scriptstyle \gamma_i:=\mathrm{inj}} & & \\
0 & \longrightarrow & G_{i-1} & \xrightarrow{\alpha_{i-1}:=\varphi\circ\mathrm{inj}} & G & \xrightarrow{\lambda_{i-1}:=\mathrm{proj}\circ\varphi^{-1}} & H_{i-1} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \beta_i:=\mathrm{inj}} & & \| & & \downarrow{\scriptstyle \nu_i:=\mathrm{proj}} & & \\
0 & \longrightarrow & G_i & \xrightarrow{\alpha_i:=\varphi\circ\mathrm{inj}} & G & \xrightarrow{\lambda_i:=\mathrm{proj}\circ\varphi^{-1}} & H_i & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \mu_i:=\mathrm{proj}} & & & & \downarrow & & \\
 & & K_i & & & & 0 & & \\
 & & \downarrow & & & & & & \\
 & & 0 & & & & & &
\end{array} \tag{67}
$$

$$
\begin{array}{ccccccccc}
 & & & & & & 0 & & \\
 & & & & & & \downarrow & & \\
 & & 0 & & & & \widehat{K_j} & & \\
 & & \downarrow & & & & \downarrow{\scriptstyle \mu_j^\star:=\mathrm{inj}} & & \\
0 & \longrightarrow & \widehat{H_j} & \xrightarrow{\lambda_j^\star:=(\varphi^\star)^{-1}\circ\mathrm{inj}} & \widehat{G} & \xrightarrow{\alpha_j^\star:=\mathrm{proj}\circ\varphi^\star} & \widehat{G_j} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \nu_j^\star:=\mathrm{inj}} & & \| & & \downarrow{\scriptstyle \beta_j^\star:=\mathrm{proj}} & & \\
0 & \longrightarrow & \widehat{H_{j-1}} & \xrightarrow{\lambda_{j-1}^\star:=(\varphi^\star)^{-1}\circ\mathrm{inj}} & \widehat{G} & \xrightarrow{\alpha_{j-1}^\star:=\mathrm{proj}\circ\varphi^\star} & \widehat{G_{j-1}} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \gamma_j^\star:=\mathrm{proj}} & & & & \downarrow & & \\
 & & \widehat{K_j} & & & & 0 & & \\
 & & \downarrow & & & & & & \\
 & & 0 & & & & & &
\end{array} \tag{68}
$$

with the canonical homomorphic sections

$$
\sigma_i := \mathrm{inj} : K_i \to G_i = \textstyle\prod_{k=1}^{i} K_k \text{ and } \widehat{\sigma}_j := \mathrm{inj} : \widehat{K_j} \to \widehat{H_{j-1}} = \textstyle\prod_{k=j}^{r} \widehat{K_k}. \tag{69}
$$

These diagrams induce the index transformations ind and $\widehat{\text{ind}}$ from Corollaries 55 and 56; indeed

$$\begin{aligned}\text{ind}((k_i)_{i=1,\cdots,r}) &= \textstyle\sum_{i=1}^r \alpha_i\sigma_i(k_i) = \varphi\left(\sum_{i=1}^r \text{inj}\circ\text{inj}(k_i)\right)\\ &= \varphi\left(\textstyle\sum_{i=1}^r (0,\cdots,0,k_i,0,\cdots,0)\right) = \varphi((k_i)_{i=1,\cdots,r}),\end{aligned}$$

and hence

$$\text{ind} = \varphi : \textstyle\prod_{i=1}^r K_i \cong G \text{ and likewise } \widehat{\text{ind}} = (\varphi^\star)^{-1} : \prod_{j=1}^r \widehat{K_j} \cong \widehat{G}. \tag{70}$$

Also, with the notation from (35), we have

$$\begin{aligned}\text{fact}_{ij}(k,\widehat{k}) &= \langle \alpha_i\sigma_i(k_i), \lambda_{j-1}^\star \widehat{\sigma}_j(\widehat{k}_j)\rangle\\ &= \langle \varphi\,\text{inj}(k_i), (\varphi^\star)^{-1}\,\text{inj}(\widehat{k}_j)\rangle = \langle \varphi^{-1}\varphi\,\text{inj}(k_i), \text{inj}(\widehat{k}_j)\rangle\\ = \langle (0,\cdots,0,k_i,0,\cdots,0), (0,\cdots,0,\widehat{k}_j,0,\cdots,0)\rangle &= \begin{cases}\langle k_i,\widehat{k}_i\rangle & \text{if } i=j,\\ 1 & \text{if } i\neq j,\end{cases} \text{ and hence}\\ \varphi_\varrho(k_\varrho;\widehat{k}_1,\cdots,\widehat{k}_\varrho) &= \langle k_\varrho,\widehat{k}_\varrho\rangle,\ \varrho=1,\cdots,r.\end{aligned}$$

Theorem 58 now implies the following theorem.

THEOREM 70. *Assume that a group isomorphism $\varphi : \prod_{i=1}^r K_i \cong G$ is given. Then the following recursive algorithm computes the Fourier transform $\widehat{a} \in K^{\widehat{G}}$ of a function $a \in K^G$ with complexity $N(e_1+\cdots+e_r-r)$, where $N := \text{ord}(G)$ and $e_i := \text{ord}(K_i)$. Inductively define functions*

$$\begin{aligned}&a_\varrho : \widehat{K_1}\times\cdots\times\widehat{K_\varrho}\times K_{\varrho+1}\times\cdots\times K_r \to K \ \text{ for } \varrho=0,\dots,r \text{ by } a_0 := a\circ\varphi \text{ and}\\ &a_\varrho(\widehat{k}_1,\cdots,\widehat{k}_\varrho,k_{\varrho+1},\cdots,k_r) := \textstyle\sum_{k_\varrho\in K_\varrho} a_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},k_\varrho,\cdots,k_r)\langle k_\varrho,\widehat{k}_\varrho\rangle \text{ or}\\ &a_\varrho(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},-,k_{\varrho+1},\cdots,k_r) := \text{Four}_{K_\varrho}\left(a_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},-,k_{\varrho+1},\cdots,k_r)\right).\\ &\text{Then } \widehat{a} = a_r\circ\varphi^\star.\end{aligned}$$

If, in particular, $G := \prod_{i=1}^r K_i$ and $\varphi = \text{id}$, then $a_0 = a$ and $\widehat{a} = a_r$.

Example 71 (Walsh–Fourier FFT). We apply the preceding theorem to $d := 2$, the group

$$G := \widehat{G} := (\mathbb{Z}/\mathbb{Z}2)^r = \{0,1\}^r \ni k = (k_1,\cdots,k_r),\ 0\le k_i\le 1,$$

of exponent 2 with the form $k\bullet l := \sum_{i=1}^r k_i l_i \in \mathbb{Z}/\mathbb{Z}2$, and a ring K in which 2 is invertible so that Assumption 29 is satisfied for $\zeta := -1$. The Walsh–Fourier DFT is given by

$$\text{Four}_G : K^G \cong K^G,\ \text{Four}_G(a)(\widehat{k}) := \widehat{a}(\widehat{k}_1,\cdots,\widehat{k}_r) = \textstyle\sum_{k\in G} a(k)(-1)^{k\bullet\widehat{k}}$$

and inductively computed with complexity $r*2^r$ by means of the algorithm

$$\begin{aligned}&a_0 := a \text{ and for } 1\le\varrho\le r\\ &a_\varrho(\widehat{k}_1,\cdots,\widehat{k}_\varrho,k_{\varrho+1},\cdots,k_r)\\ &:= a_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},0,k_{\varrho+1},\cdots,k_r) + a_{\varrho-1}(\widehat{k}_1,\cdots,\widehat{k}_{\varrho-1},1,k_{\varrho+1},\cdots,k_r)(-1)^{\widehat{k}_\varrho},\\ &\widehat{a} = a_r.\end{aligned}$$

The next example contains the prime factor algorithm according to Good.

Example 72 (the Good FFT or the prime factor algorithm [19]). In the situation of Theorem 66 assume that the numbers e_i are relatively prime. The euclidean algorithm and the Chinese remainder theorem yield representations

$$1 = r_i * e_i + s_i * d/e_i = a_i + b_i, \quad b_i := s_i * d/e_i,$$

and the group isomorphism

$$\Delta : G := \mathbb{Z}/\mathbb{Z}d \cong \prod_{i=1}^r K_i := \prod_{i=1}^r \mathbb{Z}/\mathbb{Z}e_i,\ \bar{l} \mapsto (\bar{l}, \cdots, \bar{l}).$$

The inverse map of Δ is

$$\varphi := \Delta^{-1} : \prod_{i=1}^r \mathbb{Z}/\mathbb{Z}e_i \cong \mathbb{Z}/\mathbb{Z}d,\ \varphi(\overline{k_1}, \cdots, \overline{k_r}) = \overline{\textstyle\sum_{i=1}^r k_i b_i}.$$

For the application of Theorem 70 we compute $\varphi^\star$. The equations

$$\langle \varphi(\overline{k_1}, \cdots, \overline{k_r}), \bar{l}\rangle = \zeta^{\sum_{i=1}^r k_i b_i l}$$
$$= \zeta^{\sum_{i=1}^r k_i (s_i l) d/e_i} = \prod_{i=1}^r \langle \overline{k_i}, \overline{s_i l}\rangle_{K_i} = \langle \overline{k}, (\overline{s_1 l}, \cdots, \overline{s_r l})\rangle$$

imply

$$\varphi^\star(\bar{l}) = (\overline{s_1 l}, \cdots, \overline{s_r l}) = (\overline{s_1}, \cdots, \overline{s_r})\Delta(\bar{l}) \in \prod_{i=1}^r \mathbb{Z}/\mathbb{Z}e_i.$$

Application of Theorem 70 to the preceding data now shows that the following algorithm computes $\widehat{a} \in K^G$ from $a \in K^G$ with complexity $d(e_1 + \cdots + e_r - r)$. Inductively define functions

$$a_\varrho : \prod_{i=1}^r \{0, \cdots, e_i - 1\} \to K \text{ for } \varrho = 0, \ldots, r \text{ by}$$
$$a_0(k_1, \cdots, k_r) := a\left(\overline{\textstyle\sum_{i=1}^r k_i b_i}\right),$$
$$a_\varrho(\widehat{k}_1, \cdots, \widehat{k}_\varrho, k_{\varrho+1}, \cdots, k_r) := \textstyle\sum_{k_\varrho=0}^{e_\varrho - 1} a_{\varrho-1}(\widehat{k}_1, \cdots, \widehat{k}_{\varrho-1}, k_\varrho, \cdots, k_r)\zeta^{k_\varrho \widehat{k}_\varrho d/e_\varrho}.$$ Then
$$\overline{a}(\bar{l}) = a_r(\overline{s_1 l}, \cdots, \overline{s_r l}), \quad \bar{l} \in \mathbb{Z}/\mathbb{Z}d.$$

Consider, in particular, the case of Example 67, i.e., $d = 6 = e_1 e_2 = 2 * 3$. Then

$$1 = (-1) * 2 + 1 * 6/2 = 2 * 6/3 + (-1) * 3;\ \text{hence } s_1 = 1,\ b_1 = 3,\ s_2 = 2,\ b_2 = 4.$$

The maps

$$\varphi,\ (\varphi^\star)^{-1} : \mathbb{Z}/\mathbb{Z}2 \times \mathbb{Z}/\mathbb{Z}3 = \{0,1\} \times \{0,1,2\} \to \mathbb{Z}/\mathbb{Z}6 = \{0,1,2,3,4,5\}$$

have the value table

(k_1, k_2)	$(0,0)$	$(0,1)$	$(0,2)$	$(1,0)$	$(1,1)$	$(1,2)$
$\varphi(k_1, k_2)$	0	4	2	3	1	5
$(\varphi^\star)^{-1}(k_1, k_2)$	0	2	4	3	5	1

Since the maps φ and $(\varphi^\star)^{-1}$ differ from the index maps ind and $\widehat{\text{ind}}$ from Example 67, the FFT-algorithms from Theorem 66 and Example 72 applied to the same case $d = e_1 * \cdots * e_r$ with relatively prime e_i differ, too.

7. Fast convolution. The assumptions of section 3 are in force; in particular, the Fourier transform is invertible.

The FFT also induces a fast convolution algorithm for the group algebra $K[G]$ and the polynomial algebra $K[z_1, \cdots, z_r]$. Let, more generally, A be a commmutative K-algebra with a fixed chosen basis of length N, for instance, $K[G]$ with the standard basis. The multiplication $A \times A \to A$ is K-bilinear, but not linear, and therefore requires the notion of a *bilinear or multiplicative* complexity. Several papers and books treat this type of complexity and construct fast algorithms of small multiplicative complexity [40], [3], [1], [34], [9, Def. 14.7], [30], [18]. In the present paper we do not treat these algorithms and use only the complexity of linear maps as introduced in section 4. For fixed $a \in A$ the map $A \to A,\ b \mapsto ab$, is K-linear and therefore its linear complexity (with respect to the chosen basis),

$$\mu_A(a) := \mu(A \to A,\ b \mapsto ab) \le N^2 \text{ and then } \mu_{\lin}(A) := \max_{b \in A} \mu_A(b) \le N^2, \tag{71}$$

is defined according to Definition 46. It is obvious that K^G with the argumentwise multiplication and the standard basis has the complexity

$$\mu_{\lin}(K^G) \le N := \ord(G) \tag{72}$$

since the corresponding matrices are diagonal matrices with at most N nonzero entries.

THEOREM 73 (fast convolution). *The data are as in Theorem* 63. *Let $a \in K[G]$ be an arbitrarily chosen but fixed function and consider the linear map $f : K[G] \to K[G],\ b \mapsto a * b$. Then f is the composition of the maps*

$$f : K[G] \xrightarrow{\Four_G} K^{\widehat{G}} \xrightarrow{\widehat{a}\cdot(-)} K^{\widehat{G}} \xrightarrow{N^{-1}\Four_{\widehat{G}}} K^G \xrightarrow{S_G} K^G,$$

and hence its complexity satisfies

$$\mu_{K[G]}(a) := \mu(f) \le N(1 + 2\Lambda(N)) \text{ and thus also } \mu_{\lin}(K[G]) \le N(1 + 2\Lambda(N)).$$

Proof. Let $c := a * b$; hence $\widehat{c} := \widehat{a}\widehat{b}$ by the convolution theorem. The Fourier inversion theorem implies

$$f(b) = c = S_G(N^{-1}\Four_{\widehat{G}})(\widehat{c}) = S_G(N^{-1}\Four_{\widehat{G}})(\widehat{a}\widehat{b}),$$

and f is indeed the asserted composition. According to Theorem 49 its complexity is at most the sum of the complexities of its factors. The two Fourier transforms have complexity at most $N\Lambda(N)$ according to Theorem 63 and the argumentwise multiplication with $\widehat{a}$ at most N. The complexity of the antipode is zero since it is an index transformation; see Definition and Corollary 48. The algorithm for $\Four_{\widehat{G}}$ from Theorem 61 can be adapted to the computation of $N^{-1}\Four_{\widehat{G}}$ by replacing

$$\widehat{\varphi}_r(k_r, \widehat{k}_r) = \fact_{rr}(k_r, \widehat{k}_r) = \langle k_r, \widehat{k}_r \rangle$$

in the recursion step $c_r \mapsto c_{r-1}$ by $N^{-1}\langle k_r, \widehat{k}_r \rangle$. This implies that also $N^{-1}\Four_{\widehat{G}}$ has complexity at most $N\Lambda(N)$, and therefore the complexity of $b \mapsto a * b$ and of $K[G]$ is indeed at most $N(1 + 2\Lambda(N))$. □

ALGORITHM 74 (fast convolution). *The fast algorithm for the convolution $a * b$ in the group algebra $K[G]$ consists of the following steps:*

1. *Precompute the Fourier transform $\widehat{a} \in K^{\widehat{G}}$. This computation and its complexity are not counted because $\widehat{a}$ is assumed known when f is applied.*

2. *Compute* $\widehat{b}$ *with the decimation in time FFT according to Theorems* 58 *and* 63 *with complexity* $N\Lambda(N)$.
3. *Compute* $\widehat{c} := \widehat{a}\widehat{b},\ (\widehat{a}\widehat{b})(\widehat{g}) = \widehat{a}(\widehat{g})\widehat{b}(\widehat{g})$ *with complexity at most* N.
4. *Compute* $N^{-1}\widehat{\widehat{c}}$ *with the slight modification of the decimation in frequency FFT from Theorem* 61 *with complexity* $N\Lambda(N)$ *and then apply the antipode to the result to obtain* $c = a * b$.

It suffices to compute $b_r(\widehat{k})$ *only in the first FFT-algorithm (see Theorem* 58*) and to start the second FFT-algorithm with* $c_r(\widehat{k}) = a_r(\widehat{k})b_r(\widehat{k})$*; i.e., the computation of the elements* $\widehat{g} = \widehat{\mathrm{ind}}(\widehat{k}) = \sum_{j=1}^{r} \lambda_{j-1}^{\star}\widehat{\sigma}_j(\widehat{k}_j)$ *is superfluous.*

Remark 75. If in the preceding algorithm for $a * b$ the complexity of computing $\widehat{a}$ is also counted, then the total complexity of the algorithm is $N(1 + 3\Lambda(N))$. Recall, however, that in this article we gave only a formal definition for the complexity of a linear, but not of a bilinear, map such as $a * b$ with variable a and b. Our complexity counts *all* necessary elementary computation steps for the computation of $c = a*b$ and not only the *essential* multiplications which enter into the *multiplicative complexity.*

The fast convolution also induces a fast algorithm for the multiplication of multivariate polynomials in $K[z] = K[z_1, \cdots, z_r]$. For this purpose we consider the case

$$\begin{gathered} G = \widehat{G} = \mathbb{Z}/\mathbb{Z}d_1 \times \cdots \times \mathbb{Z}/\mathbb{Z}d_r \\ \underset{\text{ident.}}{=} I(d) := \{0, \cdots, d_1 - 1\} \times \cdots \times \{0, \cdots, d_r - 1\} \ni \mu = (\mu_1, \cdots, \mu_r), \\ \mu \bullet \nu := \overline{\textstyle\sum_{i=1}^{r} \mu_i \nu_i \frac{d}{d_i}} \in \mathbb{Z}/\mathbb{Z}d,\ \langle \mu, \nu \rangle = \zeta^{\mu \bullet \nu}. \end{gathered} \tag{73}$$

The group algebra $K[G]$ has the K-basis $\delta_\mu,\ \mu \in G$. With

$$x := (x_1, \cdots, x_r),\ x_i := \delta_{(0,\cdots,0,1,0,\cdots,0)},\ 1 \text{ at the } i\text{th place},\ i = 1, \ldots, r, \text{ we get}$$
$$\delta_\mu = x^\mu \text{ and } x_i^{d_i} - 1 = 0.$$

Lemma and Definition 76. *The substitution homomorphism* $K[z] \to K[G]$, $z_i \mapsto x_i$, *induces an isomorphism*

$$K[z]/\langle z_1^{d_1} - 1, \cdots, z_r^{d_r} - 1\rangle \cong K[G],\ \overline{z^\mu} \mapsto \delta_\mu = x^\mu,\ \overline{f} \mapsto f(x). \tag{74}$$

In what follows we therefore identify *these two algebras, i.e., for*

$$f = \textstyle\sum_{\mu \in \mathbb{N}^r} f_\mu z^\mu \in K[z] :\ \overline{f} = f(x) = \sum_{\mu \in \mathbb{N}^r} f_\mu x^\mu = \sum_{\mu \in \mathbb{N}^r} f_\mu \delta_\mu.$$

In particular, we get the K*-linear isomorphism*

$$\begin{aligned} K[z]_{I(d)} &:= \{f \in K[z];\ \textit{for all } i = 1, \ldots, r :\ \deg_{z_i}(f) \le d_i - 1\} \\ &= \oplus_{\mu \in I(d)} K z^\mu \cong K[G],\ z^\mu \mapsto \delta_\mu = x^\mu. \end{aligned}$$

In other words, one can reproduce f *from* $f(x)$ *if the degree bounds* $\deg_{z_i}(f) \le d_i - 1$ *are observed.*

Proof. Induction by means of the canonical isomorphism

$$\begin{aligned} &K[z]/\langle z_1^{d_1} - 1, \cdots, z_r^{d_r} - 1\rangle \\ &\cong (K[z_1, \cdots, z_{r-1}]/\langle z_1^{d_1} - 1, \cdots, z_{r-1}^{d_{r-1}} - 1\rangle)[z_r]/\langle z_r^{d_r} - 1\rangle \end{aligned}$$

shows that this algebra has the K-basis $\overline{z^\mu},\ \mu \in I(d)$. The induced homomorphism (74) maps this K-basis onto the basis $x^\mu,\ \mu \in I(d) \underset{\text{ident.}}{=} G$ of $K[G]$, and is thus an isomorphism. □

Now let m, n, d be vectors in $\mathbb{N}^r$ with the property

$$m_i + n_i \le d_i + 1,\ i = 1, \dots, r, \text{ such that } K[z]_{I(m)} \times K[z]_{I(n)} \xrightarrow{\text{mult}} K[z]_{I(d)} \tag{75}$$

is well defined.

COROLLARY 77 (fast multiplication of polynomials). *The multiplication*

$$\begin{gathered} K[z]_{I(m)} \times K[z]_{I(n)} \xrightarrow{\text{mult}} K[z]_{I(d)},\ (P, Q) \mapsto PQ, \\ P = \textstyle\sum_{\mu \in I(m)} a_\mu z^\mu,\ Q = \sum_{\nu \in I(n)} b_\nu z^\nu, \\ PQ = \textstyle\sum_{\lambda \in I(d)} \sum_{\mu,\nu,\ \mu+\nu=\lambda} \{a_\mu b_\nu,\ \mu_j \le m_j - 1,\ \nu_j \le n_j - 1\} z^\lambda \end{gathered} \tag{76}$$

equals the composition of the maps

$$K[z]_{I(m)} \times K[z]_{I(n)} \xrightarrow{\text{inj} \times \text{inj}} K[z]_{I(d)} \times K[z]_{I(d)} \cong K[G] \times K[G] \xrightarrow{*} K[G] \cong K[z]_{I(d)}. \tag{77}$$

If the product PQ is computed according to the algorithm in (76)*, the complexity is*

$$\textstyle\prod_{i=1}^r m_i n_i.$$

If, on the other hand, (77) *is used with the fast convolution algorithm, Algorithm* 74*, then the algorithm has the complexity*

$$N(1 + 3\Lambda(N)),\ \textit{where } N = \operatorname{ord}(G) := d_1 * \cdots * d_r.$$

Note that this algorithm depends on the choice of $d_1, \cdots, d_r$.

Proof. The proof is obvious since in (77) all maps except the convolution have complexity zero. See Remark 75 for the applied complexity notion. □

In applications of the preceding algorithm (77) the degrees m_j and n_j are given in general, whereas the numbers $d_j > m_j + n_j - 2$ may be suitably chosen. We illustrate the case

$$r = 1, \quad m_1 = n_1 = m, \quad 2 * (m - 1) < d = N.$$

Examples 78.

(1) The standard choice is

$$\begin{gathered} N = 2^e, e \ge 2,\ \Lambda(2^e) = e, m \le 2^{e-1};\ \text{hence} \\ N(1 + 3\Lambda(N)) = 2^e(1 + 3e).\ \text{But} \\ 2^{e-2} \le 1 + 3e \text{ for } 2 \le e \le 6;\ \text{hence} \\ m^2 \le 2^{2(e-1)} \le 2^e(1 + 3e) = N(1 + 3\Lambda(N)) \text{ for } N = 2^e, 2 \le e \le 6. \end{gathered}$$

This signifies that for the convolution of polynomials of degree at most 31 the direct computation of complexity $m^2 = 1024$ is faster than the algorithm of (77) with $N = 2^6$ and complexity $2^6 * (1 + 3 * 6) = 1216$.

(2)

$$\begin{gathered} m := 36, N_1 := 72 = 2^3 * 3^2 < N_2 = 128 = 2^7, \\ \Lambda(N_1) = 3 * 1 + 2 * 2 = 7 = \Lambda(N_2).\ \text{Again} \\ m^2 = 1296 < N_1(1 + 3\Lambda(N_1)) = 1584 < N_2(1 + 3\Lambda(N_2) = 2816. \end{gathered}$$

Also in this case the direct computation of the product is better than the two algorithms (77) for N_1 (resp., N_2).

(3)

$$m := 70, N_1 = 144 = 2^4 * 3^2, N_2 := 2^8,$$
$$\Lambda(N_1) = 4 + 2 * 2 = 8 = \Lambda(N_2). \text{ Then}$$
$$N_1(1 + 3\Lambda(N_1)) = 3600 < m^2 = 4900 < N_2(1 + 3\Lambda(N_2)) = 6400.$$

The algorithm for N_1 is faster than the direct computation, while that for the smallest power-of-two, 2^8, which exceeds $2 * 69$ is slower. This example shows that the standard choice of the power-of-two Cooley–Tukey FFT may not work at all or may give bad results for the fast multiplication of polynomials.

(4) This example is a multivariate one with

$$r > 1, \text{ but } m_1 = \cdots = m_r = n_1 = \cdots = n_r = 2.$$

The polynomials P and Q are of degree at most one in each indeterminate z_i or contain only square-free monomials. The direct computation of PQ has the total complexity $\prod_{j=1}^{r} m_j n_j = 4^r$. The optimal choice for the d_j is

$$d_1 = \cdots = d_r = 3; \text{ hence } N = 3^r, \ \Lambda(N) = 2r.$$

The algorithm (77) for these data has the complexity

$$N(1 + 3\Lambda(N)) = 3^r(1 + 6r) < 4^r \text{ for } r \geq 16.$$

The best applicable power-of-two FFT is that with $d_1 = \cdots = d_r = 4$ and the ensuing multiplication complexity $4^r(1 + 6r)$ which is much slower than the direct multiplication.

8. Number theoretic transforms (NTT). The following considerations give interesting examples of the DFT with coefficient rings instead of fields. They are simple variants or special cases of those in [26, Chap. 8], [15], [18, Chap. 7], where also the technical significance of these transforms is discussed. We adapt our notation to that of [26] and consider $N > 0$, a commutative ring K, and a primitive Nth root of one $\zeta \in K$. Consider the groups

$$G := \widehat{G} := \mathbb{Z}/\mathbb{Z}N \underset{\text{ident.}}{=} \{0, \cdots, N-1\} \text{ with } \overline{k} \bullet \overline{l} := \overline{kl} \in \mathbb{Z}/\mathbb{Z}N, \ \langle \overline{k}, \overline{l} \rangle := \zeta^{kl},$$
$$\mu := \langle \zeta \rangle = \{\eta_0, \cdots, \eta_i := \zeta^i, \cdots, \eta_{N-1} := \zeta^{N-1}\}.$$

The Fourier transform $\text{Four} := \text{Four}_G$ on G is given as (see Example 20)

$$a, \ \widehat{a} := \text{Four}_{\mathbb{Z}/\mathbb{Z}N}(a) \in K^N, \ \widehat{a}(l) = \sum_{k=0}^{N-1} \zeta^{lk} a(k), \text{ or}$$
$$\begin{pmatrix} \widehat{a}(0) \\ \widehat{a}(1) \\ \cdots \\ \widehat{a}(N-1) \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \eta_0 & \eta_1 & \cdots & \eta_{N-1} \\ \cdots & \cdots & \cdots & \cdots \\ \eta_0^{N-1} & \eta_1^{N-1} & \cdots & \eta_{N-1}^{N-1} \end{pmatrix} \begin{pmatrix} a(0) \\ a(1) \\ \cdots \\ a(N-1) \end{pmatrix}.$$

The determinant of this Vandermonde matrix is

$$\det := \prod_{0 \leq i < j \leq N-1} (\eta_j - \eta_i) = \prod_{0 \leq i < j \leq N-1} \zeta^i (\zeta^{j-i} - 1), \tag{78}$$

whose factors are the units ζ^i and the $\eta - 1, \ 1 \neq \eta \in \mu$.

Reminder 79 (see [23, pp. 203–207]). Let $z := \exp(\frac{2\pi i}{N})$ denote a complex primitive Nth root of one and $\nu := \langle z \rangle$ the cyclic group of all complex Nth roots of one. If $d \geq 1$ is a divisor of N, the set $\nu_d := \{z^{\frac{N}{d}k}; 1 \leq k \leq d-1, \gcd(k,d) = 1\}$ consists exactly of the $\varphi(d) := \mathrm{ord}(\mathrm{U}(\mathbb{Z}/\mathbb{Z}d))$ primitive dth roots of one, and the dth *cyclotomic polynomial* $\Phi_d := \prod_{x \in \nu_d}(X - x)$ is the (irreducible) minimal polynomial of all its roots in $\mathbb{Q}[X]$ and has coefficients in $\mathbb{Z}$, the latter property being derived from the obvious product representation

$$X^N - 1 = \prod_{x \in \nu}(X - x) = \prod_{d|N} \prod_{x \in \nu_d}(X - x) = \prod_{d|N} \Phi_d. \tag{79}$$

Since $\Phi_d \in \mathbb{Z}[X]$ the value $\Phi_d(x)$ is defined for every element x of any ring.

THEOREM 80 (see [26, Thms. 8.3, 8.4], [15, Satz 2.8]). *The following assertions are equivalent.*

(1) *Assumption* 29 *is satisfied, and hence the Fourier inversion theorem, Theorem* 34*, holds for all finite abelian groups of exponent* N*, i.e.,* (i) $N \in \mathrm{U}(K)$ *and* (ii)

$$\textit{for all } d > 1,\ d \mid N,\ \eta := \zeta^{\frac{N}{d}}: \quad 1 + \eta + \cdots + \eta^{d-1} = 0.$$

(2) *The Fourier transform* $\mathrm{Four}_{\mathbb{Z}/\mathbb{Z}N}$ *is an isomorphism.*

(3) *For all* $\eta \neq 1$ *in* $\mu = \langle \zeta \rangle$ *the element* $\eta - 1$ *is a unit in* K*.*

(4) (i) $N \in \mathrm{U}(K)$*.* (ii) $\Phi_N(\zeta) = 0$*.*

Proof. (1) $\Rightarrow$ (2): This is a special case.

(2) $\Leftrightarrow$ (3): The Fourier transform is an isomorphism if and only if its (Vandermonde) determinant (78) is a unit, and this is the case if and only if all factors $\eta - 1$, $1 \neq \eta \in \mu$, of this determinant are units, the ζ^i being units by assumption.

(3) $\Rightarrow$ (1): As just shown, $\mathrm{Four}_{\mathbb{Z}/\mathbb{Z}N}$ is an isomorphism. Let $d > 1$ be a divisor of N and let $\eta := \zeta^{\frac{N}{d}}$ be the root of order $\mathrm{ord}(\eta) = d$; hence

$$0 = \eta^d - 1 = (\eta - 1)(\eta^{d-1} + \cdots + 1).$$

But

$$\frac{N}{d} < N,\ \mathrm{ord}(\zeta) = N \ \Rightarrow\ \eta \neq 1 \underset{(3)}{\Rightarrow} \eta - 1 \in \mathrm{U}(K) \ \Rightarrow\ \eta^{d-1} + \cdots + 1 = 0,$$

and this is the second condition of Assumption 29. The proof of Theorem 34 then implies that $\mathrm{Four}^2_{\mathbb{Z}/\mathbb{Z}N} = NS_{\mathbb{Z}/\mathbb{Z}N}$. Since $\mathrm{Four}_{\mathbb{Z}/\mathbb{Z}N}$ and $S_{\mathbb{Z}/\mathbb{Z}N}$ are isomorphisms, N is invertible in K.

(1), (2), (3) $\Rightarrow$ (4): Equation (79) implies

$$0 = \zeta^N - 1 = \Phi_N(\zeta) \prod_{d|N,\ 1 \leq d < N} \Phi_d(\zeta).$$

But $\Phi_d \mid X^d - 1$ and condition (3) imply that

$$\text{for all } d \text{ with } d \mid N,\ 1 \leq d < N : \Phi_d(\zeta) \in \mathrm{U}(K);$$

hence $\Phi_N(\zeta) = 0$.

(4) $\Rightarrow$ (1): Let $1 < d$ be a divisor of N and let

$$Y := X^{\frac{N}{d}};\ \text{hence } X^N - 1 = Y^d - 1 = (Y - 1)(Y^{d-1} + \cdots + 1).$$

The polynomial Φ_N is irreducible in $\mathbb{Z}[X]$ and divides X^N-1, but not $X^{\frac{N}{d}}-1$ since $d>1$; hence Φ_N divides $Y^{d-1}+\cdots+1$. But then

$$\Phi_N(\zeta)=0,\ \eta:=Y(\zeta)=\zeta^{\frac{N}{d}},\ \text{and thus } \eta^{d-1}+\cdots+1=0.$$

This is exactly the second condition of Assumption 29. □

Reminder 81 (see [23, Exercise 7, p. 73]). Let

$$p=\text{odd prime},\ m\geq 1,\ K:=\mathbb{Z}/\mathbb{Z}p^m,\ \text{can}:K\to\mathbb{Z}/\mathbb{Z}p,\ \overline{k}\mapsto\overline{k}.\ \text{The group}$$
$$\mathrm{U}(K)=\{\eta=\overline{k}\in K;\ \gcd(p,k)=1 \text{ or } \text{can}(\eta)\neq 0 \text{ or } \text{can}(\eta)\in\mathrm{U}(\mathbb{Z}/\mathbb{Z}p)\}$$

is cyclic of order $\varphi(p^m)=p^{m-1}(p-1)$. More precisely, one obtains an exact sequence

$$1\to\langle 1+\overline{p}\rangle\xrightarrow{\subseteq}\mathrm{U}(K)\underset{\sigma}{\overset{\text{can}}{\rightleftarrows}}\mathrm{U}(\mathbb{Z}/\mathbb{Z}p)\to 1,$$

where $\langle 1+\overline{p}\rangle$ is cyclic of order p^{m-1} and where σ is the unique section of can which satisfies the condition $\sigma(\lambda^p)=\sigma(\lambda)^p$; indeed, σ is the well-defined map

$$\sigma:\mathrm{U}(\mathbb{Z}/\mathbb{Z}p)\to U(K),\ \overline{k}\mapsto\overline{k^{p^{m-1}}},$$

which is a monomorphism and induces the isomorphism

$$\mathrm{U}(\mathbb{Z}/\mathbb{Z}p)\times\langle 1+\overline{p}\rangle\cong\mathrm{U}(K),\ (\lambda,\eta)\mapsto\sigma(\lambda)\eta.$$

Since $\mathrm{U}(\mathbb{Z}/\mathbb{Z}p)$ is cyclic of order $p-1$ and $\gcd(p^{m-1},p-1)=1$, the Chinese remainder theorem implies that $\mathrm{U}(K)$ is cyclic, too, and is generated by $\sigma(\lambda)(1+\overline{p})$, where λ is a primitive $(p-1)$st root of one in $\mathbb{Z}/\mathbb{Z}p$. If $p=2$ and $m\geq 3$, the group $\mathrm{U}(\mathbb{Z}/\mathbb{Z}2^m)$ is not cyclic and is uninteresting for the DFT as will be shown instantly.

LEMMA 82. *Let p be prime, $m\geq 1$, $K:=\mathbb{Z}/\mathbb{Z}p^m$, and $\zeta\in K$ a primitive Nth root of one which satisfies the equivalent conditions of Theorem* 80. *Then N divides $p-1$. In particular, if $p=2$, then $N=1$ and $\zeta=1$, and therefore the case $p=2$ is uninteresting within the context of the DFT.*

Proof. Assume $p-1<N$. By Theorem 80, $\zeta^{p-1}-1$ is a unit in K and hence so is $\text{can}(\zeta^{p-1}-1)=\text{can}(\zeta)^{p-1}-1=0$ in $\mathbb{Z}/\mathbb{Z}p$, which is a contradiction. On the other hand, N divides the order $\varphi(p^m)=p^{m-1}(p-1)$ of $\mathrm{U}(K)$, and thus N is a divisor of $p-1$. □

THEOREM 83 (see [26, Thm. 8.6], [15, Satz 2.2]). *Let $M>2$ be an odd number, $M=p_1^{m_1}*\cdots*p_s^{m_s}$ its prime factor decomposition, $K=\mathbb{Z}/\mathbb{Z}M$, and $N>0$. Then K contains an Nth root of one satisfying the equivalent conditions of Theorem* 80 *if and only if N divides* $\gcd(p_1-1,\cdots,p_s-1)$.

Proof. The Chinese remainder theorem furnishes the isomorphism

$$\Delta:K=\mathbb{Z}/\mathbb{Z}M\cong K_1\times\cdots\times K_s:=\mathbb{Z}/\mathbb{Z}p_1^{m_1}\times\cdots\times\mathbb{Z}/\mathbb{Z}p_s^{m_s},\ \overline{k}\mapsto\Delta(\overline{k})=(\overline{k},\cdots,\overline{k}).$$

Assume $\zeta\in K$ satisfies the assumptions of Theorem 80 and let

$$\Delta(\zeta)=(\zeta_1,\cdots,\zeta_s);\ \text{hence } N=\text{ord}(\zeta)=\text{lcm}(N_1,\cdots,N_s),\ N_i:=\text{ord}(\zeta_i),\ \text{and}$$
$$\Delta(\zeta^m-1)=(\zeta_1^m-1,\cdots,\zeta_s^m-1).$$

The latter element is a unit if $m:=N_i<N$, but $\zeta_i^{N_i}-1=0$; hence $N=N_1=\cdots=N_s$ and $N\mid p_i-1,\ i=1,\ldots,s$, by Lemma 82.

If, conversely, for all i the number N divides p_{i-1}, λ_i is a generator of $\mathrm{U}(\mathbb{Z}/\mathbb{Z}p_i)$ and $\sigma_i : \mathrm{U}(\mathbb{Z}/\mathbb{Z}p_i) \to \mathrm{U}(\mathrm{K}_i)$ is the section according to Reminder 81, then

$$\zeta := \Delta^{-1}\left(\sigma_1\left(\lambda_1^{\frac{p_1-1}{N}}\right), \cdots, \sigma_s\left(\lambda_1^{\frac{p_s-1}{N}}\right)\right)$$

is the asserted root of one. □

We refer the reader to [26] and [18] for the discussion of special cases of the preceding theorem, in particular, those of *Mersenne and Fermat number transforms* with $M = 2^n - 1$ (resp., $M = 2^n + 1$).

Acknowledgment. I thank the two referees for their careful reading of the article, their positive opinion, and their suggestions. These have been incorporated into the revised version to the best of my abilities. One of the referees is obviously an aesthete.

REFERENCES

[1] L. Auslander, E. Feig, and S. Winograd, *The multiplicative complexity of the discrete Fourier transform*, Adv. in Appl. Math., 5 (1984), pp. 31–55.
[2] L. Auslander and R. Tolimieri, *Is computing with the finite Fourier transform pure or applied mathematics?*, Bull. Amer. Math. Soc. (N.S.), 1 (1979), pp. 847–897.
[3] L. Auslander and S. Winograd, *The multiplicative complexity of certain semilinear systems defined by polynomials*, Adv. in Appl. Math., 1 (1980), pp. 257–299.
[4] K. G. Beauchamp, *Transforms for Engineers. A Guide to Signal Processing*, Clarendon Press, Oxford, UK, 1987.
[5] T. Beth, *Verfahren der schnellen Fouriertransformation*, Teubner, Stuttgart, Germany, 1984.
[6] N. Bourbaki, *Éléments de mathématique. Fasc.* XXXII: *Théories spectrales*, Hermann, Paris, 1967, chap. I–II.
[7] W. L. Briggs and V. E. Henson, *The DFT: An Owners' Manual for the Discrete Fourier Transform*, SIAM, Philadelphia, 1995.
[8] E. O. Brigham, *The Fast Fourier Transform*, Prentice–Hall, Englewood Cliffs, NJ, 1974.
[9] P. Bürgisser, M. Clausen, and M. A. Shokrollahi, *Algebraic Complexity Theory*, Springer, Berlin, 1997.
[10] C. S. Burrus, *Efficient Fourier transform and convolution algorithms*, in Advanced Topics in Signal Processing, J. S. Lim and A. V. Oppenheim, eds., Prentice–Hall, Englewood Cliffs, NJ, 1988, pp. 199–245.
[11] C. S. Burrus, J. H. McClellan, A. V. Oppenheim, T. W. Parks, R. W. Schafer, and H. W. Schussler, *Computer-Based Exercises in Signal Processing: Using MATLAB* 5, Prentice–Hall, Englewood Cliffs, NJ, 1998.
[12] M. Clausen and U. Baum, *Fast Fourier Transforms*, BI-Wissenschaftsverlag, Mannheim, Germany, 1993.
[13] E. Chu and A. George, *Inside the FFT Black Box*, CRC Press, Boca Raton, FL, 2000.
[14] C. K. Chui and G. Chen, *Signal Processing and Systems Theory*, Springer, Berlin, 1992.
[15] R. Creutzburg and M. Tasche, *F-Transformation und Faltung in kommutativen Ringen*, Elektron. Inform.-Verarb. Kybernetik, 21 (1985), pp. 129–149.
[16] J. C. Cooley and J. C. Tukey, *An algorithm for machine calculation of complex Fourier series*, Math. Comp., 19 (1965), pp. 297–301.
[17] D. E. Dudgeon and R. M. Mersereau, *Multidimensional Digital Signal Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1984.
[18] H. K. Garg, *Digital Signal Processing Algorithms*, CRC Press, Boca Raton, FL, 2000.
[19] I. J. Good, *The relationship between two fast Fourier transforms*, IEEE Trans. Comput., 20 (1971), pp. 310–317.
[20] L. Hörmander, *The Analysis of Linear Partial Differential Operators* I, Springer, Berlin, 1983.
[21] E. Kamen, *Introduction to Signals and Systems*, Macmillan, New York, 1987.
[22] V. P. Khavin, *Methods and structure of commutative harmonic analysis* I, in Commutative Harmonic Analysis, Encyclopaedia Math. Sci., 15, V. P. Khavin and N. K. Nikol'skij, eds., Springer, Berlin, 1991, pp. 1–111.
[23] S. Lang, *Algebra*, Addison–Wesley, Reading, MA, 1965.

[24] J. S. Lim, *Two-dimensional signal processing*, in Advanced Topics in Signal Processing, J. S. Lim and A. V. Oppenheim, eds., Prentice–Hall, Englewood Cliffs, NJ, 1988, pp. 338–415.
[25] H. D. Lüke, *Signalübertragung*, Springer, Berlin, 1990.
[26] H. J. Nussbaumer, *Fast Fourier Transform and Convolution Algorithms*, Springer, Berlin, 1981.
[27] U. Oberst, *Explizite Rekursionsformeln zur schnellen Fouriertransformation*, in Actes Sémi. Loth. de Combinatoire 18, IRMA, University of Strasbourg, Strasbourg, France, 1988, pp. 119–126.
[28] U. Oberst, *The Fast Fourier Transform*, Publ. 79, Centro Vito Volterra, Universita di Roma II, Rome, 1991.
[29] U. Oberst, *Galois Theory and the Fast Gelfand Transform*, Publ. 99, Centro Vito Volterra, Universita di Roma II, Rome, 1992.
[30] U. Oberst and S. Walch, *The Optimal Fast Fourier, Gelfand and Hartley Transforms*, in preparation.
[31] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[32] A. V. Oppenheim and A. S. Willsky, *Signals and Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1983.
[33] C. M. Rader, *Discrete Fourier transforms when the number of data points is prime*, Proc. IEEE, 56 (1968), pp. 1107–1108.
[34] V. Strassen, *Algebraic complexity theory*, in Algorithms and Complexity, J. V. Leeuwen, ed., Elsevier, Amsterdam, 1990, pp. 633–672.
[35] R. Tolimieri, *Multiplicative characters and the discrete Fourier transform*, Adv. in Appl. Math., 7 (1986), pp. 344–380.
[36] R. Tolimieri and M. An, *Lesser known FFT algorithms*, in Twentieth Century Harmonic Analysis—A Celebration, J. S. Byrnes, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 151–162.
[37] R. Unbehauen, *Systemtheorie*, Oldenbourg Verlag, München, Germany, 1990.
[38] S. Walch and U. Oberst, *Equivariant fast Gelfand and Fourier transforms: Algorithms and computer implementations*, in Mathematical Theory of Networks and Systems, Proceedings of the MTNS-98 Symposium, Il Poligrafo, Padova, Italy, 1998, pp. 899–902.
[39] S. Winograd, *On computing the discrete Fourier transform*, Math. Comp., 32 (1978), pp. 175–199.
[40] S. Winograd, *On the multiplicative complexity of the discrete Fourier transform*, Adv. in Math., 32 (1979), pp. 83–117.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 541–561

PERFORMANCE BOUNDS IN L_p-NORM FOR APPROXIMATE VALUE ITERATION*

RÉMI MUNOS†

Abstract. *Approximate value iteration* (AVI) is a method for solving large Markov decision problems by approximating the optimal value function with a sequence of value function representations V_n processed according to the iterations $V_{n+1} = \mathcal{A}\mathcal{T}V_n$, where $\mathcal{T}$ is the so-called *Bellman operator* and $\mathcal{A}$ an *approximation operator*, which may be implemented by a *supervised learning* (SL) algorithm. Usual bounds on the asymptotic performance of AVI are established in terms of the L_∞-norm approximation errors induced by the SL algorithm. However, most widely used SL algorithms (such as least squares regression) return a function (the best fit) that minimizes an empirical approximation error in L_p-norm ($p \geq 1$). In this paper, we extend the performance bounds of AVI to weighted L_p-norms, which enables us to directly relate the performance of AVI to the approximation power of the SL algorithm, hence assuring the tightness and practical relevance of these bounds. The main result is a performance bound of the resulting policies expressed in terms of the L_p-norm errors introduced by the successive approximations. The new bound takes into account a concentration coefficient that estimates how much the discounted future-state distributions starting from a probability measure used to assess the performance of AVI can possibly differ from the distribution used in the regression operation. We illustrate the tightness of the bounds on an optimal replacement problem.

Key words. Markov decision processes, dynamic programming, optimal control, function approximation, error analysis, reinforcement learning, statistical learning

AMS subject classifications. 49L20, 90C40, 90C59, 93E20

DOI. 10.1137/040614384

1. Introduction.

We consider the problem of solving large state-space *Markov decision processes* (MDPs) [29] in an infinite time horizon, discounted reward setting.

The *value iteration* algorithm is a method for computing the optimal value function V^* by processing a sequence of value function representations V_n according to the iterations $V_{n+1} = \mathcal{T}V_n$, where $\mathcal{T}$ is the so-called *Bellman operator*. Due to a contraction property—in L_∞-norm—of the Bellman operator, the iterates V_n converge to V^* as $n \to \infty$. However, this method is intractable when the number of states is so large that an exact representation of the values is impossible. We therefore need to represent the functions with a moderate number of coefficients and use methods for finding an approximate solution.

A very popular algorithm is the *approximate value iteration* (AVI) algorithm. It has long been implemented in many different settings in dynamic programming (DP) [32, 5] with online variants in the field of reinforcement learning (RL) [7, 33]. It is defined by a sequence of value function representations V_n that are processed recursively by means of the iterations

$$V_{n+1} = \mathcal{A}\mathcal{T}V_n, \tag{1.1}$$

where $\mathcal{T}$ is the *Bellman operator* and $\mathcal{A}$ an *approximation operator*, which may be sampling-based implemented by a *supervised learning* (SL) algorithm (see, e.g., [15]).

*Received by the editors September 3, 2004; accepted for publication (in revised form) December 8, 2006; published electronically May 1, 2007.
http://www.siam.org/journals/sicon/46-2/61438.html

†SequeL team, INRIA Futurs, Université de Lille, 59653 Villeneuve d'Ascq, France (remi.munos@inria.fr).

Since we will make use of different norms, let us now recall their definition: Let $u \in \mathbb{R}^N$. Its supremum (L_∞) norm is defined by $||u||_\infty := \sup_{1\leq x\leq N} |u(x)|$. Now, for μ being a probability measure on $\{1,\ldots,N\}$, the weighted L_p-(semi)norm (for $p \geq 1$)—denoted by $L_{p,\mu}$—of u is $||u||_{p,\mu} := \left[\sum_{1\leq x\leq N} \mu(x)|u(x)|^p\right]^{1/p}$. In addition, we denote by $||\cdot||_p$ the unweighted L_p-norm (i.e., when μ is uniform).

A typical implementation of AVI is *fitted value iteration*, which, given a function space $\mathcal{F}$, computes at each iteration a new value representation $V_{n+1} \in \mathcal{F}$ by projecting onto $\mathcal{F}$ the Bellman image of the current estimate V_n. For illustration, a sampling-based version of this algorithm could be defined as follows: At stage n, we draw a set of independent states $\{x_k \sim \mu\}_{1\leq k\leq K}$, where μ is some probability measure on the state space, compute the Bellman values $\{v_k := \mathcal{T}V_n(x_k)\}_{1\leq k\leq K}$ for the current approximation V_n at those states, then make a call to an SL algorithm with the data $\{(x_k, v_k)\}_{1\leq k\leq K}$ ($\{x_k\}$ being the input and $\{v_k\}$ the desired output). The SL algorithm would return a function V_{n+1} (the best fit) that minimizes some empirical loss

$$V_{n+1} := \arg\min_{g\in\mathcal{F}} \frac{1}{K} \sum_{1\leq k\leq K} l(g(x_k) - v_k),$$

where the *loss function* l is usually a square or an absolute function (or variants, such as the ϵ-insensitive loss used in support vectors [36]).

This is a sampling-based version of the minimization problem in a weighted (by μ) absolute or quadratic norm ($L_{p,\mu}$-norm with $p = 1$ or 2, respectively)

$$\arg\min_{g\in\mathcal{F}} ||g - \mathcal{T}V_n||_{p,\mu}.$$

The field of *statistical learning* analyses the difference between the minimized empirical loss $\frac{1}{K}\sum_{1\leq k\leq K} l(V_{n+1}(x_k) - v_k)$ and the corresponding $L_{p,\mu}$-norm approximation error $||V_{n+1} - \mathcal{T}V_n||_{p,\mu}$ in terms of the number of samples K and a capacity measure of the function space $\mathcal{F}$ (such as the *covering number* or the *Vapnik–Chervonenkis (VC) dimension* [28, 36] of $\mathcal{F}$).

It is therefore natural to search for bounds on the performance of AVI that rely on weighted L_p-norms ($p \geq 1$) of the approximation errors $||V_{n+1} - \mathcal{T}V_n||_{p,\mu}$.

Unfortunately, the main field of investigation so far in approximate DP makes use of the supremum norm [4, 5, 6, 29, 7, 16, 13]. For example, the asymptotic performance of the policies deduced by the AVI algorithm may be bounded in terms of the L_∞-norm of the approximation errors $||V_{n+1} - \mathcal{T}V_n||_\infty$ (see section 2). However, this bound is not very useful since this uniform approximation error is difficult to control in general and is not very practical because most currently known SL algorithms solve an empirical minimization problem in L_p-norm (like least squares regression, neural networks, support vector, and kernel regression). Since most approximation operators provide good approximations in L_p-norm but a poor performance with respect to (w.r.t.) the L_∞-norm, it would be relevant to measure the algorithm performance w.r.t. the former norm.

The purpose of this paper is to extend error bounds for AVI to L_p-norms. The performance of AVI can therefore be directly related to the approximation power of the SL algorithm.

To begin with, let us mention that of course norms are equivalent (in the case of finite-dimensional spaces) since $||\cdot||_p \leq ||\cdot||_\infty \leq N^{1/p}||\cdot||_p$ (with $p \geq 1$ and N being the number of states); thus the usual L_∞ bound for AVI (detailed in section 2) may

also be used to derive an L_p-norm bound. However, because of the $N^{1/p}$ factor, this yields a very loose bound for large scale problems.

The bounds derived here (see Theorem 5.2 in section 5) depend on a new concentration (or stability) measure of the MDP: The *concentration coefficient* $C(\nu, \mu)$ measures how much the discounted average future-state distribution starting from some distribution ν used to assess the performance of AVI (through the weighting of the L_p-norm of the algorithm's performance) can possibly diverge from the distribution μ used in the regression step (by the SL algorithm). This concentration coefficient is defined as an upper bound, taken for any nonstationary policy, of the derivative of the discounted future-state distribution (starting from ν and following a policy) w.r.t. the regression distribution μ.

This coefficient is related to the so-called *top-Lyapunov exponent*, which is commonly used to analyze the stability of stochastic processes. Further discussion about this concept in continuous spaces (where this coefficient is defined in terms of the Radon–Nikodým derivative of the related probability measures) can be found in [27].

A sufficient condition for the concentration coefficient to be small is when the MDP is "smooth" (i.e., when the transition probabilities are strongly stochastic, e.g., close to uniform distribution). Actually, we derive another bound, this time on the L_∞ performance of the AVI algorithm (but still in terms of the L_p approximation errors) using another concentration coefficient $C(\mu)$ that relates the immediate transition probabilities of the MDP to the regression distribution μ. For a uniform μ, a smooth MDP will define a small $C(\mu)$ value, and our bound will be sharp. However, for an MDP with deterministic transitions, the coefficient $C(\mu)$ could depend heavily on the number of states N, making our new bounds no more informative than a usual L_∞-norm bound. This is illustrated in the *chain walk* MDP (for which $C(\mu) = N$) described in subsection 5.5. However, even for deterministic MDPs, the concentration coefficient $C(\nu, \mu)$ may be small, and independent of N, as illustrated in the same example. For such cases, the new L_p bound is arbitrarily better than the usual L_∞ one.

The main intuition underlying this extension of usual L_∞ bounds to L_p-norms is actually simple (see the first paragraph of section 5) and is a consequence of the componentwise bounds obtained in section 4.

To the best of our knowledge, this weighted L_p-norm analysis of AVI is new. Previous L_p analyses in *approximate dynamic programming* (ADP) include *temporal difference learning* (for the evaluation of a fixed policy) with linear approximation [35] and *approximate policy iteration* [26] (and [1] in the continuous space, sampling-based case). Let us mention that there is an important body of literature in the domain of weighted L_∞-norm analysis of ADP [7, 17], especially for the linear programming approach [10]. Let us also remark that there exists an important related field concerned with stability, ergodicity, and convergence properties of future state distributions w.r.t. the invariant probability measure (in Markov chains [19] or MDPs [18, 25]). This is not the direction followed in this paper since we are interested in the discounted reward case (with a fixed discount factor) and not the average reward case.

The paper is organized as follows: In section 2, we recall some approximation results in L_∞-norm. Section 3 is a rough survey of approximation operators and SL algorithms. The main tool used in this paper is the derivation of the componentwise bounds for AVI, detailed in section 4. The performance bounds in L_p-norms are stated in section 5, and the main result of this paper is given in Theorem 5.2. A subsection

provides some intuition on these results in case the AVI algorithm would converge, which leads to bounds expressed in terms of the L_p Bellman residual. Section 6 details practical implementations of AVI (a sampling-based method using state-action value function approximation). The case of a continuous measurable state space is treated in section 7, and a numerical experiment on an optimal replacement problem is detailed.

Preliminaries. We now describe the framework of MDPs in the infinite-time horizon, discounted reward setting, considered here.

Let X be the state space, assumed to be finite with N states, and let A be a finite action space. The results given in this paper extend to infinite state spaces (either countable spaces or continuous spaces, the latter case being illustrated in section 7). Let $p(x,a,y)$ be the probability that the next state is y given that the current state is x and the action a. Let $r(x,a,y)$ be the (deterministic) reward received when a transition $(x,a) \to y$ occurs.

We call a *(Markov* or *stationary) policy* π a mapping from X to A. We denote by P^π the $N \times N$ matrix with elements $P^\pi(x,y) := p(x,\pi(x),y)$ and by r^π the N-vector with components $r^\pi(x) := \sum_y p(x,\pi(x),y) r(x,\pi(x),y)$.

For a given policy π, the *value function* V^π (considered as a vector with N components) is defined as the expected sum of discounted rewards:

$$V^\pi(x) := \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t\, r(x_t,a_t,x_{t+1}) | x_0 = x, a_t = \pi(x_t)\Big],$$

where $\gamma \in [0,1)$ is the *discount factor*. It is well known that V^π is the fixed-point of the operator $\mathcal{T}^\pi : \mathbb{R}^N \to \mathbb{R}^N$ defined, for any vector $W \in \mathbb{R}^N$, by $\mathcal{T}^\pi W := r^\pi + \gamma P^\pi W$.

The *optimal value function* $V^* := \sup_\pi V^\pi$ is the fixed-point of the Bellman operator $\mathcal{T}$ defined, for any $W \in \mathbb{R}^N$, $x \in X$, by

$$\mathcal{T}W(x) = \max_{a \in A} \sum_{y \in X} p(x,a,y)[r(x,a,y) + \gamma W(y)].$$

We say that a policy π is *greedy w.r.t.* $W \in \mathbb{R}^N$ if for all $x \in X$,

$$\pi(x) \in \arg\max_{a \in A} \sum_{y \in X} p(x,a,y)[r(x,a,y) + \gamma W(y)].$$

The goal is to find an optimal policy π^*, which is such that for all $x \in X$, $V^{\pi^*}(x) = \max_\pi V^\pi(x)$. It is easy to see that a policy greedy w.r.t. V^* is optimal. Since $\mathcal{A}$ is finite, such an optimal policy always exits.

2. Approximation results in L_∞-norm. Consider the *AVI algorithm* defined by (1.1) and define

$$\varepsilon_n := \mathcal{T}V_n - V_{n+1} \in \mathbb{R}^N \tag{2.1}$$

as the *approximation error* at stage n. In general, AVI does not converge, but nevertheless its asymptotic behavior may be analyzed. If the approximation errors are uniformly bounded $||\varepsilon_n||_\infty \le \varepsilon$, then a bound on the difference between the asymptotic performance of policies π_n greedy w.r.t. V_n and the optimal policy is (see, e.g., [7])

$$\limsup_{n \to \infty} ||V^* - V^{\pi_n}||_\infty \le \frac{2\gamma}{(1-\gamma)^2}\varepsilon. \tag{2.2}$$

Since the proof is very simple, it is recalled here.

Proof. From the triangle inequality, the γ-contraction of the Bellman operators $\mathcal{T}$ and $\mathcal{T}^{\pi_n}$, and the fact that π_n is greedy w.r.t. V_n (i.e., $\mathcal{T}^{\pi_n} V_n = \mathcal{T} V_n$), we have

$$\begin{aligned} ||V^* - V^{\pi_n}||_\infty &\leq ||\mathcal{T}V^* - \mathcal{T}^{\pi_n} V_n||_\infty + ||\mathcal{T}^{\pi_n} V_n - \mathcal{T}^{\pi_n} V^{\pi_n}||_\infty \\ &\leq \gamma ||V^* - V_n||_\infty + \gamma(||V_n - V^*||_\infty + ||V^* - V^{\pi_n}||_\infty), \end{aligned}$$

and thus

$$||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{1-\gamma} ||V^* - V_n||_\infty. \tag{2.3}$$

Moreover, $||V^* - V_{n+1}||_\infty \leq ||\mathcal{T}V^* - \mathcal{T}V_n||_\infty + ||\mathcal{T}V_n - V_{n+1}||_\infty \leq \gamma ||V^* - V_n||_\infty + \varepsilon$. Now, taking the upper limit yields $\limsup_{n\to\infty} ||V^* - V_n||_\infty \leq \varepsilon/(1-\gamma)$, which combined with (2.3) yields (2.2). □

This L_∞ bound is expressed in terms of the uniform approximation error over all states, which is difficult to guarantee, especially for large state-space problems. Moreover, it is not very useful in practice since most current approximation operators and supervised learning methods perform a minimization problem in L_1- or L_2-norm (although some exceptions of L_∞ function approximation in the framework of DP exist; see, e.g., [12, 14]).

3. Approximation operators and supervised learning algorithms. In this section we present an overview of the problem of function approximation in the context of *statistical learning* (see, e.g., [36, 15]). To illustrate, an example of a supervised learning (SL) algorithm would take as input some data $\{(x_k, v_k)\}_{1\leq k\leq K}$, where the states $\{x_k \in X\}$ are drawn according to some distribution μ on X, and the values $\{v_k \in \mathbb{R}\}$ are unbiased estimates of some (unknown) random function with mean $f(x_k)$. This SL algorithm would return a function (called the *best fit*) that minimizes (within a given class of functions $\mathcal{F}$) the empirical loss, solving

$$\inf_{g\in\mathcal{F}} \frac{1}{K} \sum_{k=1}^{K} l(v_k - g(x_k)),$$

where the loss function l is usually an absolute or a quadratic function (or variants, such as the *ϵ-insensitive loss function* used in support vectors or *Huber loss function* used for robust regression [36]).

If the unknown function is deterministic (i.e., $v_k = f(x_k)$), $\mathcal{A}$ may be considered as an approximation operator that returns a compact representation $g \in \mathcal{F}$ of an unknown function f by minimizing some empirical L_p-norm ($p = 1$ or 2) based on the data. This is a sampling-based version of a minimization problem in weighted norm $L_{p,\mu}$. Statistical learning theory establishes bounds on the error between the minimized empirical loss $\frac{1}{K}\sum_{k=1}^{K} l(f(x_k) - g(x_k))$ and the $L_{p,\mu}$-norm difference $||f - g||_{p,\mu}$ in terms of the number of samples K and the capacity (or complexity) measure of the function space $\mathcal{F}$, characterized, e.g., by the *covering number* or the *VC dimension* [28, 36] of $\mathcal{F}$.

The projection onto the span of a fixed family of functions (often called *features*) is called *linear approximation* and include *splines*, *radial basis, and Fourier* or *wavelet decomposition*. It is often the case that a better approximation is reached when choosing the features according to f (i.e., *feature selection*). This *nonlinear approximation*

is particularly efficient when f has piecewise regularities (e.g., in the adaptive wavelet basis [24] such functions are compactly represented with few nonzero coefficients). Greedy algorithms for selecting the best features among a given dictionary of functions include the *matching pursuit* and variants [9]. Approximation theory studies the approximation error in terms of the smoothness of f [11].

In statistical learning, SL algorithms include *neural network*, *locally weighted learning* and *kernel regression* [2], *support vectors*, and *reproducing kernels* [37, 36].

Hence, given the fact that we may always bound the empirical minimized error using statistical learning tools, in what follows, we will establish our bounds using the $L_{p,\mu}$-norm of the approximation errors. An extension of these results to sampling-based AVI is described in [27] and a policy iteration algorithm with Bellman residual minimization using a single sample-path is described in [1].

4. Componentwise performance bounds. In this section, we formulate componentwise performance bounds, from which L_p bounds will be derived in the next section. The L_∞ bound previously stated (2.2) is also an immediate consequence of a componentwise bound.

4.1. Performance bound for AVI. A componentwise bound on the asymptotic performance of the policies π_n greedy w.r.t. V_n is provided now.

LEMMA 4.1. *Consider the AVI algorithm defined by* (1.1) *and write* $\varepsilon_n = \mathcal{T}V_n - V_{n+1} \in \mathbb{R}^N$ *to denote the approximation error at stage* n. *Let* π_n *be a greedy policy w.r.t.* V_n. *We have*

$$\limsup_{n\to\infty} V^* - V^{\pi_n} \le \limsup_{n\to\infty} (I - \gamma P^{\pi_n})^{-1} \times \left(\sum_{k=0}^{n-1} \gamma^{n-k} \left[(P^{\pi^*})^{n-k} + P^{\pi_n} P^{\pi_{n-1}} \dots P^{\pi_{k+2}} P^{\pi_{k+1}} \right] |\varepsilon_k| \right), \tag{4.1}$$

where $|\varepsilon_k|$ *denotes the vector of absolute values of* ε_k.

In order to prove this lemma, we first need the following preliminary result.

LEMMA 4.2. *Let* A *be an invertible matrix such that all the elements of its inverse are positive. Then the solutions to the inequality* $Au \le b$ *are also solutions to* $u \le A^{-1}b$.

Proof of Lemma 4.2. Let u be a solution to $Au \le b$. This means that there exists a vector c with positive components such that $Au = b - c$; thus $u = A^{-1}b - A^{-1}c$. Since all components of $A^{-1}c$ are positive, we deduce that $u \le A^{-1}b$. □

Proof of Lemma 4.1. From the definitions of $\mathcal{T}$ and $\mathcal{T}^\pi$ we have componentwise $\mathcal{T}V_k \ge \mathcal{T}^{\pi^*}V_k$ and $\mathcal{T}V^* \ge \mathcal{T}^{\pi_k}V^*$, and thus

$$V^* - V_{k+1} = \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V_k + \mathcal{T}^{\pi^*}V_k - \mathcal{T}V_k + \varepsilon_k \le \gamma P^{\pi^*}(V^* - V_k) + \varepsilon_k,$$

$$V^* - V_{k+1} = \mathcal{T}V^* - \mathcal{T}^{\pi_k}V^* + \mathcal{T}^{\pi_k}V^* - \mathcal{T}V_k + \varepsilon_k \ge \gamma P^{\pi_k}(V^* - V_k) + \varepsilon_k,$$

where in the second line we used the definition of π_k as a greedy policy w.r.t. V_k, i.e., $\mathcal{T}^{\pi_k}V_k = \mathcal{T}V_k$. We deduce by induction

$$V^* - V_n \le \sum_{k=0}^{n-1} \gamma^{n-k-1}(P^{\pi^*})^{n-k-1}\varepsilon_k + \gamma^n (P^{\pi^*})^n (V^* - V_0), \tag{4.2}$$

$$V^* - V_n \ge \sum_{k=0}^{n-1} \gamma^{n-k-1}(P^{\pi_{n-1}} P^{\pi_{n-2}} \dots P^{\pi_{k+1}})\varepsilon_k + \gamma^n (P^{\pi_n} P^{\pi_{n-1}} \dots P^{\pi_1})(V^* - V_0). \tag{4.3}$$

Now, using again the definition of π_n and the fact that $\mathcal{T}V_n \geq \mathcal{T}^{\pi^*}V_n$, we have

$$\begin{aligned} V^* - V^{\pi_n} &= \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V_n + \mathcal{T}^{\pi^*}V_n - \mathcal{T}V_n + \mathcal{T}V_n - \mathcal{T}^{\pi_n}V^{\pi_n} \\ &\leq \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V_n + \mathcal{T}V_n - \mathcal{T}^{\pi_n}V^{\pi_n} \\ &= \gamma P^{\pi^*}(V^* - V_n) + \gamma P^{\pi_n}(V_n - V^{\pi_n}) \\ &= \gamma P^{\pi^*}(V^* - V_n) + \gamma P^{\pi_n}(V_n - V^* + V^* - V^{\pi_n}). \end{aligned}$$

Thus $(I - \gamma P^{\pi_n})(V^* - V^{\pi_n}) \leq \gamma(P^{\pi^*} - P^{\pi_n})(V^* - V_n)$. Now, since $(I - \gamma P^{\pi_n})$ is invertible and its inverse $\sum_{k\geq 0}(\gamma P^{\pi_n})^k$ has positive elements, we use Lemma 4.2 to deduce that

$$V^* - V^{\pi_n} \leq \gamma(I - \gamma P^{\pi_n})^{-1}(P^{\pi^*} - P^{\pi_n})(V^* - V_n).$$

This, combined with (4.2) and (4.3), and after taking the absolute value (note that the vector $V^* - V^{\pi_n}$ is nonnegative), yields

$$\begin{aligned} V^* - V^{\pi_n} \leq (I - \gamma P^{\pi_n})^{-1} \Bigg\{ &\sum_{k=0}^{n-1} \gamma^{n-k} \left[(P^{\pi^*})^{n-k} + (P^{\pi_n} P^{\pi_{n-1}} \dots P^{\pi_{k+1}}) \right] |\varepsilon_k| \\ &+ \gamma^{n+1} \left[(P^{\pi^*})^{n+1} + (P^{\pi_n} P^{\pi_n} P^{\pi_{n-1}} \dots P^{\pi_1}) \right] |V^* - V_0| \Bigg\}. \end{aligned}$$

We deduce (4.1) by taking the upper limit. ∎

4.2. Performance bound based on the Bellman residual. In this section, we derive a componentwise performance bound of a policy π greedy w.r.t. some function $V \in \mathbb{R}^N$ in terms of the Bellman residual of V. This result extends the L_∞ bound (see a proof in [38]):

$$||V^* - V^\pi||_\infty \leq \frac{2}{1-\gamma} ||\mathcal{T}V - V||_\infty. \tag{4.4}$$

The componentwise counterpart of this bound is stated now.

LEMMA 4.3. *Let $V \in \mathbb{R}^N$ and let π be a policy greedy w.r.t. V. Then*

$$V^* - V^\pi \leq \left[(I - \gamma P^{\pi^*})^{-1} + (I - \gamma P^\pi)^{-1} \right] |\mathcal{T}V - V|. \tag{4.5}$$

We immediately notice that (4.4) is a direct consequence of this result, since for any stochastic matrix P, $||(I - \gamma P)^{-1}||_\infty = 1/(1-\gamma)$.

Proof of Lemma 4.3. We use the fact that $\mathcal{T}V \geq \mathcal{T}^{\pi^*}V$ and the definition of π (i.e., $\mathcal{T}V = \mathcal{T}^\pi V$) to derive

$$\begin{aligned} V^* - V^\pi &= \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V + \mathcal{T}^{\pi^*}V - \mathcal{T}V + \mathcal{T}V - \mathcal{T}^\pi V^\pi \\ &\leq \gamma P^{\pi^*}(V^* - V^\pi + V^\pi - V) + \gamma P^\pi(V - V^\pi); \end{aligned}$$

hence $(I - \gamma P^{\pi^*})(V^* - V^\pi) \leq \gamma(P^{\pi^*} - P^\pi)(V^\pi - V)$. Again, since $(I - \gamma P^{\pi^*})$ is invertible and its inverse has positive elements, from Lemma 4.2, we deduce

$$V^* - V^\pi \leq \gamma(I - \gamma P^{\pi^*})^{-1}(P^{\pi^*} - P^\pi)(V^\pi - V).$$

Moreover,

$$\begin{aligned}(I-\gamma P^\pi)(V^\pi - V) &= V^\pi - V - \gamma P^\pi V^\pi + \gamma P^\pi V \\ &= r^\pi + \gamma P^\pi V - (r^\pi + \gamma P^\pi V^\pi) + V^\pi - V \\ &= \mathcal{T}^\pi V - \mathcal{T}^\pi V^\pi + V^\pi - V = \mathcal{T}V - V,\end{aligned}$$

and thus

$$\begin{aligned}V^* - V^\pi &\le \gamma(I-\gamma P^{\pi^*})^{-1}(P^{\pi^*} - P^\pi)(I-\gamma P^\pi)^{-1}(\mathcal{T}V - V) \\ &= (I-\gamma P^{\pi^*})^{-1}\Big[(I-\gamma P^\pi) - (I-\gamma P^{\pi^*})\Big](I-\gamma P^\pi)^{-1}(\mathcal{T}V - V) \\ &= \Big[(I-\gamma P^{\pi^*})^{-1} - (I-\gamma P^\pi)^{-1}\Big](\mathcal{T}V - V) \\ &\le \Big[(I-\gamma P^{\pi^*})^{-1} + (I-\gamma P^\pi)^{-1}\Big]|\mathcal{T}V - V|. \qquad \square\end{aligned}$$

5. Approximation results in L_p-norms. In this section, we generalize the previously mentioned L_∞ bounds to L_p-norms. The main intuition behind this extension is simple and relies on the componentwise results described in the previous section.

Indeed, assume that there exist two vectors u and v with positive components, such that componentwise $u \le Qv$, where Q is a stochastic matrix. Of course, we may deduce that $||u||_\infty \le ||v||_\infty$, but in addition, if ν and μ are probability measures on X such that componentwise $\nu Q \le C\mu$, where $C \ge 1$ is a constant (and using usual matrix notation with the probability measures being considered as row vectors), then we deduce that

$$||u||_{p,\nu} \le C^{1/p}||v||_{p,\mu}.$$

Indeed we have

$$\begin{aligned}||u||_{p,\nu}^p = \sum_{x\in X}\nu(x)|u(x)|^p &\le \sum_{x\in X}\nu(x)\Big[\sum_{y\in X}Q(x,y)v(y)\Big]^p \\ &\le \sum_{x\in X}\nu(x)\sum_{y\in X}Q(x,y)v(y)^p \\ &\le C\sum_{y\in X}\mu(y)|v(y)|^p = C||v||_{p,\mu}^p,\end{aligned}$$

using Jensen's inequality.

For example, if the Markov chain induced by Q has an invariant probability measure ν, then we have $||u||_{p,\nu} \le ||v||_{p,\nu}$ (i.e., the constant $C = 1$). This is the main tool used in [35] to derive an L_p-norm bound for temporal difference learning with linear function approximation, where only one policy is considered.

Now, in an MDP, there are several policies, and thus several stochastic matrices to be considered in order to relate $||u||_{p,\nu}$ to $||v||_{p,\mu}$. The next subsection defines the *concentration coefficients* $C_1(\nu,\mu)$, $C_2(\nu,\mu)$, and $C(\mu)$ that generalize the constant C used here for the case when several policies are considered.

A simple case for which the above idea may apply is the case of Bellman residual bounds: Choose $u = V^* - V^\pi$ and $v = \frac{2}{1-\gamma}|\mathcal{T}V - V|$, and notice that the L_∞ bound (4.4) is a consequence of (4.5). The above idea will yield an L_p-norm performance bound (this will be done in subsection 5.3).

This same idea also holds for deriving performance bounds for AVI. We notice that the L_∞ bound (2.2) may be deduced from the componentwise bounds (4.1), and extension to L_p-norms is possible with an adequate constant, to be defined now.

5.1. Definition of the concentration coefficients. We now define the concentration coefficients $C(\mu)$, $C_1(\nu, \mu)$, and $C_2(\nu, \mu)$, which depend on the MDP, under which the distributions ν and μ may be related. Let ν and μ be two probability measures on X.

DEFINITION 5.1. *We call* $C(\mu) \in \mathbb{R}^+ \cup \{+\infty\}$ *the* transition probabilities concentration coefficient*, defined by*

$$C(\mu) = \max_{x,y \in X,\, a \in A} \frac{p(x,a,y)}{\mu(y)}$$

(with the convention that $0/0 = 0$*, and we set* $C(\mu) = \infty$ *if* $\mu(y) = 0$ *and* $p(x,a,y) > 0$ *for some* x, y, a*). Now, let* $\pi_1, \pi_2, \ldots$ *denote any sequence of policies. For every integer* $m \geq 1$*, we define* $c(m) \in \mathbb{R}^+ \cup \{+\infty\}$ *by*

$$c(m) = \max_{\pi_1,\ldots,\pi_m,\, y \in X} \frac{(\nu P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m})(y)}{\mu(y)} \tag{5.1}$$

(with the same convention as above) and write $c(0) = 1$*. Note that these constants depend on* ν *and* μ*.*

We define $C_1(\nu, \mu)$ *and* $C_2(\nu, \mu) \in \mathbb{R}^+ \cup \{+\infty\}$*, the* first *and* second order discounted future state distribution concentration coefficients*, by*

$$C_1(\nu, \mu) := (1-\gamma) \sum_{m \geq 0} \gamma^m c(m), \tag{5.2}$$

$$C_2(\nu, \mu) := (1-\gamma)^2 \sum_{m \geq 1} m \gamma^{m-1} c(m). \tag{5.3}$$

Note that since these coefficients will appear in our bounds we are interested in the cases of finite values, for which it is sufficient that the distribution μ be strictly positive.

The transition probability concentration coefficient $C(\mu)$ was introduced in [26] to derive performance bounds for approximate policy iteration. $C(\mu)$ provides information about the relative smoothness of the immediate transition probabilities w.r.t. μ, whereas $C_1(\nu, \mu)$ and $C_2(\nu, \mu)$ give information about the worst discounted average future state distribution when starting from ν and following any policy. Informally, the future state transition is a probability measure over the state space induced by the state visitation frequency of the Markov chain resulting from the MDP when following a policy.

The coefficients $c(m)$ measure how much the future state distributions $\nu P^{\pi_1} \ldots P^{\pi_m}$ may possibly differ from the distribution μ. The definition of $C_1(\nu, \mu)$ and $C_2(\nu, \mu)$ introduces an exponential discounting (first order discounting weight of γ^m for $C_1(\nu, \mu)$, and second order discounting weight of $(m+1)\gamma^m$ for $C_2(\nu, \mu)$, where m is the horizon time). The discounting makes these coefficients small for a reasonably

large class of MDPs. For any sequence of policies $\pi_1, \ldots, \pi_m$, the (first and second order) discounted future state distributions starting from ν and using this sequence of policies (i.e., $\{x_i \sim p(x_{i-1}, \pi_i(x_{i-1}), \cdot)\}_{1 \le i \le m}$) is bounded by these coefficients ($C_1(\nu,\mu)$ and $C_2(\nu,\mu)$) times μ: for all x_0, y in X,

$$(1-\gamma) \sum_{m \ge 0} \gamma^m \Pr\big(x_m = y \big| x_0 \sim \nu, \pi_1, \ldots, \pi_m\big) \le C_1(\nu,\mu)\mu(y),$$

$$(1-\gamma)^2 \sum_{m \ge 1} m\gamma^{m-1} \Pr\big(x_m = y \big| x_0 \sim \nu, \pi_1, \ldots, \pi_m\big) \le C_2(\nu,\mu)\mu(y).$$

These coefficients are related to the so-called *top-Lyapunov exponent* Γ, which play a fundamental role in the stability analysis of stochastic processes. It turns out that the stability of a stochastic system, as related to the top-Lyapunov condition $\Gamma \le 0$ [8], is equivalent to the finiteness of the concentration coefficients. Hence, a small value of these coefficients can be interpreted as a stability condition too. Further discussion about this concept can be found in the report [27].

5.2. L_p-norm performance bounds for AVI. The next result establishes performance bounds for AVI in terms of the $L_{p,\mu}$-norm of the approximation errors $\varepsilon_n = V_{n+1} - \mathcal{T}V_n$.

THEOREM 5.2. *Let μ and ν be two probability measures on X. Consider the AVI algorithm defined by* (1.1)*, write π_n to denote a policy greedy w.r.t. V_n, and let $\varepsilon_n = V_{n+1} - \mathcal{T}V_n \in \mathbb{R}^N$ be the approximation error. Let $\varepsilon > 0$ and assume that $\mathcal{A}$ returns ε-approximations V_{n+1} in the $L_{p,\mu}$-norm ($p \ge 1$) of $\mathcal{T}V_n$, i.e., $||\varepsilon_n||_{p,\mu} \le \varepsilon$ for $n \ge 0$. Then*

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \le \frac{2\gamma}{(1-\gamma)^2} \big[C(\mu)\big]^{1/p} \varepsilon, \tag{5.4}$$

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_{p,\nu} \le \frac{2\gamma}{(1-\gamma)^2} \big[C_2(\nu,\mu)\big]^{1/p} \varepsilon. \tag{5.5}$$

Notice that the left-hand side (l.h.s.) of the first result (5.4) evaluates the performance in terms of a L_∞-norm, whereas the l.h.s. of the second result (5.5) makes use of an L_p-norm (although the right-hand side (r.h.s.) of both results is expressed in the L_p-norm). The first result does not depend on the distribution ν and may directly be compared to the L_∞ bound (2.2). Actually (5.4) directly implies (2.2) when $p \to \infty$ (for any strictly positive measure μ).

Proof of Theorem 5.2. First, notice that the coefficient $C(\mu)$ is always larger than $C_2(\nu,\mu)$ for any distribution ν. Indeed, for all $m \ge 1$, $c(m) \le C(\mu)$. Thus $C_2(\nu,\mu) \le (1-\gamma)^2 \sum_{m\ge1} m\gamma^{m-1} C(\mu) = C(\mu)$. Thus, if the bound (5.5) holds for any ν, choosing ν to be a Dirac at each state implies that (5.4) also holds. Therefore, we only need to prove (5.5). We may rewrite (4.4) as

$$V^* - V^{\pi_n} \le \frac{2\gamma(1-\gamma^{n+1})}{(1-\gamma)^2} \left[\sum_{k=0}^{n-1} \alpha_k A_k |\varepsilon_k| + \alpha_n A_n |V^* - V_0| \right],$$

with the positive coefficients $\{\alpha_k\}_{0 \le k \le n}$,

$$\alpha_k := \frac{(1-\gamma)\gamma^{n-k-1}}{1-\gamma^{n+1}} \quad \text{for } 0 \le k < n,$$

$$\alpha_n := \frac{(1-\gamma)\gamma^n}{1-\gamma^{n+1}}$$

(we notice that the sum $\sum_{k=0}^{n} \alpha_k = 1$), and the stochastic matrices $\{A_k\}_{0 \le k \le n}$,

$$A_k := \frac{1-\gamma}{2}(I - \gamma P^{\pi_n})^{-1}\left[(P^{\pi^*})^{n-k} + (P^{\pi_n} P^{\pi_{n-1}} \dots P^{\pi_{k+1}})\right] \quad \text{for } 0 \le k < n,$$

$$A_n := \frac{1-\gamma}{2}(I - \gamma P^{\pi_n})^{-1}\left[(P^{\pi^*})^{n+1} + (P^{\pi_n} P^{\pi_n} \dots P^{\pi_1})\right].$$

Since the two sides of this componentwise bound are positive, we may take the $L_{p,\nu}$-norm of those two vectors:

$$\begin{aligned} &||V^* - V^{\pi_n}||_{p,\nu}^p \\ &\quad \le \left[\frac{2\gamma(1-\gamma^{n+1})}{(1-\gamma)^2}\right]^p \sum_{x \in X} \nu(x) \left[\sum_{k=0}^{n-1} \alpha_k A_k |\varepsilon_k| + \alpha_n A_n |V^* - V_0|\right]^p (x) \\ &\quad \le \left[\frac{2\gamma(1-\gamma^{n+1})}{(1-\gamma)^2}\right]^p \sum_{x \in X} \nu(x) \left[\sum_{k=0}^{n-1} \alpha_k A_k |\varepsilon_k|^p + \alpha_n A_n |V^* - V_0|^p\right](x), \end{aligned} \tag{5.6}$$

using two times Jensen's inequality (since the coefficients $\{\alpha_k\}_{0 \le k \le n}$ sum to 1 and the matrix A_k are stochastic) (i.e., convexity of $x \to |x|^p$). The second term in the brackets disappears when taking the upper limit. Now, from the definition of the coefficients $c(m)$, $\nu A_k \le (1-\gamma) \sum_{m \ge 0} \gamma^m c(m+n-k)\mu$; thus the first term in (5.6) satisfies

$$\begin{aligned} \sum_x \nu(x) \sum_{k=0}^{n-1} \alpha_k A_k |\varepsilon_k|^p(x) &\le \sum_{k=0}^{n-1} \alpha_k (1-\gamma) \sum_{m \ge 0} \gamma^m c(m+n-k) ||\varepsilon_k||_{p,\mu}^p \\ &\le \frac{(1-\gamma)^2}{1-\gamma^{n+1}} \sum_{m \ge 0} \sum_{k=0}^{n-1} \gamma^{m+n-k-1} c(m+n-k) \varepsilon^p \\ &\le \frac{1}{1-\gamma^{n+1}} C_2(\nu, \mu) \varepsilon^p, \end{aligned}$$

where we replaced α_k by their values, and used the fact that $||\varepsilon_k||_{p,\mu} \le \varepsilon$. By taking the upper limit in (5.6), we deduce (5.5). □

What if AVI converges? We know that there is no guarantee that AVI converges. However, experimentally, we observe that in some cases convergence occurs. It is interesting to notice that in such cases, better bounds may be derived (in any norm) whenever $\gamma > 1/2$. Indeed, convergence of AVI would mean that there exists $V \in \mathbb{R}^N$ such that $\lim_{n \to \infty} V_n = V$. Thus, by taking the limit in (1.1), we deduce that V is a fixed-point of the operator $\mathcal{A}\mathcal{T}$, i.e., $V = \mathcal{A}\mathcal{T}V$, and the approximation error (2.1) tends to the residual $\mathcal{T}V - V$ of V.

We deduce that the asymptotic performance of AVI is the performance of a policy π greedy w.r.t. V, and thus may be expressed in terms of the residual $\mathcal{T}V - V$. Hence, the bounds based on the Bellman residual (the L_∞-norm bound (4.4) or the componentwise bound (4.5)), which yields a coefficient $2/(1-\gamma)$ instead of $2\gamma/(1-\gamma)^2$ (for AVI bounds), provides a better bound whenever $\gamma > 1/2$. The next subsection provides an extension of Bellman residual bounds to L_p-norms.

5.3. L_p-norm bounds based on the Bellman residual. Here, we relate the performance of a policy π greedy w.r.t. V (where $V \in \mathbb{R}^N$) in terms of the $L_{p,\mu}$-norm of its residual $\mathcal{T}V - V$.

THEOREM 5.3. *Let V be a vector of size N and π a policy greedy w.r.t. V. Let μ and ν be two probability measures on X. Then*

$$||V^* - V^\pi||_\infty \leq \frac{2}{(1-\gamma)}\big[C(\mu)\big]^{1/p}||\mathcal{T}V - V||_{p,\mu}, \tag{5.7}$$

$$||V^* - V^\pi||_{p,\nu} \leq \frac{2}{(1-\gamma)}\big[C_1(\nu,\mu)\big]^{1/p}||\mathcal{T}V - V||_{p,\mu}. \tag{5.8}$$

Here also the first result (5.7) provides an L_∞-norm bound on the performance, which may directly be compared to the L_∞ bound (4.4) (letting $p \to \infty$), whereas an L_p-norm performance bound is stated in the second result (5.8).

Proof of Theorem 5.3. We may rewrite (4.5) as

$$V^* - V^\pi \leq \frac{2}{1-\gamma} A|\mathcal{T}V - V|,$$

where A is the stochastic matrix

$$A = \frac{1-\gamma}{2}\Big[(I - \gamma P^{\pi^*})^{-1} + (I - \gamma P^\pi)^{-1}\Big].$$

Using the idea described in the introduction of this section, we have

$$\begin{aligned} ||V^* - V^\pi||_{p,\nu}^p &\leq \Big[\frac{2}{1-\gamma}\Big]^p \sum_{x\in X} \nu(x)\Big[A|\mathcal{T}V - V|\Big]^p(x) \\ &\leq \Big[\frac{2}{1-\gamma}\Big]^p \sum_{x\in X} \nu(x)\Big[A|\mathcal{T}V - V|^p\Big](x) \end{aligned} \tag{5.9}$$

from Jensen's inequality. Now, from the definition of the coefficients $c(m)$, $\nu A \leq (1-\gamma)\sum_{m\geq 0}\gamma^m c(m)\mu = C_1(\nu,\mu)\mu$, and thus

$$||V^* - V^\pi||_{p,\nu}^p \leq \Big[\frac{2}{1-\gamma}\Big]^p C_1(\nu,\mu)\mu|\mathcal{T}V - V|^p = \Big[\frac{2}{1-\gamma}\Big]^p C_1(\nu,\mu)||\mathcal{T}V - V||_{p,\mu}^p,$$

which proves (5.8). Now, since $C(\mu) \geq C_1(\nu,\mu)$ for any ν, choosing ν to be a Dirac at each state yields (5.7). □

For the purposes of intuition, the components $A(x,y)$ of the matrix A indicate a bound on the contribution of the (absolute value of the) residual at a state y to the performance error at the state x. Indeed,

$$V^*(x) - V^\pi(x) \leq \frac{2}{1-\gamma}\sum_{y\in X} A(x,y)|\mathcal{T}V - V|(y).$$

It is clear from (5.9) that if we chose $\mu = \nu A$, then the L_p bound becomes

$$||V^* - V^\pi||_{p,\nu} \leq \frac{2}{(1-\gamma)}||\mathcal{T}V - V||_{p,\mu}. \tag{5.10}$$

This bound may inspire us for solving a direct Bellman residual minimization problem, in some given function space $\mathcal{F}$,

$$\min_{V \in \mathcal{F}} ||\mathcal{T}V - V||_{p,\mu}^p,$$

where the distribution μ now depends on V, through the policy π greedy w.r.t. V, i.e., $\mu = \nu A = \frac{1-\gamma}{2}\nu[(I - \gamma P^{\pi^*})^{-1} + (I - \gamma P^{\pi})^{-1}]$. We write $\mu = (\mu^\pi + \mu^*)/2$, with $\mu^\pi = (1-\gamma)\nu(I-\gamma P^\pi)^{-1}$ being the discounted future state distribution starting from ν and following policy π, and $\mu^* = (1-\gamma)\nu(I - \gamma P^{\pi^*})^{-1}$, similarly defined from the optimal policy π^*.

Thus the $L_{p,\mu}$-norm of the residual to be minimized is composed of two contributions:

$$||\mathcal{T}V - V||_{p,\mu}^p = \frac{1}{2}\Big(||\mathcal{T}V - V||_{p,\mu^\pi}^p + ||\mathcal{T}V - V||_{p,\mu^*}^p\Big). \tag{5.11}$$

One may consider an iterative optimization method, such as a gradient method, where at each iteration an empirical residual would be computed and minimized. Minimization of the first term in (5.11) is easy to implement by designing a sampling device from μ^π (i.e., start from an initial state $x \sim \nu$ and follow transitions using the current policy π during a horizon time that is a exponential random variable with coefficient γ). The second term is more difficult to deal with because there is no sampling device from μ^* since π^* is unknown; one may consider a somewhat uniform density instead or use a discounted future state distribution using a stochastic policy (where each action has a strict positive probability to be chosen).

5.4. Some intuition about the coefficients $C(\mu)$, $C_1(\nu,\mu)$, and $C_2(\nu,\mu)$. Let us give some more insight about these coefficients in the case of a uniform distribution $\mu = (\frac{1}{N} \dots \frac{1}{N})$. In this case, from its definition, the coefficient $C(\mu)$ is always smaller than the number of states N. $C(\mu)$ equals N if there exists at least a deterministic transition (i.e., for some $x, y \in X$, $a \in A$, we have $p(x,a,y) = 1$). In that case, the L_p (say, for $p = 1$) bound (5.4) would not be better than the L_∞ one (2.2) combined with the simple norm comparison result $||\cdot||_\infty \leq N||\cdot||_1$.

Hence, the L_p bound (5.4) (resp., (5.7)) is more informative than the usual L_∞ one (2.2) (resp., (4.4)) whenever the concentration coefficient $C(\mu)$ is smaller than the number of states. An interesting case for which this happens is when the state space is continuous and the transition kernel admits a density w.r.t. μ, for which case $C(\mu)$ is the upper bound of this density. This continuous space case will be considered in section 7 and illustrated on an optimal replacement problem.

Now, consider the coefficients $C_1(\nu,\mu)$ and $C_2(\nu,\mu)$ when ν and μ are both uniform.

- Their largest possible value is obtained in an MDP where for a specific policy π all states jump to a given state—say state 1—with probability 1. Thus, for any ν, for all m, $\nu(P^\pi)^m = (1\,0 \dots 0) \leq c(m)\mu$ holds with $c(m) = N$ (with equality in state 1), and therefore $C_1(\nu,\mu) = C_2(\nu,\mu) = N$. This is the worst case because the future state distribution accumulates on a single state. In that case, the L_p bound (5.5) (resp., (5.8)) may actually be derived from the L_∞ one (2.2) (resp., (4.4)) since $||\cdot||_p \leq ||\cdot||_\infty$ and $||\cdot||_\infty \leq N^{1/p}||\cdot||_p$.
- Their lowest possible value is obtained in an MDP with uniform transition probabilities $p(x,a,y) = 1/N$ for all $x, y \in X$ and $a \in A$. When ν and μ are both uniform then $c(m) = 1$ and $C_1(\nu,\mu) = C_2(\nu,\mu) = 1$ (this is the lowest

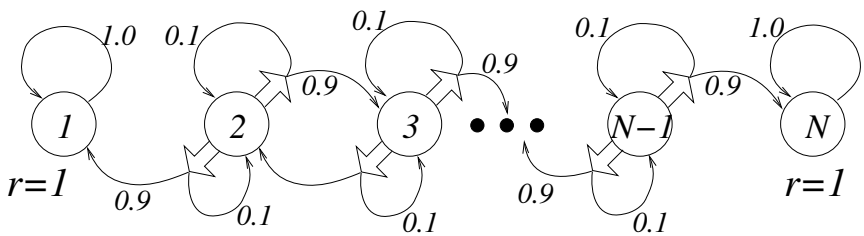


FIG. 5.1. *The chain walk MDP.*

possible value since for a uniform ν and any stochastic matrix P we have $\max_y \sum_x \nu(x)P(x,y) \geq 1/N$).

Notice, however, that any deterministic MDP would not necessarily lead to a high value of the coefficients $C_1(\nu,\mu)$ and $C_2(\nu,\mu)$ (contrarily to the case of $C(\mu)$). Indeed, in an MDP where the policies consist in permutations of the states (for which each state has a unique successor and unique predecessor), then $C(\mu) = N$ (since the transitions are deterministic, as seen previously), but $C_1(\nu,\mu) = C_2(\nu,\mu) = 1$ for uniform distributions ν and μ (since for all $m \geq 0$, $c(m) = 1$). Another example where the discounted future state distribution concentration coefficients is low (and independent of the number of states N) is provided in the chain walk MDP described in the next subsection.

The concentration coefficients $C_1(\nu,\mu)$ and $C_2(\nu,\mu)$ express how the (first and second order) discounted future state distribution, starting from the initial distribution ν, may possibly differ from μ. A low value of these coefficients means that the mass of the discounted future state distribution starting from ν does not accumulate on few specific states for which the distribution μ is low. For the purpose of obtaining low values of these coefficients (thus probably good performance for AVI), it is desirable that μ be somehow uniformly distributed (this condition was already mentioned in [22, 21, 26] to secure the policy improvement steps in approximate policy iteration).

5.5. Illustration on the chain walk MDP. We illustrate the fact that the L_p-norm bound (5.5) given in Theorem 5.2 is tighter than the L_∞-norm (2.2) (combined with the norm comparison $||\cdot||_\infty \leq N^{1/p}||\cdot||_p$) on the *chain walk* MDP defined in [23] (see Figure 5.1). This case provides an example for which the coefficient $C(\mu)$ is high (its value is the number of states N) but $C_1(\nu,\mu)$ and $C_2(\nu,\mu)$ are low (independent of N).

This is a linear chain with N states with two dead-end states: states 1 and N. On each of the interior states $2 \leq x \leq N-1$ there are two possible actions: right or left, which moves the state in the intended direction with probability 0.9, and fails with probability 0.1, leaving the state unchanged. The reward simply depends on the current state and is 1 at boundary states and 0 elsewhere: $r = (1\,0\ldots0\,1)'$.

We consider an approximation of the value function in the two-dimensional function space $\mathcal{F} := \{f_\alpha(x) = \alpha_1 + \alpha_2 x\}_{\alpha\in\mathbb{R}^2}$, where $x \in \{1,\ldots,N\}$ is the state index. Assume that the initial approximation is zero: $V_0 = (0\ldots0)'$. Then $\mathcal{T}V_0 = (1\,0\ldots0\,1)'$. The best fit (in L_∞-norm) of $\mathcal{T}V_0$ in $\mathcal{F}$ is the constant function $V_1 = (\frac{1}{2}\ldots\frac{1}{2})'$, which produces an error $||V_1 - \mathcal{T}V_0||_\infty = \frac{1}{2}$.

Let us choose uniform distributions $\nu = \mu = (\frac{1}{N}\ldots\frac{1}{N})$. In L_1-norm, the best fit of $\mathcal{T}V_0$ in $\mathcal{F}$ is $V_1 = (0\ldots0)'$ (for $N > 4$) and the resulting error is $||V_1 - \mathcal{T}V_0||_1 = \frac{2}{N}$. In L_2-norm the best fit is also constant $V_1 = (\frac{2}{N}\ldots\frac{2}{N})'$ and the error is $||V_1 - \mathcal{T}V_0||_2 = \frac{\sqrt{2N-4}}{N}$.

In these three cases, we observe by induction that the successive approximations V_n are constant; thus $\mathcal{T}V_n = r + \gamma V_n$ and the approximation errors remain the same as in the first iteration: for all $n \geq 0$, $||V_{n+1} - \mathcal{T}V_n||_\infty = \frac{1}{2}$, $||V_{n+1} - \mathcal{T}V_n||_1 = \frac{2}{N}$, and $||V_{n+1} - \mathcal{T}V_n||_2 = \frac{\sqrt{2N-4}}{N}$.

Since V_n is constant, any policy π_n is greedy w.r.t. V_n. Hence for $\pi_n = \pi^*$ the l.h.s. of (2.2) and (5.5) are equal to zero. Now, in order to compare the r.h.s. of these inequalities, let us calculate the coefficients $C(\mu)$ and $C_1(\nu, \mu)$ and $C_2(\nu, \mu)$. Since state 1 jumps to itself with probability 1, we have no better coefficient than $C(\mu) = N$.

Now, the maximum in (5.1) is reached when the mass of the future state distribution is mostly concentrated on one specific state—say state 1—which corresponds to a policy π_{Left} that chooses everywhere action left. We see that for $\nu = \mu$,

$$\nu(P^{\pi_{\text{Left}}})^m(x) \leq \nu(P^{\pi_{\text{Left}}})^m(1) \leq (1 + 0.9m)\mu(x)$$

for all $x \geq 0$, and thus $c(m) \leq 1 + 0.9m$. We deduce that the coefficients $C_1(\nu, \mu) \leq (1-\gamma)\sum_{m\geq 0} \gamma^m(1+0.9m)$ and $C_2(\nu, \mu) \leq (1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1}(1+0.9m)$ are upper bounded by a value that is *independent of the number of states* N.

Thus, if we consider the performance of AVI in L_1-norm, the bound (5.5) (for $p = 1$) provides an approximation of order $O(N^{-1})$, whereas the L_1 bound that would be obtained from the usual L_∞ result (2.2) combined with the norm comparison $||\cdot||_\infty \leq N||\cdot||_1$ would provide a $O(1)$ approximation only.

Similarly, the L_2-norm bound is of order $O(N^{-1/2})$, whereas the L_∞-norm bound (2.2) combined with $||\cdot||_\infty \leq N^{1/2}||\cdot||_2$ would only be of order $O(1)$.

Thus, if our supervised learning algorithm returns the best regression function by minimizing an approximation error in L_p-norm (which is usually the case in practice), *the bound* (5.5) *may be arbitrarily more informative than* (2.2) *for large values of* N.

6. Practical algorithms. Practical implementations of AVI depend on the amount of knowledge available on the state dynamics as well as the way the expectation operation (in the Bellman operator) may be processed.

In the case of a *complete model* (when the state transitions $p(x, a, y)$ are perfectly known) and if the expectation operation is computationally tractable, then a possible implementation of AVI has already been described in the introduction: at each stage n, we select a set of states $\{x_k \in X\}_{1\leq k\leq K}$ drawn according to some distribution μ, compute the backed-up values $\{v_k = \mathcal{T}V_n(x_k)\}_{1\leq k\leq K}$, and make a call to an SL algorithm with the data $\{(x_k; v_k)\}_{1\leq k\leq K}$, which returns an ε-approximation V_{n+1} in $L_{p,\mu}$-norm, i.e., $||V_{n+1} - \mathcal{T}V_n||_{p,\mu} \leq \varepsilon$. Of course, we need additional assumptions on the number of samples K and the complexity of the function space $\mathcal{F}$ (in terms of covering number or VC dimension) to guarantee that the empirical loss $\left(\frac{1}{K}\sum_{k=1}^K |V_{n+1}(x_k) - v_k|^p\right)^{1/p}$ is close to the norm of the approximation error $||V_{n+1} - \mathcal{T}V_n||_{p,\mu}$, but such considerations are omitted here, and we direct the interested reader to [36, 15, 30].

However, it is often the case that no explicit representation of the transition probabilities $p(x, a, y)$ is available, but there exists a sampling device that allows us to generate states y according to the distribution $p(x, a, \cdot)$ at any state x and action a of our choice. We call this a *generative model* (see [20] for a survey of several sampling models). One possible way to compute the expectation operation in the Bellman operator is to replace it by an empirical mean using this sampling device. This leads to *sampling-based fitted value iteration*, studied in [34].

Another alternative, closer in spirit to reinforcement learning (RL) [33], consists in introducing the state-action value function, or Q-function, defined for each state-action $(x,a) \in X \times A$ by

$$Q^*(x,a) := \sum_{y \in X} p(x,a,y)\big[r(x,a,y) + \gamma V^*(y)\big].$$

We have the properties that $V^*(x) = \max_{a \in A} Q^*(x,a)$, and Q^* is the fixed-point of the operator $\mathcal{R}$, mapping from the space of functions $X \times A \to \mathbb{R}$ to itself, defined for any $Q : X \times A \to \mathbb{R}$ by

$$\mathcal{R}Q(x,a) := \sum_{y \in X} p(x,a,y)\big[r(x,a,y) + \gamma \max_{b \in A} Q(y,b)\big].$$

An AVI algorithm using this representation would consist in defining successive approximations Q_n (with any initial Q_0) according to the recursion

$$Q_{n+1} = \mathcal{A}\mathcal{R}Q_n, \tag{6.1}$$

where $\mathcal{A}$ is an SL algorithm on $X \times A$. A model-free RL algorithm would collect a number of transitions of the form $\{(x_k,a_k) \stackrel{r_k}{\to} y_k\}_{1 \le k \le K}$, where a_k is an action chosen in state x_k, the next state y_k being generated according to the generative model (i.e., $y_k \sim p(x_k,a_k,\cdot)$), and $r_k = r(x_k,a_k,y_k)$ is the received reward. We then compute the back-up values $v_k = r_k + \gamma \max_{b \in A} Q_n(y_k,b)$ (which provides an unbiased estimate of $\mathcal{R}Q_n(x_k,a_k)$), and make a call to the SL algorithm with the data $\{(x_k,a_k); v_k\}_{1 \le k \le K}$ (the inputs being the couples $\{(x_k,a_k)\}$, and the desired output $\{v_k\}$), which returns the next Q-function Q_{n+1}.

An interesting case is when $\mathcal{A}$ is a linear operator *in the values* $\{v_k\}$ such as in linear approximation, memory-based learning (k-nearest neighbors, locally weighted learning [3, 15]) or support vector regression (in the case of a quadratic loss function). In that case, the approximation $\mathcal{A}$ and expectation $\mathbb{E}$ operators commute and the approximation Q_{n+1} returned by the SL algorithm is therefore an unbiased estimate of $\mathcal{A}\mathcal{R}Q_n$. Thus when K is large, such an iteration acts like a (model-based) AVI iteration, and bounds similar to those of Theorem 5.2 may be derived.

Notice that a policy π'_n derived from the approximate Q-function, $\pi'_n(x) \in \arg\max_{a \in A} Q_n(x,a)$, is different from the policy π_n greedy w.r.t. V_n, defined by $V_n(x) = \max_a Q_n(x,a)$. Indeed, the latter satisfies $\pi_n(x) \in \arg\max_{a \in A} \mathcal{R}Q_n(x,a)$. However, bounds similar to (2.2), (5.4), and (5.5) on the performance of such policies π'_n may be derived analogously. An example of such a bound in L_∞-norm is provided now. Extension to L_p bounds would follow along the same lines as in sections 4 and 5.

The performance $Q^\pi : X \times A \to \mathbb{R}$ of a policy π is defined as follows: $Q^\pi(x,a)$ is the expected sum of rewards when starting from x, choosing action a and using policy π thereafter. Q^π is also the fixed-point of the Bellman operator $\mathcal{R}^\pi$, mapping from the space of functions $X \times A \to \mathbb{R}$ to itself, defined by

$$\mathcal{R}^\pi Q(x,a) := \sum_{y \in X} p(x,a,y)\big[r(x,a,y) + \gamma Q(y,\pi(y))\big].$$

THEOREM 6.1. *Consider the AVI algorithm defined by the Q-function iteration* (6.1). *Let ε be a uniform bound on the L_∞ approximation errors of the Q-functions,*

i.e., $||Q_{n+1} - \mathcal{R}Q_n||_\infty \leq \varepsilon$. *The asymptotic performance of the policy* π'_n *(defined by* $\pi'_n(x) \in \arg\max_{a \in A} Q_n(x,a)$*) satisfies*

$$\limsup_{n \to \infty} ||Q^* - Q^{\pi'_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\varepsilon.$$

Proof of Theorem 6.1. The proof is similar to that of (2.2); it suffices to replace the V-value by the Q-values, the $\mathcal{T}$ (resp., $\mathcal{T}^\pi$) operator by the $\mathcal{R}$ (resp., Q^π) operators, and notice that $\mathcal{R}^{\pi'_n} Q_n = \mathcal{R}Q_n$. □

7. Numerical experiment in the continuous case. All previous results extend to the case of continuous measurable state spaces. We first redefine the concentration coefficients in this context and illustrate numerically the method on an optimal replacement problem, for which the coefficient $C(\mu)$ is explicitly computed.

Let us write $P(x,a,B)$ the transition probability kernel, where B is any measurable subset of X. For a stationary policy $\pi : X \to A$, we write $P^\pi(x,B) = P(x,\pi(x),B)$, which defines a right linear operator (defined on the space of bounded measurable function V with domain X): $P^\pi V(x) := \int_X V(y) P^\pi(x,dy)$, and a left-linear operator (defined on the space of probability measures μ on X): $\mu P^\pi(B) := \int_X P^\pi(x,B)\mu(dx)$. The product of two kernels P^{π_1} and P^{π_2} is defined by $P^{\pi_1}P^{\pi_2}(x,B) := \int_X P^{\pi_1}(x,dy)P^{\pi_2}(y,B)$.

7.1. Concentration coefficients. With this notation, the concentration coefficients are defined as follows: let ν and μ be two probability distributions on X.

We assume that for all $x \in X$, $a \in A$, $P(x,a,\cdot)$ is absolutely continuous w.r.t. μ and the Radon–Nikodým derivative of $P(x,a,\cdot)$ w.r.t. $\mu(\cdot)$ is bounded uniformly in x and a. Then the transition probabilities concentration coefficient $C(\mu)$ is defined by

$$C(\mu) := \sup_{x \in X, a \in A} \frac{dP(x,a,\cdot)}{d\mu}.$$

Notice that if μ is the Lebesgue measure over X, and if $P(x,a,\cdot)$ admits a uniformly bounded density, then the concentration coefficient $C(\mu)$ is equal to the upper bound of this density. This case is illustrated in the numerical experiment below. The first and second order discounted future state distribution concentration coefficients $C_1(\nu,\mu)$ and $C_2(\nu,\mu)$ are defined similarly to (5.2) and (5.3).

7.2. An optimal replacement problem. This experiment illustrates the respective tightness of the L_∞-, L_1-, and L_2-norm bounds on a continuous space control problem excerpted from [31].

A one-dimensional continuous variable $x_t \in [0, x_{\max}]$ measures the accumulated utilization (such as the odometer reading on a car) of a product. $x_t = 0$ denotes a brand-new product. At each discrete time t, there are two possible decisions: either keep ($a_t = \mathrm{K}$) or replace ($a_t = \mathrm{R}$), in which case an additional cost $C_{replace}$ (of selling the existing product and replacing it for a new one) occurs. The transition densities are exponential with parameter β with a truncated queue. Moreover, if the next state y is larger than the maximal value $x_{\max}$ (e.g., the car breaks down because it is too damaged), then a new state is immediately redrawn and a penalty $C_{dead} > C_{replace}$ occurs. The transition densities are thus defined as follows: defining

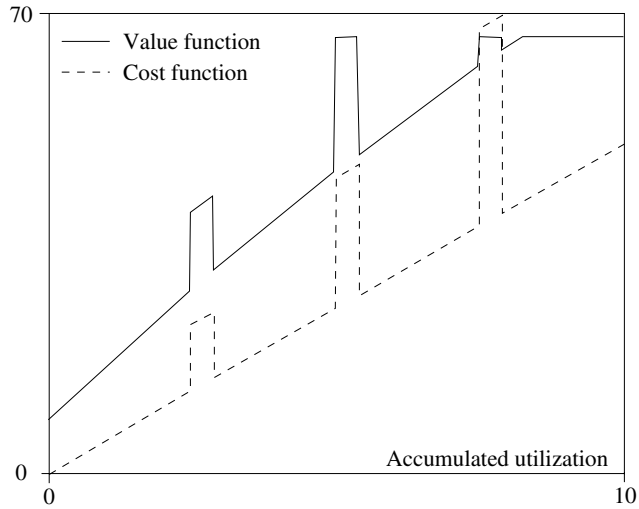


FIG. 7.1. *Cost and value functions.*

$q(x) := \beta e^{-\beta x}/(1 - e^{-\beta x_{\max}})$,

$$p(x, a = \mathrm{R}, y) = \begin{cases} q(y) & \text{if } y \in [0, x_{\max}], \\ 0 & \text{otherwise.} \end{cases}$$

$$p(x, a = \mathrm{K}, y) = \begin{cases} q(y - x) & \text{if } y \in [x, x_{\max}], \\ q(y - x + x_{\max}) & \text{if } y \in [0, x), \\ 0 & \text{otherwise.} \end{cases}$$

The current cost (opposite of a reward) $c(x)$ is the sum of a slowly increasing function (maintenance cost) and a discontinuous punctual cost (e.g., which may represent car insurance fees).

The current cost function and the optimal value function (computed by a discretization on a high resolution grid) are shown in Figure 7.1.

We choose the numerical values $\gamma = 0.6$, $\beta = 0.6$, $C_{replace} = 50$, $C_{dead} = 70$, and $x_{\max} = 10$. We consider a uniform distribution μ on the domain $[0, x_{\max}]$. We choose K points (with $K = 200$ or 2000 points) uniformly located over the domain $\{x_k := kx_{\max}/K\}_{0 \le k < K}$ to perform the L_2 minimization fitting problem at each iteration:

$$V_{n+1} = \arg\min_{f \in \mathcal{F}} \frac{1}{K} \sum_{k=1}^{K} [f(x_k) - \mathcal{T}V_n(x_k)]^2,$$

where $\mathcal{F}$ is the space spanned by a truncated cosine basis (with $M = 20$ or $M = 40$ basis functions):

$$\mathcal{F} := \left\{ f(x) = \sum_{m=1}^{M} \alpha_m \cos\left(m\pi \frac{x}{x_{\max}} \right) \right\}_{\alpha \in \mathbb{R}^M}.$$

We start with initial values $V_0 = 0$. In Figure 7.2 we show the first iteration (for the grid with $K = 200$ points): the backed-up values $\mathcal{T}V_0$ (indicated with crosses), the

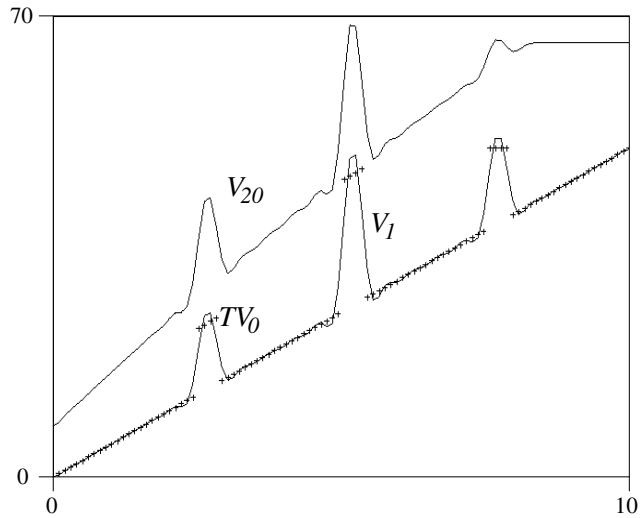


FIG. 7.2. *$\mathcal{T}V_0$ (crosses), V_1, and V_{20}.*

TABLE 7.1
Comparison of the r.h.s. of the L_∞, L_1, and L_2 bounds.

	$\|\|\varepsilon_n\|\|_\infty$	$C(\mu)\|\|\varepsilon_n\|\|_1$	$\sqrt{C(\mu)}\|\|\varepsilon_n\|\|_2$
$K = 200$, $M = 20$	12.4	0.367	1.16
$N = 2000$, $M = 40$	12.4	0.0552	0.897

corresponding approximation V_1 (best fit of $\mathcal{T}V_0$ in the cosine approximation space $\mathcal{F}$). The approximate value function computed after 20 iterations (when there are no significant improvement of the approximations) is also plotted.

The concentration coefficient $C(\mu)$ is the highest peak of the transition density w.r.t. the uniform distribution μ; thus $C(\mu) = q(0)x_{\max} = \beta x_{\max}/(1 - e^{-\beta x_{\max}}) \simeq 6$.

Table 7.1 compares the r.h.s. (up to the constant $2\gamma/(1-\gamma)^2$) of equations (2.2) and (5.4) for $p = 1$ and 2, their l.h.s. being the same since they use the same L_∞-norm. We notice that the L_1 and L_2 bounds (5.4) are much tighter than the L_∞ one (2.2). Moreover we observe that the L_1 and L_2 approximation errors tend to 0 when the number K of sampling points and the number M of basis functions go to infinity, whereas the L_∞ bound does not. Indeed, since the cost function is discontinuous, the L_∞ approximation error (using continuous function approximation such as the cosine basis used here) will never be smaller than half the value of the largest jump, even for large values of K and M. This example illustrates the fact that the L_p bound (5.4) may be arbitrarily tighter than the L_∞ one (2.2).

8. Conclusion. Theorem 5.2 provides a useful tool to bound the performance of AVI from the L_p-norm of the approximation errors, and thus in terms of the approximation power of most SL algorithms. Expressing the performance of AVI in the same norm as the norm used by the supervised learner to solve the regression problem guarantees the tightness and practical application of the bounds.

In order for these bounds to be of any use, we need to estimate an upper bound on the concentration coefficients $C(\mu)$, $C_1(\nu, \mu)$, and $C_2(\nu, \mu)$, which may be difficult

in general. We illustrate the case of low values of $C_1(\nu, \mu)$, and $C_2(\nu, \mu)$ in the chain walk MDP, and the case of a low value of $C(\mu)$ in the optimal replacement problem. Future work would consider defining classes of problems for which these coefficients may be evaluated.

Extension to other loss functions l, such as ϵ-*insensitive* (used in support vectors) or *Huber loss function* (for robust regression) [36], is straightforward (as long as l is an increasing and convex function over $\mathbb{R}^+$). Another possible extension is AVI for Markov games.

Acknowledgments. The author wishes to thank Csaba Szepesvári and the anonymous reviewers who helped significantly improve the clarity of the concepts introduced in the paper.

REFERENCES

[1] A. Antos, Cs. Szepesvari, and R. Munos, *Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path*, in Proceedings of the Conference on Learning Theory, Springer-Verlag, New York, 2006, pp. 574–588.

[2] C. G. Atkeson, A. W. Moore, and S. A. Schaal, *Locally weighted learning*, AI Rev., 11 (1997), pp. 11–73.

[3] C. G. Atkeson, A. W. Moore, and S. A. Schaal, *Locally weighted learning for control*, AI Rev., 11 (1997), pp. 75–113.

[4] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[5] R. E. Bellman and S. E. Dreyfus, *Functional approximation and dynamic programming*, Math. Tables Aids Comput., 13 (1959), pp. 247–251.

[6] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[7] D. P. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Nashua, NH, 1996.

[8] P. Bougerol and N. Picard, *Strict stationarity of generalized autoregressive processes*, Ann. Probab., 20 (1992), pp. 1714–1730.

[9] G. M. Davies, S. Mallat, and M. Avellaneda, *Adaptive greedy approximations*, J. Constr. Approx., 13 (1997), pp. 57–98.

[10] D. P. de Farias and B. Van Roy, *The linear programming approach to approximate dynamic programming*, Oper. Res., 51 (2003), pp. 850–865.

[11] R. DeVore, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150.

[12] G. Gordon, *Stable function approximation in dynamic programming*, in Proceedings of the International Conference on Machine Learning, Morgan Kaufmann, 1995, pp. 261–268.

[13] G. J. Gordon, *Approximate Solutions to Markov Decision Processes*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, 1999.

[14] C. Guestrin, D. Koller, and R. Parr, *Max-norm projections for factored MDPs*, in Proceedings of the International Joint Conference on Artificial Intelligence, Lawrence Erlbaum, 2001, pp. 673–682.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Ser. Statist., Springer-Verlag, New York, 2001.

[16] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes, Basic Optimality Criteria*, Springer-Verlag, New York, 1996.

[17] O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.

[18] O. Hernández-Lerma, R. Montes-De-Oca, and R. Cavazos-Caneda, *Recurrence conditions for Markov decision processes with Borel state space: A survey*, Ann. Oper. Res., 28 (1991), pp. 29–46.

[19] A. Hordijk and F. Spieksma, *On ergodicity and recurrence properties of a Markov chain with an application to an open Jackson network*, Adv. Appl. Probab., 24 (1992), pp. 343–376.

[20] S. M. Kakade, *On the Sample Complexity of Reinforcement Learning*, Ph.D. thesis, University College London, 2003.

[21] S. Kakade and J. Langford, *Approximately optimal approximate reinforcement learning*, in Proceedings of the 19th International Conference on Machine Learning, Morgan Kaufmann, 2002, pp. 267–274.

[22] D. Koller and R. Parr, *Policy iteration for factored MDPs*, in Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, 2000, pp. 326–334.
[23] M. Lagoudakis and R. Parr, *Least-squares policy iteration*, J. Mach. Learn. Res., 4 (2003), pp. 1107–1149.
[24] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA, 1997.
[25] S. P. Meyn, *Stability, performance evaluation, and optimization*, in Handbook of Markov Decision Processes: Methods and Applications, Kluwer Academic, Boston, MA, 2002, pp. 305–346.
[26] R. Munos, *Error bounds for approximate policy iteration*, in Proceedings of the 19th International Conference on Machine Learning, AAAI Press, 2003, pp. 560–567.
[27] R. Munos and Cs. Szepesvári, *Finite-Time Bounds for Sampling-Based Fitted Value Iteration*, Technical report, INRIA, 2006; available online from http://hal.inria.fr/inria-00120882.
[28] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, New York, 1984.
[29] M. L. Puterman, *Markov Decision Processes, Discrete Stochastic Dynamic Programming*, Wiley-Interscience, New York, 1994.
[30] E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, 2005.
[31] J. Rust, *Numerical dynamic programming in economics*, in Handbook of Computational Economics, Elsevier/North–Holland, Amsterdam, 1996, pp. 619–729.
[32] A. L. Samuel, *Some studies in machine learning using the game of checkers*, IBM J. Res. Develop., 3 (1959), pp. 210–229; reprinted in Computers and Thought, E. A. Feigenbaum and J. Feldman, eds., McGraw-Hill, New York, 1963.
[33] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Bradford Book, MIT Press, Cambridge, MA, 1998.
[34] Cs. Szepesvari and R. Munos, *Finite time bounds for sampling based fitted value iteration*, in Proceedings of the International Conference on Machine Learning, ACM, New York, 2005, pp. 881–886.
[35] J. N. Tsitsiklis and B. Van Roy, *An analysis of temporal difference learning with function approximation*, IEEE Trans. Automat. Control, 42 (1997), pp. 674–690.
[36] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.
[37] V. Vapnik, S. E. Golowich, and A. Smola, *Support vector method for function approximation, regression estimation and signal processing*, in Advances in Neural Information Processing Systems, 1997, pp. 281–287.
[38] R. J. Williams and L. C. Baird, *Tight Performance Bounds on Greedy Policies Based on Imperfect Value Functions*, Technical report NU-CCS-93-14, Northeastern University, Boston, MA, 1993.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 562–584

SIMULATION AND BISIMULATION OF NONLINEAR CONTROL SYSTEMS WITH ADMISSIBLE CLASSES OF INPUTS AND DISTURBANCES*

KEVIN A. GRASSE†

Abstract. Motivated by analogous concepts in theoretical computer science, the concepts of simulation and bisimulation relations have recently been introduced in the study of the geometric theory of nonlinear control systems. Simulation relations generalize such notions as trajectory propagation and trajectory lifting, while bisimulation relations generalize such notions as system equivalence and state-space reduction. This paper explores previously obtained necessary and sufficient conditions for the existence of (bi)simulation relations and determines to what extent those relations can be maintained if inputs and disturbances are restricted to specified "admissible" classes. We show how the results obtained here for general simulation relations extend prior results obtained for trajectory propagation and trajectory lifting. Also addressed briefly are some sufficient conditions for simulation relations to hold semiglobally in time.

Key words. nonlinear control systems, simulation, bisimulation, related systems, trajectory propagation, trajectory lifting, controlled invariance

AMS subject classifications. Primary, 93B11; Secondary, 93B29; 93C10, 93A13, 93B17

DOI. 10.1137/050638072

1. Introduction.

An effective and recurring approach to the problem of classifying nonlinear control systems entails defining a relation (or equivalence relation) between systems that affords the possibility of "complex" systems being related to "simpler" systems. The usage of the terms "complex" and "simpler" is intentionally vague here, but one can think of a system as being "complex" if, for example, its state space has large dimension or if it is highly nonlinear. "Simpler" systems, on the other hand, may evolve on a state space of reduced dimension or be either linear or mildly nonlinear. Through the use of an appropriately defined relation between systems one may hope to infer properties and behaviors of the complex system from those of the simpler system. Motivated by an analogous concept in theoretical computer science, Haghverdi, Tabuada, and Pappas [7] initiated the study of *bisimulation relations* in continuous-time dynamical and control systems. Roughly speaking, a bisimulation relation between two control systems is first and foremost a relation (i.e., a subset of the Cartesian product of their state spaces) but has the additional property that trajectories of the first system can be paired, by way of the relation, with trajectories of the second system and, conversely, trajectories of the second system can be paired, by way of the relation, with trajectories of the first system. (We will present one version of a precise definition of a bisimulation relation in section 2, based on that given in [18], but readers should be aware that there are some variations in the definitions of this notion in the references cited in this paper.) In connection with the aforementioned work [7], Pappas [11] derived detailed results characterizing bisimulation relations induced by linear surjections for discrete and continuous-time linear control systems, while Tabuada and Pappas [14] derived a characterization of bisimulation relations induced by nonlinear submersions for nonlinear continuous-time systems

*Received by the editors August 12, 2005; accepted for publication (in revised form) November 4, 2006; published electronically May 4, 2007.
http://www.siam.org/journals/sicon/46-2/63807.html
†Department of Mathematics, University of Oklahoma, Norman, OK 73019 (kgrasse@ou.edu).

that are affine in the control. Further progress in bisimulation relations as they apply to continuous-time control systems was made by van der Schaft in [17, 18], where, among other things, he derives an elegant and constructive algorithm for computing the maximal bisimulation relation between two systems.

Bisimulation relations are natural objects in control systems theory in that they nicely generalize such well-studied notions as state-space equivalence (see, e.g., [2]) and state-space reduction (see, e.g., [1]). As pointed out by various authors [5, 11, 18], these relations also have close ties to the fundamental concept of *controlled invariance* in geometric control theory [3, 8, 9]. Bisimulation relations are, by their very nature, "two-way" relations, but one can create an obvious definition of a analogous "one-way" counterpart, which is referred to as a *simulation relation* (see [18, Def. 5.1] or section 2 for the definition). Simulation relations also occur frequently in control systems theory in a variety of contexts. For example, if Φ is a mapping between the state spaces of two systems that propagates trajectories of the "domain system" to the "range system," then the graph of Φ is an example of a simulation relation; in such a case the systems are sometimes called Φ-*related* and the range system is called an *abstraction* of the domain system [12, 13, 15, 4]. Alternatively, the graph of Φ is also a simulation relation (however, in the opposite direction) if we can lift trajectories from the range system to the domain system [5, 16].

This paper refines the aforementioned work on bisimulation relations by paying somewhat closer attention to regularity properties of the controls (or, as we will see shortly, of the input-disturbance pairs) that generate the system's trajectories. Our motivation for this work rests with certain mathematical foundational issues in nonlinear control theory. When one formulates precise mathematical definitions of control systems and the related objects of controls and trajectories (see, for example, the meticulous treatment in the book of Sontag [19]), some care must be exercised to ensure that when a control and initial condition are specified the resulting trajectories exist, are unique, and depend continuously on the data. Put differently, we want our dynamical model to be *well posed*. To this end, limitations are placed on the class of controls that are deemed "admissible," so that the resulting nonautonomous ordinary differential equations meet the usual theoretical criteria for the existence, uniqueness, and continuous dependence of solutions (the so-called C^1 *Carathéodory conditions*). However, at the same time one wants the class of controls to be as large as possible to facilitate proving positive results about controllability, existence of optimal controls, etc. In particular, existence proofs for optimal controls usually require that controls be members of a specified class of (Lebesgue) measurable functions of time (say, an L^2 space). When considerations of (bi)simulation relations are juxtaposed with regularity issues of the underlying controls, one is consequently confronted with the question of the extent to which such relations will persist if one restricts controls to specified admissible classes. It turns out that for linear, time-invariant systems, matters take care of themselves rather nicely (see Remark 3.11). However, one does not have to delve very deeply into classes of nonlinear systems to see that some care is required to be assured that (bi)simulation relations are indeed maintained when controls are restricted to (even rather large) "admissible" classes (see Example 2.7). Previously, the author considered related control-admissibility issues for trajectory propagation [4] and trajectory lifting [5], and this work should be regarded as an extension of these papers. We also acknowledge the substantial influence of the paper of van der Schaft [18] in motivating the results contained herein.

The remainder of this paper is organized as follows. In section 2 we recall the definitions of the basic objects at hand (control systems and controls) and we formulate

the basic definitions of simulation and bisimulation relations with admissible classes of controls. Our definitions closely follow those given by van der Schaft [17, 18], and in particular we adopt his approach of decomposing controls into inputs and disturbances. However, we have phrased our definitions to be inherently *local* in time because of the phenomenon of finite escape time of trajectories of nonlinear systems. Section 3 contains the main results of the paper on necessary and sufficient conditions for (bi)simulation relations with admissible classes of inputs and disturbances. In section 4 the results of section 3 are shown to subsume previously obtained results on trajectory propagation (i.e., Φ-related systems) and trajectory lifting. We also consider two simple cases in which the "local-in-time" nature of (bi)simulation relations can be guaranteed to be global in time; for example, such behavior would certainly be expected of linear systems as we will show.

2. Preliminaries. By the term *differentiable manifold* we will always mean a connected, finite-dimensional, second-countable, Hausdorff, differentiable manifold of class C^k with $k \geq 2$. Given a differentiable manifold M, we use TM to denote the tangent bundle of M, $\pi_M : TM \to M$ its canonical projection onto M, and we recall that TM is a differentiable manifold of class C^{k-1}. If $\Phi : M \to N$ is a C^1 mapping of manifolds, then $d\Phi : TM \to TN$ will denote its differential, while for each $x \in M$ we use $d\Phi_x : T_xM \to T_{\Phi(x)}N$ to denote the corresponding linear mapping on the indicated tangent spaces (that is, fibers of the tangent bundles).

Given C^k differentiable manifolds M and O, and given a separable metric space Λ, we will say that a mapping $\Phi : M \times \Lambda \to O$ is *nicely C^k on M relative to Λ* if for each $\lambda \in \Lambda$ the mapping $x \mapsto \Phi(x, \lambda)$ is C^k and if the partial derivatives of Φ with respect to x up to order k are continuous on $M \times \Lambda$. In particular, a C^1 *control system with state space M and control space Λ* is a mapping $F : M \times \Lambda \to TM$ that is nicely C^1 on M relative to Λ and satisfies $(\pi_M \circ F)(x, \lambda) = x$ for every $(x, \lambda) \in M \times \Lambda$ (for more details, see [4, Def. 2.1] or [6, Def. 2.2]). Control systems will sometimes be informally specified by the notation $\dot{x} = F(x, v(t))$. *Potential controls* for such a system are members of the family $\mathcal{U}^{\Lambda}_{\text{meas}}$ of all Lebesgue-measurable mappings of $\mathbb{R}$ into Λ; that is, $v \in \mathcal{U}^{\Lambda}_{\text{meas}}$ if and only if for every open subset W of Λ the preimage $v^{-1}(W)$ is a Lebesgue-measurable subset of $\mathbb{R}$. Useful subclasses of $\mathcal{U}^{\Lambda}_{\text{meas}}$ are the family $\mathcal{U}^{\Lambda}_{\text{cpt}}$ of all Lebesgue-measurable mappings of $\mathbb{R}$ into Λ that are *essentially compact valued on compact intervals* (see [6, Ex. 2.10(2)]), the family $\mathcal{U}^{\Lambda}_{\text{simp}}$ of all Lebesgue-measurable mappings of $\mathbb{R}$ into Λ that are *simple* (that is, have finite range), and the family $\mathcal{U}^{\Lambda}_{\text{cont}}$ of all continuous mappings of $\mathbb{R}$ into Λ. Obviously, we have $\mathcal{U}^{\Lambda}_{\text{simp}} \subseteq \mathcal{U}^{\Lambda}_{\text{cpt}}$ and $\mathcal{U}^{\Lambda}_{\text{cont}} \subseteq \mathcal{U}^{\Lambda}_{\text{cpt}}$. In the case where Λ is the finite-dimensional Euclidean space $\mathbb{R}^p$, it will also be handy to designate by $L^1_{\text{loc}}(\mathbb{R}, \mathbb{R}^p) \subseteq \mathcal{U}^{\mathbb{R}^p}_{\text{meas}}$ the family of all Lebesgue-measurable mappings of $\mathbb{R}$ into $\mathbb{R}^p$ that are integrable on compact subintervals of $\mathbb{R}$. A potential control $v : \mathbb{R} \to \Lambda$ is called *admissible* for a C^1 control system $F : M \times \Lambda \to TM$ if the mapping $F_v : M \times \mathbb{R} \to TM$ defined by $F_v(x, t) = F(x, v(t))$ is such that its local representation with respect to every coordinate chart of M satisfies C^1 Carathéodory conditions (see [4, Def. 2.6] or [6, Def. 2.8]). We let $\mathcal{U}^{\Lambda}_{\text{meas}}(F)$ denote the subset of $\mathcal{U}^{\Lambda}_{\text{meas}}$ consisting of all admissible controls for the C^1 control system F. In general the family of admissible controls $\mathcal{U}^{\Lambda}_{\text{meas}}(F)$ is strongly dependent on the control system F, but one always has

$$\mathcal{U}^{\Lambda}_{\text{simp}} \subseteq \mathcal{U}^{\Lambda}_{\text{meas}}(F),\ \mathcal{U}^{\Lambda}_{\text{cpt}} \subseteq \mathcal{U}^{\Lambda}_{\text{meas}}(F),\ \mathcal{U}^{\Lambda}_{\text{cont}} \subseteq \mathcal{U}^{\Lambda}_{\text{meas}}(F)$$

for an arbitrary C^1 control system F (see [6, Ex. 2.10(2)]). Standard results from the theory of ordinary differential equations guarantee that for every $x_0 \in M$ and for

every $v \in \mathcal{U}^{\Lambda}_{\text{meas}}(F)$ there exist an open interval $J \subseteq \mathbb{R}$ containing 0 and a unique mapping $\psi : J \to M$ such that $\psi(0) = x_0$, ψ is absolutely continuous on every compact subinterval of J, and

$$\dot{\psi}(t) = F(\psi(t), v(t)) \quad \text{for almost every} \quad t \in J$$

(by "almost every" we mean except on a set of Lebesgue measure zero; henceforth we will use the standard abbreviation a.e.). We call $t \mapsto \psi(t)$ a *trajectory* of F corresponding to the initial condition x_0 and the control v. Often we will use the notation

$$\psi(t) \stackrel{\text{def}}{=} \psi(t, x_0, v)$$

to make explicit the dependence of the trajectory on the initial condition and control, in which case we will refer to the function ψ of time, the initial state, and the control as the *trajectory mapping* of F.

The formal definitions of simulation and bisimulation relations will be cast in the somewhat more general framework of control systems with *inputs*, *disturbances*, and *outputs*. Roughly speaking, we view the control as comprising all external factors that can affect the system, and then partition the control into an input (or *deterministic*) component that we may think of as being within our domain of influence, and a disturbance (or *nondeterministic*) component that we may think of as being outside of our domain of influence. As noted by Pappas and his coauthors (see [7, 11, 14]), the initial introduction of simulation and bisimulation relations in control systems theory was motivated by analogous concepts in theoretical computer science which apply to nondeterministic automata. The presence of the disturbance component in the control systems formulation of these relations allows the evolutionary behavior of the control systems under consideration to mirror the nondeterministic aspect of the transitions allowed in nondeterministic automata. We will comment further on interpretations of the disturbance component in Remark 2.4.

DEFINITION 2.1.

(i) *A* C^1 input-disturbance *(ID) system is a* C^1 *control system* $F : M \times \Lambda \to TM$ *whose control space* Λ *is a Cartesian product* $\Lambda = \Omega \times \Delta$. *We refer to* Ω *as the input space and* Δ *as the disturbance space, and we note that each of* Ω *and* Δ *inherits the structure of a separable metric space from* Λ.

(ii) *A* C^1 input-disturbance-output *(IDO) system is a pair* (F, h), *where* $F : M \times \Omega \times \Delta \to TM$ *is a* C^1 *ID system and* $h : M \times \Omega \to O$ *is a continuous mapping of* $M \times \Omega$ *into a topological space* O; *we call* h *the* output mapping *and* O *the* output space.

(iii) *If* $\mathcal{U} \subseteq \mathcal{U}^{\Omega}_{\text{meas}}$ *and* $\mathcal{D} \subseteq \mathcal{U}^{\Delta}_{\text{meas}}$ *are chosen to satisfy* $\mathcal{U} \times \mathcal{D} \subseteq \mathcal{U}^{\Omega\times\Delta}_{\text{meas}}(F)$, *then we refer to the four-tuple* $(F, h, \mathcal{U}, \mathcal{D})$ *as an* IDO system with admissible inputs $\mathcal{U}$ and admissible disturbances $\mathcal{D}$.

A C^1 IDO system $(F, h, \mathcal{U}, \mathcal{D})$ can also be designated by the more informal notation

$$\begin{aligned} \dot{x} &= F(x, u(t), d(t)), \quad u \in \mathcal{U}, \quad d \in \mathcal{D}, \\ y &= h(x, u(t)), \end{aligned}$$

where $u(\cdot)$ is the input function and $d(\cdot)$ is the disturbance function.

The following definitions of simulation and bisimulation relations between two IDO systems are patterned after the definition of bisimulation given by van der Schaft

in [18, Def. 2.1], but we will be a bit more explicit about the assumed regularity properties of the inputs and disturbances.

DEFINITION 2.2. *Let M and N be differentiable manifolds, let O be a topological space, let Ω, Δ, and E be separable metric spaces, and suppose that we are given a pair of C^1 IDO systems with admissible inputs and admissible disturbances*

$$F : M \times \Omega \times \Delta \to TM, \quad h : M \times \Omega \to O, \quad u \in \mathcal{U}, \quad d \in \mathcal{D}$$

and

$$\widetilde{F} : N \times \Omega \times E \to TN, \quad \tilde{h} : N \times \Omega \to O, \quad u \in \mathcal{U}, \quad e \in \mathcal{E},$$

which have the common input space Ω, common output space O, and common family of admissible inputs $\mathcal{U}$, where

$$\mathcal{U} \subseteq \mathcal{U}_{\text{meas}}^{\Omega}, \quad \mathcal{D} \subseteq \mathcal{U}_{\text{meas}}^{\Delta}, \quad \mathcal{E} \subseteq \mathcal{U}_{\text{meas}}^{E}$$

are such that

$$\mathcal{U} \times \mathcal{D} \subseteq \mathcal{U}_{\text{meas}}^{\Omega \times \Delta}(F), \quad \mathcal{U} \times \mathcal{E} \subseteq \mathcal{U}_{\text{meas}}^{\Omega \times E}(\widetilde{F}).$$

A nonempty subset $\mathcal{R} \subseteq M \times N$ is called a simulation relation of $(F, h, \mathcal{U}, \mathcal{D})$ by $(\widetilde{F}, \tilde{h}, \mathcal{U}, \mathcal{E})$ *if for every $(x_0, z_0) \in \mathcal{R}$, for every $u \in \mathcal{U}$, and for every $d \in \mathcal{D}$ there exist $e \in \mathcal{E}$ and a compact interval I containing 0 in its interior such that for every $t \in I$ both $\psi(t, x_0, u, d)$ and $\tilde{\psi}(t, z_0, u, e)$ are defined (here ψ and $\tilde{\psi}$ are the trajectory mappings of F and $\widetilde{F}$, respectively), and we have*

$$t \in I \;\Rightarrow\; (\psi(t, x_0, u, d), \tilde{\psi}(t, z_0, u, e)) \in \mathcal{R} \tag{2.1}$$

and

$$t \in I \;\Rightarrow\; h(\psi(t, x_0, u, d), u(t)) = \tilde{h}(\tilde{\psi}(t, z_0, u, e), u(t)). \tag{2.2}$$

Remark 2.3. Relations (2.1) and (2.2) are required to hold only on a compact interval containing 0 in its interior because of the possibility of finite escape time for the trajectories of the nonlinear systems under consideration. Naturally, in cases where the systems are *complete* one would like to strengthen these relations to hold for *all* $t \in \mathbb{R}$, a point we will address later (see section 4). We further note that if one chooses to identify inputs and disturbances that agree almost everywhere, then relation (2.2) should be required to hold only for a.e. $t \in I$, but we will not make such identifications in this paper.

Remark 2.4. We offer a few comments on Definition 2.2 as it relates to the presence of the disturbance component.

(i) The concept of simulation relation as defined here (which again we attribute to van der Schaft [18]) guarantees only that the simulating system $(\widetilde{F}, \tilde{h}, \mathcal{U}, \mathcal{E})$ has the capability of mimicking the *class* of all input-output "behaviors" of the simulated system $(F, h, \mathcal{U}, \mathcal{D})$ in the manner dictated by the simulation relation $\mathcal{R}$. However, it does not guarantee that the simulating system can be made to mimic specific individual behaviors of the simulated system because we have no assurance that the disturbance affecting the simulating system is the one whose existence is assured by the definition. Nevertheless, there are useful control problems that can be formulated

with such a nondeterministic interpretation of the disturbances. For example, if one can design a state feedback controller for the simulating system that drives the output to zero (say, as $t \to \infty$) for all disturbances in a specified class, then one can seek conditions under which such a controller will exist for the simulated system.

(ii) Other interpretations of the disturbance component are also possible and of potential utility in the control-systems formulation of simulation relations. For example, it may happen that the disturbance affecting the simulated system is nondeterministic, but is also measurable (the wind velocity affecting the glide path of an aircraft may not be predictable in advance of its occurrence, but it can be measured as it occurs). To force the simulating system to mimic the simulated system, one could try to synthesize a "disturbance" that generates a desired trajectory of the simulating system from the states and common input of both systems, as well as the measured values of the disturbance affecting the simulated system. In this context, the term "disturbance" as applied to the simulating system has a connotation that differs somewhat from a purely random effect, and a certain amount of care must be exercised.

DEFINITION 2.5. *Let* $(F,h,\mathcal{U},\mathcal{D})$ *and* $(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$ *be as Definition* 2.2. *A nonempty subset* $\mathcal{R} \subseteq M \times N$ *is called a* bisimulation relation between $(F,h,\mathcal{U},\mathcal{D})$ and $(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$ *if* $\mathcal{R}$ *is both a simulation relation of* $(F,h,\mathcal{U},\mathcal{D})$ *by* $(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$ *and a simulation relation of* $(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$ *by* $(F,h,\mathcal{U},\mathcal{D})$.

Reference [18] contains a statement of the following proposition, which provides a necessary and sufficient condition for a submanifold of $M \times N$ to be a bisimulation relation between two IDO systems (see, specifically, [18, Prop. 7.1 and Rem. 7.4]; we have altered the notation of [18] to be consistent with the notation used here).

PROPOSITION 2.6. *Let* $\mathcal{R}$ *be a* C^2 *submanifold of* $M \times N$. *Then* $\mathcal{R}$ *is a bisimulation relation between two IDO systems* $(F,h,\mathcal{U},\mathcal{D})$ *and* $(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$ *if and only if for every* $(x,z) \in M \times N$ *and* $\omega \in \Omega$ *the following conditions are met:*

(i) *For every* $\delta \in \Delta$ *there exists* $\varepsilon \in E$ *such that*

$$\big(F(x,\omega,\delta), \widetilde{F}(z,\omega,\varepsilon)\big) \in T_{(x,z)}\mathcal{R}, \tag{2.3}$$

and conversely, for every $\varepsilon \in E$ *there exists* $\delta \in \Delta$ *such that* (2.3) *holds;*

(ii)

$$h(x,\omega) = \tilde{h}(z,\omega). \tag{2.4}$$

Regularity issues of inputs and disturbances are deliberately understated in [18] because most of the presentation therein focuses on linear, time-invariant systems, where such regularity issues are not of critical significance (see Remark 3.11). However, the next example shows that regularity of inputs and disturbances can be a significant factor in nonlinear IDO systems.

Example 2.7. Consider the IDO system with $M = \Delta = O = \mathbb{R}$ given by

$$F(x,\delta) = 1 + x^2\delta, \quad h(x) = x,$$

and consider the IDO system with $N = \mathbb{R}^2$ and $E = O = \mathbb{R}$ given by

$$\widetilde{F}(z_1,z_2,\varepsilon) = (1 + z_1^3\varepsilon, \varepsilon), \quad \tilde{h}(z_1,z_2) = z_1$$

(observe that in both cases the input space Ω is absent). Because these systems are affine in their disturbances, we will use in each case the largest reasonable class of disturbances

$$\mathcal{D} = \mathcal{E} = L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}).$$

If we set

$$\mathcal{R} = \{(x, z_1, z_2) \in \mathbb{R}^3 \mid x = z_1\},$$

then $\mathcal{R}$ is a closed vector subspace (and, in particular, a closed submanifold) of $M \times N = \mathbb{R}^3$, and it is easy to check that conditions (2.3) and (2.4) are satisfied. However, we claim that $\mathcal{R}$ cannot be a bisimulation relation as defined in Definition 2.5. To see this, select $(x_0, z_{10}, z_{20}) = (0, 0, 0) \in \mathcal{R}$ and $d \in \mathcal{D},\ d(t) \equiv 1$. It is easily seen that the corresponding state trajectory of F is

$$t \mapsto \psi(t, 0, d) = \tan t \quad (|t| < \pi/2).$$

Suppose that $e \in \mathcal{E}$ and I is a compact interval containing 0 such that relation (2.1) holds. Then the corresponding state trajectory

$$t \mapsto \tilde{\psi}(t, (0,0), e) = \big(\tilde{\psi}_1(t, (0,0), e), \tilde{\psi}_2(t, (0,0), e)\big)$$

of $\widetilde{F}$ must satisfy for a.e. $t \in I$

$$\begin{aligned} & \tan t = \tilde{\psi}_1(t, (0,0), e) \\ \Rightarrow \quad & \frac{d}{dt} \tan t = \frac{d}{dt} \tilde{\psi}_1(t, (0,0), e) = 1 + \tilde{\psi}_1(t, (0,0), e)^3 e(t) \\ \Rightarrow \quad & \sec^2 t = 1 + \tan^2 t = 1 + (\tan^3 t) e(t), \end{aligned}$$

so it must be the case that $e(t) = 1/\tan t$ for a.e. t. It follows that $e(\cdot)$ is not integrable on I. Furthermore, even if we attempted to press ahead with this choice of e, the second component of the trajectory would have to satisfy

$$\frac{d}{dt} \tilde{\psi}_2(t, (0,0), e) = \frac{1}{\tan t},$$

which precludes $t \mapsto \tilde{\psi}_2(t, (0,0), e)$ being absolutely continuous on any compact subinterval of $\mathbb{R}$ containing 0. Roughly speaking, the source of the difficulty is that the disturbance vector fields

$$x \mapsto x^2, \quad (z_1, z_2) \mapsto \begin{bmatrix} z_1^3 \\ 1 \end{bmatrix}$$

are "rank deficient" modulo the tangent space to the purported bisimulation relation $\mathcal{R}$ at the origin (see condition (i) of Proposition 3.6 for a precise statement of the rank condition that is violated here).

3. Main results. Our main result deals with nonlinear IDO systems that are affine in the disturbance, a concept that we define next.

DEFINITION 3.1. *A C^1 IDO system $F : M \times \Omega \times \Delta \to TM$ is said to be* affine in the disturbance *if the disturbance space Δ is a finite-dimensional Euclidean space $\mathbb{R}^p$ and F has the form*

$$F(x, \omega, \delta) = f(x, \omega) + \sum_{i=1}^{p} \delta_i g_i(x) \tag{3.1}$$

for $(x, \omega, \delta) \in M \times \Omega \times \mathbb{R}^p$, where $f : M \times \Omega \to TM$ is a C^1 control system, $g_1, \ldots, g_p$ are C^1 vector fields on M, and $\delta = (\delta_1, \ldots, \delta_p) \in \mathbb{R}^p$.

We will often write (3.1) in the abbreviated form

$$F(x,\omega,\delta) = f(x,\omega) + G(x)\delta, \tag{3.2}$$

where $G(x) = [g_1(x), \ldots, g_p(x)]$.

Remark 3.2. For the disturbance affine ID system F having the form (3.1) or (3.2), it is easy to check that

$$\mathcal{U}^{\Omega}_{\text{meas}}(f) \times L^1_{\text{loc}}(\mathbb{R}, \mathbb{R}^p) \subseteq \mathcal{U}^{\Omega\times\mathbb{R}^p}_{\text{meas}}(F).$$

Notation 3.3. Given finite-dimensional (real) vector spaces $\mathcal{V}$ and $\mathcal{W}$, we will let $\mathcal{L}(\mathcal{V},\mathcal{W})$ denote the family of all linear mappings from $\mathcal{V}$ into $\mathcal{W}$. Given differentiable manifolds M, N and points $x \in M$, $z \in N$, we will make the canonical identification

$$T_{(x,z)}(M \times N) \cong T_xM \times T_zN,$$

and often write tangent vectors $v \in T_{(x,z)}(M \times N)$ in the stacked form

$$v = \begin{bmatrix} \bar{v} \\ \tilde{v} \end{bmatrix}, \quad \text{where } \bar{v} \in T_xM \text{ and } \tilde{v} \in T_zN.$$

Given C^1 vector fields $\tilde{g}_1, \ldots, \tilde{g}_q$ on N and $z \in N$, we set $\widetilde{G}(z) = [\tilde{g}_1(z), \ldots, \tilde{g}_q(z)]$ and use the notation

$$\operatorname{Im} \begin{bmatrix} 0 \\ \widetilde{G}(z) \end{bmatrix}$$

to stand for the vector subspace of the tangent space $T_xM \times T_zN$ spanned by the tangent vectors

$$\begin{bmatrix} 0 \\ \tilde{g}_1(z) \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ \tilde{g}_q(z) \end{bmatrix}.$$

An analogous interpretation applies to the notation

$$\operatorname{Im} \begin{bmatrix} G(x) \\ 0 \end{bmatrix},$$

where $G(x) = [g_1(x), \ldots, g_p(x)]$ and $g_1, \ldots, g_p$ are C^1 vector fields on M.

The following two lemmas will play a key role in the proof of our main result.

LEMMA 3.4. *Let $\mathcal{V}$ be a finite-dimensional (real) vector space with inner product $\langle\cdot,\cdot\rangle$, let $\mathcal{W} \subseteq \mathcal{V}$ be a vector subspace having codimension $0 < \sigma < \dim \mathcal{V}$, and let $\eta_1, \ldots, \eta_\sigma \in \mathcal{V}$ be a basis for the orthogonal complement $\mathcal{W}^\perp$. If $v_1, \ldots, v_\lambda \in \mathcal{V}$ is a finite collection of vectors, if $\mathcal{Z} = \operatorname{span}\{v_1, \ldots, v_\lambda\}$, and if*

$$\ell = \dim(\mathcal{Z} + \mathcal{W})/\mathcal{W},$$

then the $\sigma \times \lambda$ matrix

$$\big[\langle \eta_i, v_j \rangle\big]_{1\le i\le\sigma, 1\le j\le\lambda}$$

has rank ℓ.

Proof. This is a straightforward exercise in linear algebra. □

LEMMA 3.5. *Let r, p, q be positive integers, let $\mathcal{X}$ be a C^k differentiable manifold of dimension r, and let $\Sigma : \mathcal{X} \to \mathcal{L}(\mathbb{R}^p, \mathbb{R}^q)$ be a C^k mapping such that $\zeta \mapsto \operatorname{rank} \Sigma(\zeta)$ is constant for $\zeta \in \mathcal{X}$. Then there exists a C^k mapping $\Theta : \mathcal{X} \to \mathcal{L}(\mathbb{R}^q, \mathbb{R}^p)$ such that*

$$\zeta \in \mathcal{X} \quad \Rightarrow \quad \Sigma(\zeta)\Theta(\zeta)\Sigma(\zeta) = \Sigma(\zeta).$$

Proof. See the proof of Claim 1 in [3, Thm. 3.11] and [3, Rem. 3.14]. □

In the next proposition we have isolated some of the technical details that are required in the proof of our main result.

PROPOSITION 3.6. *Let*

$$F : M \times \Omega \times \Delta \to TM, \quad \widetilde{F} : N \times \Omega \times \mathbb{R}^q \to TN$$

be two C^1 ID systems that have common input space Ω, and suppose that $\widetilde{F}$ is affine in its disturbance; that is,

$$(z, \omega, \varepsilon) \in N \times \Omega \times \mathbb{R}^q \;\Rightarrow\; \widetilde{F}(z, \omega, \varepsilon) = \tilde{f}(z, \omega) + \widetilde{G}(z)\varepsilon, \tag{3.3}$$

where $\widetilde{G}(z) = [\tilde{g}_1(z), \ldots, \tilde{g}_q(z)]$ and $\tilde{g}_1, \ldots, \tilde{g}_q$ are C^1 vector fields on N. Let $\mathcal{R} \subseteq M \times N$ be a C^2 immersed submanifold of $M \times N$ with the following properties:

(i) *For $(x, z) \in \mathcal{R}$ the vector subspace*

$$\widetilde{\mathcal{V}}_{(x,z)} = T_{(x,z)}\mathcal{R} + \operatorname{Im} \begin{bmatrix} 0 \\ \widetilde{G}(z) \end{bmatrix} \tag{3.4}$$

of $T_{(x,z)}(M \times N)$ has a constant dimension as (x, z) varies over $\mathcal{R}$.

(ii) *For every $(x, z) \in \mathcal{R}$, $\omega \in \Omega$, and $\delta \in \Delta$, there exists $\varepsilon \in \mathbb{R}^q$ such that*

$$\begin{bmatrix} F(x, \omega, \delta) \\ \tilde{f}(z, \omega) + \widetilde{G}(z)\varepsilon \end{bmatrix} \in T_{(x,z)}\mathcal{R}. \tag{3.5}$$

Then there exists a mapping $\Upsilon : \mathcal{R} \times \Omega \times \Delta \to \mathbb{R}^q$ that is nicely C^1 in $(x, z) \in \mathcal{R}$ relative to $(\omega, \delta) \in \Omega \times \Delta$ and satisfies

$$((x, z), \omega, \delta) \in \mathcal{R} \times \Omega \times \Delta \;\Rightarrow\; \begin{bmatrix} F(x, \omega, \delta) \\ \tilde{f}(z, \omega) + \widetilde{G}(z)\Upsilon(x, z, \omega, \delta) \end{bmatrix} \in T_{(x,z)}\mathcal{R}. \tag{3.6}$$

Proof. We will show that the desired mapping Υ exists locally (relative to $\mathcal{R}$) in a neighborhood of an arbitrary point (x_0, z_0) in $\mathcal{R}$, after which we will glue the local versions of Υ together via a partition of unity to obtain the desired globally defined mapping (again, relative to $\mathcal{R}$). Thus we fix a point $(x_0, z_0) \in \mathcal{R}$ and appeal to a standard result in the theory of differentiable manifolds (see, for example, [20, p. 28]) which guarantees the existence of an open neighborhood $\mathcal{S}$ of (x_0, z_0) relative to $\mathcal{R}$ such that $\mathcal{S}$ is a *slice* of a coordinate neighborhood $\mathcal{N}$ of (x_0, z_0) in $M \times N$ (the terminology is as in [20]). We note that $\mathcal{S}$ may not be an open subset of $\mathcal{R}$ in the relative topology that $\mathcal{R}$ inherits from $M \times N$ due to the fact that we are assuming only that $\mathcal{R}$ is an immersed submanifold; in particular, we are not asserting that $\mathcal{S} = \mathcal{R} \cap \mathcal{N}$, but only that $\mathcal{S}$ is a subset of $\mathcal{N}$ on which certain of the coordinate functions on $\mathcal{N}$ are constant. Let $m = \dim M$, $n = \dim N$, and let σ be the codimension of $\mathcal{R}$ in $M \times N$. Since $\mathcal{S}$ is a slice of the coordinate neighborhood $\mathcal{N}$, it is easy to obtain a family of $m + n$ C^1 vector fields defined on $\mathcal{N}$,

$$\xi_1, \ldots, \xi_{m+n-\sigma}, \eta_1, \ldots, \eta_\sigma : \mathcal{N} \to T(M \times N),$$

such that

$$(x,z)\in\mathcal{S}\ \Rightarrow\ T_{(x,z)}\mathcal{S}=T_{(x,z)}\mathcal{R}=\operatorname{span}_{\mathbb{R}}\{\xi_1(x,z),\dots,\xi_{m+n-\sigma}(x,z)\} \tag{3.7}$$

and

$$(x,z)\in\mathcal{S}\ \Rightarrow\ \langle\eta_i(x,z),\xi_j(x,z)\rangle=0\quad(i=1,\dots,\sigma,j=1,\dots,m+n-\sigma), \tag{3.8}$$

where $\langle\cdot,\cdot\rangle$ is any conveniently chosen Riemannian metric on $M\times N$. In particular, for $(x,z)\in\mathcal{S}$ and $v\in T_{(x,z)}(M\times N)$ we have

$$v\in T_{(x,z)}\mathcal{S}=T_{(x,z)}\mathcal{R}\ \Leftrightarrow\ \langle\eta_i(x,z),v\rangle=0\text{ for every }i=1,\dots,\sigma. \tag{3.9}$$

Without loss of generality we may further assume that $\mathcal{N}=U\times V$, where U is a coordinate neighborhood of x_0 in M that is diffeomorphic to $\mathbb{R}^m$ and V is a coordinate neighborhood of z_0 in N that is diffeomorphic to $\mathbb{R}^n$. We can then identify the ID system F with its local representation

$$F:\mathbb{R}^m\times\Omega\times\Delta\to\mathbb{R}^m$$

relative to U. Similarly we can identify the ID system $\widetilde{F}$ with its local representation

$$\widetilde{F}:\mathbb{R}^n\times\Omega\times\mathbb{R}^q\to\mathbb{R}^n,\quad \widetilde{F}(z,\omega,\varepsilon)=\tilde{f}(z,\omega)+\widetilde{G}(z)\varepsilon,$$

relative to V, where $\tilde{f}:\mathbb{R}^n\times\Omega\to\mathbb{R}^n$ is nicely C^1 in $z\in\mathbb{R}^n$ relative to $\omega\in\Omega$ and $\widetilde{G}:\mathbb{R}^n\to\mathcal{L}(\mathbb{R}^q,\mathbb{R}^n)$ is C^1. Moreover, the C^1 vector fields $\xi_1,\dots,\xi_{m+n-\sigma},\eta_1,\dots,\eta_\sigma$ can be identified with their local representations

$$\xi_1,\dots,\xi_{m+n-\sigma},\eta_1,\dots,\eta_\sigma:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}^m\times\mathbb{R}^n.$$

If we identify the slice $\mathcal{S}\subseteq\mathcal{N}$ with its image in $\mathbb{R}^m\times\mathbb{R}^n$ under the coordinate diffeomorphism $\mathcal{N}\cong\mathbb{R}^m\times\mathbb{R}^n$, then $\mathcal{S}$ is an immersed submanifold of $\mathbb{R}^m\times\mathbb{R}^n$ for which (3.7) continues to hold and for which (3.8) can be replaced with

$$(x,z)\in\mathcal{S}\ \Rightarrow\ \eta_i(x,z)^\tau\xi_j(x,z)=0\quad(i=1,\dots,\sigma,j=1,\dots,m+n-\sigma),$$

where the superscript τ denotes the matrix transpose, elements of $\mathbb{R}^m\times\mathbb{R}^n$ are viewed as column vectors, and we select the Riemannian metric induced by the standard inner product on $\mathbb{R}^m\times\mathbb{R}^n$.

Likewise, (3.9) can be replaced with

$$v\in T_{(x,z)}\mathcal{S}\ \Leftrightarrow\ \eta_i(x,z)^\tau v=0\text{ for every }i=1,\dots,\sigma.$$

Next we select columns $\tilde{g}_{j_1},\dots,\tilde{g}_{j_\ell}$ from the $n\times q$-matrix $\widetilde{G}=[\tilde{g}_1,\dots,\tilde{g}_q]$ such that the family of $m+n$-dimensional vectors

$$\xi_1(x,z),\dots,\xi_{m+n-\sigma}(x,z),\begin{bmatrix}0\\ \tilde{g}_{j_1}(z)\end{bmatrix},\dots,\begin{bmatrix}0\\ \tilde{g}_{j_\ell}(z)\end{bmatrix} \tag{3.10}$$

is a basis of the vector space $\widetilde{\mathcal{V}}_{(x_0,z_0)}$ defined in (3.4) when $x=x_0$ and $z=z_0$. By continuity the vector fields (3.10) will be linearly independent in some open neighborhood $\mathcal{S}_0\subseteq\mathcal{S}$ of (x_0,z_0) (open relative to the submanifold topology of $\mathcal{R}$), and since

by assumption the dimension of $\widetilde{\mathcal{V}}_{(x,z)}$ is constant for $(x,z) \in \mathcal{R}$, we infer that the vectors (3.10) will form a basis for $\widetilde{\mathcal{V}}_{(x,z)}$ for every $(x,z) \in \mathcal{S}_0$. Because

$$T_{(x,z)}\mathcal{R} = \operatorname{span}_{\mathbb{R}}\{\xi_1(x,z), \ldots, \xi_{m+n-\sigma}(x,z)\},$$

we see that

$$(x,z) \in \mathcal{S}_0 \;\Rightarrow\; \dim \widetilde{\mathcal{V}}_{(x,z)}/T_{(x,z)}\mathcal{R} = \ell. \tag{3.11}$$

Since the vectors $\eta_1(x,z), \ldots, \eta_\sigma(x,z)$ form a basis for the orthogonal complement of $T_{(x,z)}\mathcal{R}$, if we let $L(x,z)$ denote the $\sigma \times (m+n)$-matrix

$$L(x,z) = \begin{bmatrix} \eta_1(x,z)^\tau \\ \vdots \\ \eta_\sigma(x,z)^\tau \end{bmatrix}, \tag{3.12}$$

then relation (3.11) and Lemma 3.4 imply that

$$(x,z) \in \mathcal{S}_0 \;\Rightarrow\; \operatorname{rank} L(x,z) \begin{bmatrix} 0 \\ \widetilde{G}(z) \end{bmatrix} = \ell. \tag{3.13}$$

It will also be convenient to partition the matrix $L(x,z)$ defined by (3.12) as

$$L(x,z) = \big[\Lambda(x,z), \widetilde{\Lambda}(x,z)\big], \tag{3.14}$$

where $\Lambda(x,z)$ has dimensions $\sigma \times m$, $\widetilde{\Lambda}(x,z)$ has dimensions $\sigma \times n$, and both have entries that are C^1 functions of $(x,z) \in \mathcal{S}_0$. From relation (3.13) and the partitioned form of $L(x,z)$ given by (3.14), we obtain

$$(x,z) \in \mathcal{S}_0 \;\Rightarrow\; \operatorname{rank} \widetilde{\Lambda}(x,z)\widetilde{G}(z) = \ell.$$

Thus by Lemma 3.5 there exists a C^1 mapping $\Theta : \mathcal{S}_0 \to \mathcal{L}(\mathbb{R}^\sigma, \mathbb{R}^q)$ such that

$$(x,z) \in \mathcal{S}_0 \;\Rightarrow\; \widetilde{\Lambda}(x,z)\widetilde{G}(z)\Theta(x,z)\widetilde{\Lambda}(x,z)\widetilde{G}(z) = \widetilde{\Lambda}(x,z)\widetilde{G}(z) \tag{3.15}$$

(for later reference we point out that up to this point in the proof we have not yet invoked assumption (ii)).

With the aid of the mapping Θ just obtained, we will now demonstrate the (local) existence of the desired mapping Υ. For $(x,z) \in \mathcal{S}_0$ and $(\omega,\delta) \in \Omega \times \mathbb{R}^p$ assumption (ii) implies the existence of $\varepsilon \in \mathbb{R}^q$ such that (3.5) is satisfied. Clearly ε depends on the variables (x,z,ω,δ), that is, $\varepsilon = \varepsilon(x,z,\omega,\delta)$, but it is not yet apparent that this dependence has the regularity properties claimed in the statement of the proposition. Because the rows of the matrix L defined by (3.12) span the orthogonal complement of $T_{(x,z)}\mathcal{S}_0 = T_{(x,z)}\mathcal{R}$, we see that

$$\begin{aligned} \mathbb{R}^\sigma \ni 0 &= L(x,z) \begin{bmatrix} F(x,\omega,\delta) \\ \tilde{f}(z,\omega) + \widetilde{G}(z)\varepsilon(x,z,\omega,\delta) \end{bmatrix} \\ &= \big[\Lambda(x,z), \widetilde{\Lambda}(x,z)\big] \begin{bmatrix} F(x,\omega,\delta) \\ \tilde{f}(z,\omega) + \widetilde{G}(z)\varepsilon(x,z,\omega,\delta) \end{bmatrix} \\ &= \Lambda(x,z)F(x,\omega,\delta) + \widetilde{\Lambda}(x,z)\big[\tilde{f}(z,\omega) + \widetilde{G}(z)\varepsilon(x,z,\omega,\delta)\big]. \end{aligned} \tag{3.16}$$

Define a mapping $\Gamma : \mathcal{S}_0 \times \Omega \times \Delta \to \mathbb{R}^\sigma$ by

$$\Gamma(x,z,\omega,\delta) = -\Lambda(x,z)F(x,\omega,\delta) - \widetilde{\Lambda}(x,z)\tilde{f}(z,\omega). \tag{3.17}$$

It is clear that Γ is nicely C^1 in the variables $(x,z) \in \mathcal{S}_0$ relative to $(\omega,\delta) \in \Omega \times \Delta$ and relation (3.16) yields

$$\Gamma(x,z,\omega,\delta) = \widetilde{\Lambda}(x,z)\widetilde{G}(z)\varepsilon(x,z,\omega,\delta), \tag{3.18}$$

so from (3.18) and (3.15) we obtain

$$\begin{aligned} &\widetilde{\Lambda}(x,z)\widetilde{G}(z)\Theta(x,z)\Gamma(x,z,\omega,\delta) \\ &\qquad = \widetilde{\Lambda}(x,z)\widetilde{G}(z)\Theta(x,z)\widetilde{\Lambda}(x,z)\widetilde{G}(z)\varepsilon(x,z,\omega,\delta) \\ &\qquad = \widetilde{\Lambda}(x,z)\widetilde{G}(z)\varepsilon(x,z,\omega,\delta) \\ &\qquad = \Gamma(x,z,\omega,\delta). \end{aligned} \tag{3.19}$$

If we define $\Upsilon : \mathcal{S}_0 \times \Omega \times \Delta \to \mathbb{R}^q$ by

$$\Upsilon(x,z,\omega,\delta) = \Theta(x,z)\Gamma(x,z,\omega,\delta), \tag{3.20}$$

then Υ is nicely C^1 in the variables $(x,z) \in \mathcal{R}$ relative to $(\omega,\delta) \in \Omega \times \Delta$ because Γ has this property and Θ is C^1. Moreover, from (3.19) and (3.17) we obtain

$$\begin{aligned} \widetilde{\Lambda}(x,z)\widetilde{G}(z)\Upsilon(x,z,\omega,\delta) &= \Gamma(x,z,\omega,\delta) \\ &= -\Lambda(x,z)F(x,\omega,\delta) - \widetilde{\Lambda}(x,z)\tilde{f}(z,\omega), \end{aligned}$$

which can be rewritten in the form

$$\begin{aligned} 0 &= \left[\Lambda(x,z), \widetilde{\Lambda}(x,z)\right] \begin{bmatrix} F(x,\omega,\delta) \\ \tilde{f}(z,\omega) + \widetilde{G}(z)\Upsilon(x,z,\omega,\delta) \end{bmatrix} \\ &= L(x,z) \begin{bmatrix} F(x,\omega,\delta) \\ \tilde{f}(z,\omega) + \widetilde{G}(z)\Upsilon(x,z,\omega,\delta) \end{bmatrix}. \end{aligned}$$

Since the rows of the matrix $L(x,z)$ span the orthogonal complement of $T_{(x,z)}\mathcal{R}$, we immediately obtain relation (3.6) for every $((x,z),\omega,\delta) \in \mathcal{S}_0 \times \Omega \times \Delta$. Consequently, the existence of the desired mapping Υ has been established in some neighborhood of an arbitrary point $(x_0,z_0) \in \mathcal{R}$.

To obtain the globally defined version of Υ, we appeal to the local result just obtained and a partition-of-unity argument. By the local result and by the second countability of $\mathcal{R}$ in its submanifold topology, there exist a countable open cover $\{\mathcal{S}_i \mid i \in \mathbb{N}\}$ of $\mathcal{R}$ and mappings

$$\Upsilon_i : \mathcal{S}_i \times \Omega \times \Delta \to \mathbb{R}^q$$

that are nicely C^1 in $(x,z) \in \mathcal{S}_i$ relative to $(\omega,\delta) \in \Omega \times \Delta$ and satisfy (3.6) for $((x,z),\omega,\delta) \in \mathcal{S}_i \times \Omega \times \Delta$. If we select a C^1 partition of unity $\{\psi_i : \mathcal{R} \to \mathbb{R} \mid i \in \mathbb{N}\}$ subordinate to the open cover $\{\mathcal{S}_i \mid i \in \mathbb{N}\}$, then the mapping

$$\Upsilon : \mathcal{R} \times \Omega \times \Delta \to \mathbb{R}^q, \quad \Upsilon((x,z),\omega,\delta) = \sum_{i=1}^{\infty} \psi_i(x,z)\Upsilon_i((x,z),\omega,\delta)$$

is easily seen to satisfy (3.6) everywhere on $\mathcal{R} \times \Omega \times \Delta$ while being nicely C^1 in the variables (x,z) relative to (ω,δ). This completes the proof. □

Remark 3.7.

1. The constant-dimension assumption (i) in Proposition 3.6 is fairly common in results of this type; see, for example, [10, Prop. 11.2].

2. If F is affine in its disturbance variable (say, with $\Delta = \mathbb{R}^p$), then Υ can also be chosen to be affine in the disturbance variable δ in the sense that there exist mappings

$$P : \mathcal{R} \times \Omega \to \mathbb{R}^q, \quad Q : \mathcal{R} \times \Omega \to \mathcal{L}(\mathbb{R}^p, \mathbb{R}^q),$$

both of which are nicely C^1 in (x, z) relative to $\omega \in \Omega$ and for which

$$\Upsilon((x, z), \omega, \delta) = P((x, z), \omega) + Q((x, z), \omega)\delta.$$

This follows directly from the formulas (3.20) and (3.17), which specify Υ locally, and the fact that the affine structure is preserved under the pasting process by which the globally defined version of Υ is obtained.

We are now in a position to prove the main results of this paper.

THEOREM 3.8. *Let (F, h) and $(\widetilde{F}, \tilde{h})$ be two C^1 IDO systems that satisfy the assumptions of Proposition* 3.6 *(in particular, they have common input space Ω, common output space O, and $\widetilde{F}$ is affine in its disturbance of the form* (3.3)*). Let $\mathcal{R} \subseteq M \times N$ be a C^2 immersed submanifold of $M \times N$ for which constant-dimension assumption* (i) *of Proposition* 3.6 *holds, and consider the following "simulation condition":*

(*SC*) *For every $(x, z) \in \mathcal{R}$ and $\omega \in \Omega$ we have $h(x, \omega) = \tilde{h}(z, \omega)$. Furthermore, for every $(x, z) \in \mathcal{R}$, $\omega \in \Omega$, and $\delta \in \Delta$, there exists $\varepsilon \in \mathbb{R}^q$ such that*

$$\begin{bmatrix} F(x, \omega, \delta) \\ \tilde{f}(z, \omega) + \widetilde{G}(z)\varepsilon \end{bmatrix} \in T_{(x,z)}\mathcal{R}. \tag{3.21}$$

Suppose that $\mathcal{U} \subseteq \mathcal{U}^{\Omega}_{\text{meas}}(\tilde{f})$, $\mathcal{D} \subseteq \mathcal{U}^{\Delta}_{\text{meas}}$ are such that $\mathcal{U} \times \mathcal{D} \subseteq \mathcal{U}^{\Omega\times\Delta}_{\text{meas}}(F)$, and let $\mathcal{E} \subseteq L^1_{\text{loc}}(\mathbb{R}, \mathbb{R}^q)$. Then the following statements hold:

1. *If $\mathcal{R}$ is a simulation relation of $(F, h, \mathcal{U}, \mathcal{D})$ by $(\widetilde{F}, \tilde{h}, \mathcal{U}, \mathcal{E})$ and if $\mathcal{U}$ and $\mathcal{D}$ contain all of the constant mappings in their respective ranges, then (SC) holds.*

2. *If (SC) holds, then $\mathcal{R}$ is a simulation relation of $(F, h, \mathcal{U}, \mathcal{D})$ by $(\widetilde{F}, \tilde{h}, \mathcal{U}, L^1_{\text{loc}}(\mathbb{R}, \mathbb{R}^q))$.*

Proof. To prove the first statement we first observe that our assumptions on $\mathcal{U}$ and $\mathcal{E}$ yield

$$\mathcal{U} \times \mathcal{E} \subseteq \mathcal{U}^{\Omega}_{\text{meas}}(\tilde{f}) \times L^1_{\text{loc}}(\mathbb{R}, \mathbb{R}^q) \subseteq \mathcal{U}^{\Omega\times\mathbb{R}^q}_{\text{meas}}(\widetilde{F})$$

by Remark 3.2. Assume that $\mathcal{R}$ is a simulation relation of $(F, h, \mathcal{U}, \mathcal{D})$ by $(\widetilde{F}, \tilde{h}, \mathcal{U}, \mathcal{E})$, and let $(x_0, z_0) \in \mathcal{R}$, $\omega_0 \in \Omega$, and $\delta_0 \in \mathbb{R}^p$ be given. If we let $u : \mathbb{R} \to \Omega$ and $d : \mathbb{R} \to \mathbb{R}^p$ be the constant mappings $u(t) \equiv \omega_0$ and $d(t) \equiv \delta_0$, then by assumption $u \in \mathcal{U}$ and $d \in \mathcal{D}$, so by Definition 2.2 there exist $e \in \mathcal{E}$ and a compact interval I containing 0 in its interior such that (we will use the abbreviated notations $\psi(t) = \psi(t, x_0, u, d)$ and $\tilde{\psi}(t) = \tilde{\psi}(t, z_0, u, e)$ for the trajectories)

$$t \in I \;\Rightarrow\; (\psi(t), \tilde{\psi}(t)) \in \mathcal{R} \tag{3.22}$$

and

$$t \in I \;\Rightarrow\; h(\psi(t), \omega_0) = \tilde{h}(\tilde{\psi}(t), \omega_0).$$

Thus we can set $t = 0$ to obtain $h(x_0, \omega_0) = \tilde{h}(z_0, \omega_0)$. Furthermore, the trajectories are absolutely continuous as functions of t, so we can differentiate (3.22) to obtain

$$\big(\dot{\psi}(t), \dot{\tilde{\psi}}(t)\big) \in T_{(\psi(t),\tilde{\psi}(t))}\mathcal{R} \quad \text{for a.e. } t \in I. \tag{3.23}$$

If the disturbance $t \mapsto e(t)$ is at least continuous, then (3.23) will hold for every $t \in I$ and we can then set $t = 0$ and immediately obtain (3.21) for $x = x_0$, $z = z_0$, $\omega = \omega_0$, $\delta = \delta_0$, and $\varepsilon = e(0)$. However, if e is measurable only as a function of t, then there is no guarantee that $t = 0$ is a Lebesgue point of the mapping $t \mapsto \tilde{\psi}(t)$ and further reasoning is required. The argument of Proposition 3.6 yields an open neighborhood $\mathcal{S}_0$ of (x_0, z_0) in $\mathcal{R}$ and C^1 mapping $L : \mathcal{S}_0 \to \mathcal{L}(\mathbb{R}^{m+n}, \mathbb{R}^\sigma)$ such that for $(x, z) \in \mathcal{S}_0$ we have

$$v \in T_{(x,z)}\mathcal{S}_0 = T_{(x,z)}\mathcal{R} \quad \Leftrightarrow \quad L(x, z)v = 0 \in \mathbb{R}^\sigma \tag{3.24}$$

(as before, σ is the codimension of $\mathcal{R}$ in $M \times N$, and since we are working locally we have passed to local coordinates $M \cong \mathbb{R}^m$, $N \cong \mathbb{R}^n$). If we partition L as in (3.14), then we also obtain a C^1 mapping $\Theta : \mathcal{S}_0 \to \mathcal{L}(\mathbb{R}^\sigma, \mathbb{R}^q)$ satisfying (3.15). Choose $\epsilon > 0$ such that $(-\epsilon, \epsilon) \subseteq I$ and $(\psi(t), \tilde{\psi}(t)) \in \mathcal{S}_0$ for $t \in (-\epsilon, \epsilon)$. Then (3.23) and (3.24) yield

$$\begin{aligned} \mathbb{R}^\sigma \ni 0 &= L(\psi(t), \tilde{\psi}(t)) \begin{bmatrix} F(\psi(t), \omega_0, \delta_0) \\ \tilde{f}(\tilde{\psi}(t), \omega_0) + \widetilde{G}(\tilde{\psi}(t))e(t) \end{bmatrix} \\ &= \Lambda(\psi(t), \tilde{\psi}(t))F(\psi(t), \omega_0, \delta_0) \\ &\qquad + \widetilde{\Lambda}(\psi(t), \tilde{\psi}(t))\big[\tilde{f}(\tilde{\psi}(t), \omega_0) + \widetilde{G}(\tilde{\psi}(t))e(t)\big] \\ &\qquad\qquad \text{for a.e. } t \in (-\epsilon, \epsilon). \end{aligned} \tag{3.25}$$

The argument used in the proof of Proposition 3.6 shows that if $\Upsilon : \mathcal{S}_0 \times \Omega \times \Delta \to \mathbb{R}^q$ is the mapping defined in (3.20), then relation (3.25) will continue to hold with $e(t)$ replaced by $\Upsilon(\psi(t), \tilde{\psi}(t), \omega_0, \delta_0)$. However, $t \mapsto \Upsilon(\psi(t), \tilde{\psi}(t), \omega_0, \delta_0)$ is continuous because Υ, ψ, and $\tilde{\psi}$ are continuous, so we obtain

$$|t| < \epsilon \Rightarrow 0 = L(\psi(t), \tilde{\psi}(t)) \begin{bmatrix} F(\psi(t), \omega_0, \delta_0) \\ \tilde{f}(\tilde{\psi}(t), \omega_0) + \widetilde{G}(\tilde{\psi}(t))\Upsilon(\psi(t), \tilde{\psi}(t), \omega_0, \delta_0) \end{bmatrix}. \tag{3.26}$$

In particular, we can set $t = 0$ in (3.26) and use relation (3.24) to infer that (3.21) holds when the variables x, z, ω, δ, and ε are replaced with $x_0, z_0, \omega_0, \delta_0$, and $\Upsilon(x_0, z_0, \omega_0, \delta_0)$, respectively, and the first statement is proved.

To prove the second statement we assume that (SC) holds and fix $(x_0, z_0) \in \mathcal{R}$, $u \in \mathcal{U}$, and $d \in \mathcal{D}$. The hypotheses of Proposition 3.6 are satisfied, so we infer the existence of a mapping $\Upsilon : \mathcal{R} \times \Omega \times \Delta \to \mathbb{R}^q$ that is nicely C^1 in the variables $(x, z) \in \mathcal{R}$ relative to $(\omega, \delta) \in \Omega \times \Delta$ and for which relation (3.6) holds. It follows that the mapping

$$\mathcal{F} : \mathcal{R} \times \Omega \times \Delta \to T(M \times N)$$

given by

$$\mathcal{F}((x, z), \omega, \delta) = \begin{bmatrix} F(x, \omega, \delta) \\ \tilde{f}(z, \omega) + \widetilde{G}(z)\Upsilon(x, z, \omega, \delta) \end{bmatrix} \tag{3.27}$$

takes values in the tangent bundle $T\mathcal{R}$ of the submanifold $\mathcal{R}$ ($T\mathcal{R}$ is, of course, a subbundle of $T(M \times N)$), and as such defines a C^1 control system on $\mathcal{R}$ with control space $\Omega \times \Delta$. A straightforward (though not completely obvious) check of the construction of Υ in Proposition 3.6 shows that

$$\mathcal{U} \times \mathcal{D} \subseteq \mathcal{U}_{\text{meas}}^{\Omega \times \mathbb{R}^p}(\mathcal{F}),$$

so for the given initial point $(x_0, z_0) \in \mathcal{R}$ and the given controls $u \in \mathcal{U}$, $d \in \mathcal{D}$, the corresponding state trajectory

$$t \mapsto \Psi(t, (x_0, z_0), u, d)$$

of the system $\mathcal{F}$ is defined on a compact interval I containing 0 in its interior and takes values in $\mathcal{R}$. If we define $e : \mathbb{R} \to \mathbb{R}^q$ by

$$e(t) = \begin{cases} \Upsilon(\Psi(t, (x_0, z_0), u, d), u(t), d(t)) & \text{if } t \in I, \\ 0 & \text{otherwise,} \end{cases} \tag{3.28}$$

then the form of Υ derived in Proposition 3.6 and the assumptions imposed on u and d force $e(\cdot)$ to be Lebesgue integrable on $\mathbb{R}$. One can also easily verify that

$$t \in I \;\Rightarrow\; \Psi(t, (x_0, z_0), u, d) = \big(\psi(t, x_0, u, d), \tilde{\psi}(t, z_0, u, e)\big),$$

where ψ and $\tilde{\psi}$ are the trajectory mappings for F and $\widetilde{F}$, respectively, so, in particular,

$$t \in I \;\Rightarrow\; \big(\psi(t, x_0, u, d), \tilde{\psi}(t, z_0, u, e)\big) \in \mathcal{R},$$

and, moreover,

$$t \in I \;\Rightarrow\; h(\psi(t, x_0, u, d), u(t)) = \tilde{h}(\tilde{\psi}(t, z_0, u, e), u(t))$$

by the assumption on h and $\tilde{h}$ in (SC), so the proof is complete. □

Remark 3.9. One can easily prove variations of assertion (2) in Theorem 3.8 to obtain simulation results for other (restricted) classes of inputs and disturbances. For example, a straightforward modification of the proof of Theorem 3.8.2 shows that if $\mathcal{U} \subseteq \mathcal{U}_{\text{cont}}^{\Omega}$ and $\mathcal{D} \subseteq \mathcal{U}_{\text{cont}}^{\Delta}$, then (SC) implies that $\mathcal{R}$ is a simulation relation of $(F, h, \mathcal{U}, \mathcal{D})$ by $(\widetilde{F}, \tilde{h}, \mathcal{U}, \mathcal{U}_{\text{cont}}^{\mathbb{R}^q})$. One simply checks that the requisite disturbance e is synthesized by way of the mapping Υ given by Proposition 3.6, and Υ is continuous (the definition of e in (3.28) needs to be adjusted slightly by extending e from I to $\mathbb{R}$ continuously instead of by zero). Similar results can be obtained for piecewise continuous inputs and disturbance, or (piecewise) differentiable inputs and disturbances if the input and disturbance spaces are assumed to be differentiable manifolds and the control systems are jointly differentiable in their state and control variables.

Two applications of Theorem 3.8 immediately yield the following bisimulation result.

THEOREM 3.10. *Let (F, h) and $(\widetilde{F}, \tilde{h})$ be two C^1 IDO systems that have common input space Ω, common output space O, and are affine in their disturbances (so F has the form (3.2) and $\widetilde{F}$ has the form (3.3)). Let $\mathcal{R} \subseteq M \times N$ be a C^2 immersed submanifold of $M \times N$ for which the following rank condition holds:*

(RC) For $(x, z) \in \mathcal{R}$ the vector subspaces

$$\mathcal{V}_{(x,z)} = T_{(x,z)}\mathcal{R} + \operatorname{Im}\begin{bmatrix} G(x) \\ 0 \end{bmatrix}, \quad \widetilde{\mathcal{V}}_{(x,z)} = T_{(x,z)}\mathcal{R} + \operatorname{Im}\begin{bmatrix} 0 \\ \widetilde{G}(z) \end{bmatrix}$$

of $T_{(x,z)}(M \times N)$ have constant—but possibly unequal—dimensions as (x, z) varies over $\mathcal{R}$ (in particular, these vector spaces need not coincide).

Consider the following "bisimulation condition":

(BC) For every $(x,z)\in\mathcal{R}$ *and* $\omega\in\Omega$ *we have* $h(x,\omega)=\tilde{h}(z,\omega)$. *Furthermore, for every* $(x,z)\in\mathcal{R}$, $\omega\in\Omega$, *and* $\delta\in\mathbb{R}^p$, *there exists* $\varepsilon\in\mathbb{R}^q$ *such that*

$$\begin{bmatrix} f(x,\omega)+G(x)\delta \\ \tilde{f}(z,\omega)+\widetilde{G}(z)\varepsilon \end{bmatrix} \in T_{(x,z)}\mathcal{R}, \tag{3.29}$$

and conversely for every $(x,z)\in\mathcal{R}$, $\omega\in\Omega$, *and* $\varepsilon\in\mathbb{R}^q$, *there exists* $\delta\in\mathbb{R}^p$ *such that* (3.29) *holds.*

Then the following statements hold:

1. *Let* $\mathcal{U}\subseteq\mathcal{U}^{\Omega}_{\mathrm{meas}}(f)\cap\mathcal{U}^{\Omega}_{\mathrm{meas}}(\tilde{f})$, $\mathcal{D}\subseteq L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^p)$, $\mathcal{E}\subseteq L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$ *be specified classes of inputs and disturbances that contain all of the constant mappings of* $\mathbb{R}$ *in their respective ranges. If* $\mathcal{R}$ *is a bisimulation relation between* $(F,h,\mathcal{U},\mathcal{D})$ *and* $(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$, *then* (BC) *holds.*

2. *If* $\mathcal{U}\subseteq\mathcal{U}^{\Omega}_{\mathrm{meas}}(f)\cap\mathcal{U}^{\Omega}_{\mathrm{meas}}(\tilde{f})$ *and if* (BC) *holds, then* $\mathcal{R}$ *is a bisimulation relation between* $(F,h,\mathcal{U},L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^p))$ *and* $(\widetilde{F},\tilde{h},\mathcal{U},L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q))$.

Remark 3.11. Of some interest is the special case of the previous theorem when the systems F and $\widetilde{F}$ are linear. That is, we suppose that the systems' state, input, and disturbance spaces are Euclidean (say, $M=\mathbb{R}^m$, $N=\mathbb{R}^n$, $\Omega=\mathbb{R}^r$, $\Delta=\mathbb{R}^p$, $E=\mathbb{R}^q$), F and $\widetilde{F}$ have the form

$$F(x,\omega,\delta)=Ax+B\omega+G\delta,\quad (x,\omega,\delta)\in\mathbb{R}^m\times\mathbb{R}^r\times\mathbb{R}^p,$$

and

$$\widetilde{F}(z,\omega,\varepsilon)=\widetilde{A}z+\widetilde{B}\omega+\widetilde{G}\varepsilon,\quad (z,\omega,\varepsilon)\in\mathbb{R}^n\times\mathbb{R}^r\times\mathbb{R}^q$$

for constant matrices A, B, G, $\widetilde{A}$, $\widetilde{B}$, and $\widetilde{G}$ of the appropriate dimensions, and $\mathcal{R}$ is a vector subspace of $M\times N=\mathbb{R}^m\times\mathbb{R}^n$. Then (RC) holds trivially (note that $T_{(x,z)}\mathcal{R}=\mathcal{R}$ in this case), so for any $\mathcal{U}\subseteq L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^r)$, we can infer that $\mathcal{R}$ is a bisimulation relation between $(F,h,\mathcal{U},L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^p))$ and $(\widetilde{F},\tilde{h},\mathcal{U},L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q))$ if the following bisimulation condition holds:

(BC′) For every $(x,z)\in\mathcal{R}$ and $\omega\in\mathbb{R}^r$ we have $h(x,\omega)=\tilde{h}(z,\omega)$. Furthermore, for every $(x,z)\in\mathcal{R}$, $\omega\in\mathbb{R}^r$, and $\delta\in\mathbb{R}^p$, there exists $\varepsilon\in\mathbb{R}^q$ such that

$$(Ax+B\omega+G\delta,\widetilde{A}z+\widetilde{B}\omega+\widetilde{G}\varepsilon)\in\mathcal{R}, \tag{3.30}$$

and conversely for every $(x,z)\in\mathcal{R}$, $\omega\in\mathbb{R}^r$, and $\varepsilon\in\mathbb{R}^q$, there exists $\delta\in\mathbb{R}^p$ such that (3.30) holds.

4. Discussion and related results. A particularly important example of simulation relations arises from the notion of systems "related" by smooth mappings [4, 12, 13]. Specifically, suppose we are given a pair of C^1 IDO systems

$$F:M\times\Omega\times\Delta\to TM,\quad h:M\times\Omega\to O,\quad u\in\mathcal{U},\quad d\in\mathcal{D}$$

and

$$\widetilde{F}:N\times\Omega\times E\to TN,\quad \tilde{h}:N\times\Omega\to O,\quad u\in\mathcal{U},\quad e\in\mathcal{E},$$

which have the common input space Ω, the common output space O, the common family of admissible inputs $\mathcal{U}$, and admissible families of disturbances $\mathcal{D}$ and $\mathcal{E}$, respectively, and let $\Phi:M\to N$ be a C^2 mapping. We wish to explore conditions

under which these IDO systems are "related" by the mapping Φ according to either of the following definitions (as before, ψ and $\tilde{\psi}$ denote the trajectory mappings of F and $\widetilde{F}$, respectively).

DEFINITION 4.1. *We say that* Φ maps trajectories of $(F, h, \mathcal{U}, \mathcal{D})$ to trajectories of $(\widetilde{F}, \tilde{h}, \mathcal{U}, \mathcal{E})$ while preserving outputs *if for every* $x_0 \in M$, $u \in \mathcal{U}$, *and* $d \in \mathcal{D}$ *there exist* $e \in \mathcal{E}$ *and a compact interval* I *containing* 0 *in its interior such that both of the trajectories* $t \mapsto \psi(t, x_0, u, d)$ *and* $t \mapsto \tilde{\psi}(t, z_0, u, e)$ *(here* $z_0 = \Phi(x_0)$*) are defined for every* $t \in I$ *and satisfy*

$$t \in I \;\Rightarrow\; \Phi\big(\psi(t, x_0, u, d)\big) = \tilde{\psi}(t, z_0, u, e) \tag{4.1}$$

and

$$t \in I \;\Rightarrow\; h(\psi(t, x_0, u, d), u(t)) = \tilde{h}(\tilde{\psi}(t, z_0, u, e), u(t)). \tag{4.2}$$

DEFINITION 4.2. *We say that* Φ lifts trajectories of $(\widetilde{F}, \tilde{h}, \mathcal{U}, \mathcal{E})$ to trajectories of $(F, h, \mathcal{U}, \mathcal{D})$ while preserving outputs *if for every* $z_0 \in N$, $u \in \mathcal{U}$, $e \in \mathcal{E}$, *and* $x_0 \in \Phi^{-1}(z_0)$ *there exist* $d \in \mathcal{D}$ *and a compact interval* I *containing* 0 *in its interior such that both of the trajectories* $t \mapsto \psi(t, x_0, u, d)$ *and* $t \mapsto \tilde{\psi}(t, z_0, u, e)$ *are defined for every* $t \in I$ *and satisfy relations* (4.1) *and* (4.2).

Remark 4.3. We note in passing that one can handle systems without outputs under this framework by simply specifying that the output mappings h and $\tilde{h}$ are constant, since in this case the output relation (4.2) is trivially satisfied.

As one might expect, the notions of mapping and lifting trajectories are subsumed by the notion of simulation of one system by another if we associate to Φ its graph

$$\mathcal{R} = \mathrm{Graph}(\Phi) = \{(x, \Phi(x)) \mid x \in M\}.$$

Observe that $\mathcal{R}$ is first and foremost a relation in $M \times N$, but it is also a C^2 (in this case closed and imbedded) submanifold of $M \times N$ that is C^2 diffeomorphic to M. The tangent spaces to $\mathcal{R}$ are easily described as follows:

$$T_{(x,\Phi(x))}\mathcal{R} = \left\{ \begin{bmatrix} v \\ d\Phi_x(v) \end{bmatrix} \mid v \in T_x M \right\} \quad (x \in M) \tag{4.3}$$

(we use the stacked representation of tangent vectors in $T\mathcal{R} \subseteq T(M \times N)$ introduced in Notation 3.3). Since $z_0 = \Phi(x_0)$ and relation (4.1) are equivalent to $(x_0, z_0) \in \mathcal{R}$ and

$$t \in I \;\Rightarrow\; \big(\psi(t, x_0, u, d), \tilde{\psi}(t, z_0, u, e)\big) \in \mathcal{R},$$

we immediately obtain the following propositions.

PROPOSITION 4.4. *The* C^2 *mapping* $\Phi : M \to N$ *maps trajectories of* $(F, h, \mathcal{U}, \mathcal{D})$ *to trajectories of* $(\widetilde{F}, \tilde{h}, \mathcal{U}, \mathcal{E})$ *while preserving outputs if and only if the relation* $\mathcal{R} = \mathrm{Graph}(\Phi)$ *is a simulation relation of* $(F, h, \mathcal{U}, \mathcal{D})$ *by* $(\widetilde{F}, \tilde{h}, \mathcal{U}, \mathcal{E})$.

PROPOSITION 4.5. *The* C^2 *mapping* $\Phi : M \to N$ *lifts trajectories of* $(\widetilde{F}, \tilde{h}, \mathcal{U}, \mathcal{E})$ *to trajectories of* $(F, h, \mathcal{U}, \mathcal{D})$ *while preserving outputs if and only if the relation* $\mathcal{R} = \mathrm{Graph}(\Phi)$ *is a simulation relation of* $(\widetilde{F}, \tilde{h}, \mathcal{U}, \mathcal{E})$ *by* $(F, h, \mathcal{U}, \mathcal{D})$.

For illustrative purposes, we show how the previous propositions, when combined with the results of section 3, allow us to recover previously obtained results on trajectory propagation and trajectory lifting.

THEOREM 4.6 (see [4, Thm. 3.11]). *Let (F,h) and $(\widetilde{F},\tilde{h})$ be two C^1 IDO systems that satisfy the assumptions of Proposition* 3.6 *(in particular, they have common input space Ω, common output space O, and $\widetilde{F}$ is affine in its disturbance of the form* (3.3)*). Let $\Phi : M \to N$ be a C^2 mapping such that*

(i) $(x,\omega) \in M \times \Omega \;\Rightarrow\; \tilde{h}(\Phi(x),\omega) = h(x,\omega)$*;*

(ii) *for every $(x,\omega) \in M \times \Omega$ we have*

$$\{d\Phi_x F(x,\omega,\delta) \mid \delta \in \Delta\} \subseteq \{\tilde{f}(\Phi(x),\omega) + \widetilde{G}(\Phi(x))\varepsilon \mid \varepsilon \in \mathbb{R}^q\};$$

(iii) *the vector space $\operatorname{Im}\tilde{G}(\Phi(x))$ has constant dimension as x varies over M.*

If $\mathcal{U} \subseteq \mathcal{U}^{\Omega}_{\mathrm{meas}}(\tilde{f})$, $\mathcal{D} \subseteq \mathcal{U}^{\Delta}_{\mathrm{meas}}$, and $\mathcal{U}\times\mathcal{D} \subseteq \mathcal{U}^{\Omega\times\Delta}_{\mathrm{meas}}(F)$, then Φ maps trajectories of $(F,h,\mathcal{U},\mathcal{D})$ to trajectories of $(\widetilde{F},\tilde{h},\mathcal{U},L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q))$ while preserving outputs.

Proof. If $(x,z) \in \operatorname{Graph}(\Phi)$ and $\omega \in \Omega$, then $z = \Phi(x)$ and assumption (i) yields $h(x,\omega) = \tilde{h}(z,\omega)$. Moreover, for $(x,\omega) \in M \times \Omega$ and $\delta \in \Delta$ assumption (ii) yields a $\varepsilon \in \mathbb{R}^q$ for which

$$d\Phi_x F(x,\omega,\delta) = \tilde{f}(z,\omega) + \widetilde{G}(z)\varepsilon,$$

and this, along with the description of the tangent spaces of $\mathcal{R} = \operatorname{Graph}(\Phi)$ given by (4.3), implies

$$\begin{bmatrix} F(x,\omega,\delta) \\ \tilde{f}(z,\omega) + \widetilde{G}(z)\varepsilon \end{bmatrix} \in T_{(x,z)}\mathcal{R},$$

so we obtain the simulation condition (SC) of Theorem 3.8. For $(x,z) \in \mathcal{R}$ we also have (here I_m denotes the $m \times m$ identity matrix)

$$\begin{aligned}
\widetilde{\mathcal{V}}_{(x,z)} = T_{(x,z)}\mathcal{R} + \operatorname{Im}\begin{bmatrix} 0 \\ \widetilde{G}(z) \end{bmatrix} &= T_{(x,\Phi(x))}\mathcal{R} + \operatorname{Im}\begin{bmatrix} 0 \\ \widetilde{G}(\Phi(x)) \end{bmatrix} \\
&= \left\{ \begin{bmatrix} v \\ d\Phi_x(v) + \widetilde{G}(\Phi(x))\varepsilon \end{bmatrix} \;\middle|\; v \in T_xM,\ \varepsilon \in \mathbb{R}^q \right\} \\
&= \left\{ \begin{bmatrix} I_m & 0 \\ d\Phi_x & \widetilde{G}(\Phi(x)) \end{bmatrix} \begin{bmatrix} v \\ \varepsilon \end{bmatrix} \;\middle|\; v \in T_xM,\ \varepsilon \in \mathbb{R}^q \right\}.
\end{aligned}$$

It follows that

$$\dim \widetilde{\mathcal{V}}_{(x,\Phi(x))} = m + \operatorname{rank}\widetilde{G}(\Phi(x)),$$

from which assumption (iii) implies that $\dim \widetilde{\mathcal{V}}_{(x,z)}$ is constant as (x,z) varies over $\mathcal{R} = \operatorname{Graph}(\Phi)$. Thus we infer from Theorem 3.8.2 that $\mathcal{R}$ is a simulation relation of $(F,h,\mathcal{U},\mathcal{D})$ by $(\widetilde{F},\tilde{h},\mathcal{U},L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q))$, and the theorem is now a direct consequence of Proposition 4.4. ∎

Remark 4.7. In the literature (see, for example, [4, 12, 13]) the systems F and $\widetilde{F}$ are called Φ-*related* if assumption (ii) of Theorem 4.6 holds. Theorem 4.6 is also closely related to (and generalizes somewhat) a result of Elkin [1, Thm. 2.1].

THEOREM 4.8 (see [5, Thm. 3.2]). *Let*

$$F : M \times \Omega \times \mathbb{R}^p \to TM, \quad h : M \times \Omega \to O$$

and

$$\widetilde{F} : N \times \Omega \times E \to TN, \quad \tilde{h} : N \times \Omega \to O$$

be two C^1 IDO systems that have common input space Ω, common output space O, and suppose that F is affine in its disturbance of the form (3.2). *Let $\Phi : M \to N$ be a C^2 mapping, let $\mathcal{R} = \mathrm{Graph}(\Phi) \subseteq M \times N$, and suppose that the following conditions are met:*

(i) $(x,\omega) \in M \times \Omega \;\Rightarrow\; \tilde{h}(\Phi(x),\omega) = h(x,\omega)$;

(ii) *for every $(z,\omega) \in N \times \Omega$ and for every $x \in \Phi^{-1}(z)$ we have*

$$\{d\Phi_x[f(x,\omega) + G(x)\delta] \mid \delta \in \Delta\} \supseteq \{\widetilde{F}(z,\omega,\varepsilon) \mid \varepsilon \in \mathbb{R}^q\};$$

(iii) *the vector space $\operatorname{Im} d\Phi_x \circ G(x)$ has constant dimension as x varies over M. If $\mathcal{U} \subseteq \mathcal{U}^{\Omega}_{\mathrm{meas}}(f)$ and $\mathcal{E} \subseteq \mathcal{U}^{E}_{\mathrm{meas}}$ are such that $\mathcal{U} \times \mathcal{E} \subseteq \mathcal{U}^{\Omega\times E}_{\mathrm{meas}}(\widetilde{F})$, then Φ lifts trajectories of $(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$ to trajectories of $(F,h,\mathcal{U},L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^p))$ while preserving outputs.*

Proof. As in the proof of the previous theorem, we immediately see that $(x,z) \in \mathcal{R}$ and $\omega \in \Omega \;\Rightarrow\; h(x,\omega) = \tilde{h}(z,\omega)$. Moreover, for $(x,z) \in \mathcal{R} = \mathrm{Graph}(\Phi)$, $\omega \in \Omega$, and $\varepsilon \in E$ we have $x \in \Phi^{-1}(z)$, and assumption (ii) yields $\delta \in \mathbb{R}^p$ such that

$$d\Phi_x[f(x,\omega) + G(x)\delta] = \widetilde{F}(z,\omega,\varepsilon),$$

and this, along with the description of the tangent spaces of $\mathcal{R} = \mathrm{Graph}(\Phi)$ given by (4.3), implies

$$\begin{bmatrix} f(x,\omega) + G(x)\delta \\ \widetilde{F}(z,\omega,\varepsilon) \end{bmatrix} \in T_{(x,z)}\mathcal{R}.$$

Thus we have established the analogue of the simulation condition (SC) in Theorem 3.8 obtained by reversing the roles of F and $\widetilde{F}$. For $(x,z) \in \mathcal{R}$ we also have

$$\begin{aligned}
\mathcal{V}_{(x,z)} = T_{(x,z)}\mathcal{R} + \operatorname{Im}\begin{bmatrix} G(x) \\ 0 \end{bmatrix} &= T_{(x,\Phi(x))}\mathcal{R} + \operatorname{Im}\begin{bmatrix} G(x) \\ 0 \end{bmatrix} \\
&= \left\{ \begin{bmatrix} v + G(x)\delta \\ d\Phi_x(v) \end{bmatrix} \;\middle|\; v \in T_xM,\ \delta \in \mathbb{R}^p \right\} \\
&= \left\{ \begin{bmatrix} I_m & G(x) \\ d\Phi_x & 0 \end{bmatrix} \begin{bmatrix} v \\ \delta \end{bmatrix} \;\middle|\; v \in T_xM,\ \delta \in \mathbb{R}^p \right\}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\dim \mathcal{V}_{(x,z)} &= \operatorname{rank}\begin{bmatrix} I_m & G(x) \\ d\Phi_x & 0 \end{bmatrix} \\
&= \operatorname{rank}\begin{bmatrix} I_m & 0 \\ -d\Phi_x & I_n \end{bmatrix}\begin{bmatrix} I_m & G(x) \\ d\Phi_x & 0 \end{bmatrix} \\
&= \operatorname{rank}\begin{bmatrix} I_m & G(x) \\ 0 & -d\Phi_x G(x) \end{bmatrix},
\end{aligned}$$

where the second equality is due to the fact that we are premultiplying by an invertible matrix. Consequently,

$$\dim \mathcal{V}_{(x,\Phi(x))} = m + \operatorname{rank} d\Phi_x G(\Phi(x)),$$

from which assumption (iii) implies that $\dim \mathcal{V}_{(x,z)}$ is constant as (x,z) varies over $\mathcal{R} = \mathrm{Graph}(\Phi)$. Thus Theorem 3.8.2 implies that $\mathcal{R}$ is a simulation relation of

$(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$ by $(F,h,\mathcal{U},L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^p))$. The theorem is now a direct consequence of Proposition 4.5. □

We conclude by making some remarks concerning finite escape time of trajectories. As pointed out in Remark 2.3, our definition of bisimulation relation (Definition 2.2) is local in time (i.e., holds only on a compact interval containing the initial time 0) because the trajectories of the "simulating system" $\widetilde{F}$ may not be guaranteed to exist for all times for which the corresponding trajectories of the "original system" F are defined. However, it is desirable to identify situations in which the simulating trajectory is guaranteed to exist over any compact interval on which the corresponding trajectory of the original system is defined. Such a situation is addressed by the following definition.

DEFINITION 4.9. *Let $(F,h,\mathcal{U},\mathcal{D})$ and $(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$ be a pair of C^1 IDO systems with admissible inputs and disturbances, and let $\mathcal{R}\subseteq M\times N$ be a simulation relation of $(F,h,\mathcal{U},\mathcal{D})$ by $(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$ as given by Definition* 2.2. *We say that $\mathcal{R}$ is* semiglobal in time *if for every $(x_0,z_0)\in\mathcal{R}$, for every $u\in\mathcal{U}$, for every $d\in\mathcal{D}$, and for every compact interval I containing* 0 *on which the trajectory $t\mapsto\psi(t,x_0,u,d)$ is defined there exists $e\in\mathcal{E}$ such that the trajectory $t\mapsto\tilde{\psi}(t,z_0,u,e)$ is defined for every $t\in I$ and relations* (2.1) *and* (2.2) *hold for every $t\in I$.*

We note in passing that usage the term "semiglobal" instead of "global" is intended to emphasize the fact that we do not require the stronger property that our systems' trajectories be defined for all $t\in\mathbb{R}$. Indeed, the realization of trajectories that are defined for all $t\in\mathbb{R}$ (or all $t\geq 0$) would necessitate further restrictions on the class of systems under consideration and the families of admissible inputs and disturbances (at the very least, local L^1 boundedness would have to be replaced with L^1 boundedness), which we will not pursue here.

As a first step in realizing simulation relations that are semiglobal in time, one could impose the condition of completeness on $\widetilde{F}$. However, it is not obvious that such a condition would, in and of itself, be sufficient to guarantee the desired behavior because additional nonlinearities in the state variables may be introduced by the feedback mapping Υ through which the simulating control e is synthesized. Here we will identify two fairly simple situations in which the simulating trajectory is guaranteed to have the same interval of definition as the original trajectory, but we do not claim that this discussion is a definitive resolution of the matter. Our first result in this direction deals with compact simulation relations.

PROPOSITION 4.10. *Assume that $(F,h,\mathcal{U},\mathcal{D})$ and $(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$ are IDO systems with admissible classes of inputs and disturbances that satisfy the assumptions of Theorem* 3.8 *relative to the C^2 immersed submanifold $\mathcal{R}$ of $M\times N$, assume that condition (SC) holds, and suppose that $\mathcal{E}=L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$. Suppose further that $\mathcal{R}$ is a compact subset of $M\times N$. Then $\mathcal{R}$ is a simulation relation of $(F,h,\mathcal{U},\mathcal{D})$ by $(\widetilde{F},\tilde{h},\mathcal{U},\mathcal{E})$ that is semiglobal in time.*

Proof. Fix $(x_0,z_0)\in\mathcal{R}$, $u\in\mathcal{U}$, $d\in\mathcal{D}$, and let I be any compact subinterval of $\mathbb{R}$ containing 0 on which $\psi(\cdot,x_0,u,d)$ is defined (as before, ψ is the trajectory mapping of F). The proof of Theorem 3.8.2 shows that $\mathcal{F}:\mathcal{R}\times\Omega\times\Delta\to T(M\times N)$ defined by (3.27) is tangent to the submanifold $\mathcal{R}$ and thus defines a C^1 control system on $\mathcal{R}$ with control space $\Omega\times\Delta$. Since $\mathcal{R}$ is compact, standard continuation results in the theory of ordinary differential equations guarantee that the trajectory $t\mapsto\Psi(t,(x_0,z_0),u,d)$ of $\mathcal{F}$ with initial condition (x_0,z_0) corresponding to the input u and disturbance d is defined for all $t\in I$ and takes values in $\mathcal{R}$. As pointed out in the proof of Theorem 3.8, if $\tilde{\psi}$ is the trajectory mapping of $\widetilde{F}$, then $\psi(\cdot,x_0,u,d)$ is the

first component of $\Psi(\cdot,(x_0,z_0),u,d)$ and $\tilde{\psi}(\cdot,z_0,u,e)$ is the second component, where e is defined by (3.28). A routine check of the assumptions in force shows that e is in fact a member of $L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$, so the conclusion follows directly from Theorem 3.8.2. □

To obtain sufficient conditions for noncompact simulation relations that are global or semiglobal in time, one must impose rather special conditions on the systems under consideration. We will be content to present one illustrative result in this direction.

PROPOSITION 4.11. *Suppose that the system state spaces and the input space are Euclidean (i.e., $M=\mathbb{R}^m$, $N=\mathbb{R}^n$, and $\Omega=\mathbb{R}^r$), and let*

$$F:\mathbb{R}^m\times\mathbb{R}^r\times\Delta\to\mathbb{R}^m,\quad h:\mathbb{R}^m\times\mathbb{R}^r\to O$$

and

$$\widetilde{F}:\mathbb{R}^n\times\mathbb{R}^r\times\mathbb{R}^q\to\mathbb{R}^n,\quad \tilde{h}:\mathbb{R}^n\times\mathbb{R}^r\to O$$

be two C^1 IDO systems that have common input space $\mathbb{R}^r$, common output space O, and suppose that $\widetilde{F}$ is linear; that is,

$$(z,\omega,\varepsilon)\in\mathbb{R}^n\times\mathbb{R}^r\times\mathbb{R}^q\ \Rightarrow\ \widetilde{F}(z,\omega,\varepsilon)=\widetilde{A}z+\widetilde{B}\omega+\widetilde{G}\varepsilon$$

for constant matrices $\widetilde{A}$, $\widetilde{B}$, and $\widetilde{G}$ of the appropriate dimensions. Let $\mathcal{R}$ be a vector subspace of $M\times N=\mathbb{R}^m\times\mathbb{R}^n$ for which the following modified simulation condition holds:

(SC′) For every $(x,z)\in\mathcal{R}$ and $\omega\in\mathbb{R}^r$ we have $h(x,\omega)=\tilde{h}(z,\omega)$. Furthermore, for every $(x,z)\in\mathcal{R}$, $\omega\in\mathbb{R}^r$, and $\delta\in\Delta$, there exists $\varepsilon\in\mathbb{R}^q$ such that

$$\begin{bmatrix}F(x,\omega,\delta)\\ \widetilde{A}z+\widetilde{B}\omega+\widetilde{G}\varepsilon\end{bmatrix}\in\mathcal{R}.$$

Further suppose that $\mathcal{U}\subseteq L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^r)$, $\mathcal{D}\subseteq\mathcal{U}^{\Delta}_{\mathrm{meas}}$ are such that $\mathcal{U}\times\mathcal{D}\subseteq\mathcal{U}^{\mathbb{R}^r\times\Delta}_{\mathrm{meas}}(F)$. Then $\mathcal{R}$ is a simulation relation of $(F,h,\mathcal{U},\mathcal{D})$ by $(\widetilde{F},\tilde{h},\mathcal{U},L^1_{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q))$ that is semiglobal in time.

Proof. We first observe that assumption (i) of Proposition 3.6 is trivially satisfied in this case because the vector space

$$\mathcal{V}_{(x,z)}=\mathcal{R}+\operatorname{Im}\begin{bmatrix}0\\ \widetilde{G}\end{bmatrix}$$

is now constant as (x,z) varies over $\mathcal{R}$ (itself a vector space). Letting σ be the codimension of the vector subspace $\mathcal{R}$ in $\mathbb{R}^m\times\mathbb{R}^n$ and letting N be any $\sigma\times(m+n)$ matrix whose rows span the orthogonal complement of $\mathcal{R}$, we see that the matrix functions Λ, $\widetilde{\Lambda}$, and Θ of Proposition 3.6 can be taken to be constant. Thus the mapping Υ given by Proposition 3.6 will have the form (see (3.20) and (3.17))

$$\Upsilon(x,z,\omega,\delta)=\Theta\big[-\Lambda F(x,\omega,\delta)-\widetilde{\Lambda}(\widetilde{A}z+\widetilde{B}\omega)\big].$$

The proof of Proposition 3.6 shows that the control system

$$\mathcal{F}:\mathcal{R}\times\mathbb{R}^r\times\Delta\to\mathbb{R}^m\times\mathbb{R}^n \tag{4.4}$$

given by

$$\mathcal{F}((x,z),\omega,\delta) = \begin{bmatrix} F(x,\omega,\delta) \\ (I_n - \widetilde{G}\Theta\widetilde{\Lambda})(\widetilde{A}z + \widetilde{B}\omega) - \widetilde{G}\Theta\Lambda F(x,\omega,\delta) \end{bmatrix} \tag{4.5}$$

is actually tangent to the subspace $\mathcal{R}$ (i.e., takes values in $\mathcal{R}$), and $\mathcal{R}$ is closed as a subset of $\mathbb{R}^m \times \mathbb{R}^n$, so the trajectories of $\mathcal{F}$ will stay in $\mathcal{R}$ if they are initialized in $\mathcal{R}$. Fix $(x_0, z_0) \in \mathcal{R}$, $u \in \mathcal{U}$, $d \in \mathcal{D}$, and let I be any compact subinterval of $\mathbb{R}$ containing 0 on which $\psi(\cdot, x_0, u, d)$ is defined (as before, ψ is the trajectory mapping of F). Write $\psi(t) = \psi(t, x_0, u, d)$ to lighten the notation and consider the linear, time-varying ordinary differential equation on $\mathbb{R}^n$ given by

$$\dot{z} = (I_n - \widetilde{G}\Theta\widetilde{\Lambda})(\widetilde{A}z + \widetilde{B}u(t)) - \widetilde{G}\Theta\Lambda F(\psi(t), u(t), d(t)). \tag{4.6}$$

Because u is locally integrable on $\mathbb{R}$ and, by our definitions, $t \mapsto F(\psi(t), u(t), d(t))$ is integrable on I, the linearity of (4.6) guarantees that its unique solution, call it $t \mapsto \eta(t)$, for the initial condition $\eta(0) = z_0$ is defined for all $t \in I$. Furthermore, it is clear that the mapping $t \mapsto (\psi(t), \eta(t))$ is a trajectory of the control system $\mathcal{F}$ defined by (4.5) with initial condition (x_0, z_0) corresponding to the input u and the disturbance d. Since $(x_0, z_0) \in \mathcal{R}$ and $\mathcal{F}$ is tangent to $\mathcal{R}$, we infer that

$$t \in I \quad \Rightarrow \quad (\psi(t), \eta(t)) \in \mathcal{R}. \tag{4.7}$$

If we define $e : \mathbb{R} \to \mathbb{R}^q$ by

$$e(t) = \begin{cases} -\Theta\widetilde{\Lambda}\widetilde{A}\eta(t) - \Theta\widetilde{\Lambda}\widetilde{B}u(t) - \Theta\Lambda F(\psi(t), u(t), d(t)) & \text{if } t \in I, \\ 0 & \text{otherwise,} \end{cases}$$

then e is integrable on $\mathbb{R}$ (because it is integrable on I) and we have

$$\dot{\eta}(t) = \widetilde{A}\eta(t) + \widetilde{B}u(t) + \widetilde{G}e(t) \quad \text{for a.e. } t \in I.$$

Consequently we see that $\eta(t) = \tilde{\psi}(t, z_0, u, e)$ for $t \in I$, where $\tilde{\psi}$ is the trajectory mapping of $\widetilde{F}$. When combined with (4.7) and the fact that $\psi(t) = \psi(t, x_0, u, d)$, this immediately yields (2.1). Relation (2.2) now follows from this and the assumption that $h(x, \omega) = \tilde{h}(z, \omega)$ for every $(x, z) \in \mathcal{R}$ and $\omega \in \mathbb{R}^r$, so the proof is complete. ∎

REFERENCES

[1] V. I. Elkin, *Reduction of Nonlinear Control Systems*, Kluwer, Dordrecht, The Netherlands, 1999.

[2] B. Jakubczyk, *Equivalence and invariants of nonlinear control systems*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel-Dekker, New York, 1990, pp. 177–218.

[3] K. A. Grasse, *On controlled invariance for fully nonlinear systems*, Internat. J. Control, 56 (1992), pp. 1121–1137.

[4] K. A. Grasse, *Admissibility of trajectories for control systems related by smooth mappings*, Math. Control Signals Systems, 16 (2003), pp. 120–140.

[5] K. A. Grasse, *Lifting of trajectories of control systems related by smooth mappings*, Systems Control Lett., 54 (2005), pp. 195–205.

[6] K. A. Grasse and H. J. Sussmann, *Global controllability by nice controls*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel-Dekker, New York, 1990, pp. 33–79.

[7] E. Haghverdi, P. Tabuada, and G. Pappas, *Bisimulation relations for dynamical, control, and hybrid systems*, Theoret. Comput. Sci., 342 (2005), pp. 229–261.

[8] A. Isidori, A. J. Krener, C. Gori-Giorgi, and S. Monaco, *Locally (f, g)-invariant distributions*, Systems Control Lett., 1 (1981/82), pp. 12–15.
[9] H. Nijmeijer, *Controlled invariance for affine control systems*, Internat. J. Control, 34 (1981), pp. 825–833.
[10] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[11] G. J. Pappas, *Bisimilar linear systems*, Automatica J. IFAC, 39 (2003), pp. 2035–2047.
[12] G. J. Pappas, G. Lafferriere, and S. Sastry, *Hierarchically consistent control systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1144–1160.
[13] G. J. Pappas and S. Simić, *Consistent abstractions of affine control systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 745–756.
[14] P. Tabuada and G. J. Pappas, *Bisimilar control affine systems*, Systems Control Lett., 52 (2004), pp. 49–58.
[15] P. Tabuada and G. J. Pappas, *Quotients of fully nonlinear control systems*, SIAM J. Control Optim., 43 (2005), pp. 1844–1866.
[16] P. Tabuada and G. J. Pappas, *Hierarchical trajectory refinement for a class of nonlinear systems*, Automatica J. IFAC, 41 (2005), pp. 701–708.
[17] A. J. van der Schaft, *Bisimulation of dynamical systems*, in Hybrid Systems: Computation and Control: 7th International Workshop Proceedings, Lecture Notes in Comput. Sci. 2993, Springer-Verlag, Heidelberg, pp. 555–569.
[18] A. J. van der Schaft, *Equivalence of dynamical systems by bisimulation*, IEEE Trans. Automat. Control, 50 (2005), pp. 286–298.
[19] E. D. Sontag, *Mathematical Control Theory*, 2nd ed., Springer-Verlag, New York, 1998.
[20] F. W. Warner, *Foundations of Differentiable Manifolds and Lie Groups*, Scott, Foresman and Company, Glenview, IL, 1971.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 585–603

RELATIVELY OPTIMAL CONTROL: A STATIC PIECEWISE-AFFINE SOLUTION*

FRANCO BLANCHINI† AND FELICE ANDREA PELLEGRINO‡

Abstract. A relatively optimal control is a stabilizing controller that, without initialization nor feedforwarding and tracking the optimal trajectory, produces the optimal (constrained) behavior for the nominal initial condition of the plant. In a previous work, for discrete-time linear systems, we presented a linear *dynamic* relatively optimal control. Here we provide a *static* solution, namely a deadbeat piecewise-affine state-feedback controller based on a suitable partition of the state space into polyhedral sets. The vertices of the polyhedra are the states of the optimal trajectory; hence a bound for the complexity of the controller is known in advance. We also show how to obtain a controller that is not deadbeat by removing the zero terminal constraint while guaranteeing stability. Finally, we compare the proposed static compensator with the dynamic one.

Key words. optimal control, linear systems, discrete-time systems, invariance

AMS subject classifications. 93B52, 93B51, 93C05, 49N35

DOI. 10.1137/050643180

1. Introduction.

It is known that, unless for very special cases, determining an optimal control in a feedback form under output or input constraints is a computationally hard task. The problem can be addressed in a receding horizon fashion, but in this case an optimization problem must be solved online at each time interval. Explicit (piecewise-affine) solutions exist [1, 2] but are limited to quadratic or 1-norm cost and linear constraints. However, for those systems that are explicitly built to perform specific operation through a specific trajectory with known initial and final states, the request for optimality from *any* initial state can be relaxed, requiring optimality only from a *specific* initial condition. The relatively optimal control (ROC) [5] is defined as *a stabilizing controller that guarantees optimality of the trajectory and constraint satisfaction from a given (or a set of given) initial condition(s) without the involvement of any feedforward action.* In [5] it has been proved that a controller enjoying these properties is linear dynamic and its order is equal to the length of the optimal trajectory minus the order of the plant. In [6] the zero terminal constraint was removed in order to assign a characteristic polynomial to the closed-loop system, and the problem of output feedback was addressed. Here, a *static* ROC is constructed by partitioning the state space into polyhedral sets whose vertices are the states of the optimal trajectory and their opposite.

The main contribution of the paper can be summarized in the following points.

- It is shown that for discrete-time linear systems with convex constraints and cost, it is always possible to construct a static ROC by means of a proper partition of the state space into polyhedral sets (a procedure to construct it is provided).

*Received by the editors October 20, 2005; accepted for publication (in revised form) November 6, 2006; published electronically May 4, 2007. This research was supported by MURST, Italy.
http://www.siam.org/journals/sicon/46-2/64318.html

†Dipartimento di Matematica e Informatica, Università di Udine, 33100 Udine, Italy (blanchini@uniud.it).

‡Dipartimento di Elettrotecnica, Elettronica e Informatica, University of Trieste, via A. Valerio, 10-34127 Trieste, Italy (fapellegrino@units.it).

- If the constraints and/or the cost are not convex, a sufficient condition on the optimal trajectory that guarantees that the static ROC can be constructed is provided.
- The proposed controller is a deadbeat piecewise-affine state-feedback controller. The vertices of each of the polyhedral sets are the states of the optimal trajectory and their opposite. The control at each vertex is the corresponding control of the optimal sequence, while the control at a generic state is given by a convex combination of the controls corresponding to the vertices of the polyhedron to which the state belongs.
- An upper bound on the number of polyhedral sets as a function of the order of the system and the length of the optimal trajectory is provided.
- By removing the zero state terminal constraint and requiring the final state of the optimal trajectory to belong to a controlled invariant set, it is possible to obtain a nondeadbeat controller.
- The proposed static controller is compared with the dynamic one previously introduced.

2. Problem statement. We give the discrete-time reachable system

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k), \\ y(k) &= Cx(k) + Du(k), \end{aligned} \tag{1}$$

where $x(k) \in \mathbb{R}^n$, $u(k) \in \mathbb{R}^m$, $y(k) \in \mathbb{R}^q$ and A, B, C, D are matrices of appropriate dimensions. For this system we consider the locally bounded convex cost functions of the output

$$g(y), \quad l_i(y), \quad i = 1, 2, \ldots, s \tag{2}$$

(we assume they are 0-symmetric, i.e., $g(y) = g(-y)$ and $l_i(y) = l_i(-y)$) with assigned initial condition

$$\bar{x} \neq 0 \tag{3}$$

and the constraint

$$y(k) \in \mathcal{Y}, \tag{4}$$

where $\mathcal{Y}$ is a convex, closed, and 0-symmetric set. Then we consider the following problem (consistently with [5] and with no loss of generality, we assume $k = 1$ is initial time):

$$J_{opt}(\bar{x}) = \min \sum_{k=1}^{N} g(y(k)) \tag{5}$$

subject to

$$x(k+1) = Ax(k) + Bu(k), \ k = 1, \ldots, N, \tag{6}$$

$$y(k) = Cx(k) + Du(k), \ k = 1, \ldots, N, \tag{7}$$

$$\sum_{k=1}^{N} l_i(y(k)) \leq \mu_i, \quad i = 1, 2, \ldots, s, \tag{8}$$

$$y(k) \in \mathcal{Y}, \quad k = 1, \ldots, N, \tag{9}$$

$$x(1) = \bar{x}, \tag{10}$$

$$x(N+1) = 0, \tag{11}$$

$$N \geq 1 \quad \text{assigned (or free).} \tag{12}$$

In the extremely general formulation of the problem we have considered the option of N free, in order to also consider the special case of minimum-time control. The choice of N depends on the circumstances and has to guarantee the feasibility of the above open-loop optimal control problem. Note that the cost and the constraint achieved, assuming g and l_i only depending on y, are quite general since we can include cost and pointwise or integral constraints depending on both x and u by suitable choices of C and D. Finding an open-loop solution for the above problem is well known as a convex problem [10] which can be solved by means of standard convex programming algorithms. Here we are interested in a feedback static solution; more precisely the problem we consider is the following.

PROBLEM 1. *Find a static state-feedback compensator of the form $u = \Phi(x)$ which is stabilizing and such that for $x(1) = \bar{x}$ the control and state trajectories are the optimal ones.*

Any solution of the above problem will be referred to as a *static relatively optimal controller*. We stress that in the ROC framework, the constraints (8) and (9) represent *design specification "soft" constraints*. Hence their violation implies a performance loss only and is allowed for nonnominal initial conditions. In the following we will construct a solution to Problem 1 in two steps: First, a relatively optimal controller that is only locally stabilizing (being defined in a convex subset of the state space, containing the origin) will be constructed. We will refer to this controller as the *local relatively optimal controller*. Then the local controller will be extended to the whole state space, obtaining a *global relatively optimal controller*.

Remark 2.1. Any initial state $\bar{x}$ for which the problem is feasible (hence the constraints can be satisfied) is suitable as a nominal initial condition; there are no further restrictions.

3. Main results. We now assume that the optimal trajectory starting from the assigned initial condition $\bar{x}$ does exist and has been computed (offline). We consider the following assumption.

ASSUMPTION 1. *The optimal trajectory is such that the residual cost is strictly decreasing, i.e.,*

$$\sum_{k=h}^{N} g(y(k)) < \sum_{k=h+1}^{N} g(y(k)) \quad \forall h = 1, \ldots, N-1.$$

Assumption 1 is absolutely reasonable and avoids trivialities (it is obviously true, for instance, if g is positive definite with respect to y). The way we solve Problem 1 can be explained as follows: Based on points of the optimal trajectory and their opposite (connected by the solid line in Figure 1), we partition the state space into disjoint regions. The convex hull of the points of the optimal trajectory and their opposite (the shaded hexagon in Figure 1) represents a region that can be divided into simplices, in each of which the control is affine. This region includes the nominal initial state $\bar{x}$ (possibly in its interior). The external part is divided into cones, centered in the origin, and "truncated to keep the outer part," in each of which the controller is linear. The control is Lipschitz continuous. To formally state the main result we need to introduce some notations. The inequality $p \leq 0$, if p is a vector, has to be interpreted componentwise. Let us denote by $\bar{1}$ the vector (the dimension depending on the context) having all components equal to 1:

$$\bar{1} = [\ 1 \ \ 1 \ \ \ldots \ \ 1\ ]^T \tag{13}$$

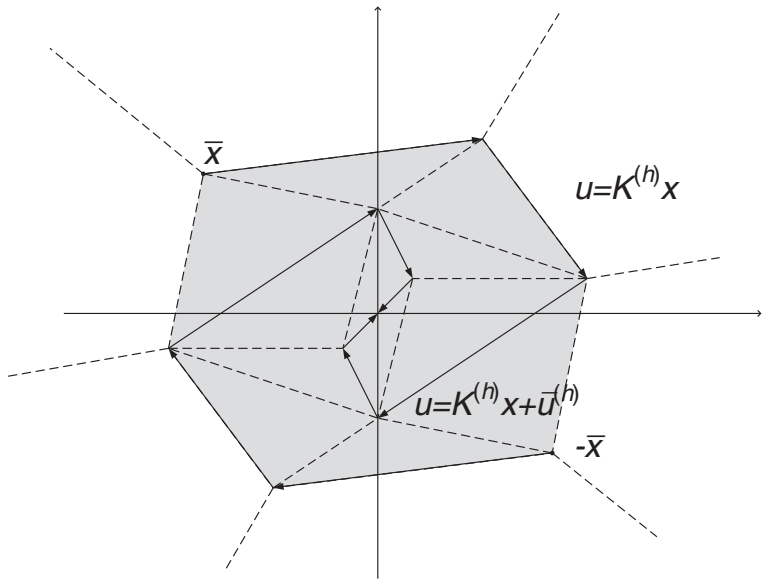


FIG. 1. *The state space partition.*

(note that the expression $\bar{1}^T p$ is the sum of the components of vector p). Given an $n \times (n+1)$ full row rank matrix X, a *simplex* in $\mathbb{R}^n$ is a set of the form

$$\mathcal{S}(X) = \{x = Xp : \quad p \geq 0, \quad \bar{1}^T p = 1\}. \tag{14}$$

Given an $n \times n$ full rank matrix X, a *simplicial cone* in $\mathbb{R}^n$ is a set of the form

$$\mathcal{C}(X) = \{x = Xp : \quad p \geq 0\}. \tag{15}$$

Note that a simplicial cone is always generated by a simplex having the origin among its corners. Together with these standard notations we need to consider the complement (the outer part) of the unit sector in a simplicial cone, which is the closure of the complement in $\mathcal{C}$,

$$\tilde{\mathcal{C}}(X) = \{x = Xp : \quad p \geq 0, \quad \bar{1}^T p \geq 1\}. \tag{16}$$

If $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a function and X is an $n \times m$ matrix, we denote by $\Phi(X)$ the following vector:

$$\Phi(X) = [\ \Phi(x_1) \quad \Phi(x_2) \quad \dots \quad \Phi(x_m)\ ]^T. \tag{17}$$

The next theorem states that there exists a control which is optimal for $x(1) = \bar{x}$ and locally stabilizing.

THEOREM 3.1. *There exists a convex and compact polyhedron $\mathcal{P}$, including the origin in its interior, which is partitioned into simplices $\mathcal{S}^{(h)}$, each generated by an $n \times (n+1)$ matrix $X^{(h)}$ whose columns are vectors properly chosen from among the states of the optimal trajectory and their opposite:*

$$\mathcal{P} = \bigcup \mathcal{S}^{(h)} = \bigcup \mathcal{S}(X^{(h)}) \tag{18}$$

such that each pair of simplices has an intersection with an empty interior,

$$\text{int}\{\mathcal{S}^{(h)} \bigcap \mathcal{S}^{(k)}\} = \emptyset, \quad h \neq k, \tag{19}$$

and such that $\bar{x} \in \mathcal{P}$. *To each simplex* $\mathcal{S}^{(h)}$ *we can associate an* $m \times (n+1)$ *matrix* $U^{(h)}$ *whose columns are vectors properly chosen from among the inputs of the optimal trajectory and their opposite. The piecewise-affine static controller*

$$u = \Phi_{\mathcal{P}}(x) = K^{(h)}x + \bar{u}^{(h)} = U^{(h)} \begin{bmatrix} X^{(h)} \\ \bar{1}^T \end{bmatrix}^{-1} \begin{bmatrix} x \\ 1 \end{bmatrix} \quad \textit{for} \;\; x \in \mathcal{S}^{(h)} \tag{20}$$

is Lipschitz continuous and relatively optimal inside $\mathcal{P}$*; more precisely it is stabilizing with a domain of attraction* $\mathcal{P}$ *and for* $x(1) = \bar{x}$ *produces the optimal trajectory. Moreover, for each* $x(1) \in \mathcal{P}$, *the constraints are satisfied and the transient cost is bounded as*

$$J(x(1)) \leq \max_{i=1,\ldots,n+1} J_{opt}(x_{k_i}), \tag{21}$$

where $x_{k_1}, \; x_{k_2}, \ldots, x_{k_{n+1}}$ *are the vertices of the simplex* $\mathcal{S} \ni x(1)$ *and* $J_{opt}(x_{k_i})$ *is the optimal cost associated with the initial condition* x_{k_i}.

The next theorem states that the same control can be globally extended over $\mathbb{R}^n$.

THEOREM 3.2. *The control* (20) *can be extended onto* $\mathbb{R}^n$ *as follows. The complement of the polytope* $\mathcal{P}$ *can be partitioned into complements of simplices inside a cone*

$$\tilde{\mathcal{C}}^{(h)} = \tilde{\mathcal{C}}(X^{(h)}), \tag{22}$$

each generated by a square invertible matrix $X^{(h)}$, *having intersection with empty interior*

$$\text{int}\{\tilde{\mathcal{C}}^{(h)} \bigcap \tilde{\mathcal{C}}^{(k)}\} = \emptyset, \quad h \neq k, \tag{23}$$

and intersection with empty interior with $\mathcal{P}$,

$$\text{int}\{\tilde{\mathcal{C}}^{(h)} \bigcap \mathcal{P}\} = \emptyset, \tag{24}$$

such that

$$\mathcal{P} \bigcup \left[\bigcup_h \tilde{\mathcal{C}}^{(h)} \right] = \mathbb{R}^n. \tag{25}$$

To each set $\tilde{\mathcal{C}}^{(h)}$ *can be associated an* $m \times n$ *matrix* $U^{(h)}$ *whose columns are vectors properly chosen from among the inputs of the optimal trajectory, obtaining a control*

$$u = \Phi(x) = K^{(h)}x = U^{(h)} \left[X^{(h)} \right]^{-1} x \; . \tag{26}$$

The extended control obtained in this way is globally Lipschitz continuous and relatively optimal.

Theorems 3.1 and 3.2 will be proved constructively in sections 4 and 5, respectively.

4. Construction of a local relatively optimal control. Denote by

$$\bar{x}(1), \dots, \bar{x}(N)$$

the optimal state trajectory from the initial condition $\bar{x} = \bar{x}(1)$, obtained by solving (5)–(12). We introduce the notation (basically inverting time)

$$x_1 = \bar{x}(N), \quad x_2 = \bar{x}(N-1), \dots, \quad x_N = \bar{x}(1), \tag{27}$$

and

$$u_1 = \bar{u}(N), \quad u_2 = \bar{u}(N-1), \dots, \quad u_N = \bar{u}(1), \tag{28}$$

and we coherently assume $x_0 = 0$; hence we have that $x_{i-1} = Ax_i + Bu_i$, $i = 1, \dots, N$. We also denote by x_{-i}, $i = 1, \dots, N$, the opposite of x_i. Then we introduce the following assumption, which simplifies considerably the proof of Theorem 3.1 but is not essential (in fact it can be easily removed as we will show later on).

ASSUMPTION 2. *The matrix $X_n = [x_1\ x_2\ \dots\ x_n]$, formed by the last n states of the optimal trajectory, is invertible.*

Let us consider the polyhedral set

$$\mathcal{P}_n = \{x = X_n p : \|p\|_1 \le 1\}. \tag{29}$$

Such a set is the convex hull of the last n states of the optimal trajectory and their opposite. It contains the origin in its interior and is 0-symmetric. An example for $n = 2$ is shown in Figure 2: $\mathcal{P}_n$ (the darkest area) is the convex hull of the last two states of the optimal trajectory (connected by the solid line) and their opposite (connected by the dashed line). Thanks to Assumption 2 the following lemma holds.

LEMMA 4.1. *The linear control*

$$u(x) = U_n X_n^{-1} x, \tag{30}$$

where $U_n = [u_1\ u_2\ \dots\ u_n]$, renders positively invariant the set $\mathcal{P}_n$ satisfying the constraints for all initial conditions inside the set. In particular, it is deadbeat and steers the state to zero in at most n steps.

Proof. The control law $u(x) = U_n X_n^{-1} x$ is a control-at-the-vertices strategy. All $x \in \mathcal{P}_n$ can be written in a unique way as a linear combination of the columns of X_n, namely, the last n states of the optimal trajectory:

$$x = X_n p. \tag{31}$$

Since X_n is invertible, it follows that

$$p(x) = X_n^{-1} x; \tag{32}$$

hence the control law $u(x) = U_n X_n^{-1} x$ basically computes a control which is a linear combination of the controls at the vertices of $\mathcal{P}_n$ according to the coefficients $p(x)$. Positive invariance is a consequence of the fact that, by construction, the control at each vertex keeps the state inside the set [4]. The satisfaction of the constraints is guaranteed for all initial conditions inside the set because the input and state constraints are convex and 0-symmetric. To prove that the control is deadbeat, note that if at time k we have

$$x(k) = x_n p_n + \dots + x_2 p_2 + x_1 p_1, \tag{33}$$

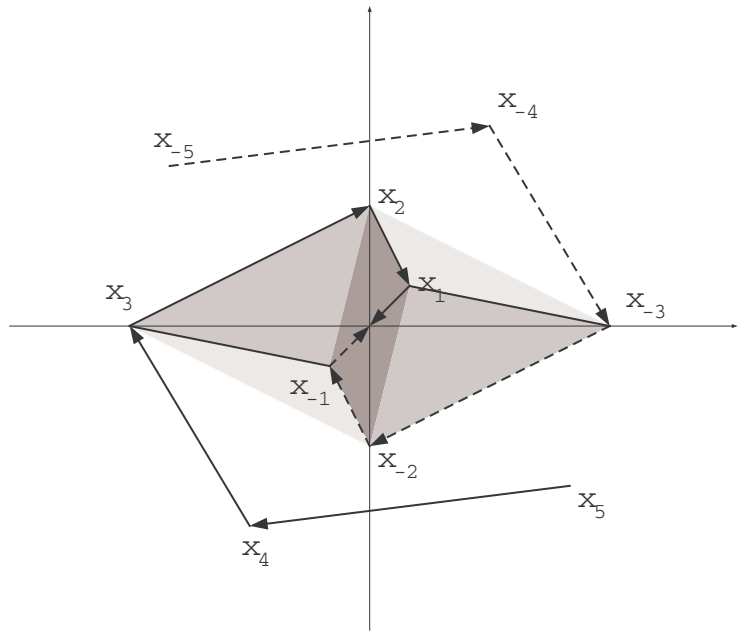


FIG. 2. *Considering x_3 and its opposite x_{-3}, we can construct four simplices starting from $\mathcal{P}_2$ (the darkest area).*

then the computed control will be

$$u(k) = u_n p_n + \cdots + u_2 p_2 + u_1 p_1. \tag{34}$$

Since $x_{i-1} = Ax_i + Bu_i$, we obtain, by linearity,

$$x(k+1) = x_{n-1} p_n + \cdots + x_1 p_2 + 0 p_1, \tag{35}$$

and at the next step we will have, reasoning in the same way,

$$x(k+2) = x_{n-2} p_n + \cdots + x_1 p_3 + 0 p_2 + 0 p_1, \tag{36}$$

and so on; therefore, we immediately verify that after at most n steps the system will reach the origin. □

Remark 4.1. The control law defined above is such that inside $\mathcal{P}_n$, at each step, the state is a convex combination of points with decreasing index and 0.

Note that if the system reaches the state $x_n = \bar{x}(N - n + 1) \in \mathcal{P}_n$, it starts following the last n steps of the optimal trajectory. Note also that $\mathcal{P}_n$ (which will be the first element of a sequence of sets) is affine to a diamond and thus can be partitioned into simplices. The next sets of the sequence are computed as follows.

Consider the state x_{n+1} (corresponding to the state x_3 in the example of Figure 2). Since x_{n+1} and its opposite $x_{-(n+1)}$ are outside $\mathcal{P}_n$ (as will be shown later), they can be connected with a certain number of vertices of $\mathcal{P}_n$ without crossing such a set; thus simplices are formed by some vertices of $\mathcal{P}_n$ and the two points x_{n+1} and $x_{-(n+1)}$ (in the example of Figure 2, such simplices are the triangles (x_3, x_2, x_{-1})

and (x_3, x_{-1}, x_{-2}) and their symmetric). Denoting by $\mathcal{S}^j_{n+1}$, $j = 1, \ldots, m_{n+1}$, the simplices having x_{n+1} as a vertex and by $\mathcal{S}^j_{n+1}$, $j = -m_{n+1}, \ldots, -1$, those having $x_{-(n+1)}$ as a vertex, we can define the set $\mathcal{P}_{n+1}$ as follows:

$$\mathcal{P}_{n+1} = \bigcup_{j=\pm 1,\ldots,\pm m_{n+1}} \mathcal{S}^j_{n+1} \cup \mathcal{P}_n. \tag{37}$$

The procedure goes on exactly in the same manner to generate the sequence of sets $\mathcal{P}_k$, $k = n+1, n+2, \ldots, N$, ordered by inclusion and the corresponding simplicial partition: If we define the set

$$\mathcal{P}_k = \operatorname{conv}\{x_1, x_2, \ldots, x_k, x_{-1}, x_{-2}, \ldots, x_{-k}\}, \quad k < N, \tag{38}$$

we can consider the vector x_{k+1} and form a new simplicial partition for $\mathcal{P}_{k+1}$ by adding new simplices. It is fundamental to note that each new simplicial partition of $\mathcal{P}_{k+1}$ preserves all the simplices forming the simplicial partition for $\mathcal{P}_k$. To prove that the construction is well defined we need the following lemma.

LEMMA 4.2. *The vector x_{k+1} in the construction is outside $\mathcal{P}_k$.*

Proof. Assume by contradiction that $x_{k+1} \in \mathcal{P}_k$. Then x_{k+1} could be written as a convex combination of the vertices of $\mathcal{P}_k$. So if we take x_{k+1} as an initial state, since we considered convex constraints, then x_{k+1} could be driven to zero in a time not exceeding k at a cost not exceeding the maximum cost of all vertices of $\mathcal{P}_k$. This is in contradiction with Assumption 1. □

Therefore the procedure is such that $\{\mathcal{P}_k\}$ is a strictly increasing (in the sense of inclusion) sequence of sets, each of which preserves the simplicial partition of the former. This construction terminates once $\mathcal{P} \doteq \mathcal{P}_N$ is constructed.

Note that the innermost set $\mathcal{P}_n$ can be partitioned into simplices $\mathcal{S}^j_n$, each having the origin as a vertex. The remaining n vertices are any independent subset of $x_{\pm k}$, $k = 1, \ldots, n$, namely, the last n steps of the optimal trajectory and their opposite. It is easy to recognize that in this case the control law (30) takes the form (20), i.e.,

$$u = \Phi(x) = K^j_n x = U^j_n \left[X^j_n\right]^{-1} x\ , \tag{39}$$

where j denotes the simplex, while X^j_n and U^j_n are matrices whose columns are a subset of the states of the optimal trajectory (and their opposite) and the corresponding optimal control values (and their opposite).

The next step is to show how to associate a control with each simplex. With each of the simplices $\mathcal{S}^j_k$,

1. associate a matrix X^j_k whose columns are the vertices (in arbitrary order).
2. associate a matrix U^j_k whose columns are the controls corresponding to the vertices (in the same order as they appear in X^j_k). If the vertex belongs to the optimal trajectory, take the corresponding control; if it belongs to the opposite of the optimal trajectory, take the opposite of the corresponding control.

Now, the control strategy is as follows: Given $x \in \mathcal{P}$, if $x \in \mathcal{S}^j_k$, then

$$\Phi_{\mathcal{P}}(x) = U^j_k p, \tag{40}$$

where $p \geq 0$ is the (unique) vector such that

$$x = X^j_k\ p, \quad \bar{1}^T p = 1. \tag{41}$$

Note that p is such that

$$\begin{bmatrix} X_k^j \\ \bar{1}^T \end{bmatrix} p = \begin{bmatrix} x \\ 1 \end{bmatrix}, \tag{42}$$

so that u is of the form (20).

Remark 4.2. Given a simplex $\mathcal{S}_k^j$, the vector p of (42) is the vector of the barycentric (normalized) coordinates of x with respect to the vertices of the simplex (the columns of X_k^j). Barycentric coordinates, as well as simplicial partitions, are well known in the context of finite element analysis [7].

To show the properties of the control we need to introduce the index $\mathrm{In}(\mathcal{S})$ of a sector $\mathcal{S}$ as the maximum of the absolute values of the indices of its generating vectors. Formally, if $\mathcal{S}$ is generated by corners $x_{k_1}, x_{k_2}, \ldots, x_{k_n}$, then

$$\mathrm{In}(\mathcal{S}) = \max\{|k_1|, |k_2|, \ldots, |k_n|\}. \tag{43}$$

For reasons that will be clear soon, $\mathrm{In}(\mathcal{S})$ will be referred to as the *distance* of $\mathcal{S}$ from 0.

Remark 4.3. The notion of "sector" deserves some comment. Indeed, we now consider possibly degenerate simplices that can have an empty interior formed by some points x_k and with the origin repeatedly considered. For instance, $\mathcal{S}$ could be generated by $[0\ 0\ x_1\ x_2]$, representing a two-dimensional degenerate simplex in the three-dimensional space. Note also that $\mathrm{In}(\mathcal{S}) \leq k$ for all sectors inside $\mathcal{P}_k$.

The next lemma shows that, with the proposed control, if the system state is inside a sector, then it jumps to another one closer to zero.

LEMMA 4.3. *The proposed strategy is such that if $x \in \mathcal{S}$, a sector of $\mathcal{P}_k$, then $Ax + Bu(x) \in \mathcal{S}'$ with*

$$In(\mathcal{S}') < In(\mathcal{S}), \tag{44}$$

as long as $In(\mathcal{S}) \neq 0$, and therefore if $x(1) \in \mathcal{P}_k$, the control steers the system to zero in at most k steps.

Proof. As a first step we note that, according to Lemma 4.1 and Remark 4.1, the jump to sector closer to 0 occurs for all $x \in \mathcal{P}_n$. Now we proceed by induction. Assume that $x \in \mathcal{P}_{n+1}$. If $x \in \mathcal{P}_n$, there is nothing to prove; otherwise x is necessarily in a sector $\mathcal{S}$ generated by x_{n+1} or its opposite $x_{-(n+1)}$ and other vertices of smaller indices

$$x = \sum_{i=1}^{n+1} x_{k_i} p_i, \quad \sum_{i=1}^{n+1} p_i = 1, \quad p_i \geq 0, \tag{45}$$

with $|k_i| \leq n$, $i = 1, 2, \ldots, n$, and $|k_{n+1}| = n + 1$. Then we have, by construction,

$$Ax + B\Phi_{\mathcal{P}}(x) = A\left[\sum_{i=1}^{n+1} x_{k_i} p_i\right] + B\left[\sum_{i=1}^{n+1} u_{k_i} p_i\right] = \sum_{i=1}^{n+1} p_i \underbrace{[Ax_{k_i} + Bu_{k_i}]}_{\in \mathcal{P}_n} \in \mathcal{P}_n. \tag{46}$$

Therefore, necessarily $Ax + B\Phi_{\mathcal{P}}(x)$ is in a sector with index $\mathrm{In} \leq n$. The rest of the proof proceeds in the same way. Any point x in $\mathcal{P}_{k+1}$, if not in $\mathcal{P}_k$, is included in a sector $\mathcal{S}$ with index $\mathrm{In}(\mathcal{S}) = k + 1$ and, by means of the same machinery, we can show that $Ax + B\Phi_{\mathcal{P}}(x) \in \mathcal{S}'$ with $\mathrm{In}(\mathcal{S}') \leq k$. The fact that if $x(1) \in \mathcal{P}_k$, the state converges to 0 in at most k steps is an immediate consequence. ∎

The procedure for partitioning the state space and constructing the region $\mathcal{P}_N \supset \mathcal{P}_n$ and the associated local controller can be summarized as follows.

PROCEDURE 4.1. *We give the system* (1) *and the optimal open-loop trajectory, computed by solving* (5)–(12)*, which satisfies Assumption* 2.

1. *Let the set* $\mathcal{P}_n = \{x : x = X_n p, \|p\|_1 \leq 1\}$*, where* $X_n = [x_1\ x_2\ \dots\ x_n]$*, be the convex hull of the last* n *states of the optimal trajectory and their opposite.*
2. *Let* $U_n = [u_1\ u_2\ \dots\ u_n]$ *be the matrix whose columns are the control vectors corresponding to the last* n *states of the optimal trajectory.*
3. *Take* $i = n + 1$.
4. *Construct the simplices* $\mathcal{S}_i^j,\ j = \pm 1, \dots, \pm m_i$*, by connecting* x_i *and* x_{-i} *to the vertices of* $\mathcal{P}_{i-1}$ *without crossing such a set. This is always possible since* $x_i, x_{-i} \notin \mathcal{P}_{i-1}$.
5. *Let* X_i^j *be the matrix whose columns are the vertices of* $\mathcal{S}_i^j$ *in an arbitrary order and* U_i^j *be the controls corresponding to the vertices in the same order. For vertices belonging to the opposite of the optimal trajectory, take the opposite of the control.*
6. *Let* $\mathcal{P}_i = \bigcup_j \mathcal{S}_i^j \cup \mathcal{P}_{i-1}$.
7. *Increase* i.
8. *If* $i \leq N$*, go back to step* 4.

Note that, by construction, the sets $\mathcal{P}_i$, $i = n, \dots, N$, are convex, 0-symmetric, and such that $\mathcal{P}_i \subset \mathcal{P}_{i+1}$. The set $\mathcal{P}_{i+1} \setminus \mathcal{P}_i$, the difference between $\mathcal{P}_{i+1}$ and $\mathcal{P}_i$, is composed of simplices S_i^j, each of which has all vertices but one (precisely x_{i+1} or $x_{-(i+1)}$) belonging to $\mathcal{P}_i$.

In order to prove Theorem 3.1 we must provide the following lemmas.

LEMMA 4.4. *The proposed control* $\Phi_{\mathcal{P}}(x)$ *is Lipschitz continuous inside* $\mathcal{P} = \mathcal{P}_N$.

Proof. Since the cardinality of the partition is finite, it is sufficient to prove continuity. Inside each of the simplices, the control action (40) is a linear combination of the control at each vertex, with the weights being the components of p. Now, the proof follows immediately from a well-known result of finite element analysis, namely, the fact that using barycentric coordinates as weights in a triangular (or, in general, simplicial) mesh guarantees interelement continuity [7]. ∎

LEMMA 4.5. *For any state* $x(1) = x \in \mathcal{S} \subset \mathcal{P}_N$ *the proposed control* $\Phi_{\mathcal{P}}(x)$ *satisfies the constraints and it ensures a cost* $J(x)$ *bounded as*

$$J(x) \leq \max_{i=1,\dots,n+1} J_{opt}(x_{k_i}), \tag{47}$$

where $\mathcal{S}$ *is generated by the points* $x_{k_1},\ x_{k_2}, \dots, x_{k_{n+1}}$.

Proof. The constraints are convex and 0-symmetric and, by construction, they are satisfied by each of the vertices of the convex set $\mathcal{P}_N$. Hence they are satisfied by any state belonging to $\mathcal{P}_N$. Since $\mathcal{P}_N$ is positively invariant under the control law $\Phi_{\mathcal{P}}(x)$, any trajectory originating in $\mathcal{P}_N$ satisfies the constraints. It follows from Lemmas 4.2 and 4.3 that the cost achieved from a given initial condition x is bounded by the maximum cost achieved from the vertices of the sector $\mathcal{S} \ni x$. Consider the cost function $\hat{g}(x) = g(x, \Phi_{\mathcal{P}}(x))$. Since $\hat{g}(x)$ is convex and 0-symmetric, it is easy to recognize that

$$\hat{g}(x) \leq \hat{g}(x_{In(\mathcal{S}(x))}), \tag{48}$$

where $\mathcal{S}(x)$ denotes the sector $\mathcal{S} \ni x$ and $x_{In(\mathcal{S}(x))}$ belongs to the optimal trajectory. From Lemma 4.2 and from the fact that $\hat{g}(x)$ is convex and 0-symmetric, it follows

that

$$\hat{g}(x_i) \leq \hat{g}(x_j) \tag{49}$$

for $0 < i < j \leq N$. Therefore, the maximum in the right-hand side of (47) is obtained for $k_i = \mathrm{In}(\mathcal{S}(x))$, i.e., it is the (optimal) cost from the vertex $x_{In(\mathcal{S}(x))}$. Let us now compare the cost achieved from x to that achieved from $x_{In(\mathcal{S}(x))}$. We recall that the control law $\Phi_{\mathcal{P}}(x)$ steers the system from $x \in \mathcal{S}(x) \subset \mathcal{P}_N$ to zero in at most $In(\mathcal{S}(x))$ steps. Denoting

$$f(x) = Ax + B\Phi_{\mathcal{P}}(x) \tag{50}$$

and

$$f^i(x) = f(f(\dots f(x)\dots)), \tag{51}$$

we can rewrite (48) as

$$\hat{g}(f^i(x)) \leq \hat{g}(x_{In(\mathcal{S}(f^i(x)))}), \tag{52}$$

for all $i = 0, \dots, \mathrm{In}(\mathcal{S}(x)) - 1$. On the other hand, Lemma 4.3 states that the sequence of indices corresponding to a trajectory originating in $\mathcal{P}_N$ is strictly decreasing. It follows that

$$\mathrm{In}(\mathcal{S}(f^i(x))) \leq \mathrm{In}(\mathcal{S}(x)) - i \tag{53}$$

for all $i = 0, \dots, \mathrm{In}(\mathcal{S}(x)) - 1$. Therefore, from (49) we can write

$$\hat{g}(x_{In(\mathcal{S}(f^i(x)))}) \leq \hat{g}(x_{In(\mathcal{S}(x))-i}) \tag{54}$$

and, by (52),

$$\hat{g}(f^i(x)) \leq \hat{g}(x_{In(\mathcal{S}(x))-i}). \tag{55}$$

Finally, by summing over $i = 0, \dots, \mathrm{In}(\mathcal{S}(x)) - 1$, we obtain

$$J(x) \leq J(x_{In(\mathcal{S}(x))}). \qquad \square \tag{56}$$

Now we show how to remove Assumption 2. If Assumption 2 does not hold, the construction of the regions is basically the same. The only difference is that now we must start the construction from the beginning (i.e., $\mathcal{P}_1$, $\mathcal{P}_2, \dots$) until we construct the region $\mathcal{P}_r$, where $r > n$ is the smallest value for which $[x_1\ x_2\ \dots\ x_r]$ has full rank (and then $\mathcal{P}_r$ is a neighborhood of the origin). In forming the sets $\mathcal{P}_k$, $k < r$, we construct a partition of "degenerate polytopes" in subspaces having the same properties mentioned above. When we add the vertex x_r (and its opposite $-x_r$), we reach full dimension and can construct a (nondegenerate) simplex partition of $\mathcal{P}_r$ in which each simplex has x_r as a vertex. Then the construction proceeds as already mentioned, with the difference that the control is not ultimately linear since, in general, we cannot associate a linear control with $\mathcal{P}_r$. If such an r does not exist (i.e., the whole optimal trajectory belongs to a proper subspace of $\mathbb{R}^n$), we can extend the trajectory backward, i.e., adding points x_{N+1}, x_{N+2} to reach the full rank. Clearly, optimality is ensured only from $\bar{x} = x_N$. Using the same trick of extending the trajectory backward, we can arbitrarily enlarge the domain of attraction.

We are now in the position of proving relative optimality with local stability of the control.

Proof of Theorem 3.1. The constructed simplicial partition and the control are of the form (20), which is Lipschitz as proved in Lemma 4.4. If we assume $x(1) = \bar{x}$, then the trajectory is the optimal constrained one by construction. The fact that the control is stabilizing follows from Lemma 4.3. The satisfaction of the constraints and the cost bound (21) follow easily from Lemma 4.5. □

TABLE 1
Upper bound for the number of simplices given the number N of steps of the optimal trajectory and the order n of the system.

N, n	3	4	8	12	16
4	33	39	-	-	-
8	133	207	1425	-	-
12	297	503	11965	54257	-
16	525	927	47497	592013	$2.1\ 10^6$
20	817	1479	132085	$3.2\ 10^6$	$2.8\ 10^7$

An important question is whether the complexity of the controller (i.e., the number of simplices obtained by partitioning the state space according to Procedure 4.1) is known in advance. For $n = 2$, the number of simplices (triangles) is exactly $4N - k$, where k is the number of the vertices of the convex hull of the optimal trajectory and its opposite [12]. For $n > 2$, since such simplices form a *triangulation* [8] of a point set, their number N_s is bounded according to the expression [13]

$$N_s \leq \begin{pmatrix} 2N + 2 - \left\lceil \frac{n+1}{2} \right\rceil \\ \left\lfloor \frac{n+1}{2} \right\rfloor \end{pmatrix} + \begin{pmatrix} 2N + 1 - \left\lceil \frac{n}{2} \right\rceil \\ \left\lfloor \frac{n}{2} \right\rfloor \end{pmatrix} - (n+1), \tag{57}$$

where $\lfloor x \rfloor$ denotes the maximum integer less than or equal to x, $\lceil x \rceil$ denotes the minimum integer greater than or equal to x, and $\binom{a}{b}$ denotes the binomial coefficient. Table 1 reports such an upper bound for some pairs of N and n. Upper bound (57) resembles that provided in [2], in the context of the explicit linear $1/\infty$-norm regulator for constrained systems, with the substantial difference that *for the static ROC the upper bound does not depend on the number of constraints*, since the controller is computed based on the optimal trajectory only. In other words, the number of constraints does not influence directly the complexity of the controller.

Remark 4.4. As shown above, the convexity of the constraints and the cost guarantees that

$$x_i \notin \mathcal{P}_{i-1} \ \forall i = n+1, \ldots, N. \tag{58}$$

However, as long as condition (58) on the optimal trajectory is satisfied, the ROC can be constructed independently of the convexity of the optimization problem. In other words, a sufficient condition for constructing the static ROC is that each of the points of the optimal trajectory does not belong to the convex hull of the subsequent points and their opposite. Obviously, the satisfaction of the constraints is guaranteed for all the trajectories originating in $\mathcal{P} = \mathcal{P}_N$ only if the constraints are convex and 0-symmetric.

5. Construction of a global controller. For $x \in \mathcal{P} = \mathcal{P}_N$, the controller described above is a solution for Problem 1. However, the control law is not defined for $x \notin \mathcal{P}$. A possible way to extend the control outside $\mathcal{P}$ is to "immerse" $\mathcal{P}$ in

the maximal invariant set $\mathcal{X}_{max}$, namely, the set of all states which can be brought to the origin in finitely many steps without state or input constraint violations (note that $\mathcal{P} \subseteq \mathcal{X}_{max}$). Then, for $x \notin \mathcal{P}$, one can apply the control law derived from $\mathcal{X}_{max}$ (many algorithms have been proposed to find $\mathcal{X}_{max}$ and an associated control law; see, for example, [9]). By definition, the constraints are satisfied and the convergence is guaranteed if and only if $x(1) \in \mathcal{X}_{max}$.

A different strategy can be derived as the *natural extension of the controller computed within* $\mathcal{P}$. In this way we have basically two advantages:

- the obtained controller is globally Lipschitz;
- the state behavior outside the set $\mathcal{P}$ resembles the internal one and therefore the system performs "reasonably well" outside $\mathcal{P}$.

The set $\mathcal{P}$ is a polytope including the origin in its interior. This means that the state space can be partitioned in simplicial cones, each having a center in the origin and generated by n vertices of $\mathcal{P}$. These cones $\mathcal{C}^{(h)}$ have a nonempty interior, have intersections with empty interior, and cover $\mathbb{R}^n$:

$$\text{int}\{\mathcal{C}^{(h)}\} \neq \emptyset, \tag{59}$$

$$\text{int}\{\mathcal{C}^{(h)} \bigcap \mathcal{C}^{(k)}\} = \emptyset, \quad h \neq k, \tag{60}$$

$$\bigcup_h \mathcal{C}^{(h)} = \mathbb{R}^n. \tag{61}$$

For each cone generated by a square matrix $X^{(h)}$, we consider the complement with respect to $\mathcal{P}$,

$$\tilde{\mathcal{C}}^{(h)}; \tag{62}$$

therefore the union of the complements and the simplices forming $\mathcal{P}$ cover $\mathbb{R}^n$. For each cone generated by an invertible $X^{(h)}$ we consider the corresponding input matrix $U^{(h)}$[1] and the control

$$\tilde{\Phi}(x) = U^{(h)}[X^{(h)}]^{-1}x. \tag{63}$$

Such a control is Lipschitz [3].

In principle, continuity is not an issue in discrete-time systems. In practice, it avoids chattering. Thus we state the next lemma.

LEMMA 5.1. *Consider the following extension outside* $\mathcal{P}$ *of the control* $\Phi_{\mathcal{P}}(x)$*:*

$$\Phi(x) \doteq \begin{cases} \Phi_{\mathcal{P}}(x) & \textit{for } x \in \mathcal{P}, \\ \tilde{\Phi}(x) & \textit{for } x \notin \mathcal{P}. \end{cases} \tag{64}$$

Such a control is globally Lipschitz.

Proof. Since the control is piecewise affine and the cardinality of the partition is finite, we need only prove global continuity. Then the Lipschitz constant for each component of Φ is given by the maximum value of the norm of its gradient. Note also that $\Phi_{\mathcal{P}}$ is continuous inside $\mathcal{P}$ and $\tilde{\Phi}$ is continuous outside $\mathcal{P}$. As a consequence, we need only prove that the extended control is continuous in $\partial\mathcal{P}$. Consider $\hat{x} \in (\mathcal{P} \cap \tilde{\mathcal{C}}^{(h)}) \subset \partial\mathcal{P}$. By construction, $\tilde{\mathcal{C}}^{(h)}$ and $\mathcal{P}$ have a facet in common, precisely

[1]The matrix whose columns are the (optimal) control vectors associated with the columns in $X^{(h)}$, elements of the optimal trajectory.

the facet whose vertices are the generator vectors of $\tilde{\mathcal{C}}^{(h)}$. Since $\hat{x}$ lies in the common facet, it can be expressed as a linear combination of those vectors in a unique way. From (40) and (63) it follows that $\Phi_{\mathcal{P}}(\hat{x})$ and $\tilde{\Phi}(\hat{x})$ are the linear combination of the controls associated with the same vectors according to the same coefficients; then $\tilde{\Phi}(\hat{x}) = \Phi_{\mathcal{P}}(\hat{x})$ for all $\hat{x} \in \mathcal{P} \cap \tilde{\mathcal{C}}^{(h)}$, i.e., the extended control is continuous in $\partial\mathcal{P}$. □

Now the problem is to show that the extended control $\Phi(x)$ is globally stabilizing. Then we consider as a candidate Lyapunov function the Minkowski function of $\mathcal{P}$, that is, the norm whose unit ball is $\mathcal{P}$:

$$\Psi(x) = \min\{\lambda \geq 0 : \quad x \in \lambda\mathcal{P}\}. \tag{65}$$

We have the following preliminary lemma.

LEMMA 5.2. *The function $\Psi(x(k))$ is nonincreasing as long as $x(k) \notin \mathcal{P}$.*

Proof. Since $x(k+1) = Ax + B\Phi(x(k))$, we must prove that

$$\Psi(Ax + B\Phi(x)) \leq \Psi(x) \quad \forall x \notin \mathcal{P}. \tag{66}$$

As shown above, the extended control $\Phi(x)$ is globally continuous. Furthermore, it is linear inside each of the $\tilde{\mathcal{C}}^{(h)}$. As a consequence, outside the interior of $\mathcal{P}$, the control can be expressed as

$$\Phi(x) = \Psi(x)\Phi_{\mathcal{P}}(\bar{x}), \quad x \notin \text{int}(\mathcal{P}), \tag{67}$$

where the vector

$$\bar{x} = \frac{x}{\Psi(x)} \tag{68}$$

belongs to the boundary of $\mathcal{P}$. Consider a generic $x \notin \mathcal{P}$ and its "projection" $\bar{x}$ onto $\partial\mathcal{P}$. Since $\Phi_{\mathcal{P}}$ renders invariant the set $\mathcal{P}$, it follows that

$$A\bar{x} + B\Phi_{\mathcal{P}}(\bar{x}) \in \mathcal{P} \tag{69}$$

and, multiplying by $\Psi(x)$,

$$A\Psi(x)\bar{x} + B\Psi(x)\Phi_{\mathcal{P}}(\bar{x}) \in \Psi(x)\mathcal{P}. \tag{70}$$

From (67) and (70) and by substituting $x = \Psi(x)\bar{x}$ we obtain

$$Ax + B\Phi(x) \in \Psi(x)\mathcal{P}, \tag{71}$$

which implies, by the definition of the Minkowski function, that

$$\Psi(Ax + B\Phi(x)) \leq \Psi(x), \quad x \notin \mathcal{P}. \qquad \square \tag{72}$$

Lemma 5.2 proves the boundedness of the state but not convergence to 0. To prove convergence, since the function $\Psi(x)$ is only nonincreasing, we must use a trick. Define

$$x(k+1) = f(x) = Ax + B\Phi(x) \tag{73}$$

and consider the N steps forward system defined as the composition of f:

$$x(k+N) = f^N(x(k)) = f(f(\dots f(x)\dots)) \doteq F(x). \tag{74}$$

By means of this system we can show the following.

LEMMA 5.3. *The function* $\Psi(x(k))$*, as long as* $x(k) \notin int(\mathcal{P})$*, is strictly decreasing along the trajectory of the system* (74)*, precisely*

$$\Delta\Psi(x) \doteq \Psi(F(x)) - \Psi(x) < 0 \quad \forall x \notin \mathrm{int}(\mathcal{P}). \tag{75}$$

Proof. Consider a generic $x \notin \mathrm{int}(\mathcal{P})$ and its projection $\bar{x}$ onto $\partial\mathcal{P}$. As a first step we observe that there exists $1 \le h \le N$ such that

$$\bar{x}, f(\bar{x}), \dots, f^{h-1}(\bar{x}) \in \partial\mathcal{P}, \tag{76}$$

$$f^h(\bar{x}) \in \mathrm{int}(\mathcal{P}). \tag{77}$$

This is an immediate consequence of the fact that the control steers the system to zero in at most N steps starting from any $x \in \mathcal{P}$ and, in particular, from any $\bar{x} \in \partial\mathcal{P}$. By definition, we can express $x \notin \mathrm{int}(\mathcal{P})$ as the product of its projection $\bar{x}$ onto $\partial\mathcal{P}$ and the Minkowski function $\Psi(x)$:

$$x = \bar{x}\Psi(x). \tag{78}$$

By substituting in $f(x)$ we get

$$f(x) = Ax + B\Phi(x) = A\Psi(x)\bar{x} + B\Psi(x)\Phi_{\mathcal{P}}(\bar{x}) = \Psi(x)(A\bar{x} + B\Phi_{\mathcal{P}}(\bar{x})) = \Psi(x)f(\bar{x}), \tag{79}$$

where the last equality holds since $\Phi_{\mathcal{P}}(\bar{x}) = \Phi(\bar{x})\ \forall \bar{x} \in \partial\mathcal{P}$. Similarly it can be shown that $f^i(x) = \Psi(x)f^i(\bar{x})\ \forall i = 2, \dots, N$. Now, by multiplying (76) and (77) by $\Psi(x)$ it follows that

$$x, f(x), \dots, f^{h-1}(x) \in \partial(\Psi(x)\mathcal{P}), \tag{80}$$

$$f^h(x) \in \mathrm{int}(\Psi(x)\mathcal{P}); \tag{81}$$

then we obtain

$$\Psi(f^h(x)) < \Psi(x). \tag{82}$$

Thanks to Lemma 5.2, during the next $N - h$ steps, the state cannot escape from the region $\mathrm{int}(\Psi(x)\mathcal{P})$; hence

$$\Psi(F(x)) = \Psi(f^N(x)) \le \Psi(f^h(x)). \tag{83}$$

Finally, from these and (82) it follows that

$$\Psi(F(x)) < \Psi(x). \qquad \square \tag{84}$$

Now we are in the position of proving global stability.

Proof of Theorem 3.2. The considered control is Lipschitz and piecewise affine. We need to prove global asymptotic stability. To prove this fact we show that for any initial state $x(1) = x^* \notin \mathcal{P}$ there exists a finite M such that $x(M) \in \mathcal{P}$. Once $\mathcal{P}$ is reached, the state converges to zero as already proved.

This requires standard Lyapunov arguments. Indeed, the composed function $F(x)$ and the candidate Lyapunov function $\Psi(x)$ are continuous, and thus the function $\Delta\Psi(x)$ is continuous. Consider the compact set

$$\mathcal{H} = \{x :\ 1 \le \Psi(x) \le \Psi(x^*)\}. \tag{85}$$

In such a set, $\Delta\Psi(x)$ admits a negative maximum $-\mu$ with $\mu > 0$. Then we have

$$\Psi(x(k+N)) - \Psi(x(k)) = \Delta\Psi(x(k)) \leq -\mu. \tag{86}$$

This means that

$$\Psi(x(hN)) \leq \Psi(x^*) - h\mu. \tag{87}$$

But this means that, in finite time, the condition $\Psi(x(k)) < 1$ occurs, and therefore $x(k)$ reaches $\mathcal{P}$. □

Remark 5.1. The constraint (11) may be relaxed as follows:

$$x(N+1) \in \mathcal{X}_{fin}, \tag{88}$$

where $\mathcal{X}_{fin}$ is a 0-symmetric controlled-invariant polyhedron (that is, there exists a local control that renders $\mathcal{X}_{fin}$ positively invariant and such that the constraints are satisfied for all initial conditions inside the set). Then one can construct the ROC by positing $\mathcal{P}_n = \mathcal{X}_{fin}$ and following steps 3–7 of Procedure 4.1. As a result, a dual control strategy may be adopted: Apply the control $\Phi(x)$ for $x(k) \notin \mathcal{X}_{fin}$ and switch to the local control as soon as the condition $x(k) \in \mathcal{X}_{fin}$ is satisfied.

6. Example. Consider the double integrator

$$x(k+1) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} x(k) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(k), \tag{89}$$

under the constraints $|x(k)| \leq 5, |u(k)| \leq 3$. Given the initial state $x(1) = [-2 \quad 5]^T$, the horizon $N = 5$, the final state $x(N+1) = 0$, and the cost function $J = \sum_{i=1}^{N} u(k)^2$, the optimal (open-loop) control and trajectory, found by solving a quadratic-programming problem are, respectively,

$$\bar{U} = \begin{bmatrix} -3 & -2.9 & -1.3 & 0.3 & 1.9 \end{bmatrix} \tag{90}$$

and

$$\bar{X} = \begin{bmatrix} -2 & 3 & 5 & 4.1 & 1.9 \\ 5 & 2 & -0.9 & -2.2 & -1.9 \end{bmatrix}. \tag{91}$$

The optimal trajectory is reported in Figure 3. By means of Procedure 4.1, the triangulation reported in Figure 4 is obtained; the number of triangles is 12 (including the four triangles in which the darkest region, i.e., $\mathcal{P}_2$, can be split). The piecewise-affine control law obtained by applying a control-at-the-vertices strategy inside each of the triangles, as stated above, is relatively optimal, and hence is optimal for the nominal initial condition and guarantees convergence and constraint satisfaction for the other initial conditions. In Figure 4, the trajectories from three nonnominal initial conditions are reported. Note that the number of steps required to reach the origin depends on the triangle to which the initial state belongs. Figure 5 shows the effectiveness of the *extended* control, reporting some trajectories starting from outside $\mathcal{P}$; the dash-dotted lines represent the boundaries between the simplicial cones $\mathcal{C}^{(h)}$ in the complement of $\mathcal{P}$.

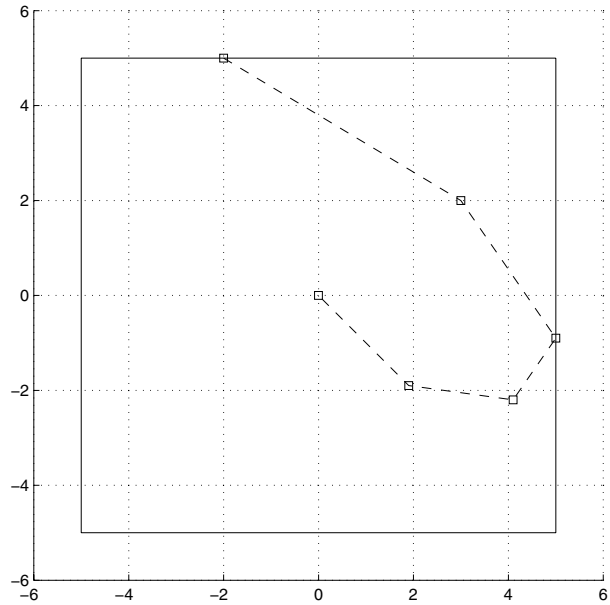


FIG. 3. *The optimal trajectory.*

7. Comparison with the dynamic ROC. Some significant differences between the dynamic ROC [5, 6] and the static one are highlighted in the following points:

1. Since the static ROC is nonlinear, the trajectory originating from $\lambda\bar{x}$ is not proportional, in general, to the one originating from $\bar{x}$ as with the dynamic ROC. However, by construction, opposite initial conditions generate opposite trajectories.
2. The dynamic ROC allows for the optimization from a set of n linearly independent initial conditions, while the static version described in this paper is thought of for a *single initial condition*. Extending the results to more than one initial condition for the static ROC is a matter of further investigation.
3. The dynamic ROC cannot guarantee the satisfaction of the constraints for initial conditions different from the nominal one. Hence it is suitable only in those cases when the constraints are actually *soft* constraints (constraints whose violation causes a performance loss only). On the contrary, by immersing the set $\mathcal{P}$ in the maximal invariant set as outlined in the beginning of section 5, the piecewise-affine solution can deal effectively with *hard* constraints.
4. The dynamic ROC is a linear system of order $N - n$, and hence is of low complexity. By looking at Table 1 it is clear that the complexity of the static ROC is much higher and depends strongly on the dimension of the state space (although the table shows only an upper bound for the number of simplices). As a consequence, the implementation of the static ROC may be difficult for high-order systems. However, almost all the approaches based on partitions of the state space are prone to this problem.

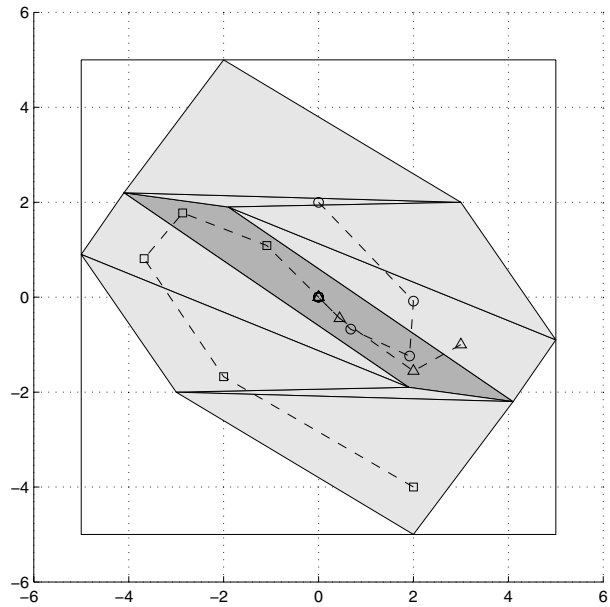


FIG. 4. *The triangulation induced by the optimal trajectory and the trajectories from three nonnominal initial conditions inside* $\mathcal{P}$.

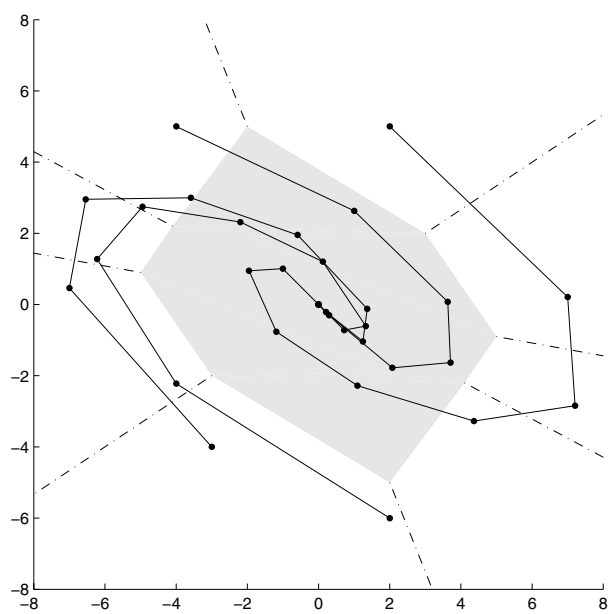


FIG. 5. *Some trajectories starting from outside* $\mathcal{P}$.

8. Conclusions. In this paper, a static version of the relatively optimal control (ROC) [5, 6] is proposed. The proposed controller is a deadbeat piecewise-affine state-feedback controller, based on a triangulation of the points of the optimal trajectory (computed offline). An upper bound on the number of polyhedral sets (whose vertices are the states of the optimal trajectory and their opposite) as a function of the order of the system and the length of the optimal trajectory is provided. The control at each vertex is the corresponding control vector of the optimal sequence, while the control at a generic state is given by a convex combination of the controls corresponding to the vertices of the set to which the state belongs. By removing the zero state terminal constraint and requiring the final state of the optimal trajectory to belong to a controlled invariant set, it is possible to obtain a nondeadbeat controller. The proposed control can deal effectively with hard constraints (a significant advantage with respect to the dynamic one previously introduced). We point out that the 0-symmetry of the constraint set is not necessary for the construction of the ROC, namely, for solving Problem 1; however, it guarantees the additional property that the constraints are satisfied for all initial conditions inside $\mathcal{P}_N$. Further work includes extending the results to more than one initial condition and exploiting the particular structure of the triangulation in order to obtain a tighter bound on the number of simplices.

REFERENCES

[1] A. Bemporad, M. Morari, V. Dua, and E. N. Pistikopoulos, *The explicit linear quadratic regulator for constrained systems*, Automatica J. IFAC, 38 (2002), pp. 3–20.

[2] A. Bemporad, F. Borrelli, and M. Morari, *Model predictive control based on linear programming—the explicit solution*, IEEE Trans. Automat. Control, 47 (2002), pp. 1974–1985.

[3] F. Blanchini, *Nonquadratic Lyapunov functions for robust control*, Automatica J. IFAC, 31 (1995), pp. 451–561.

[4] F. Blanchini, *Set invariance in control*, Automatica J. IFAC, 35 (1999), pp. 1747–1767.

[5] F. Blanchini and F. A. Pellegrino, *Relatively optimal control and its linear implementation*, IEEE Trans. Automat. Control, 48 (2003), pp. 2151–2162.

[6] F. Blanchini and F. A. Pellegrino, *Relatively optimal control with characteristic polynomial assignment and output feedback*, IEEE Trans. Automat. Control, 51 (2006), pp. 183–191.

[7] D. S. Burnett, *Finite Element Analysis*, Addison-Wesley, Reading, MA, 1987.

[8] J. A. DeLoera, J. Rambau, and F. Santos, *Triangulations of Point Sets*, in preparation.

[9] P. O. Gutman and M. Cwikel, *An algorithm to find maximal state constraint sets for discrete-time linear dynamical systems with bounded control and states*, IEEE Trans. Automat. Control, 32 (1987), pp. 251–254.

[10] D. Q. Mayne, *Control of constrained dynamic systems,* European J. Control, 7 (2001), pp. 87–99.

[11] A. Rantzer and M. Johansson, *Piecewise linear quadratic optimal control*, IEEE Trans. Automat. Control, 45 (2000), pp. 629–637.

[12] K. Imai, *Structures of triangulations of points*, IEICE Trans. Inform. Systems, E83-D (2000), pp. 428–437.

[13] G. M. Ziegler, *Lectures on Polytopes*, Springer-Verlag, New York, 1995.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 604–629

OPTIMAL CONTROL OF THE STATIONARY NAVIER–STOKES EQUATIONS WITH MIXED CONTROL-STATE CONSTRAINTS*

J. C. DE LOS REYES† AND F. TRÖLTZSCH‡

Abstract. In this paper we consider the distributed optimal control of the Navier–Stokes equations in the presence of pointwise mixed control-state constraints. After deriving a first order necessary condition, the regularity of the mixed constraint multiplier is investigated. Second order sufficient optimality conditions are studied as well. In the last part of the paper, a semismooth Newton method is applied for the numerical solution of the control problem. The convergence of the method is proved and numerical experiments are carried out.

Key words. optimal control, Navier–Stokes equations, mixed control-state constraints, semismooth Newton methods

AMS subject classifications. 49K20, 76D05, 65J15

DOI. 10.1137/050646949

1. Introduction.

Continuing our efforts in the investigation of optimal control problems governed by the Navier–Stokes equations in the presence of pointwise control and state constraints (cf. [7, 8, 9, 10, 25]), we consider the following mixed control-state constrained problem:

$$\begin{cases} \min\ J(y,u) = \frac{1}{2}\int_\Omega |y - z_d|^2\ dx + \frac{\alpha}{2}\int_\Omega |u|^2\ dx \\ \text{subject to} \\ -\nu\Delta y + (y\cdot\nabla)y + \nabla p = u, \\ \operatorname{div} y = 0, \\ y|_\Gamma = g, \\ a \le \varepsilon u + y \le b \text{ a.e.}, \end{cases} \tag{1.1}$$

where $\alpha > 0$ and $\varepsilon > 0$. Due to the mixed nature of the pointwise constraints, expressed by the last relation of (1.1), the corresponding Lagrange multiplier is expected to be more regular than in the state constrained case (cf. [8]). In fact, such a constraint can be introduced as a way of regularizing the state constrained case, and it is expected that, as ε tends to zero, the solutions converge to the optimal solution of the state constrained problem (see [21]).

Optimal control of partial differential equations in the presence of state constraints is a very challenging research field, mainly due to the difficult structure of the Lagrange multiplier associated to the state constraints (see [2, 3, 4]). In the case of Navier–Stokes control, the problem has been investigated in [8], where the measure structure of the multiplier was studied.

This paper is a contribution to the numerical analysis of optimal control problems of the Navier–Stokes equations with pointwise state constraints.

*Received by the editors December 6, 2005; accepted for publication (in revised form) December 17, 2006; published electronically May 4, 2007. This research was supported by DFG Sonderforschungsbereich 557 "Control of complex turbulent shear flows."

http://www.siam.org/journals/sicon/46-2/64694.html

†Department of Mathematics, EPN, Quito, Ecuador and Institut für Mathematik, TU Berlin, Berlin, Germany (reyes@math.tu-berlin.de).

‡Institut für Mathematik, TU Berlin, Berlin, Germany (troeltzsch@math.tu-berlin.de).

The unconstrained and control constrained optimal control problems of the Navier–Stokes equations have been studied in many papers (see [1, 5, 7, 14, 15, 16, 18, 19, 25, 26]), where optimality conditions and/or numerical methods were discussed. Moreover, we refer the reader to the detailed references in [13].

In contrast to this, only a few papers consider associated problems with state constraints. To our best knowledge, in flow control, only [8, 10, 11, 27] deal with state constraints. In [8] and [11, 27] necessary optimality conditions are derived for the stationary and time dependent problems, respectively. In [10] the numerical solution utilizing a penalized problem together with a semismooth Newton method has been studied.

The novelty of our paper consists of a Lavrentiev-type regularization of the state constraints. Here we follow an approach introduced in [21, 22] to approximate the state constraints by mixed control-state constraints. This approach permits us to work with regular functions rather than with measures, which are unavoidable for pure pointwise state constraints. In this way, we are able to show regularity of Lagrange multipliers and to derive second order sufficient optimality conditions. An additional novelty is the consideration of semismooth Newton methods in this context. We set up a semismooth Newton algorithm for the numerical solution of the control problem and prove local superlinear convergence of the method. All these issues have not yet been considered in the literature.

The outline of the paper is as follows. In section 2, the optimal control problem is stated and existence of a global optimal solution is proved. In section 3, the problem is reformulated as a control constrained optimal control problem and first order necessary optimality conditions are obtained. Sufficient conditions of second order type are the topic of section 4. In section 5, a semismooth Newton algorithm is stated and the superlinear convergence of the method is proved. Reports on numerical experiments are summarized in section 6.

2. Problem statement and existence of solution. Consider a bounded regular domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$. Our objective is to characterize and find a solution $(u^*, y^*) \in \mathbf{L}^2(\Omega) \times \mathbf{H}^1(\Omega)$ to the following optimal control problem:

$$\begin{cases} \min\ J(y,u) = \frac{1}{2}\int_\Omega |y - z_d|^2\ dx + \frac{\alpha}{2}\int_\Omega |u|^2\ dx \\ \text{subject to} \\ -\nu\Delta y + (y\cdot\nabla)y + \nabla p = u, \\ \text{div } y = 0, \\ y|_\Gamma = g, \\ a \le \varepsilon u + y \le b \text{ a.e.}, \end{cases} \tag{2.1}$$

where $\alpha > 0$, $\varepsilon > 0$, z_d is the desired state, $a \le b \in \mathbf{L}^2(\Omega)$, and $g \in \mathbf{H}_0^{1/2}(\Gamma)$, with $\mathbf{H}_0^{1/2}(\Gamma) := \{v \in \mathbf{H}^{1/2}(\Gamma) : \int_\Gamma v\cdot\vec{n}\ d\Gamma = 0\}$ are given. The inequalities in the last line of (2.1) have to be understood componentwise. We denote by $(\cdot,\cdot)_X$ the inner product in the Hilbert space X and by $\|\cdot\|_X$ the associated norm. The subindex is suppressed if the L^2-inner product or norm is meant. Hereafter, the bold notation stands for the product of spaces. Additionally, we introduce the solenoidal space $V = \{v \in \mathbf{H}_0^1(\Omega) : \text{div } v = 0\}$, the closed subspace $\mathbf{H} := \{v \in \mathbf{H}^1(\Omega) : \text{div } v = 0\}$, and the trilinear form $c : \mathbf{H}\times\mathbf{H}\times\mathbf{H} \to \mathbb{R}$ defined by

$$c(u,w,v) = ((u\cdot\nabla)w, v). \tag{2.2}$$

Considering a force term $f \in V'$, the weak formulation of the Navier–Stokes equations is then given by

$$\nu(\nabla y, \nabla v) + c(y, y, v) = \langle f, v\rangle_{V',V} \text{ for all } v \in V, \tag{2.3}$$

$$\gamma_0 y = g, \tag{2.4}$$

where $\nabla y = \begin{pmatrix} \partial_1 y_1 & \dots & \partial_d y_1 \\ \vdots & \ddots & \vdots \\ \partial_1 y_d & \dots & \partial_d y_d \end{pmatrix}$, $(\nabla y, \nabla v) := \sum_{i=1}^d \sum_{j=1}^d (\partial_i y_j, \partial_i v_j)_{L^2(\Omega)}$, and $\gamma_0 : \mathbf{H}^1(\Omega) \to \mathbf{H}^{1/2}(\Gamma)$ stands for the trace operator. It is now standard to show existence of a solution for (2.3)–(2.4). Also an appropriate estimate and uniqueness, for ν sufficiently large or f sufficiently small, are obtained. The main results are summarized in the following theorem.

THEOREM 2.1. *Let $\Omega \in \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded regular domain, and $f \in \mathbf{H}^{-1}(\Omega)$ and $g \in \mathbf{H}_0^{1/2}(\Gamma)$. Then, there exists at least one solution for the nonhomogeneous problem* (2.3)–(2.4) *that satisfies the estimate*

$$\|y - \hat{y}\|_V \le \frac{2}{\nu} \|F\|_{V'}, \tag{2.5}$$

where $\hat{y} \in \mathbf{H}$ is a function such that $\gamma_0 \hat{y} = g$ and $F = f + \nu\Delta\hat{y} - (\hat{y} \cdot \nabla)\hat{y}$. Moreover, if $\|\hat{y}\|_{\mathbf{H}}$ is sufficiently small, such that

$$|c(v, \hat{y}, v)| \le \frac{\nu}{2} \|v\|_V^2 \text{ for all } v \in V$$

and $\nu^2 > 4\mathcal{N}\|F\|_{V'}$, with $\mathcal{N} = \sup_{u,v,w \in V} \frac{|c(u,v,\phi)|}{\|u\|_V \|v\|_V \|w\|_V}$, then the solution is unique.

Proof. For the proof we refer the reader to [23, pp. 178–180]. ∎

Next, we verify the existence of an optimal solution for our control problem. For that purpose let us define the set of admissible solutions

$$\mathcal{T}_{ad} = \{(y, u) \in \mathbf{H} \times \mathbf{L}^2(\Omega) : (y, u) \text{ satisfies the restrictions in (2.1)}\}.$$

THEOREM 2.2. *If $\mathcal{T}_{ad}$ is nonempty, then there exists an optimal solution for* (2.1).

Proof. Assuming that there is at least one feasible pair for our problem, we take a minimizing sequence $\{(y_n, u_n)\}$ in $\mathbf{L}^2(\Omega) \times \mathbf{H}^1(\Omega)$ and, considering the quadratic nature of the cost functional, we get that $\{u_n\}$ is uniformly bounded in $\mathbf{L}^2(\Omega)$.

From estimate (2.5) it follows that the sequence $\{y_n\}$ is also uniformly bounded in $\mathbf{H}^1(\Omega)$. Therefore, we may extract a weakly convergent subsequence, also denoted by $\{(y_n, u_n)\}$, such that $u_n \rightharpoonup u^*$ in $\mathbf{L}^2(\Omega)$ and $y_n \rightharpoonup y^*$ in $\mathbf{H}^1(\Omega)$. Due to the weak sequential continuity of the nonlinear form (cf. [12, p. 286]), it follows that $c(y_n, y_n, v) \to c(y^*, y^*, v)$. Consequently, due also to the linearity and continuity of the other terms involved, the limit (y^*, u^*) satisfies the state equations.

Since the set $C := \{v \in \mathbf{L}^2(\Omega) : a \le v \le b \text{ a.e.}\}$ is closed and convex, it is weakly closed. Hence, from the linearity and continuity of the mapping $(y, u) \to \varepsilon u + y$, it follows that $\varepsilon u^* + y^* \in C$. Taking into consideration that $J(y, u)$ is weakly lower semicontinuous, the result follows in a standard way. ∎

3. First order necessary optimality conditions. Let us consider the set

$$U = \{u \in \mathbf{L}^2(\Omega) : \|u\| < \left(\nu^2 - 4\mathcal{N}\,\|\nu\Delta\hat{y} - (\hat{y}\cdot\nabla)\hat{y}\|_{V'}\right)/(4\mathcal{N}\hat{c})\},$$

where $\hat{c}$ denotes the embedding constant of $\mathbf{L}^2(\Omega)$ into V' and $\hat{y}$ is a suitable velocity profile from Theorem 2.1. According to Theorem 2.1 there exists, for each $u \in U$ on the right-hand side of (2.3), a unique solution of the Navier–Stokes equations. Introducing the control-to-state operator $G : U \to \mathbf{H}$ that assigns to each $u \in U \subset \mathbf{L}^2(\Omega)$ the corresponding Navier–Stokes solution $y(u)$, and using the composite mapping $\mathcal{G} = \mathcal{I} \circ G$, where $\mathcal{I} : \mathbf{H} \to \mathbf{L}^2(\Omega)$ stands for the continuous compact injection, problem (2.1) can be expressed in a reduced form as

$$(\mathcal{P}) \qquad \begin{cases} \min_{u\in U}\ J(u) = \frac{1}{2}\int_\Omega |\mathcal{G}u - z_d|^2\,dx + \frac{\alpha}{2}\int_\Omega |u|^2\,dx \\ \text{subject to} \\ a \le \varepsilon u + \mathcal{G}u \le b \text{ a.e. in } \Omega. \end{cases}$$

Since U is open, we cannot expect in general that $(\mathcal{P})$ admits a global solution. However, in what follows, we concentrate on certain local solutions rather than consider exclusively global ones. Therefore, we are justified to assume $u^* \in U$ below.

In what follows we will frequently utilize the condition

$$\nu > \mathcal{M}(y^*), \tag{3.1}$$

with $\mathcal{M}(y) := \sup_{v\in V} \frac{|c(v,y,v)|}{\|v\|_V^2}$, which is responsible for the ellipticity of the linearized equations (see Lemma 3.1 below). Condition (3.1) is immediately satisfied for all pairs $(y(u), u)$ that fulfill the hypotheses of Theorem 2.1 (see [7, Remark 3.1]). In particular, it holds for all pairs $(y(u), u)$ with $u \in U$.

LEMMA 3.1. *Let $u \in U$ and $y := G(u)$. The control-to-state operator G is twice Fréchet differentiable at u and its derivatives $w := G'(u)h$ and $z := G''(u)[h]^2$ are given by the unique solutions of the systems*

$$\begin{aligned} -\nu\Delta w + (w\cdot\nabla)y + (y\cdot\nabla)w + \nabla\pi &= h, \\ \operatorname{div} w &= 0, \\ w|_\Gamma &= 0 \end{aligned} \tag{3.2}$$

and

$$\begin{aligned} -\nu\Delta z + (z\cdot\nabla)y + (y\cdot\nabla)z + \nabla\varrho &= -2(w\cdot\nabla)w, \\ \operatorname{div} z &= 0, \\ z|_\Gamma &= 0, \end{aligned} \tag{3.3}$$

respectively.

Proof. Let us begin by considering system (3.2). Its variational formulation is given by

$$a_1(w,\phi) := \nu(\nabla w, \nabla\phi) + c(w,y,\phi) + c(y,w,\phi) = (h,\phi)$$

for all $\phi \in V$. Since for all pairs (y,u) with $u \in U$ condition (3.1) holds (see [7, Remark 3.1]), coercivity of $a_1(\cdot,\cdot)$ and, consequently, the existence and uniqueness of the solution w, follow.

Let us denote the increment by $\bar{y} := y_{u+h} - y_u$, where $y_u := G(u)$. Considering that

$$c(y_{u+h}, y_{u+h}, \phi) - c(y_u, y_u, \phi) = c(\bar{y}, \bar{y}, \phi) + c(y_u, \bar{y}, \phi) + c(\bar{y}, y_u, \phi), \tag{3.4}$$

it can be verified that $\bar{y}$ is a solution of

$$\nu(\nabla \bar{y}, \nabla \phi) + c(\bar{y}, \bar{y}, \phi) + c(\bar{y}, y_u, \phi) + c(y_u, \bar{y}, \phi) = (h, \phi) \text{ for all } w \in V. \tag{3.5}$$

Taking $\phi = \bar{y}$ as a test function in (3.5) yields

$$(h, \bar{y}) = \nu \|\bar{y}\|_V^2 + c(\bar{y}, y_u, \bar{y}) \geq \nu \|\bar{y}\|_V^2 - \mathcal{M}(y_u) \|\bar{y}\|_V^2$$

and therefore

$$\|\bar{y}\|_V \leq \kappa \sigma(y) \|h\|, \tag{3.6}$$

where κ denotes the Poincaré inequality constant and $\sigma(y) := \frac{1}{\nu - \mathcal{M}(y)}$. Considering now $\tilde{y} = y_{u+h} - y_u - w$, we obtain the following equation:

$$\begin{aligned} \nu(\nabla \tilde{y}, \nabla \phi) + c(y_{u+h}, y_{u+h}, \phi) - c(y_u, y_u, \phi) \\ - c(w, y_u, \phi) - c(y_u, w, \phi) = 0 \text{ for all } w \in V. \end{aligned} \tag{3.7}$$

Using (3.4) and choosing $\tilde{y}$ as a test function in (3.7) we get that

$$\nu \|\tilde{y}\|_V^2 - c(\tilde{y}, \tilde{y}, y_u) = -c(\bar{y}, \bar{y}, \tilde{y}),$$

which, together with (3.6) and condition (3.1), yields

$$\|\tilde{y}\|_V \leq \mathcal{N} \kappa^2 \sigma^3(y) \|h\|^2. \tag{3.8}$$

Hence, the Fréchet differentiability follows. Moreover, since condition (3.1) holds, existence and uniqueness of solutions for (3.2) are verified. Therefore, the inverse operator exists for all $u \in U$ as a linear continuous operator and, from the implicit function theorem, the operator G is of class C^2 from U to $\mathbf{H}$. Taking the derivative on both sides of (3.2) yields (3.3) (see [6, p. 14]). □

The idea now consists in reformulating problem $(\mathcal{P})$ in a new variable $v := \varepsilon u + \mathcal{G}(u)$ and treating it as a control constrained optimal control problem. In order to express u as a function of v we consider the operator

$$\begin{aligned} F : \mathbf{L}^2(\Omega) \times \mathbf{L}^2(\Omega) &\to \mathbf{L}^2(\Omega), \\ (v, u) &\mapsto \varepsilon u + \mathcal{G}(u) - v \end{aligned}$$

and the solvability of the equation

$$F(v, u) = 0.$$

To justify existence and uniqueness of u for each $v \in \mathbf{L}^2(\Omega)$, we will consider an $\mathbf{L}^2$ neighborhood of the optimal control u^* contained in U. From the implicit function theorem (cf. [28]), since $F(v, u)$ is clearly defined in a neighborhood of u^* and $v^* = \varepsilon u^* + \mathcal{G}(u^*)$, it suffices to verify existence and continuity of the mapping $F_u(v^*, u^*)^{-1}$ from $\mathbf{L}^2(\Omega)$ to $\mathbf{L}^2(\Omega)$.

From the open mapping theorem, existence and continuity of $F_u(v^*, u^*)^{-1}$ hold if the operator $F_u(v^*, u^*) = \varepsilon + \mathcal{G}'(u^*)$ is bijective. Let us therefore consider the equation

$$(\varepsilon + \mathcal{G}'(u^*))h = \varphi, \tag{3.9}$$

with $\varphi \in \mathbf{L}^2(\Omega)$. It is easy to see that $\mathcal{G}'(u^*) = \mathcal{I} \circ G'(u^*)$ is compact due to the embedding $\mathcal{I} : \mathbf{H}^1(\Omega) \to \mathbf{L}^2(\Omega)$. Since $\varepsilon > 0$ and $\nu > \mathcal{M}(y^*)$, it can be verified that $\ker(\varepsilon + \mathcal{G}'(u^*)) = \{0\}$ and consequently ε is not an eigenvalue of $-\mathcal{G}'(u^*)$. Applying Fredholm's alternative, we get the existence of a unique solution $h \in \mathbf{L}^2(\Omega)$ for (3.9) and consequently the existence and continuity of $F_u(v^*, u^*)^{-1}$.

Therefore, there are constants $r, r_0 > 0$ such that for each $v \in \mathbf{L}^2(\Omega)$ with $\|v - v^*\| \le r_0$, there exists a unique $u := K(v)$ with $\|u - u^*\| \le r$ and

$$\varepsilon K(v) + \mathcal{G}(K(v)) = v. \tag{3.10}$$

Moreover, since F is twice continuously Fréchet differentiable, the implicit function theorem (cf. [28]) also implies that K is twice continuously Fréchet differentiable. Let us denote by $K''(v)[\xi, \eta]$ the second derivative of K in directions ξ and η and introduce $K''(v)[\xi]^2 := K''(v)[\xi, \xi]$. Taking the first and second derivatives on both sides of (3.10) in direction ξ yields

$$(\varepsilon + \mathcal{G}'(K(v)))K'(v)\xi = \xi, \tag{3.11}$$

$$(\varepsilon + \mathcal{G}'(K(v)))K''(v)[\xi]^2 = -\mathcal{G}''(K(v))[K'(v)\xi]^2, \tag{3.12}$$

which implies that

$$K'(v) = (\varepsilon + \mathcal{G}'(K(v)))^{-1}$$

and

$$K''(v)[\xi]^2 = -(\varepsilon + \mathcal{G}'(K(v)))^{-1}\mathcal{G}''(K(v))[K'(v)\xi]^2,$$

respectively.

Locally around u^*, our control problem can therefore be equivalently formulated as

$$(\mathcal{P}_r) \qquad \begin{cases} \min\ \mathcal{J}(v) =: J(y(K(v)), K(v)) \\ \text{subject to} \\ a \le v \le b \text{ a.e.}, \\ v \in B_{r_0}(v^*). \end{cases}$$

THEOREM 3.2. *Let u^* be a local optimal solution of $(\mathcal{P})$. Then there exist Lagrange multipliers $\lambda \in V$, $q \in L^2_0(\Omega)$ and μ_a, $\mu_b \in \mathbf{L}^2(\Omega)$ such that*

$$\begin{aligned} -\nu\Delta y^* + (y^* \cdot \nabla)y^* + \nabla p &= u^*, \\ \text{div } y^* &= 0, \\ y^*|_\Gamma &= g, \end{aligned} \tag{3.13}$$

$$-\nu\Delta\lambda - (y^* \cdot \nabla)\lambda + (\nabla y^*)^T\lambda + \nabla q = z_d - y^* + \mu_a - \mu_b,$$

$$\text{div } \lambda^* = 0, \tag{3.14}$$

$$\lambda^*|_\Gamma = 0,$$

$$\lambda - \alpha u^* = \varepsilon(\mu_b - \mu_a), \tag{3.15}$$

$$a \le \varepsilon u + y^* \le b,$$

$$\mu_a,\ \mu_b \ge 0, \tag{3.16}$$

$$(\mu_{a_i}, a_i - \varepsilon u_i^* - y_i^*) = (\mu_{b_i}, b_i - \varepsilon u_i^* - y_i^*) = 0 \text{ for } i = 1, 2$$

hold in the variational sense.

Proof. Since u^* is a locally optimal solution of $(\mathcal{P})$, we get for some $r > 0$

$$J(y^*, u^*) \le J(y(u), u)$$

for all $u \in B_r(u^*)$ with $a \le \varepsilon u + y(u) \le b$. Equivalently, since $u = K(v)$ holds locally,

$$\mathcal{J}(v^*) \le \mathcal{J}(v)$$

for all $v \in B_{r_0}(v^*)$ with $a \le v \le b$ and for an appropriate constant $r_0 > 0$.

Therefore, the following first order necessary condition follows:

$$\mathcal{J}'(v^*)(v - v^*) \ge 0 \text{ for all } a \le v \le b. \tag{3.17}$$

Applying the chain rule, the derivative of $\mathcal{J}(v^*)$ in any direction $\xi \in \mathbf{L}^2(\Omega)$ is given by

$$(\mathcal{J}'(v^*), \xi) = (y^* - z_d, \mathcal{G}'(u^*)K'(v^*)\xi) + \alpha(u^*, K'(v^*)\xi), \tag{3.18}$$

which, by $h := K'(v^*)\xi$, yields

$$(\mathcal{J}'(v^*), \xi) = (y^* - z_d, \mathcal{G}'(u^*)h) + \alpha(u^*, h).$$

Denoting by $\mu \in \mathbf{L}^2(\Omega)$ the Riesz representative of $-\mathcal{J}'(v^*)$ and using explicitly the derivative of K we obtain

$$(\mu, \xi) = (\mu, (\varepsilon + \mathcal{G}'(u^*))h) = \varepsilon(\mu, h) + (\mu, \mathcal{G}'(u^*)h).$$

Therefore, (3.18) is equivalent to

$$(y^* - z_d + \mu, \mathcal{G}'(u^*)h) + (\alpha u^* + \varepsilon\mu, h) = 0. \tag{3.19}$$

We now introduce the adjoint system of equations

$$-\nu\Delta\lambda - (y^* \cdot \nabla)\lambda + (\nabla y^*)^T\lambda + \nabla q = z_d - y^* - \mu,$$

$$\text{div } \lambda^* = 0, \tag{3.20}$$

$$\lambda^*|_\Gamma = 0.$$

Since, by hypothesis $\nu > \mathcal{M}(y^*)$, the ellipticity of the adjoint operator can be easily verified and, therefore, for $z_d - y^* - \mu \in \mathbf{L}^2(\Omega)$, there exists a unique solution $\lambda \in V$ for the adjoint system.

Consequently, (3.19) can be rewritten as

$$\lambda - \alpha u^* = \varepsilon\mu. \tag{3.21}$$

Utilizing the decomposition $\mu = \mu_b - \mu_a$, with

$$\mu_b := \mu_+ = \frac{1}{2}(\mu + |\mu|),$$

$$\mu_a := \mu_- = \frac{1}{2}(-\mu + |\mu|),$$

where $|\mu| = (|\mu_1|, |\mu_2|)^T$, the optimality condition (3.17) can be rewritten as

$$(\mathcal{J}'(v^*), v^*) = \min_{a \le v \le b}(\mu_a - \mu_b, v) = \min_{a \le v \le b}\{(\mu_{a,1}, v_1) - (\mu_{b,1}, v_1) + (\mu_{a,2}, v_2) - (\mu_{b,2}, v_2)\}.$$

By fixing the second component of the new control variable $v_2 = v_2^*$ and considering the mutual disjoint sets $\{x : \mu_{a,1}(x) > 0\}$ and $\{x : \mu_{b,1}(x) > 0\}$, we obtain that

$$(\mathcal{J}'(v^*), v^*) = (\mu_{a,1}, a_1) - (\mu_{b,1}, b_1) + (\mu_{a,2}, v_2^*) - (\mu_{b,2}, v_2^*)$$

and, consequently,

$$(\mu_{a,1}, a_1 - \varepsilon u_1^* - y_1^*) - (\mu_{b,1}, b_1 - \varepsilon u_1^* - y_1^*) = 0.$$

Fixing now the first component of v and proceeding in a similar manner we get that

$$(\mu_{a,2}, a_2 - \varepsilon u_2^* - y_2^*) - (\mu_{b,2}, b_2 - \varepsilon u_2^* - y_2^*) = 0.$$

Taking into account that, by definition, $\mu_a,\ \mu_b \ge 0$ componentwise, the complementarity system (3.16) follows. □

Remark 3.3. Notice that the existence of μ_a, μ_b cannot be deduced in a standard way from Kuhn–Tucker theorems in Banach spaces, since the cone of nonnegative functions in $\mathbf{L}^2(\Omega)$ has an empty interior and we work just in this constraint space.

4. Second order sufficient condition. Next, we turn to second order sufficient optimality conditions for problem $(\mathcal{P})$. Following [22], the idea consists again in utilizing the second order sufficient optimality properties of the pure control constrained problem $(\mathcal{P}_r)$ and translate them to the original setting.

We begin by verifying the relation between the Lagrangian

$$\mathcal{L}(y, u, \lambda) = \frac{1}{2}\|y^* - z_d\|^2 + \frac{\alpha}{2}\|u\|^2 + \nu(\nabla\lambda, \nabla y) + c(y, y, \lambda) - (\lambda, u)$$

and the second derivative of the reduced functional $\mathcal{J}$.

LEMMA 4.1. *The second derivative of the reduced cost functional in direction ξ satisfies*

$$\mathcal{J}''(v^*)[\xi]^2 = \mathcal{L}''(y^*, u^*, \lambda)(w, h)^2, \tag{4.1}$$

where $h = K'(v^*)\xi$ and w is the solution to (3.2) *with h on the right-hand side.*

Proof. Considering the reduced cost functional and differentiating it twice in direction ξ we get

$$\begin{aligned}
&\mathcal{J}''(v^*)[\xi]^2 \\
&\quad = J''(K(v^*))[K'(v^*)\xi]^2 + J'(K(v^*))\ K''(v^*)[\xi]^2 \\
&\quad = \|\mathcal{G}'(K(v^*))K'(v^*)\xi\|^2 + (y(K(v^*)) - z_d, \mathcal{G}''(K(v^*))[K'(v^*)\xi]^2) \\
&\qquad + (y(K(v^*)) - z_d, \mathcal{G}'(K(v^*))K''(v^*)\xi^2) + \alpha\,\|K'(v^*)\xi\|^2 + \alpha(K(v^*), K''(v^*)\xi^2),
\end{aligned}$$

which, by the relations $h = K'(v^*)\xi,\ u^* = K(v^*),\ y^* = y(K(v^*)),\ w = \mathcal{G}'(u^*)h$, and $z = \mathcal{G}''(u^*)[h]^2$, yields

$$\begin{aligned}
(4.2)\quad \mathcal{J}''(v^*)[\xi]^2 &= \|w\|^2 + (y^* - z_d, z) \\
&\quad + (y^* - z_d, \mathcal{G}'(u^*)K''(v)\xi^2) + \alpha\,\|h\|^2 + \alpha(u^*, K''(v^*)\xi^2).
\end{aligned}$$

From the optimality condition (3.19) we get

$$(y^* - z_d, \mathcal{G}'(u^*)K''(v^*)\xi^2) + \alpha(u^*, K''(v^*)\xi^2) = -(\mu, (\varepsilon + \mathcal{G}'(u^*))K''(v^*)\xi^2),$$

which implies that

$$\mathcal{J}''(v^*)[\xi]^2 = \|w\|^2 + \alpha\,\|h\|^2 + (y^* - z_d, z) - (\mu, (\varepsilon + \mathcal{G}'(u^*))K''(v^*)\xi^2).$$

Additionally, by (3.12) we find

$$-(\mu, (\varepsilon + \mathcal{G}'(u^*))K''(v^*)\xi^2) = (\mu, z).$$

From (3.14), using integration by parts and (3.3), we get that

$$\begin{aligned}
(y^* - z_d, z) - (\mu, (\varepsilon + \mathcal{G}'(u^*))K''(v^*)[\xi]^2) &= \nu(\Delta z, \lambda) - c(y^*, z, \lambda) - c(z, y^*, \lambda) \\
&= 2c(w, w, \lambda).
\end{aligned}$$

We thus obtain

$$\mathcal{J}''(v^*)[\xi]^2 = \|w\|^2 + \alpha\,\|h\|^2 + 2((w \cdot \nabla)w, \lambda).$$

On the other hand, computing the first and second derivatives of the Lagrangian yields

$$\begin{aligned}
\mathcal{L}'(y^*, u^*, \lambda)(w, h) &= (y^* - z_d, w) + \alpha(u^*, h) + \nu(\nabla\lambda, \nabla w) \\
&\quad + c(y^*, w, \lambda) + c(w, y^*, \lambda) - (\lambda, h), \\
\mathcal{L}''(y^*, u^*, \lambda)(w, h)^2 &= \|w\|^2 + \alpha\,\|h\|^2 + 2c(w, w, \lambda),
\end{aligned}$$

and consequently

$$(4.3)\qquad \mathcal{J}''(v^*)[\xi]^2 = \mathcal{L}''(y^*, u^*, \lambda)(w, h)^2 = \|w\|^2 + \alpha\,\|h\|^2 + 2c(w, w, \lambda),$$

where w is a solution of (3.2) with h on the right-hand side. ∎

Let us now introduce the set of strongly active constraints

$$\mathcal{A}_{\tau,i} := \{x \in \Omega : |\mu_i(x)| \geq \tau\},\ i = 1, \dots, d,$$

and the critical cone

$$\tilde{C}_\tau = \left\{ \begin{array}{c} \\ v \in \mathbf{L}^2(\Omega) : \\ \\ \end{array} \begin{array}{c} v_i(x) = 0 \text{ if } x \in \mathcal{A}_{\tau,i}, \\ v_i(x) \geq 0 \text{ if } v_i^*(x) = a_i, x \notin \mathcal{A}_{\tau,i}, \\ v_i(x) \leq 0 \text{ if } v_i^*(x) = b_i, x \notin \mathcal{A}_{\tau,i} \end{array} \right\}.$$

For the investigation of optimality for a given stationary pair (y^*, u^*) let us hereafter assume that the following second order condition holds: there exist $\tau > 0$, $\delta > 0$ such that

$$\mathcal{L}''(y^*, u^*, \lambda)(w, h)^2 \geq \delta \|h\|^2 \tag{SSC}$$

for all $(w, h) \in C_\tau$, where C_τ consists of all pairs $(w, h) \in V \times \mathbf{L}^2(\Omega)$ such that $\varepsilon h + w \in \tilde{C}_\tau$ and

$$\begin{aligned} -\nu\Delta w + (w \cdot \nabla)y^* + (y^* \cdot \nabla)w + \nabla\pi &= h, \\ \text{div } w &= 0, \\ w|_\Gamma &= 0. \end{aligned} \tag{4.4}$$

THEOREM 4.2. *If u^* is a stationary point of $(\mathcal{P})$ and* (SSC) *holds for some $\delta > 0$, $\tau > 0$, then there exist constants $\rho > 0$ and $\sigma > 0$ such that*

$$J(y, u) \geq J(y^*, u^*) + \sigma \|u - u^*\| \tag{4.5}$$

for all pairs (y, u) such that $y = G(u), a \leq \varepsilon u + y \leq b$, and $\|u - u^*\| \leq \rho$.

Proof. Utilizing (3.11), (4.1), and (SSC) it follows that

$$\mathcal{J}''(v^*)[\xi]^2 \geq \delta \left\|(\varepsilon + \mathcal{G}'(u^*))^{-1}\xi\right\|^2,$$

which using the continuity of the mapping $(\varepsilon + \mathcal{G}'(u^*))$ yields

$$\mathcal{J}''(v^*)[\xi]^2 \geq \delta \left(\frac{1}{\|\varepsilon + \mathcal{G}'(u^*)\|} \|\xi\|\right)^2 = \delta \|\varepsilon + \mathcal{G}'(u^*)\|^{-2} \|\xi\|^2 = \tilde{\delta} \|\xi\|^2.$$

Using the second order sufficient conditions for the reduced problem (cf. [24, p. 190]), we get the existence of constants $\tilde{\rho} > 0,\ \tilde{\sigma} > 0$ such that

$$\mathcal{J}(v) \geq \mathcal{J}(v^*) + \tilde{\sigma} \|v - v^*\|^2$$

for all $a \leq v \leq b,\ \|v - v^*\| \leq \tilde{\rho}$.

By the implicit function theorem there exist constants $r,\ r_0 > 0$ such that for all $v \in \mathbf{L}^2(\Omega)$ with $\|v - v^*\| \leq r_0$, there is a $u = K(v)$ which satisfies $\|u - u^*\| \leq r$.

Taking $\hat{\rho} = \min(\tilde{\rho}, r_0)$ we have that $\|u - u^*\| \leq r$ and

$$J(u) \geq J(u^*) + \tilde{\sigma} \|v - v^*\|^2 \tag{4.6}$$

$$= J(u^*) + \tilde{\sigma} \|\varepsilon(u - u^*) + \mathcal{G}(u) - \mathcal{G}(u^*)\|^2. \tag{4.7}$$

From the quadratic nature of the Navier–Stokes nonlinear term we obtain, using Taylor expansion, that

$$\mathcal{G}(u) - \mathcal{G}(u^*) = \mathcal{G}'(u^*)(u-u^*) + \frac{1}{2}\mathcal{G}''(u^*)[u-u^*]^2,$$

which, considering (4.7), implies that

$$J(u) \geq J(u^*) + \tilde{\sigma}\left\|(\varepsilon + \mathcal{G}'(u^*))(u-u^*) + \frac{1}{2}\mathcal{G}''(u^*)(u-u^*)^2\right\|^2 \tag{4.8}$$

$$\geq J(u^*) + \tilde{\sigma}\left(\|(\varepsilon + \mathcal{G}'(u^*))(u-u^*)\| - \left\|\frac{1}{2}\mathcal{G}''(u^*)(u-u^*)^2\right\|\right)^2. \tag{4.9}$$

Since the operator $(\varepsilon + \mathcal{G}'(u^*))^{-1}$ is linear and continuous we get that

$$\begin{aligned}\|u-u^*\| &= \left\|(\varepsilon + \mathcal{G}'(u^*))^{-1}(\varepsilon + \mathcal{G}'(u^*))(u-u^*)\right\| \\ &\leq \left\|(\varepsilon + \mathcal{G}'(u^*))^{-1}\right\| \|(\varepsilon + \mathcal{G}'(u^*))(u-u^*)\|,\end{aligned}$$

which implies that

$$\|(\varepsilon + \mathcal{G}'(u^*))(u-u^*)\| \geq \frac{1}{\|(\varepsilon + \mathcal{G}'(u^*))^{-1}\|}\|u-u^*\| = \bar{C}\|u-u^*\|.$$

Additionally, possibly by reducing r,

$$\left\|\frac{1}{2}\mathcal{G}''(u^*)[u-u^*]^2\right\| \leq \frac{\bar{C}}{2}\|u-u^*\|.$$

Therefore, we get that

$$J(u) \geq J(u^*) + \tilde{\sigma}\left(\bar{C}\|u-u^*\| - \frac{\bar{C}}{2}\|u-u^*\|\right)^2 = J(u^*) + \sigma\|u-u^*\|^2$$

with $\sigma := \frac{\tilde{\sigma}\bar{C}^2}{4}$ and, consequently, the local optimality of u^* and the quadratic rate follow. □

Remark 4.3. For the analysis of second order numerical methods, a stronger condition is needed (see [19, 24]): there exist constants $\tau > 0,\ \delta > 0$ such that

$$\mathcal{L}''(y^*, u^*, \lambda)(w,h)^2 \geq \delta\|h\|^2 \tag{$\overline{SSC}$}$$

for all pairs $(w,h) \in V \times \mathbf{L}^2(\Omega)$ that solve (4.4) and satisfy $\varepsilon h_i + w_i = 0$ on $\mathcal{A}_{\tau,i}$ for $i = 1, \ldots, d$.

5. Semismooth Newton method. In this section we propose a semismooth Newton method for the numerical solution of $(\mathcal{P})$. The infinite dimensional method is applied to the optimality system (3.13)–(3.16) and superlinear convergence is proved. Additionally, the close relationship between semismooth Newton and primal-dual active set methods (see [17]) allows a practical formulation of the algorithm in terms of active and inactive sets.

We begin by reformulating the complementarity system (3.16) as the operator equation

$$\mu = \max(0, \mu + c(v-b)) + \min(0, \mu + c(v-a)) \tag{5.1}$$

for all $c > 0$. Equation (5.1) suggests an update strategy based on active and inactive sets information.

DEFINITION 5.1. *Let X and Z be Banach spaces and $D \subset X$ an open subset. The mapping $F : D \to Z$ is called Newton differentiable in the open subset $U \subset D$ if there exists a mapping $\Psi : U \to L(X, Z)$ such that*

$$\lim_{h\to 0} \frac{1}{\|h\|} \|F(x+h) - F(x) - \Psi(x+h)h\| = 0$$

for every $x \in U$.

Since the $\max(0,\cdot)$ and $\min(0,\cdot)$ functions are Newton differentiable (see [7, 17]) from $L^p(\Omega) \to L^q(\Omega)$, with $q < p$, the application of the semismooth Newton method is justified with the special choice $c := \alpha/\varepsilon^2$ (see Theorem 5.5 below). The complete algorithm is defined through the following steps.

ALGORITHM 5.2.

1. *Initialize the variables $u_0 \in \mathbf{L}^2(\Omega)$, $y_0 \in V$, $\mu_0 = 0$, and set $k = 1$.*
2. *Until a stopping criterion is satisfied, set, for $i = 1, \dots, d$,*

$$\mathcal{A}_{b_i}^n = \left\{ x : \mu_i^{n-1} + \frac{\alpha}{\varepsilon^2}\left(\varepsilon u_i^{n-1} + y_i^{n-1} - b_i\right) > 0 \right\},$$

$$\mathcal{A}_{a_i}^n = \left\{ x : \mu_i^{n-1} + \frac{\alpha}{\varepsilon^2}(\varepsilon u_i^{n-1} + y_i^{n-1} - a_i) < 0 \right\},$$

$$\mathcal{I}_i^n = \Omega \backslash (\mathcal{A}_{b_i}^n \cup \mathcal{A}_{a_i}^n),$$

and find the solution (y, p, λ, q) of

(5.2)

$$-\nu\Delta y_i + y_1^{n-1}\partial_1 y_i + y_2^{n-1}\partial_2 y_i + y_1\partial_1 y_i^{n-1} + y_2\partial_2 y_i^{n-1} + \partial_i p = y_1^{n-1}\partial_1 y_i^{n-1} + y_2^{n-1}\partial_2 y_i^{n-1} + \begin{cases} \frac{1}{\varepsilon}(b_i - y_i) & \text{on } \mathcal{A}_{b_i}^n, \\ \frac{\lambda_i}{\alpha} & \text{on } \mathcal{I}_i^n, \\ \frac{1}{\varepsilon}(a_i - y_i) & \text{on } \mathcal{A}_{a_i}^n, \end{cases}$$

$$\operatorname{div} y = 0,$$

$$y|_\Gamma = g,$$

(5.3)

$$-\nu\Delta\lambda_i + \frac{1}{\varepsilon}\lambda_i - y_1\partial_1\lambda_i^{n-1} - y_2\partial_2\lambda_i^{n-1} - y_1^{n-1}\partial_1\lambda_i - y_2^{n-1}\partial_2\lambda_i + \lambda_1\partial_i y_1^{n-1} + \lambda_2\partial_i y_2^{n-1} + \lambda_1^{n-1}\partial_i y_1 + \lambda_2^{n-1}\partial_i y_2 + \partial_i q = z_{d,i} - y_i - y_1^{n-1}\partial_1\lambda_i^{n-1} - y_2^{n-1}\partial_2\lambda_i^{n-1} + \lambda_1^{n-1}\partial_i y_1^{n-1} + \lambda_2^{n-1}\partial_i y_2^{n-1} + \begin{cases} \frac{\alpha}{\varepsilon^2}(b_i - y_i) & \text{on } \mathcal{A}_{b_i}^n, \\ \frac{\lambda_i}{\varepsilon} & \text{on } \mathcal{I}_i^n, \\ \frac{\alpha}{\varepsilon^2}(a_i - y_i) & \text{on } \mathcal{A}_{a_i}^n, \end{cases}$$

$$\operatorname{div} \lambda = 0,$$

$$\lambda|_\Gamma = 0.$$

Set

$$(y^n,p^n,\lambda^n,q^n)=(y,p,\lambda,q),u_i^n=\begin{cases}\frac{1}{\varepsilon}(b_i-y_i^n) & \text{on } \mathcal{A}_{b_i}^n,\\ \frac{\lambda_i^n}{\alpha} & \text{on } \mathcal{I}_i^n,\\ \frac{1}{\varepsilon}(a_i-y_i^n) & \text{on } \mathcal{A}_{a_i}^n,\end{cases}\quad \mu^n=\frac{1}{\varepsilon}(\lambda^n-\alpha u^n),$$

and go to step 2.

Note that the system to be solved in step 2 corresponds to the optimality system of the optimal control problem

(5.4)

$$\begin{cases}\min\limits_{\delta_x\in V\times\tilde{C}^n} \frac{1}{2}\langle\mathcal{L}''(x^{n-1},\lambda^{n-1})\delta_x,\delta_x\rangle+\langle\mathcal{L}'(x^{n-1},\lambda^{n-1}),\delta_x\rangle\\ \qquad +\frac{\alpha}{2\varepsilon^2}\sum\limits_{i=1}^d\int_{\mathcal{A}_{b_i}^n}\left|b_i-y_i^{n-1}-\delta_{y_i}\right|^2dx+\frac{\alpha}{2\varepsilon^2}\sum\limits_{i=1}^d\int_{\mathcal{A}_{a_i}^n}\left|a_i-y_i^{n-1}-\delta_{y_i}\right|^2dx\\ \qquad +\frac{1}{\varepsilon}\sum\limits_{i=1}^d\int_{\mathcal{A}_{b_i}^n}\lambda_i^{n-1}\cdot\delta_{y_i}\,dx+\frac{1}{\varepsilon}\sum\limits_{i=1}^d\int_{\mathcal{A}_{a_i}^n}\lambda_i^{n-1}\cdot\delta_{y_i}\,dx\\ \text{subject to}\\ -\nu\Delta\delta_{y_i}+y_1^{n-1}\partial_1\delta_{y_i}+y_2^{n-1}\partial_2\delta_{y_i}+\delta_{y_1}\partial_1y_i^{n-1}+\delta_{y_2}\partial_2y_i^{n-1}+\partial_ip^n\\ \qquad =\nu\Delta y_i^{n-1}-y_1^{n-1}\partial_1y_i^{n-1}-y_2^{n-1}\partial_2y_i^{n-1}+\begin{cases}\frac{1}{\varepsilon}(b_i-y_i^{n-1}-\delta_{y_i}) & \text{on } \mathcal{A}_{b_i}^n,\\ u_i^{n-1}+\delta_{u_i} & \text{on } \mathcal{I}_i^n,\\ \frac{1}{\varepsilon}(a_i-y_i^{n-1}-\delta_{y_i}) & \text{on } \mathcal{A}_{a_i}^n,\end{cases}\\ \operatorname{div}\delta_y=0,\\ \delta_y|_\Gamma=-y^{n-1}|_\Gamma+g,\end{cases}$$

where $x^n=(y^n,u^n)$, $\delta_x=x^n-x^{n-1}$, and

$$\tilde{C}^n:=\left\{h\in\mathbf{L}^2(\Omega):h_i(x)=0 \text{ for } x\in\mathcal{A}_{b_i}^n\cup\mathcal{A}_{a_i}^n,\ i=1,\dots,d\right\}.$$

Problem (5.4) corresponds to a quadratic control problem with affine constraints. Existence and uniqueness of a solution, as well as existence of Lagrange multipliers, will be verified next.

THEOREM 5.3. *Let $u^*\in U$ be a stationary point of $(\mathcal{P})$ that satisfies the second order condition $(\overline{SSC})$. If $\mathcal{I}_i^n\subset\mathcal{I}_{\tau,i}$, with $\mathcal{I}_{\tau,i}:=\Omega\backslash\mathcal{A}_{\tau,i}$, $i=1,\dots,d$, and $\|y^{n-1}-y^*\|_V$, $\|\lambda^{n-1}-\lambda^*\|_V$ are sufficiently small, then there exists a unique solution for system* (5.2)–(5.3).

Proof. Existence of Lagrange multipliers for (5.4) follows from the satisfaction of the regular point condition (see [20]), which in the present case is fulfilled if there exists a unique weak solution $w\in V$ of

$$\begin{aligned}-\nu\Delta w+(w\cdot\nabla)y^{n-1}+(y^{n-1}\cdot\nabla)w+\nabla\pi&=h,\\ \operatorname{div}w&=0,\\ w|_\Gamma&=0\end{aligned}\tag{5.5}$$

with $\varepsilon h+w\in\tilde{C}^n$. Multiplying both sides of (5.5) by w, existence and uniqueness follow from the Lax–Milgram theorem if the coercivity condition $\nu>\mathcal{M}(y^{n-1})$ holds.

From the definition of $\mathcal{M}(\cdot)$ we get that

$$\begin{aligned}\nu - \mathcal{M}(y^{n-1}) &= \nu - \sup_{w\in V} \frac{|c(w, y^{n-1}, w)|}{\|w\|_V^2} \\ &\geq \nu - \sup_{w\in V} \frac{|c(w, y^{n-1} - y^*, w)|}{\|w\|_V^2} - \mathcal{M}(y^*) \\ &\geq \nu - \mathcal{M}(y^*) - \mathcal{N}\left\|y^{n-1} - y^*\right\|_V.\end{aligned}$$

Choosing $\|y^{n-1} - y^*\|_V \leq \frac{\nu - \mathcal{M}(y^*)}{2\mathcal{N}}$ yields

$$\nu - \mathcal{M}(y^{n-1}) \geq \frac{\nu - \mathcal{M}(y^*)}{2} > 0,$$

and, thus, each solution of (5.4) satisfies the optimality system (5.2)–(5.3).

On the other hand, to see that a solution to (5.2)–(5.3) corresponds to the solution of (5.4) a second order condition has to hold. Denoting by $L(\delta_y, \delta_u)$ the Lagrangian of (5.4), the second order condition can be stated as follows: there exists a constant $\rho > 0$ such that

$$L''(\delta_y, \delta_u)(w, h)^2 \geq \rho \|h\|^2 \tag{5.6}$$

for all $(w, h) \in V \times \mathbf{L}^2(\Omega)$ that solve (5.5) and satisfy $\varepsilon h + w \in \tilde{C}^n$. Taking such a (w, h) arbitrary but fixed, we introduce the decomposition $(w, h) = (\xi, \bar{h}) + (\psi, \underline{h})$ with $\xi \in V$ a weak solution of

$$\begin{aligned}-\nu\Delta\xi + (\xi\cdot\nabla)y^* + (y^*\cdot\nabla)\xi + \nabla\pi_1 &= \bar{h}, \\ \operatorname{div}\xi &= 0, \\ \xi|_\Gamma &= 0,\end{aligned} \tag{5.7}$$

with

$$\bar{h}_i := \begin{cases} -\frac{1}{\varepsilon}\xi_i & \text{on } \mathcal{A}^n_{b_i}, \\ h_i & \text{on } \mathcal{I}^n_i, \\ -\frac{1}{\varepsilon}\xi_i & \text{on } \mathcal{A}^n_{a_i} \end{cases}$$

for $i = 1, \ldots, d$, and $\psi \in V$ a weak solution of

$$\begin{aligned}-\nu\Delta\psi + (\psi\cdot\nabla)y^{n-1} + (y^{n-1}\cdot\nabla)\psi + \nabla\pi_2 \\ = -((y^{n-1} - y^*)\cdot\nabla)\xi - (\xi\cdot\nabla)(y^{n-1} - y^*) + \underline{h}, \\ \operatorname{div}\psi = 0, \\ \psi|_\Gamma = 0\end{aligned} \tag{5.8}$$

with

$$\underline{h}_i = \begin{cases} -\frac{1}{\varepsilon}\psi_i & \text{on } \mathcal{A}^n_{b_i}, \\ 0 & \text{on } \mathcal{I}^n_i, \\ -\frac{1}{\varepsilon}\psi_i & \text{on } \mathcal{A}^n_{a_i} \end{cases}$$

for $i = 1, \dots, d$. We therefore get that $(\xi, \bar{h})$ solves (5.7) and satisfies $\varepsilon\bar{h} + \xi \in \tilde{C}^n$. From (4.3) and using Cauchy–Schwarz we thus obtain

$$L''(\delta_y, \delta_u)(w,h)^2 \geq \|\xi\|^2 + \alpha \left\|\bar{h}\right\|^2 - 2\|\xi\|\,\|\psi\| - 2\alpha \left\|\bar{h}\right\| \|\underline{h}\| + 2c(w, w, \lambda^{n-1}),$$

which implies, using the properties of the trilinear form, that

$$\begin{aligned} L''(\delta_y, \delta_u)(w,h)^2 &\geq \mathcal{L}''(y^*, u^*, \lambda^*)(\xi, \bar{h})^2 - 2\|\xi\|\,\|\psi\| - 2\alpha \left\|\bar{h}\right\| \|\underline{h}\| \\ &\quad - 2\mathcal{N}\|\xi\|_V^2 \left\|\lambda^{n-1} - \lambda^*\right\|_V - 4\mathcal{N}\|\xi\|_V \|\psi\|_V \left\|\lambda^{n-1}\right\|_V - 2\mathcal{N}\|\psi\|_V^2 \left\|\lambda^{n-1}\right\|_V. \end{aligned}$$

From (5.7) and (5.8) it can be verified that the following estimates hold:

$$\|\xi\|_V \leq \kappa\sigma \left\|\bar{h}\right\|, \tag{5.9}$$

$$\|\psi\|_V \leq 4\mathcal{N}\kappa\sigma^2 \left\|y^{n-1} - y^*\right\|_V \left\|\bar{h}\right\|, \tag{5.10}$$

with $\sigma := (\nu - \mathcal{M}(y^*))^{-1}$.

Since by hypothesis u^* satisfies $(\overline{SSC})$ and $\mathcal{I}_i^n \subset \mathcal{I}_{\tau,i}$, it follows that $\varepsilon\bar{h} + \xi = 0$ on $\mathcal{A}_{\tau,i}$ and, using estimates (5.9), (5.10),

$$\begin{aligned} &L''(\delta_y, \delta_u)(w,h)^2 \\ &\quad \geq \delta \left\|\bar{h}\right\|^2 - 8\mathcal{N}\kappa^4\sigma^3 \left\|y^{n-1} - y^*\right\|_V \left\|\bar{h}\right\|^2 - \frac{8}{\varepsilon}\alpha\mathcal{N}\kappa^2\sigma^2 \left\|y^{n-1} - y^*\right\|_V \left\|\bar{h}\right\|^2 \\ &\quad\quad - 2\mathcal{N}\kappa^2\sigma^2 \left\|\lambda^{n-1} - \lambda^*\right\|_V \left\|\bar{h}\right\|^2 - 16\mathcal{N}^2\kappa^2\sigma^3 \left\|\lambda^{n-1}\right\|_V \left\|y^{n-1} - y^*\right\|_V \left\|\bar{h}\right\|^2 \\ &\quad\quad - 32\mathcal{N}^3\kappa^2\sigma^4 \left\|\lambda^{n-1}\right\|_V \left\|y^{n-1} - y^*\right\|_V^2 \left\|\bar{h}\right\|^2. \end{aligned}$$

Choosing $\|y^{n-1} - y^*\|_V$ and $\|\lambda^{n-1} - \lambda^*\|_V$ sufficiently small such that

$$\begin{aligned} \rho := \delta &- 2\mathcal{N}\kappa^2\sigma^2 \left\|\lambda^{n-1} - \lambda^*\right\|_V - 8\mathcal{N}\kappa^2\sigma^2 \left\|y^{n-1} - y^*\right\|_V \big[\kappa^2\sigma + \alpha/\varepsilon \\ &+ 2\mathcal{N}\sigma \left\|\lambda^{n-1}\right\|_V + 4\mathcal{N}^2\sigma^2 \left\|\lambda^{n-1}\right\|_V \left\|y^{n-1} - y^*\right\|_V\big] > 0, \end{aligned}$$

condition (5.6) is satisfied.

Therefore, system (5.2)–(5.3) is uniquely solvable since it corresponds to the solution of a linear quadratic control problem with convex objective. ∎

Remark 5.4. From the definition of the inactive sets, it can be verified that the condition $\mathcal{I}_i^n \subset \mathcal{I}_{\tau,i}$ holds for $\|y^{n-1} - y^*\|_V$ and $\|\lambda^{n-1} - \lambda^*\|_V$ sufficiently small.

By considering the state variable y and the newly defined control variable v, the optimal control problem $(\mathcal{P})$ can locally also be expressed as the following control constrained optimal control problem:

(5.11)
$$\begin{cases} \min\ J(y,v) = \frac{1}{2}\int_\Omega |y - z_d|^2\,dx + \frac{\alpha}{2\varepsilon^2}\int_\Omega |v|^2\,dx - \frac{\alpha}{\varepsilon^2}\int_\Omega v\ y\,dx + \frac{\alpha}{2\varepsilon^2}\int_\Omega |y|^2\,dx \\ \text{subject to} \\ -\nu\Delta y + \frac{1}{\varepsilon}y + (y\cdot\nabla)y + \nabla p = \frac{1}{\varepsilon}v, \\ \operatorname{div} y = 0, \\ y|_\Gamma = g, \\ a \leq v \leq b \text{ a.e.} \end{cases}$$

The presence of the mixed term $\frac{\alpha}{\varepsilon^2}\int_\Omega v\ y\ dx$ in the cost functional is responsible for a different problem structure which does not allow the application of already known results about convergence of the semismooth Newton method for control constrained optimal control problems (see [9, 17]).

In the next theorem sufficient conditions for the local superlinear convergence of the semismooth Newton method are stated.

THEOREM 5.5. *Let $u^* \in U$ be a stationary point of $(\mathcal{P})$ that satisfies $(\overline{SSC})$. If $\|\lambda^*\|_V < \frac{\alpha^{1/2}}{4\varepsilon}(\nu - \mathcal{M}(y^*))(\frac{\varepsilon(\alpha+\varepsilon^2)-\alpha}{\varepsilon^{1/2}(\alpha+\varepsilon^2)^{1/2}+\alpha^{1/2}})$ and $\|y^0 - y^*\|_V$, $\|\lambda^0 - \lambda^*\|_V$ are sufficiently small, then the sequence $\{(y^n, v^n, \lambda^n, \mu^n)\}$ generated by the algorithm converges superlinearly in $\mathbf{H}^1(\Omega)\times\mathbf{L}^2(\Omega)\times V\times\mathbf{L}^2(\Omega)$ to $(y^*, u^*, \lambda^*, \mu^*)$. Moreover, there exists a constant $C > 0$ such that*

$$\begin{aligned}(5.12)\quad &\big\|(v^{n+1}-v^*, y^{n+1}-y^*, \lambda^{n+1}-\lambda^*)\big\|_{\mathbf{L}^2\times V\times V}\\ &\qquad\le C\left(\|y^n-y^*\|_V^2 + \|\lambda^n-\lambda^*\|_V^2\right) + o(\|(y^n-y^*, \lambda^n-\lambda^*)\|_{\mathbf{L}^p\times\mathbf{L}^p}).\end{aligned}$$

Proof. By considering system (3.13)–(3.16) and the system in step 2 of Algorithm 5.2, it can be verified that the increments $\delta_y = y^{n+1}-y^*$, $\delta_\lambda = \lambda^{n+1}-\lambda^*, \delta_u, \delta_\mu$, and δ_ϕ satisfy the system

$$\begin{aligned}(5.13)\quad &\nu(\nabla\delta_y, \nabla\phi) + \frac{1}{\varepsilon}(\delta_y, \phi) + c(y^n, \delta_y, \phi) + c(\delta_y, y^n, \phi)\\ &\qquad = \frac{1}{\varepsilon}(\delta_v, \phi) + (((y^n-y^*)\cdot\nabla)(y^n-y^*), \phi) \text{ for all } \phi\in V,\end{aligned}$$

$$\begin{aligned}(5.14)\quad &\nu(\nabla\delta_\lambda, \nabla\phi) - c(y^n, \delta_\lambda, \phi) - c(\delta_y, \lambda^n, \phi) + c(w, y^n, \delta_\lambda) + c(w, \delta_y, \lambda^n)\\ &\qquad = ((\nabla(y^n-y^*))^T(\lambda^n-\lambda^*) - ((y^n-y^*)\cdot\nabla)(\lambda^n-\lambda^*), \phi) - (\delta_y+\delta_\mu, \phi)\\ &\qquad \text{for all } \phi\in V.\end{aligned}$$

Introducing the auxiliary variable $\varphi := \varepsilon\mu + \frac{\alpha}{\varepsilon}v$ and considering (3.15) and (5.1) together with the semismooth Newton update for u^n and μ^n, we also obtain that

$$(5.15)\qquad \delta_\lambda - \frac{\alpha}{\varepsilon}\delta_v + \frac{\alpha}{\varepsilon}\delta_y = \varepsilon\delta_\mu,$$

$$(5.16)\qquad \delta_\varphi = \varepsilon\delta_\mu + \frac{\alpha}{\varepsilon}\delta_v,$$

$$(5.17)\qquad \delta_\varphi - \frac{\alpha}{\varepsilon}\delta_v = G^n_{\max}(\delta_\varphi) + G^n_{\min}(\delta_\varphi) + R,$$

where

$$G^n_{\max,i}(\phi) = \begin{cases}\phi \text{ on } \mathcal{A}^{n+1}_{b_i},\\ 0 \text{ on } \Omega\backslash\mathcal{A}^{n+1}_{b_i},\end{cases} \quad\text{and}\quad G^n_{\min,i}(\phi) = \begin{cases}\phi \text{ on } \mathcal{A}^{n+1}_{a_i},\\ 0 \text{ on } \Omega\backslash\mathcal{A}^{n+1}_{a_i}\end{cases}$$

and

$$\begin{aligned}R = &\max\left(0, \varphi^* + (\varphi^n-\varphi^*) - \frac{\alpha}{\varepsilon}b\right) - \max\left(0, \varphi^* - \frac{\alpha}{\varepsilon}b\right) - G^n_{\max}(\varphi^n-\varphi^*)\\ &+ \min\left(0, \varphi^* + (\varphi^n-\varphi^*) - \frac{\alpha}{\varepsilon}a\right) - \min\left(0, \varphi^* - \frac{\alpha}{\varepsilon}a\right) - G^n_{\min}(\varphi^n-\varphi^*).\end{aligned}$$

Due to Newton differentiability of the $\max(0,\cdot)$ and $\min(0,\cdot)$ functions (cf. [17]) from $L^p(\Omega) \to L^q(\Omega)$, with $q < p$, we therefore obtain that

$$\|R\|_{\mathbf{L}^2} = o(\|\varphi^n - \varphi^*\|_{\mathbf{L}^p}), \tag{5.18}$$

with $p > 2$.

Multiplying (5.17) by δ_v, we get that

$$-(R, \delta_v) = (G^n_{\max}(\delta_\varphi) + G^n_{\min}(\delta_\varphi), \delta_v) - (\delta_\varphi, \delta_v) + \frac{\alpha}{\varepsilon} \|\delta_v\|^2 . \tag{5.19}$$

Additionally, from the definition of $G^n_{\max}$ and $G^n_{\min}$,

$$(G^n_{\max}(\delta_\varphi) + G^n_{\min}(\delta_\varphi), \delta_v) - (\delta_\varphi, \delta_v) = (\delta_\varphi, \delta_v)_{\mathcal{I}^n}, \tag{5.20}$$

where $(v, w)_{\mathcal{I}^n} := \int_{\mathcal{I}^n} v \cdot w \, dx$.

On the other hand, substituting (5.15) into (5.14) and multiplying by δ_y, we get that

$$\begin{aligned} &\nu(\nabla\delta_\lambda, \nabla\delta_y) + \frac{1}{\varepsilon}(\delta_\lambda, \delta_y) - c(y^n, \delta_\lambda, \delta_y) - c(\delta_y, \lambda^n, \delta_y) \\ &\quad + c(\delta_y, y^n, \delta_\lambda) + c(\delta_y, \delta_y, \lambda^n) = ((\nabla(y^n - y^*))^T(\lambda^n - \lambda^*) \\ &\quad - ((y^n - y^*) \cdot \nabla)(\lambda^n - \lambda^*), \delta_y) - \|\delta_y\|^2 + \frac{\alpha}{\varepsilon^2}(\delta_v, \delta_y) - \frac{\alpha}{\varepsilon^2} \|\delta_y\|^2 , \end{aligned} \tag{5.21}$$

which, utilizing (5.13), multiplied by δ_λ yields

$$\begin{aligned} &\frac{1}{\varepsilon}(\delta_v, \delta_\lambda) + (((y^n - y^*) \cdot \nabla)(y^n - y^*), \delta_\lambda) - c(y^n, \delta_y, \delta_\lambda) - c(\delta_y, y^n, \delta_\lambda) \\ &\quad = ((\nabla(y^n - y^*))^T(\lambda^n - \lambda^*) - ((y^n - y^*) \cdot \nabla)(\lambda^n - \lambda^*), \delta_y) - \frac{\alpha}{\varepsilon^2} \|\delta_y\|^2 \\ &+ \frac{\alpha}{\varepsilon^2}(\delta_v, \delta_y) - \|\delta_y\|^2 + c(y^n, \delta_\lambda, \delta_y) + c(\delta_y, \lambda^n, \delta_y) - c(\delta_y, y^n, \delta_\lambda) - c(\delta_y, \delta_y, \lambda^n). \end{aligned} \tag{5.22}$$

Consequently, utilizing the properties of the trilinear form,

$$\begin{aligned} &\left(\frac{\alpha + \varepsilon^2}{\varepsilon^2}\right) \|\delta_y\|^2 + \frac{1}{\varepsilon}(\delta_v, \delta_\lambda) + \frac{\alpha}{\varepsilon^2}(\delta_v, \delta_y) - 2c(\delta_y, \lambda^n, \delta_y) = ((\nabla(y^n - y^*))^T(\lambda^n - \lambda^*) \\ &\quad - ((y^n - y^*) \cdot \nabla)(\lambda^n - \lambda^*), \delta_y) - (((y^n - y^*) \cdot \nabla)(y^n - y^*), \delta_\lambda) + \frac{2\alpha}{\varepsilon^2}(\delta_v, \delta_y) \end{aligned} \tag{5.23}$$

and therefore

$$\begin{aligned} \frac{1}{\varepsilon}(\delta_v, \delta_\varphi) \le \frac{2\alpha}{\varepsilon^2}(\delta_v, \delta_y) + 2\mathcal{N} \|\lambda^n\|_V \|\delta_y\|_V^2 - \left(\frac{\alpha + \varepsilon^2}{\varepsilon^2}\right) \|\delta_y\|^2 \\ + ((\nabla(y^n - y^*))^T(\lambda^n - \lambda^*) - ((y^n - y^*) \cdot \nabla)(\lambda^n - \lambda^*), \delta_y) \\ - (((y^n - y^*) \cdot \nabla)(y^n - y^*), \delta_\lambda). \end{aligned} \tag{5.24}$$

Let us now consider the increment equation (5.13) and multiply it by δ_y. We get the estimate

$$\nu \|\delta_y\|_V^2 + \frac{1}{\varepsilon}\|\delta_y\|^2 - \mathcal{M}(y^n)\|\delta_y\|_V^2 \le \frac{1}{\varepsilon}(\delta_v,\delta_y) + \mathcal{N}\|y^n - y^*\|_V^2 \|\delta_y\|_V , \tag{5.25}$$

which, by considering a y^* neighborhood such that

$$\nu - \mathcal{M}(y^n) \ge \frac{1}{2}(\nu - \mathcal{M}(y^*)) > 0 \tag{5.26}$$

and the Poincaré inequality, implies that

$$\frac{1}{2}(\nu - \mathcal{M}(y^*))\|\delta_y\|_V^2 + \frac{1}{\varepsilon}\|\delta_y\|^2 \le \frac{1}{\varepsilon}(\delta_v,\delta_y) + \mathcal{N}\|y^n - y^*\|_V^2 \|\delta_y\|_V . \tag{5.27}$$

Consequently, we obtain the estimate

$$\|\delta_y\|_V \le 2\sigma\left(\frac{\kappa}{\varepsilon}\|\delta_v\| + \mathcal{N}\|y^n - y^*\|_V^2\right), \tag{5.28}$$

with $\sigma := (\nu - \mathcal{M}(y^*))^{-1}$.

Using (5.27) in (5.24) and grouping terms yield

$$\begin{aligned}\frac{1}{\varepsilon}(\delta_v,\delta_\varphi) \le{}& \frac{2\alpha + 4\sigma\varepsilon\|\lambda^n\|_V}{\varepsilon^2}(\delta_v,\delta_y) - \left(\frac{\alpha + \varepsilon^2 + 4\sigma\varepsilon\|\lambda^n\|_V}{\varepsilon^2}\right)\|\delta_y\|^2 \\ &+ 4\mathcal{N}\sigma\|\lambda^n\|_V\|y^n - y^*\|_V^2\|\delta_y\|_V - (((y^n - y^*)\cdot\nabla)(y^n - y^*),\delta_\lambda) \\ &+ ((\nabla(y^n - y^*))^T(\lambda^n - \lambda^*) - ((y^n - y^*)\cdot\nabla)(\lambda^n - \lambda^*),\delta_y).\end{aligned} \tag{5.29}$$

Since

$$\begin{aligned}\left\| c\delta_v - \left(\frac{\alpha + \varepsilon^2 + 4\sigma\varepsilon\|\lambda^n\|_V}{\varepsilon^2}\right)^{1/2}\delta_y\right\|^2 ={}& c^2\|\delta_v\|^2 \\ &- \frac{2c}{\varepsilon}\left(\alpha + \varepsilon^2 + 4\sigma\varepsilon\|\lambda^n\|_V\right)^{1/2}(\delta_v,\delta_y) + \left(\frac{\alpha + \varepsilon^2 + 4\sigma\varepsilon\|\lambda^n\|_V}{\varepsilon^2}\right)\|\delta_y\|^2,\end{aligned} \tag{5.30}$$

we obtain, by choosing $c = \frac{\alpha + 2\sigma\varepsilon\|\lambda^n\|_V}{\varepsilon\sqrt{\alpha + \varepsilon^2 + 4\sigma\varepsilon\|\lambda^n\|_V}}$, that

$$\begin{aligned}\frac{1}{\varepsilon}(\delta_v,\delta_\varphi) \le{}& \frac{(\alpha + 2\sigma\varepsilon\|\lambda^n\|_V)^2}{\varepsilon^2(\alpha + \varepsilon^2 + 4\sigma\varepsilon\|\lambda^n\|_V)}\|\delta_v\|^2 + 4\mathcal{N}\sigma\|\lambda^n\|_V\|y^n - y^*\|_V^2\|\delta_y\|_V \\ &+ ((\nabla(y^n - y^*))^T(\lambda^n - \lambda^*) - ((y^n - y^*)\cdot\nabla)(\lambda^n - \lambda^*),\delta_y) - (((y^n - y^*)\cdot\nabla)(y^n - y^*),\delta_\lambda).\end{aligned} \tag{5.31}$$

Consequently, from (5.19)–(5.20) and (5.31) we therefore obtain

$$\begin{aligned}|(R,\delta_v)| \ge{}& \frac{\alpha}{\varepsilon}\|\delta_v\|^2 - \Bigg|\frac{(\alpha + 2\sigma\varepsilon\|\lambda^n\|_V)^2}{\varepsilon^2(\alpha + \varepsilon^2 + 4\sigma\varepsilon\|\lambda^n\|_V)}\|\delta_v\|_{\mathcal{I}^n}^2 \\ &+ ((\nabla(y^n - y^*))^T(\lambda^n - \lambda^*) - ((y^n - y^*)\cdot\nabla)(\lambda^n - \lambda^*),\delta_y)_{\mathcal{I}^n} \\ &- (((y^n - y^*)\cdot\nabla)(y^n - y^*),\delta_\lambda)_{\mathcal{I}^n} + 4\mathcal{N}\sigma\|\lambda^n\|_V\|y^n - y^*\|_V^2\|\delta_y\|_V\Bigg|,\end{aligned}$$

which, by considering a (y^*, λ^*) neighborhood such that

$$\|\lambda^n\|_V \leq 2\,\|\lambda^*\|_V\,, \tag{5.32}$$

implies that

$$\begin{aligned}|(R,\delta_v)| \geq \frac{\alpha}{\varepsilon}\,\|\delta_v\|^2 - \frac{(\alpha+4\sigma\varepsilon\,\|\lambda^*\|_V)^2}{\varepsilon^2(\alpha+\varepsilon^2)}\,\|\delta_v\|^2 \\ - \mathcal{N}\,\|y^n-y^*\|_V^2\,(8\sigma\,\|\lambda^*\|_V\,\|\delta_y\|_V + \|\delta_\lambda\|_V) \\ - 2\mathcal{N}\,\|y^n-y^*\|_V\,\|\lambda^n-\lambda^*\|_V\,\|\delta_y\|_V\,.\end{aligned}$$

Since, by hypothesis, $\|\lambda^*\|_V < \frac{\alpha^{1/2}}{4\varepsilon}(\nu - \mathcal{M}(y^*))\frac{\varepsilon(\alpha+\varepsilon^2)-\alpha}{\varepsilon^{1/2}(\alpha+\varepsilon^2)^{1/2}+\alpha^{1/2}}$, we get that

$$\beta := \frac{\alpha\varepsilon(\alpha+\varepsilon^2) - (\alpha+4\sigma\varepsilon\,\|\lambda^*\|_V)^2}{\varepsilon^2(\alpha+\varepsilon^2)} > 0, \tag{5.33}$$

and therefore

$$\begin{aligned}\beta\,\|\delta_v\|^2 \leq \|R\|\,\|\delta_v\| + \mathcal{N}\,\|y^n-y^*\|_V^2\,(8\sigma\,\|\lambda^*\|_V\,\|\delta_y\|_V + \|\delta_\lambda\|_V) \\ + 2\mathcal{N}\,\|y^n-y^*\|_V\,\|\lambda^n-\lambda^*\|_V\,\|\delta_y\|_V\,.\end{aligned} \tag{5.34}$$

On the other hand, by multiplying (5.14) by δ_λ we get that

$$\begin{aligned}\nu\,\|\delta_\lambda\|_V^2 + \frac{1}{\varepsilon}\,\|\delta_\lambda\|^2 - \mathcal{M}(y^n)\,\|\delta_\lambda\|_V^2 \leq 2\mathcal{N}\,\|\delta_y\|_V\,\|\lambda^n\|_V\,\|\delta_\lambda\|_V \\ + 2\mathcal{N}\,\|y^n-y^*\|_V\,\|\lambda^n-\lambda^*\|_V\,\|\delta_\lambda\|_V + \Big(\frac{\alpha}{\varepsilon^2}+1\Big)\,\|\delta_y\|\,\|\delta_\lambda\| + \frac{\alpha}{\varepsilon^2}\,\|\delta_v\|\,\|\delta_\lambda\|\,,\end{aligned} \tag{5.35}$$

which, considering (5.32) and (5.26), implies that

$$\begin{aligned}\frac{1}{2}(\nu - \mathcal{M}(y^*))\,\|\delta_\lambda\|_V \leq \left(\frac{\kappa^2\alpha}{\varepsilon^2} + \kappa^2 + 4\mathcal{N}\,\|\lambda^*\|_V\right)\left(\frac{\alpha+\varepsilon^2}{\varepsilon^2}\right)\|\delta_y\|_V \\ + \frac{\alpha}{\varepsilon^2}\,\|\delta_v\| + 2\mathcal{N}\,\|\lambda^n-\lambda^*\|_V\,\|y^n-y^*\|_V\,.\end{aligned} \tag{5.36}$$

Therefore, utilizing estimate (5.28), there exists a constant $\bar{C}$ such that

$$\|\delta_\lambda\|_V \leq \bar{C}(\|\delta_v\| + \|\lambda^n-\lambda^*\|_V^2 + \|y^n-y^*\|_V^2). \tag{5.37}$$

From the definition of φ and (5.18) we therefore obtain that

$$\|R\| = o(\|(y^n-y^*,\lambda^n-\lambda^*)\|_{\mathbf{L}^p\times\mathbf{L}^p}),$$

which, considering estimates (5.28) and (5.37) in (5.34), implies the existence of a constant C such that

$$\begin{aligned}\|(\delta_v,\delta_y,\delta_\lambda)\|_{\mathbf{L}^2\times V\times V} \leq C\left(\|y^n-y^*\|_V^2 + \|\lambda^n-\lambda^*\|_V^2\right) \\ + o(\|(y^n-y^*,\lambda^n-\lambda^*)\|_{\mathbf{L}^p\times\mathbf{L}^p}).\end{aligned} \tag{5.38}$$

Consequently, the superlinear convergence is verified. ∎

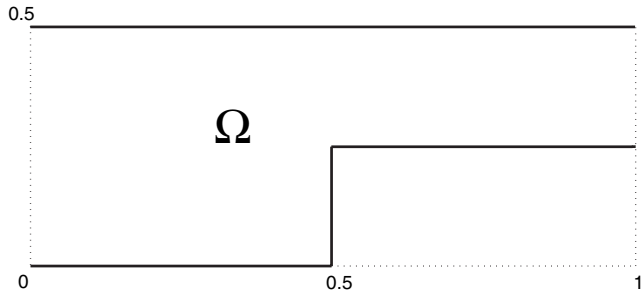


FIG. 6.1. *Forward facing step channel.*

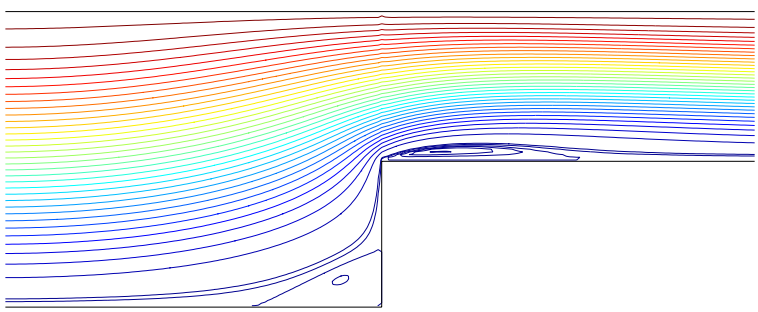


FIG. 6.2. *Streamlines of the uncontrolled state.*

6. Numerical results. For the numerical tests, a "forward facing step channel" was utilized (see Figure 6.1). The fluid flows from left to right with a parabolic inflow condition with the maximum value equal to one and "do nothing" outflow condition. In the remaining boundary parts a homogeneous Dirichlet condition was imposed. The geometry was discretized using a staggered grid and an upwinding finite differences scheme was applied. The behavior of the uncontrolled fluid flow with Reynolds number $Re = 1000$ is depicted in Figure 6.2. Two main recirculation zones, which increase their size together with the Reynolds number, can be clearly identified from the graphics.

The target of our control problem is to properly diminish the recirculations of interest by considering, together with the tracking-type cost functional, adequate pointwise control-state constraints.

For the solution of the optimality system, Algorithm 5.2 was utilized. The semismooth Newton algorithm stops when the $\mathbf{L}^2$-norm of the state increment is lower than 10^{-4}. Unless otherwise specified, the mesh step size $h = 1/240$ was considered. For the solution of the linear systems, MATLAB's exact solver was utilized.

6.1. Example 1. In this first experiment, we consider the elimination of bubbles in the channel by imposing the constraint $y_1 + \varepsilon u_1 \geq -10^{-7}$. For ε sufficiently small, this constraint avoids backward flow in the channel and thus possible recirculations. Additionally, the tracking-type component of the cost functional is responsible for a more linear behavior of the flow field. The remaining parameter data utilized are $h = 1/240$, $Re = 1000$, $\varepsilon = 10^{-4}$, and $\alpha = 0.1$. The semismooth Newton method (SSN) stops after 9 iterations, with the final active set containing 28 grid points. The cost functional takes the final value $J(y^*, u^*) = 0.00445224$ and the nonlinear

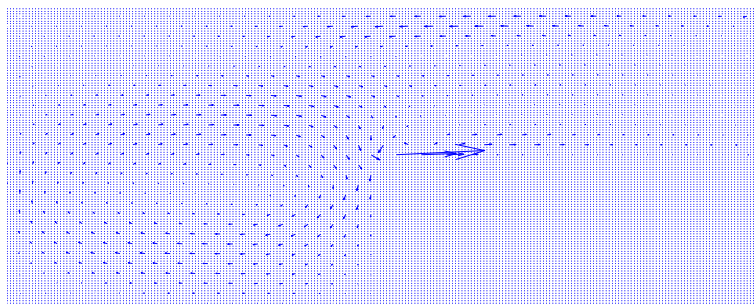


FIG. 6.3. *Example* 1: *Control vector field with tracking component.*

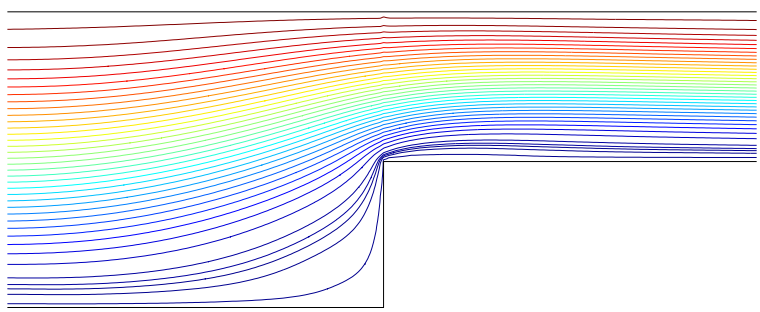


FIG. 6.4. *Example* 1: *Streamlines of the controlled state with tracking component.*

TABLE 6.1
Example 1: $h = 1/240$, $tol = 10^{-4}$.

ε	SSN iter.	$J(y^*, u^*)$	$\lvert\mathcal{A}^a \cup \mathcal{A}^b\rvert$
10^{-1}	5	0.00399972	33
10^{-2}	6	0.00410360	42
10^{-3}	8	0.00438273	29
10^{-4}	9	0.00445224	28
10^{-5}	9	0.00445989	32

complementarity function residuum the value 2.2737×10^{-9}. The optimal control field is depicted in Figure 6.3, where the concentration of the control action on the recirculation zones can be observed. The desired recirculation diminishing effect of the control can be visualized from the plot of the reached controlled state streamlines in Figure 6.4. In Table 6.1 the number of SSN iterations, the final cost functional value, and the size of the active set are registered for different ε values. It can be observed that as ε tends to 0, the problem becomes harder to solve and more SSN iterations are required.

Subsequently we consider the limit case where the tracking-type part of the cost functional is dismissed. We aim to find the control of minimum norm that allows the satisfaction of the state constraint $y_1 + \varepsilon u_1 \geq 10^{-7}$ over the domain of interest. As before, the constraint takes care that no important backward flow arises. By considering the constraint on the whole domain, i.e., $\Omega_S = \Omega$, both recirculations before and after the step are diminished (see Figure 6.5). From Figure 6.5 it can also be observed that the behavior of the fluid flow, mainly before the step, is not close

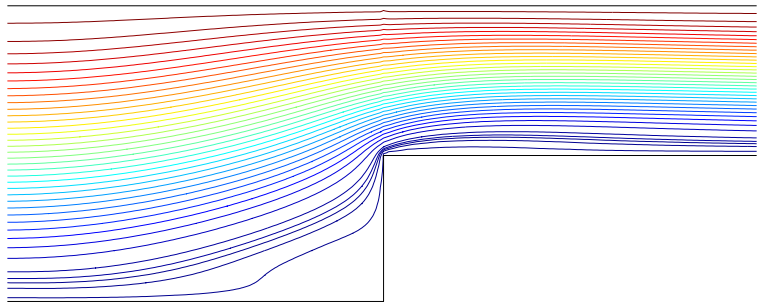


FIG. 6.5. *Example* 1: *Streamlines of the controlled state without tracking component.*

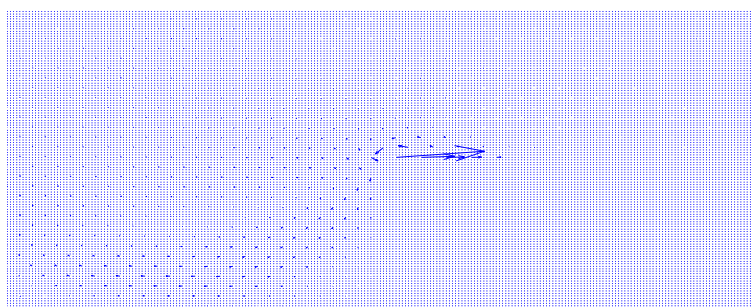


FIG. 6.6. *Example* 1: *Control vector field without tracking component.*

to the Stokes flow, as is the case when the tracking-type component is present. From the control vector plot (see Figure 6.6) it can be observed that the control action in this case is even more concentrated on the recirculations zones. The parameter values for this case are $Re = 1000$, $\varepsilon = 10^{-4}$, and $\alpha = 0.1$. The number of SSN iterations needed is 29 and the cost functional takes the final value 8.99816×10^{-4}.

In many practical cases, the recirculation reduction or elimination on the whole domain is not necessary, if not undesirable. In such cases the state constraint may be imposed in the sectors where the bubble to be diminished is localized. In the case of our geometry the essential recirculation to be diminished is the one after the step. By considering the state constraint on the subdomain $\Omega_S := [0.5, 0.75] \times [0.25, 0.5]$, this elimination is attached with the cost functional value 8.98898×10^{-4} in 6 SSN iterations. The final controlled state is shown in Figure 6.7, where it can be clearly seen that the recirculation after the step is numerically eliminated, although the one before the step becomes bigger than in the uncontrolled case.

6.2. Example 2. As an alternative strategy for the reduction of the recirculation after the step, we consider in this example a state constraint that guarantees a homogeneous outflow velocity. The constraint imposed is $y_1 + \varepsilon u_1 \leq 1.7$ and the remaining parameter values are $Re = 1000$, $\varepsilon = 10^{-3}$, and $\alpha = 0.01$. In this case, the SSN algorithm stops after 15 iterations and the resulting active set contains 2283 grid points. The cost functional takes the final value $J(y^*, u^*) = 0.003470768$. The controlled state is depicted in Figure 6.8, where an important reduction of the recirculations can be visualized.

Since the outgoing velocity is the quantity of interest, it is natural to consider the case where the constraint is imposed only in the last part of the channel. By

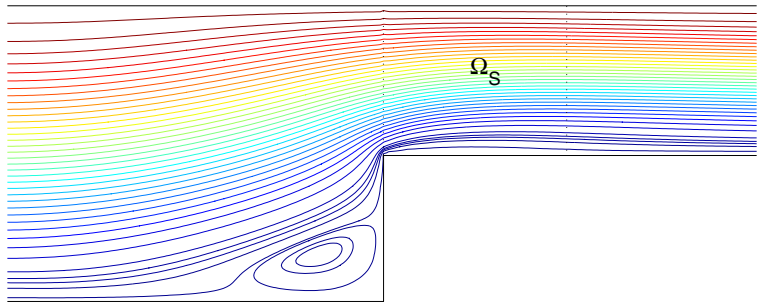


FIG. 6.7. *Example* 1: *Streamlines of the controlled state without tracking component; state constraint subdomain.*

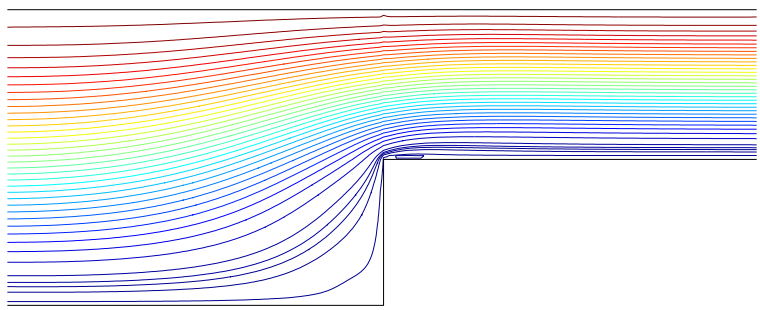


FIG. 6.8. *Example* 2: *Streamlines of the controlled state.*

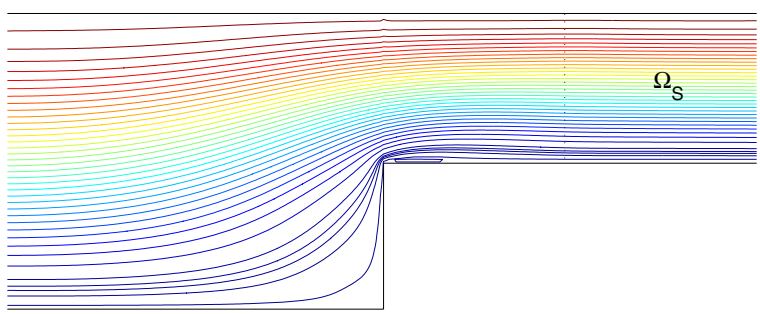


FIG. 6.9. *Example* 2: *Streamlines of the controlled state; state constraint subdomain.*

considering the domain $\Omega_S := [0.5, 0.75] \times [0.25, 0.5]$, the recirculation diminishing effect does also take place (see Figure 6.9), but with a lower final cost functional value $J(y^*, u^*) = 0.0031112131$. The SSN algorithm stops after 10 iterations with a final active set containing 906 active points. The remaining parameter values are the same as in the case $\Omega_S = \Omega$.

Finally, in order to visualize the structure of the control-state constraint multiplier, we modify the Reynolds number to 500 and impose the homogeneous outgoing velocity constraint $y_1 + \varepsilon u_1 \leq 1.7$. The evolution of the multiplier as ε decreases can be observed in Figure 6.10. In Table 6.2 the evolution of the SSN is registered. The algorithm stops after 7 iterations with the final active set containing 2465 grid points. As expected from the theoretical results, local superlinear convergence can be

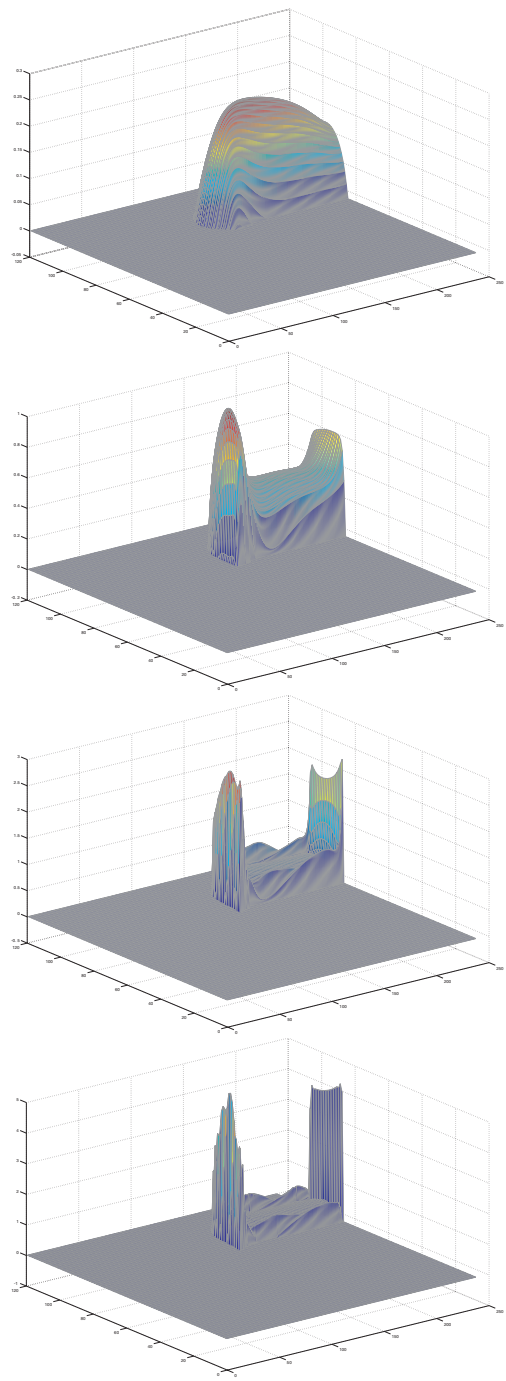


FIG. 6.10. *State constraint multiplier;* $\varepsilon = 10^{-1},\ \varepsilon = 10^{-2},\ \varepsilon = 10^{-3},\ \varepsilon = 10^{-4}$.

TABLE 6.2
Example 2: $h = \frac{1}{240}$, $\varepsilon = 10^{-3}$, $Re = 500$.

Iteration	$\mid \mathcal{A}_n \mid$	$J(y,u)$	$\|y_n - y_{n-1}\|$	$\frac{\|y_n - y_{n-1}\|}{\|y_{n-1} - y_{n-2}\|}$	NCP
1	0	0.00156432	9.4321	-	29.43065
2	2743	0.00349897	12.40964	-	4.531425
3	2571	0.003355	1.05301	0.0077	1.663621
4	2494	0.0033477	0.2134005	0.201	0.397106
5	2469	0.00334765	0.0151623	0.07079	0.052505
6	2465	0.00334765	$5.55 \cdot 10^{-4}$	0.03634	$2.22 \cdot 10^{-14}$
7	2465	0.00334765	$2.048 \cdot 10^{-8}$	$3.86 \cdot 10^{-5}$	$2.22 \cdot 10^{-14}$

observed from the data. Let us point out that, although no monotonic behavior of the cost functional along the iterations occurs, a monotonic decrease of the nonlinear complementarity function and of the size of the active set can be observed.

REFERENCES

[1] F. ABERGEL AND R. TEMAM, *On some control problems in fluid mechanics*, Theoretical and Computational Fluid Mechanics, 1 (1990), pp. 303–325.

[2] M. BERGOUNIOUX AND K. KUNISCH, *On the structure of Lagrange multipliers for state-constrained optimal control problems*, Systems Control Lett., 48 (2003), pp. 169–176.

[3] E. CASAS, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.

[4] E. CASAS, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[5] E. CASAS, *Optimality conditions for some control problems of turbulent flows*, in Flow Control (Minneapolis, MN, 1992), IMA Vol. Math. Appl. 68, Springer-Verlag, New York, 1995, pp. 127–147.

[6] E. CASAS AND F. TRÖLTZSCH, *Second-order necessary and sufficient optimality conditions for optimization problems and applications to control theory*, SIAM J. Optim., 13 (2002), pp. 406–431.

[7] J. C. DE LOS REYES, *A primal-dual active set method for bilaterally control constrained optimal control of the Navier–Stokes equations*, Numer. Funct. Anal. Optim., 25 (2005), pp. 657–683.

[8] J. C. DE LOS REYES AND R. GRIESSE, *State Constrained Optimal Control of the Stationary Navier–Stokes Equations*, Preprint 22-2005, Institute of Mathematics, TU-Berlin, Berlin, Germany, 2005.

[9] J. C. DE LOS REYES AND K. KUNISCH, *A semi-smooth Newton method for control constrained boundary optimal control of the Navier–Stokes equations*, Nonlinear Anal., 62 (2005), pp. 1289–1316.

[10] J. C. DE LOS REYES AND K. KUNISCH, *A semi-smooth Newton method for regularized state constrained optimal control of the Navier–Stokes equations*, Computing, 78 (2006), pp. 287–309.

[11] H. O. FATTORINI AND S. S. SRITHARAN, *Optimal control problems with state constraints in fluid mechanics and combustion*, Appl. Math. Optim., 38 (1998), pp. 159–192.

[12] V. GIRAULT AND P. A. RAVIART, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.

[13] M. D. GUNZBURGER, *Perspectives in Flow Control and Optimization*, Adv. Des. Control 5, SIAM, Philadelphia, 2002.

[14] M. D. GUNZBURGER, L. HOU, AND T. P. SVOBODNY, *Analysis and finite element approximation of optimal control problems for the stationary Navier–Stokes equations with distributed and Neumann controls*, Math. Comput., 57 (1991), pp. 123–151.

[15] M. D. GUNZBURGER AND S. MANSERVISI, *Analysis and approximation of the velocity tracking problem for Navier–Stokes flows with distributed control*, SIAM J. Numer. Anal., 37 (2000), pp. 1481–1512.

[16] M. Heinkenschloss, *Formulation and analysis of sequential quadratic programming method for the optimal Dirichlet boundary control of Navier–Stokes flow*, in Optimal Control (Gainesville, FL, 1997), Kluwer, Dordrecht, The Netherlands, 1998, pp. 178–203.

[17] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.

[18] M. Hintermüller and M. Hinze, *A SQP-semismooth Newton-type algorithm applied to control of the instationary Navier–Stokes system subject to control constraints*, SIAM J. Optim., 16 (2006), pp. 1177–1200.

[19] M. Hinze and K. Kunisch, *Second order methods for optimal control of time-dependent fluid flow*, SIAM J. Control Optim., 40 (2001), pp. 925–946.

[20] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Stud., 14 (1981), pp. 163–177.

[21] C. Meyer, A. Rösch, and F. Tröltzsch, *Optimal control of PDEs with regularized pointwise state constraints*, Comput. Optim. Appl., 33 (2006), pp. 209–228.

[22] C. Meyer and F. Tröltzsch, *On an elliptic optimal control problem with pointwise mixed control-state constraints*, in Recent Advances in Optimization. Proceedings of the 12th French-German-Spanish Conference on Optimization, Lecture Notes in Econom. and Math. Systems 563, Springer-Verlag, New York, 2006, pp. 187–204.

[23] R. Temam, *Navier Stokes Equations: Theory and Numerical Analysis*, North–Holland, Amsterdam, 1979.

[24] F. Tröltzsch, *Optimalsteuerung bei partiellen Differentialgleichungen*, Vieweg Verlag, Wiesbaden, Germany, 2005.

[25] F. Tröltzsch and D. Wachsmuth, *Second order sufficient optimality conditions for the optimal control of Navier–Stokes equations*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 93–119.

[26] M. Ulbrich, *Constrained optimal control of Navier–Stokes flow by semismooth Newton methods*, Systems Control Lett., 48 (2003), pp. 297–311.

[27] G. Wang, *Optimal controls of* 3*-dimensional Navier–Stokes equations with state constraints*, SIAM J. Control Optim., 41 (2002), pp. 583–606.

[28] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, *Vol.* 1, Springer-Verlag, New York, 1986.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 630–654

A HARMONIC FRAMEWORK FOR CONTROLLABILITY IN LINEAR CONTINUOUS-TIME PERIODIC SYSTEMS*

JUN ZHOU†

Abstract. Controllability of linear continuous-time periodic systems is dealt with via a harmonic analysis approach for the first time in this paper. This approach reveals that controllability of continuous-time periodic systems can be connected to necessary and sufficient conditions expressed explicitly with Fourier coefficients of the system matrices, which can be interpreted in a way similar to what we have seen in linear time-invariant cases. These controllability conditions shed new light upon structural characteristics of continuous-time periodic systems that are hard to know by means of existing time-domain controllability criteria in the literature. Controllability canonical decomposition of linear continuous-time periodic systems is revisited through state coordinate transforms of strong analytic features. The results are heuristic and significant for examining structural characteristics of continuous-time periodic systems and extending controllability-related techniques that are widely employed in linear time-invariant systems to linear continuous-time periodic systems.

Key words. controllability, periodic systems, harmonic analysis, structural decomposition

AMS subject classifications. 43A32, 65N12, 65F40, 93D20

DOI. 10.1137/050639934

1. Introduction.

Controllability is one of the structural characteristics in dynamical systems, and plays an irreplaceable role in analysis techniques and synthesis algorithms in control theory. For example, in linear time-invariant (LTI) control systems, zero/pole structural algebra [30], [32], geometric theory [39], frequency-domain techniques for multivariable control [36], linear matrix inequalities [6], H_2/H_∞ robust performance designs [13], [22], [44], and so on cannot proceed without controllability. This is also the case for engineering applications involving finite-dimensional linear continuous-time periodic (FDLCP) systems [20], [18], [21], [25], among which typical problems include stabilization of helicopter rotors and ships in waves, and reduction of electromechanical oscillations in electricity generators [9], [10], [14], [28], [29]. As a matter of fact, controllability is a prerequisite to many problems involving FDLCP modelings, for instance, in periodic Riccati [1] and Lyapunov differential equations [5]. Unfortunately, however, due to the time-varying feature of FDLCP systems, controllability and its relevant structural properties in FDLCP systems still need further scrutiny, compared with the situation in LTI systems, though various controllability concepts and corresponding criteria are introduced and lasting efforts have been made [1], [7], [11], [15], [17], [33], [34] over the last dozens of years.

Now we simply review results about controllability in FDLCP systems to motivate our study. It is well known that in LTI systems the Gramian, rank, and Popov–Belevitch–Hautus (PBH) criteria [39], [30], [44] are frequently employed for controllability, and have brought fruitful results about the structure characteristics of LTI systems [30], [32]. For instance, the structural decomposition has greatly deepened our understanding about such structural properties as zero/pole cancellation in feedback control, which in turn pave the way for optimal linear quadratic regulation

*Received by the editors September 9, 2005; accepted for publication (in revised form) September 25, 2006; published electronically May 7, 2007.
http://www.siam.org/journals/sicon/46-2/63993.html

†Department of Electrical Engineering, Kyoto University, Kyotodaigaku-Katsura, Nishikyo-ku, Kyoto 615-8510, Japan (zhouj@kuee.kyoto-u.ac.jp).

problems [36], H_2/H_∞ performance designs [13], [44], etc. In contrast, it is relatively hard to tackle controllability in the FDLCP field. There seem to be few controllability criteria for FDLCP systems that possess time-invariant expressions of the system matrices and can be implemented without the state transition matrix knowledge of FDLCP systems. Usually one must face some conservative assumptions about system matrices when utilizing such controllability criteria as that of [33], or one must base controllability testing on the state transition matrix [1], [2], [3], [4], [25], whose computation is itself another thorny and interesting problem in the FDLCP field [24], [26], [35]. In view of this, it would be fair to say that controllability and its corresponding properties in FDLCP systems have not been well addressed for the purpose of carrying over LTI techniques into FDLCP systems through controllability/observability concepts. Thus, it is natural for us to develop controllability criteria which have more explicit expressions with system matrices. Such controllability criteria are the major achievements of this paper.

Although more than a dozen kinds of controllability are described in the literature, three kinds are generally known: K-controllability proposed by Kalman [20], which leads to the Gramian controllability criterion; Kalman–Weiss–Brunovsky (KWB)-controllability [17], which is stated via linear independence of the state transition and input matrices; and H-controllability defined by Hewer [2], [17], which is stated via characteristic-multiplier/
eigenvector relations of the monodromy and input matrices of the FDLCP system concerned. The first two are defined for general dynamical systems, while the third one is introduced specifically for FDLCP systems. It has been claimed in [17] that K-controllability is equivalent to KWB-controllability. It should be noted that some arguments about validity of results in [17] are raised through a counter example in [1]. Equivalence between K-controllability and H-controllability is explicated in [2]. Bearing these in mind, we concentrate our attention in what follows only on K-controllability, and thus all the prefixes will be removed from terminologies pertaining to controllability. As an interesting side note related closely to controllability (and observability), it is worth mentioning that various characterizations of stabilizability and detectability of linear continuous-time periodic systems are given and their equivalences are scrutinized in [4].

The paper is outlined as follows. Section 2 collects preliminaries to FDLCP systems, such as the Floquet theorem, the harmonic Floquet similarity transformation formulas, the controllability definition, and the Gramian criterion and its various interpretations. In section 3, harmonic controllability conditions for FDLCP systems and their implementing algorithms are proved. Controllability of approximate modelings and controllability canonical decomposition via analytic state coordinate transforms are dealt with in section 4. There are numeric examples to illustrate the main results in section 5. Conclusions are summarized in section 6.

2. Preliminaries to FDLCP systems. Section 2.1 collects facts on FDLCP systems. Section 2.2 moves to discussions about exponential operators defined in FDLCP systems. Section 2.3 is devoted to the controllability definition, the Gramian controllability criterion, and its various interpretations. The results of section 2.3 play a key role in developing a harmonic framework for controllability.

To facilitate the statements we list notation used in the paper. $||\cdot||$ is the Euclidean norm and the norm of a matrix induced by it. l_2 is the set of all infinite-dimensional vectors $\underline{x}$ such that $||\underline{x}||_{l_2}^2 = \sum_{k=-\infty}^{+\infty} ||x_k||^2 < \infty$, where x_k is the kth (vector) entry of $\underline{x}$. $L_\infty[0,h]$ is the set of all measurable functions x defined on $[0,h)$ such that

$||x(\cdot)||_\infty = \text{ess sup}_{t\in[0,h)} ||x(t)|| < \infty$. $L_{\rm PCD}[0,h]$ is the set of all piecewise continuous functions that are differentiable almost everywhere in $[0,h)$. $L_{\rm CAC}[0,h]$ denotes the set of all continuous functions whose Fourier series are absolutely convergent. Obviously, $L_{\rm PCD}[0,h] \subset L_\infty[0,h]$ and $L_{\rm CAC}[0,h] \subset L_\infty[0,h]$. $F(t) \in L_\infty[0,h]$ means that $F(\cdot)$ is a matrix function, each element of which is h-periodic and belongs to $L_\infty[0,h]$ when its domain is restricted to $[0,h)$. $\mathcal{C}$ is the set of all complex numbers. $\mathcal{Z}$ is the ring of all integers. $*(l,k)$ denotes an $l\times k$ matrix, the exact evaluation of whose entries is not needed, while $0(l,k)$ is an $l\times k$ zero matrix.

Let $\sum_{\mu=-\infty}^{+\infty} A_\mu e^{j\mu\omega_h t}$ with $\omega_h := 2\pi/h$ be the Fourier series of $A(t) \in L_\infty[0,h]$. The Toeplitz transformation $\mathcal{T}\{A(t)\}$ is a Toeplitz operator [38] (or block Laurent operator [12]) given by

$$\mathcal{T}\{A(t)\} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \iddots \\ \cdots & A_0 & A_{-1} & A_{-2} & \cdots \\ \cdots & A_1 & A_0 & A_{-1} & \cdots \\ \cdots & A_2 & A_1 & A_0 & \cdots \\ \iddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} =: \underline{A},$$

where $A(t)$ is called the defining function of $\underline{A}$ by the terminology of [12].

2.1. Floquet theorem and its interpretations. Consider the FDLCP system given by

$$\dot{x} = A(t)x + B(t)u, \tag{1}$$

where $A(t)$ and $B(t)$ are h-periodically time-varying $n\times n$ and $n\times m$ matrices, respectively. By the Floquet theorem [10], [16], [23], [24], if the entries of $A(t)$ are piecewise continuous in t, the state transition matrix $\Phi(t,0)$ of (1) can be expressed in a Floquet factorization $\Phi(t,0) = P(t,0)e^{Qt}$, where $P(t,0)$ is absolutely continuous in t and nonsingular and h-periodic in t, and Q is constant but probably complex. Conventionally, $\Phi(h,0)$ is called the monodromy of (1), whose eigenvalues are called characteristic multipliers, while the eigenvalues of Q are called characteristic exponents.

Define the fundamental strip $\mathcal{C}_f$ on the complex place $\mathcal{C}$ as follows:

$$\mathcal{C}_f := \{z \in \mathcal{C} : -\omega_h/2 < \text{Im}(z) \le \omega_h/2\}, \quad \omega_h = 2\pi/h.$$

In a Floquet factorization $\Phi(t,0) = P(t,0)e^{Qt}$, if all characteristic exponents belong to $\mathcal{C}_f$, then $P(t,0)e^{Qt}$ is called a Floquet simplex. Lemma 2.1 claims that Floquet simplices exist in general FDLCP systems. A proof for Lemma 2.1 is given in Appendix A. Floquet simplices play a key role in establishing the main results of this paper.

LEMMA 2.1. *In the* FDLCP *system* (1)*, let the entries of* $A(t)$ *be piecewise continuous in* t*.*

(a) *The state transition matrix* $\Phi(t,0)$ *of* (1) *can always be expressed in a Floquet simplex; that is,* $\Phi(t,0) = P(t,0)e^{Qt}$ *with* $\lambda(Q) \subset \mathcal{C}_f$*;*

(b) *Suppose that* $\Phi(t,0)$ *possesses a Floquet factorization* $\tilde{P}(t,0)e^{\tilde{Q}t}$ *satisfying*

$$\tilde{P}(t,0) = \sum_{|k|\le N_p} \tilde{P}_k e^{jk\omega_h t}, \quad \lambda(\tilde{Q}) \subset \bigcup_{|k|\le N_q} \{\mathcal{C}_f + jk\omega_h\} \tag{2}$$

for some integers $N_p \geq 0$ and $N_q \geq 0$; that is, the Fourier series of $\tilde{P}(t,0)$ contains finitely many harmonic waves, and the eigenvalues of $\tilde{Q}$ belong to a horizontally bounded strip in the complex plane formed along $\mathcal{C}_f$. Then, there exists a Floquet simplex $\Phi(t,0) = P(t,0)e^{Qt}$ such that

$$P(t,0) = \sum_{|k| \leq N_p + N_q} P_k e^{jk\omega_h t}, \quad \lambda(Q) \subset \mathcal{C}_f. \tag{3}$$

Here $\tilde{P}_k$ and P_k denote the Fourier coefficients of $\tilde{P}(t,0)$ and $P(t,0)$, respectively. In the above, $\lambda(\cdot)$ is the set of all the eigenvalues of a matrix $(\cdot)$.

By the first assertion of Lemma 2.1, an interesting point about a Floquet simplex $\Phi(t,0) = P(t,0)e^{Qt}$ is that for any two eigenvalues of Q, $\mathrm{Im}(\lambda_i(Q)) - \mathrm{Im}(\lambda_k(Q)) \neq r\omega_h$ for $i \neq k$, where r is a nonzero integer. This is a key result for establishing the harmonic framework for controllability. The second assertion of Lemma 2.1 plays a role in deriving a numerically implementable controllability criterion from infinite-dimensional harmonic controllability criteria.

Now we recall the harmonic Floquet similarity transformation formula [41], [42], which has an important role in defining exponential operators of the unbounded harmonic state operators [43] in Proposition 2.3 and establishing controllability criteria in Theorems 3.2 and 3.3. In what follows, we write $\underline{B} := \mathcal{T}\{B(t)\}$, $\underline{P} := \mathcal{T}\{P(t,0)\}$, $\underline{\hat{B}} := \mathcal{T}\{P^{-1}(t,0)B(t)\}$, $\underline{Q} := \mathcal{T}\{Q\}$, and

$$\underline{E}(j0) := \mathrm{diag}[\cdots, -j2\omega_h I, -j\omega_h I, 0, j\omega_h I, j2\omega_h I, \cdots],$$

where 0-block is at the center of the infinite-dimensional matrix $\underline{E}(j0)$, which is unbounded on l_2. As an appropriate domain for $\underline{E}(j0)$, we define the set $l_E := \{\underline{x} \in l_2 : \underline{E}(j0)\underline{x} \in l_2\}$. It is shown [41] that l_E is a proper subset of l_2 and dense in l_2.

LEMMA 2.2. *In the* FDLCP *system* (1)*, let $A(t) \in L_{\mathrm{PCD}}[0,h]$ and $\Phi(t,0) = P(t,0)e^{Qt}$ be a Floquet factorization. Then, $\underline{P}$ is invertible both on l_E and l_2, and the inverse of $\underline{P}$ on l_E is that of $\underline{P}$ on l_2 restricted to l_E. Also, l_E is $\underline{P}$-, $\underline{P}^{-1}$-, $\underline{P}^H$-, and $\underline{P}^{-H}$-invariant. The unbounded operators $\underline{P}(\underline{E}(j0) - \underline{Q})\underline{P}^{-1}$ and $\underline{E}(j0) - \underline{A}$ are densely defined on l_2 and coincide on l_E with each other:*

$$\underline{P}(\underline{Q} - \underline{E}(j0))\underline{P}^{-1} = \underline{A} - \underline{E}(j0).$$

Moreover, if $B(t) \in L_{\mathrm{CAC}}[0,h]$, then $\hat{B}(t) = P^{-1}(t,0)B(t) \in L_{\mathrm{CAC}}[0,h]$ and $\underline{\hat{B}} = \underline{P}^{-1}\underline{B}$.

Furthermore, let $\Lambda(\cdot)$ denote the collection of all eigenvalues of an operator $(\cdot)$. Then

$$\Lambda(\underline{A} - \underline{E}(j0)) = \Lambda(\underline{Q} - \underline{E}(j0)) = \{\lambda(Q) + j\mu\omega_h : \mu \in \mathcal{Z}\} =: \Lambda,$$

and for each eigenvalue $\lambda \in \Lambda$ there exists an associated eigenvector of $\underline{A} - \underline{E}(j0)$ (or $\underline{Q} - \underline{E}(j0)$) that belongs to the space l_E.

2.2. Exponential operators defined on $\underline{A} - \underline{E}(j0)$ and $\underline{Q} - \underline{E}(j0)$. For the subsequent arguments, we introduce the following two infinite-dimensional exponential operators about the unbounded operators $\underline{A} - \underline{E}(j0)$ and $\underline{Q} - \underline{E}(j0)$:

$$\begin{cases} \underline{e}(\underline{A} - \underline{E}(j0), t) := \displaystyle\sum_{k=0}^{\infty} \frac{1}{k!}(\underline{A} - \underline{E}(j0))^k t^k, \\ \underline{e}(\underline{Q} - \underline{E}(j0), t) := \displaystyle\sum_{k=0}^{\infty} \frac{1}{k!}(\underline{Q} - \underline{E}(j0))^k t^k. \end{cases} \tag{4}$$

Since $\underline{A}-\underline{E}(j0)$ and $\underline{Q}-\underline{E}(j0)$ are unbounded on the Hilbert space l_2 and their infinite powers are involved in the definition, the domains of $\underline{A}-\underline{E}(j0)$ and $\underline{Q}-\underline{E}(j0)$ must be restricted appropriately in l_E to guarantee that $\underline{e}(\underline{A}-\underline{E}(j0),t)$ and $\underline{e}(\underline{Q}-\underline{E}(j0),t)$ are well defined.

To this end, we denote the eigenspace of an eigenvalue $\lambda_i \in \Lambda$ of $\underline{A}-\underline{E}(j0)$ and $\underline{Q}-\underline{E}(j0)$, respectively, by $\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{A})$ and $\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{Q})$, where $\underline{I}=\mathcal{T}\{I\}$. It is not hard to show that for any nonnegative integer k, $(\underline{A}-\underline{E}(j0))^k$ and $(\underline{Q}-\underline{E}(j0))^k$ can be restricted to $\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{A})$ and $\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{Q})$, respectively. Simple algebra yields that $\underline{e}(\underline{A}-\underline{E}(j0),t):\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{A})\to\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{A})$ and $\underline{e}(\underline{Q}-\underline{E}(j0),t):\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{Q})\to\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{Q})$ are bounded for each fixed $\lambda_i\in\Lambda$ and $t\in[0,\infty)$. Based on Lemma 2.2 one can readily assert the results in Proposition 2.3.

PROPOSITION 2.3. *In the* FDLCP *system* (1)*, let* $A(t)\in L_{\mathrm{PCD}}[0,h]$ *and* $\Phi(t,0)=P(t,0)e^{Qt}$ *be a Floquet factorization. Then,* $\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{A})\subset l_E$ *and* $\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{Q})\subset l_E$ *satisfying* $\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{A})=\underline{P}\,\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{Q})$*. Furthermore, for each fixed* $\lambda_i\in\Lambda$*, it holds on* $\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{A})$ *that* $\underline{e}(\underline{A}-\underline{E}(j0),t)=\underline{P}\,\underline{e}(\underline{Q}-\underline{E}(j0),t)\underline{P}^{-1}$.

2.3. Controllability and the Gramian criterion. As argued in the introduction, there are many ways to define controllability in general dynamic systems. Here we recall only the controllability definition given by Kalman [20], which has been refined over the years in the FDLCP setting.

DEFINITION 2.4. *Let* J *be a time interval. We call a state* $x(t_0)$ *of a dynamic system considered at time* $t_0\in J$ *controllable if there exist some time* $t_1>t_0$ *and an integrable input* $u(\cdot)$ *defined on* (t_0,t_1) *which transfers the state* $x(t_0)$ *to* $x(t_1)=0$.

If for a fixed time $t_0\in J$ *every state* $x(t_0)$ *is controllable in the above sense, we say that the system is completely controllable at* $t_0\in J$.

If for every $t_0\in J$ *the system is completely controllable, we say that the system is completely controllable over* J. *If* $J=[0,\infty)$*, we say simply that the system is completely controllable.*

If the system (1) is completely controllable, we also say that the pair $(A(\cdot),B(\cdot))$ is completely controllable. Based on the Gramian criterion, it can be proved that controllability of the FDLCP system (1) over $[0,\infty)$ can be reduced to that of the same system over $[0,kh)$, where k is a positive integer defined in Lemma 2.5. Regarding this controllability interval reduction a short but extensive survey can be found in [26]. The results [2], [7], [15] directly related to our later discussions are summarized in Proposition 2.6, which is proved in Appendix B. To claim these results in the current fashion may facilitate the reader's understanding about the subsequent arguments. This is especially true for Lemma 2.5 and Proposition 2.7.

To facilitate our statements in Propositions 2.6 and 2.7, we define

$$W_c[t_0,t_1]:=\int_{t_0}^{t_1}\Phi(t_0,\tau)B(\tau)B^T(\tau)\Phi^T(t_0,\tau)d\tau. \tag{5}$$

Clearly, $W_c[t_0,t_1]$ is at least positive semidefinite. If $W_c[t_0,t_1]$ is positive semidefinite, we simply write $W_c[t_0,t_1]\geq 0$ or, equivalently, $W_c[t_0,t_1]$ is singular or not of full rank. If $W_c[t_0,t_1]$ is strictly positive definite, we simply write $W_c[t_0,t_1]>0$ or, equivalently, $W_c[t_0,t_1]$ is nonsingular or of full rank. The Gramian criterion states that the state $x(t_0)$ is completely controllable at t_0 if and only if there exists some $t_1>t_0$ such that $W_c[t_0,t_1]>0$. The following lemma, whose proof is given in Appendix A, plays a key role in establishing Propositions 2.6 and 2.7.

LEMMA 2.5. *In the* FDLCP *system* (1)*, assume that the entries of* $A(t)$ *are piecewise continuous with respect to* t*,* $B(t)$ *belongs to* $L_\infty[0,h]$*, and* $\Phi(t,0) = P(t,0)e^{Qt}$ *is a Floquet factorization. If* $W_c[t_0, t_0+kh] \geq 0$ *with* $k \geq 1$ *being an integer larger than or equal to the degree of the minimal polynomial of the monodromy matrix* $\Phi(h,0) = e^{Qh}$*, there exists a left eigenvector* $\mathbf{a}^H$ *of* Q *such that for any integer* $v \geq 1$*,* $\mathbf{a}^H W_c[t_0, t_0+vh]\mathbf{a} = 0$*.*

PROPOSITION 2.6. *In the* FDLCP *system* (1)*, assume that the entries of* $A(t)$ *are piecewise continuous with respect to* t *and* $B(t)$ *belongs to* $L_\infty[0,h]$*. Let* $W_c[0,kh]$ *be defined as in* (5) *with the integer* k *defined in Lemma* 2.5*. Then the following three assertions are equivalent to each other.*

(a) *The system is completely controllable;*

(b) *The system is completely controllable over* $[0,kh)$*;*

(c) $W_c[0,kh]$ *is positive definite; that is,* $W_c[0,kh] > 0$*.*

Combining Lemma 2.5 and Proposition 2.6, the following result follows readily. Proposition 2.7 is important in deriving a controllability criterion based on Floquet factorizations.

PROPOSITION 2.7. *In the* FDLCP *system* (1)*, assume that the entries of* $A(t)$ *are piecewise continuous in* t*,* $B(t)$ *belongs to* $L_\infty[0,h]$*, and* $\Phi(t,0) = P(t,0)e^{Qt}$ *is a Floquet factorization. Then the system* (1) *is not completely controllable if and only if there exists a left eigenvector* $\mathbf{a}^H$ *of* Q *such that* $\mathbf{a}^H W_c[0,h]\mathbf{a} = 0$*. Here,* $W_c[0,h]$ *is given in* (5) *as appropriate.*

3. Harmonic framework for controllability in FDLCP systems. Now we establish controllability criteria through harmonic analysis on the system matrices of FDLCP systems. These results can be viewed as counterparts to the controllability criteria we frequently employ in LTI continuous-time systems. In view of this similarity, the results can be useful in clarifying structural characteristics of FDLCP systems, which have not been well examined by means of the conventional time-domain approaches [27].

3.1. Controllability criteria via Floquet factorizations. Now we state a controllability criterion by use of Floquet factorizations of FDLCP systems.

THEOREM 3.1. *In the* FDLCP *system* (1)*, suppose that* $A(t) \in L_{\mathrm{PCD}}[0,h]$ *and* $B(t) \in L_{\mathrm{CAC}}[0,h]$*. Also assume that* $\Phi(t,0) = P(t,0)e^{Qt}$ *is a Floquet factorization. Then the system* (1) *is completely controllable if and only if for any* $s \in \mathcal{C}$

$$\mathbf{a}^H[\ sI - Q \ \big|\ \cdots, \hat{B}_{-2}, \hat{B}_{-1}, \hat{B}_0, \hat{B}_1, \hat{B}_2, \cdots\ ] \neq 0 \quad \forall \mathbf{a} \neq 0 \in \mathcal{C}^n, \tag{6}$$

where n *is the dimension of the state matrix* $A(t)$ *and* $\{\hat{B}_\mu\}_{\mu=-\infty}^{\infty}$ *is the Fourier coefficient sequence of the matrix* $\hat{B}(t)(= P^{-1}(t,0)B(t))$*.*

Proof. The assumptions on $A(t)$ and $B(t)$ guarantee by Lemma 2.2 that $\hat{B}(t) \in L_{\mathrm{CAC}}[0,h]$. This in particular means that the Fourier series $\sum_\mu \hat{B}_\mu e^{j\mu\omega_h t}$ of $\hat{B}(t)$ uniformly converges to $\hat{B}(t)$ so that it is meaningful to replace $\hat{B}(t)$ with its Fourier series in the following.

(Necessity) Assume that the system (1) is completely controllable but the condition (6) fails. That is, for some $s_0 \in \mathcal{C}$, there exists a vector $\mathbf{a} \neq 0 \in \mathcal{C}^n$ such that

$$\mathbf{a}^H[\ s_0 I - Q \ \big|\ \cdots, \hat{B}_{-2}, \hat{B}_{-1}, \hat{B}_0, \hat{B}_1, \hat{B}_2, \cdots\ ] = 0.$$

It is easy to see that s_0 is an eigenvalue of Q with $\mathbf{a}^H$ being a corresponding left eigenvector. Taking into account that $\mathbf{a}^H \hat{B}_\mu = 0$ for all μ, we observe the deductions

$$\mathbf{a}^H e^{-Qt}\hat{B}(t) = \mathbf{a}^H e^{-s_0 t}\hat{B}(t) = \left\{ \sum_\mu \mathbf{a}^H \hat{B}_\mu e^{j\mu\omega_h t} \right\} e^{-s_0 t} = 0 \quad \forall t \geq 0,$$

which lead to $\mathbf{a}^H W_c[0, kh]\mathbf{a} = 0$ by the definition (5) with k given in Lemma 2.5. Then Proposition 2.6 implies that the system is not completely controllable. This yields a contradiction.

(Sufficiency) Assume that the condition (6) is true but the system is not completely controllable. Proposition 2.7 tells us that there exist an eigenvalue s_0 of Q and a left eigenvector $\mathbf{a}^H \neq 0$ such that $\mathbf{a}^H W_c[0, h]\mathbf{a} = 0$. This implies that $\mathbf{a}^H e^{-Qt}\hat{B}(t) = \mathbf{a}^H e^{-s_0 t}\hat{B}(t) = 0$ for all $t \in [0, h)$, or $\mathbf{a}^H \hat{B}(t) = 0$ for all $t \in [0, h)$. Again, replacing $\hat{B}(t)$ in the previous equation with its Fourier series $\sum_\mu \hat{B}_\mu e^{j\mu\omega_h t}$, it is trivial to derive that $\mathbf{a}^H \hat{B}_\mu = 0$ for all μ. This, together with the fact that $\mathbf{a}^H$ is a left eigenvector of Q, implies that (6) fails at s_0. ∎

Unfortunately, however, there is still an obstacle in implementing Theorem 3.1. That is, one must know the Floquet factorization of the state transition matrix of the FDLCP system. How to determine Floquet factorizations in closed form for general FDLCP systems is still an open problem, although there are numerous numeric procedures [26] for the derivation of the Floquet factorizations; for instance, directly via piecewise constant approximation [10] or indirectly but approximately analytic via Chebyshev polynomial expansion [35]. Therefore, a natural question is: can we develop any controllability criteria without the state transition matrices of FDLCP systems? This question is answered in the next section.

3.2. Controllability criteria via Fourier analysis. Based on Theorem 3.1, let us consider how to express the condition (6) without involving the state transition matrix of the FDLCP system considered.

THEOREM 3.2. *In the* FDLCP *system* (1)*, suppose that* $A(t) \in L_{\mathrm{PCD}}[0, h]$ *and* $B(t) \in L_{\mathrm{CAC}}[0, h]$*. Then the system* (1) *is completely controllable if and only if for each* $s \in \mathcal{C}_f$

$$\underline{a}^H [\ s\underline{I} + \underline{E}(j0) - \underline{A} \mid \underline{B}\ ] \neq 0 \quad \forall \underline{a} \neq 0 \in l_E, \tag{7}$$

where $\underline{A} = \mathcal{T}\{A(t)\}$ *and* $\underline{B} = \mathcal{T}\{B(t)\}$*, respectively. Also,* $\underline{I} = \mathcal{T}\{I\}$*.*

Proof. Assume in the proof discussion that a Floquet simplex for the state transition matrix $\Phi(t, 0)$ of the system (1) is meant whenever Floquet factorizations of $\Phi(t, 0)$ are mentioned. This will cause no loss of generality by Lemma 2.1 in the following arguments.

(Necessity) Assume that the system (1) is completely controllable but the condition (7) fails. Thus, there exist some $s_0 \in \mathcal{C}_f$ and a nonzero vector $\underline{a} \in l_E$ such that $\underline{a}^H [\ s_0\underline{I} + \underline{E}(j0) - \underline{A} \mid \underline{B}\ ] = 0$. By the harmonic Floquet similarity formulas of Lemma 2.2, the equation can be rewritten as

$$\underline{a}^H \underline{P} [\ s_0\underline{I} + \underline{E}(j0) - \underline{Q} \mid \underline{P}^{-1}\underline{B}\ ] \begin{bmatrix} \underline{P}^{-1} & 0 \\ 0 & \underline{I} \end{bmatrix} = 0.$$

Since the last operator matrix in the above equation is invertible on $l_2 \oplus l_2$, we obtain simply

$$\underline{a}^H \underline{P} [\ s_0\underline{I} + \underline{E}(j0) - \underline{Q} \mid \hat{\underline{B}}\ ] = 0. \tag{8}$$

Now, for our purpose, define $0 \neq \underline{a}^H \underline{P} =: [\cdots, a_{-1}^H, a_0^H, a_1^H, \cdots]$ with $a_k \in \mathcal{C}^n$. Apparently, there is at least one (vector) entry $a_k \neq 0$ since $\underline{a} \neq 0$ and $\underline{P}$ is invertible on l_E by Lemma 2.2. Note that $s_0\underline{I} + \underline{E}(j0) - \underline{Q}$ is block-diagonal. Then from (8) it follows that

$$a_k^H[(s_0 + jk\omega_h)I - Q] = 0 \quad \forall k \in \mathcal{Z}. \tag{9}$$

Revoking the Floquet simplex assumption that all eigenvalues of Q are located in $\mathcal{C}_f$ and the comment just after Lemma 2.1, we can conclude that for any integer $k \neq 0$, the matrix $(s_0 + jk\omega_h)I - Q$ must be invertible at any $s_0 \in \mathcal{C}_f$. This, together with (9), gives that $a_k = 0$ for any $k \neq 0$. Note again that not all a_k are zeros. Then the above observation means that $a_0 \neq 0$ but $a_k = 0$ for any $k \neq 0$. Using this fact back to (8), we obtain immediately after simple multiplications that $a_0^H[\ s_0 I - Q \ \big|\ \cdots, \hat{B}_{-1}, \hat{B}_0, \hat{B}_1, \cdots\ ] = 0$, $a_0 \neq 0$, which says from Theorem 3.1 that the system (1) is not completely controllable. However, this is contradictory to the assumption that the system is completely controllable.

(Sufficiency) Assume that the condition (7) holds but the system (1) is not completely controllable. By Theorem 3.1, there exist a scalar $s_0 \in \mathcal{C}$ and a nonzero vector $\alpha_0 \in \mathcal{C}^n$ satisfying

$$\alpha_0^H[\ s_0 I - Q \ \big|\ \cdots, \hat{B}_{-2}, \hat{B}_{-1}, \hat{B}_0, \hat{B}_1, \hat{B}_2, \cdots\ ] = 0.$$

However, by the Floquet simplex assumption this cannot be true unless $s_0 \in \mathcal{C}_f$. Hence, there is no need to consider any s_0 that is located beyond the fundamental strip $\mathcal{C}_f$.

Now we define the infinite-dimensional vector $\underline{\alpha}^H := [\cdots, 0^H, \alpha_0^H, 0^H, \cdots]$. Clearly, $\underline{\alpha}$ is nonzero and $\underline{\alpha} \in l_E$. Based on the matrix expressions of $s\underline{I} + \underline{E}(j0) - \underline{Q}$ and $\hat{\underline{B}}$, it is straightforward to see that $\underline{\alpha}^H[\ s_0\underline{I} + \underline{E}(j0) - \underline{Q} \ \big|\ \hat{\underline{B}}\ ] = 0$, which can be written by the harmonic Floquet similarity formulas of Lemma 2.2 as $\underline{a}^H[s_0\underline{I} + \underline{E}(j0) - \underline{A} \,|\, \underline{B}] = 0$ with $\underline{a} := \underline{P}^{-H}\underline{\alpha}$. By Lemma 2.2, l_E is $\underline{P}^{-H}$-invariant so that $\underline{a} \neq 0$ belongs to l_E. However, this means that the condition (7) does not hold at $s_0 \in \mathcal{C}_f$ with $\underline{a}$ thus defined. This is contradictory to the assumption on (7). □

THEOREM 3.3. *In the* FDLCP *system* (1)*, suppose that* $A(t) \in L_{\mathrm{PCD}}[0, h]$ *and* $B(t) \in L_{\mathrm{CAC}}[0, h]$. *Then the system* (1) *is completely controllable if and only if for each* $\lambda_i \in \Lambda$ *and any eigenvector* $\underline{a} \in \mathcal{N}(\lambda_i\underline{I} + \underline{E}(j0) - \underline{A}) \subset l_E$*, one of the following two conditions is satisfied.*

(a) $\underline{a}^H \underline{e}(\underline{A} - \underline{E}(j0), t)\underline{B} \neq 0$ *for all* $t \in [0, \infty)$*;*

(b) $\underline{a}^H[\underline{B}, (\underline{A} - \underline{E}(j0))\underline{B}, (\underline{A} - \underline{E}(j0))^2\underline{B}, \cdots] \neq 0$.

Proof. The proof can be given by repeating arguments similar in form to those in LTI continuous-time systems, based on Theorem 3.2. Here only a sketched proof for (a) is provided.

(Necessity) Assume that the system is completely controllable but the condition (a) fails. That is, there exist $\lambda_i \in \Lambda$ and an associated eigenvector $\underline{a} \in \mathcal{N}(\lambda_i\underline{I} + \underline{E}(j0) - \underline{A})$ such that $\underline{a}^H \underline{e}(\underline{A} - \underline{E}(j0), t)\underline{B} = 0$ for all $t \in [0, \infty)$. This, in particular, says that $\underline{a}^H\underline{B} = 0$ if we note that $\underline{e}(\underline{A} - \underline{E}(j0), 0) = \underline{I}$. These facts yield that $\underline{a}^H[\lambda_i\underline{I} + \underline{E}(j0) - \underline{A} \,|\, \underline{B}] = 0$, which means that the system is not completely controllable by Theorem 3.2. This brings us a contradiction.

(Sufficiency) Assume that the condition (a) is satisfied but the system is not completely controllable. Theorem 3.2 says that $\lambda_i \in \Lambda$ and an associated eigenvector $\underline{a} \in \mathcal{N}(\lambda_i\underline{I} + \underline{E}(j0) - \underline{A})$ such that $\underline{a}^H[\lambda_i\underline{I} + \underline{E}(j0) - \underline{A} \,|\, \underline{B}] = 0$, from which it follows

that $\underline{a}^H(\underline{A}-\underline{E}(j0))^i\underline{B}=0$ for all $i=0,1,2,\ldots$. By the definition of $\underline{e}(\underline{A}+\underline{E}(j0),t)$, it follows immediately that $\underline{a}^H\underline{e}(\underline{A}-\underline{E}(j0),t)\underline{B}=0$ for all $t\in[0,\infty)$. We are led to a contradiction. □

Combining Theorem 3.3 with Proposition 2.3, it is obvious that under the assumptions of Theorem 3.3, the system (1) is completely controllable if and only if for each $\lambda_i\in\Lambda$ and any eigenvector $\underline{a}\in\mathcal{N}(\lambda_i\underline{I}+\underline{E}(j0)-\underline{Q})\subset l_E$, one of the following two conditions is satisfied: (a) $\underline{a}^H\underline{e}(\underline{Q}-\underline{E}(j0),t)\hat{\underline{B}}\neq 0$ for all $t\in[0,\infty)$; (b) $\underline{a}^H[\hat{\underline{B}},(\underline{Q}-\underline{E}(j0))\hat{\underline{B}},(\underline{Q}-\underline{E}(j0))^2\hat{\underline{B}},\cdots]\neq 0$.

Remark 1. Theorems 3.2 and 3.3 can be viewed as operator-valued counterparts in the FDLCP setting to the controllability criteria in the LTI continuous-time systems; for instance, Theorem 3.2 is the harmonic version of the famous PBH criterion. In this sense we say that FDLCP systems are essentially LTI whenever controllability is considered. Moreover, some intuitive observations indicate readily that installing an h-periodic state feedback in an FDLCP system does not affect controllability between the open- and closed-loop systems. Though this conclusion is already well known in the FDLCP field, it is interesting to notice that such a controllability invariance may follow more easily by the harmonic framework.

The significance of the controllability conditions suggested in Theorem 3.3 is mainly theoretical. This is especially true for condition (b). It is worth mentioning that condition (a) of Theorem 3.3 is useful in tackling what we called the harmonic Lyapunov equation [42]. To avoid distracting the reader's attention, we do not pursue this topic in this paper.

3.3. Corollaries of Theorem 3.2. Similar to what we have encountered in testing the controllability conditions of Theorem 3.1, it is hard to numerically implement the results of Theorem 3.2 if we mention that $s\underline{I}+\underline{E}(j0)-\underline{A}$ is infinite-dimensional and unbounded, even though no state transition matrix is involved in the criterion. As we will see from Corollaries 3.4 and 3.5, the controllability conditions in Theorem 3.2 can reduce to some numerically implementable conditions if $A(t)$, $B(t)$, and/or the Floquet factorizations of the FDLCP system (1) have specific features.

To simplify our statements, we denote the submatrix of $[s\underline{I}+\underline{E}(j0)-\underline{A}\,|\,\underline{B}]$ consisting of the central $(2N+1)$ blockwise rows in $[s\underline{I}+\underline{E}(j0)-\underline{A}\,|\,\underline{B}]$ by $[s\underline{I}+\underline{E}(j0)-\underline{A}\,|\,\underline{B}]_N$.

3.3.1. An equivalent statement of Theorem 3.2.

COROLLARY 3.4. *In the* FDLCP *system* (1)*, suppose that* $A(t)\in L_{\mathrm{PCD}}[0,h]$ *and* $B(t)\in L_{\mathrm{CAC}}[0,h]$. *Then the system* (1) *is completely controllable if and only if for each* $s\in\mathcal{C}_f$ *and each* $N=0,1,2,\ldots$

$$\mathbf{a}^H[\ s\underline{I}+\underline{E}(j0)-\underline{A}\ \mid\ \underline{B}\ ]_N\neq 0\quad\forall\mathbf{a}\neq 0\in\mathcal{C}^{(2N+1)n},$$

where the integer n *is the dimension of the state matrix* $A(t)$.

In Corollary 3.4, the controllability criteria claimed on infinite-dimensional operators are converted into a group of infinitely many finite-dimensional conditions. At first glance, the latter may be equally difficult to test. However, in some cases due to the skew strip matrix expressions of $\underline{E}(j0)-\underline{A}$ and $\underline{B}$ only finitely many finite-dimensional conditions need to be tested.

3.3.2. A special case of Theorem 3.2. If Floquet factorizations of the FDLCP system (1) possesses some specific features, the criteria of Theorem 3.2 may be expressed in finite-dimensional forms.

COROLLARY 3.5. *In the* FDLCP *system* (1), *suppose that* $A(t) \in L_{\mathrm{PCD}}[0,h]$ *and* $B(t) \in L_{\mathrm{CAC}}[0,h]$. *Assume that* $\Phi(t,0) = \tilde{P}(t,0)e^{\tilde{Q}t}$ *is a Floquet factorization satisfying*

$$\tilde{P}(t,0) = \sum_{|k|\le N_p} \tilde{P}_k e^{jk\omega_h t}, \quad \lambda(\tilde{Q}) \subset \bigcup_{|k|\le N_q} \{\mathcal{C}_f + jk\omega_h\}$$

for some integers $N_p \ge 0$ *and* $N_q \ge 0$. *Then, the* FDLCP *system* (1) *is completely controllable if and only if for each* $s \in \mathcal{C}$

$$\mathbf{a}^H[\ s\underline{I} + \underline{E}(j0) - \underline{A} \mid \underline{B}\ ]_{N_p+N_q} \ne 0 \quad \forall \mathbf{a} \ne 0 \in \mathcal{C}^{(2(N_p+N_q)+1)n}, \tag{10}$$

where the integer n *is the dimension of the state matrix* $A(t)$.

Proof. Since $A(t) \in L_{\mathrm{CPD}}[0,h] \subset L_{\mathrm{CAC}}[0,h]$ and $B(t) \in L_{\mathrm{CAC}}[0,h]$, Theorem 3.2 applies and it follows that the system (1) is completely controllable if and only if for any $s \in \mathcal{C}_f$

$$\underline{a}^H[\ s\underline{I} + \underline{E}(j0) - \underline{A} \mid \underline{B}\ ] \ne 0 \quad \forall \underline{a} \ne 0 \in l_E. \tag{11}$$

Again note that (11) holds over $s \in \mathcal{C}_f$ if and only if it is true over $s \in \mathcal{C}$. The necessity is obvious. In the following, we turn to show the sufficiency.

First, let us verify that under the assumptions about the Floquet factorization $\tilde{P}(t,0)e^{\tilde{Q}t}$, any eigenvector associated with an eigenvalue of $\underline{E}(j0) - \underline{A}$ has at most $2(N_p + N_q) + 1$ consecutive nonzero entries while all other entries are zeros.

To this end, we write $\Phi(t,0) = P(t,0)e^{Qt}$ to be the Floquet simplex derived via the technique suggested in the proof of Lemma 2.1 so that (3) holds. With this Floquet simplex, it follows that for each eigenvalue λ of $\underline{Q} - \underline{E}(j0)$, any associated eigenvector must be of the form $\underline{x} = [\cdots, x^T, \cdots]^T \in l_E$, with $x \ne 0 \in \mathcal{C}^n$. By Lemma 2.2, one can assert that λ is also an eigenvalue of $\underline{A} - \underline{E}(j0)$ with an associated eigenvector $\underline{P}\,\underline{x} \in l_E$. Since $P(t,0)$ contains at most $N_p + N_q$ harmonics, it follows that only entries in the $2(N_p+N_q)+1$ skew lines along the main diagonal line of $\underline{P}$ are possibly nonzero. This, together with the specific structure of $\underline{x}$, leads to the desired assertion.

Second, we mention that (11) fails only if $\underline{a}$ is an eigenvector of $\underline{A} - \underline{E}(j0)$. In other words, we can examine (11) by testing it over all eigenvectors of $\underline{A} - \underline{E}(j0)$.

Finally, taking into account the structure features of eigenvectors of $\underline{A} - \underline{E}(j0)$ under the assumptions about the Floquet factorization, we see that for each eigenvector of $\underline{A} - \underline{E}(j0)$, only the $2(N_p + N_q) + 1$ consecutive blockwise rows in $\underline{A} - \underline{E}(j0)$ corresponding to the $2(N_p + N_q) + 1$ consecutive nonzero entries in the eigenvector are significant with regard to linear independence of vectors. Based on the above arguments, if (10) is satisfied, then we can conclude that for each eigenvector of the operator $\underline{A} - \underline{E}(j0)$, (11) is true. ∎

It can be understood from the proof of Corollary 3.5 that the exact formula for Floquet factorizations involved in Corollary 3.5 is not necessary, except for the integers N_p and N_q. Indeed, N_q always exists. Therefore, if $\tilde{P}(t,0)$ contains only finitely many harmonics, it is always possible to fix an integer N satisfying $N \ge N_p + N_q$. If for each $s \in \mathcal{C}$, $\mathbf{a}^H[s\underline{I} + \underline{E}(j0) - \underline{A} \mid \underline{B}]_N \ne 0$ for all $\mathbf{a} \ne 0 \in \mathcal{C}^{(2N+1)n}$, then (10) holds. In this sense, we must point out that the Floquet factorization assumption is not a big obstacle in implementing the controllability criterion of Corollary 3.5.

3.3.3. Noncontrollability criterion derived from Theorem 3.2. Corollaries 3.4 and 3.5 are claimed in the sense of controllability of FDLCP systems. In contrast, the corollary given below provides sufficient conditions for noncontrollability.

COROLLARY 3.6. *In the* FDLCP *system* (1)*, suppose that* $A(t) \in L_{\mathrm{PCD}}[0,h]$ *and* $B(t) \in L_{\mathrm{CAC}}[0,h]$*. If for some* $s \in \mathcal{C}_f$ *there exists a vector sequence* $\{\mathbf{a}_\mu\}_{\mu=-\infty}^{\infty}$ *with* $\mathbf{a}_\mu \in \mathcal{C}^n$ *and* $\mathbf{a}_\mu \neq 0$ *for some* μ *such that*

$$\mathbf{a}_\mu^H[\ (s+j\mu\omega_h)\mathbf{I}-\mathbf{A}\ \big|\ \mathbf{B}\ ]=0 \quad \forall \mu \in \mathcal{Z}, \tag{12}$$

then the system (1) *is not completely controllable. In the above,* $\mathbf{I}$ *and* $\mathbf{A}$ *are defined as follows:*

$$\mathbf{I}=[\cdots,0,I,0,\cdots], \quad \mathbf{A}=[\cdots,A_{-1},A_0,A_1,\cdots],$$

where $\{A_\mu\}_{\mu=-\infty}^{\infty}$ *is the Fourier coefficient sequence of* $A(t)$*.* $\mathbf{B}$ *is defined similarly to* $\mathbf{A}$ *but in terms of the Fourier coefficients of* $B(t)$*.*

Proof. By the assumption on $\{\mathbf{a}_\mu\}_{\mu=-\infty}^{\infty}$, there exists at least one vector, say $\mathbf{a}_\mu$, satisfying $\mathbf{a}_\mu \neq 0$. Let us define an infinite-dimensional vector $\underline{a}^H := [\cdots,\mathbf{a}_\mu^H,\cdots]$ with $\mathbf{a}_\mu$ being situated at the μth position in $\underline{a}$. Obviously, $\underline{a}$ is nonzero and belongs to l_E. Bearing this in mind, arranging the conditions (12) according to $\underline{a}$, it follows readily that $\underline{a}^H[s\underline{I}-\underline{A}\,|\,\underline{B}]=0$. This implies by Theorem 3.2 that the system is not completely controllable. □

In less rigorous words, Corollary 3.6 says that some linear dependence among the Fourier coefficients of $A(t)$ and $B(t)$ of the FDLCP system (1) can be a structural reason why the system loses controllability, as is already known in LTI cases. Needless to say, other coefficients' linear dependence may also result in non-controllability. The linear dependence among the Fourier coefficients of $A(t)$ and $B(t)$ described in Corollary 3.6 is just one of the simplest cases. In other words, linear independence of the n infinite-dimensional row vectors in $[s\mathbf{I}-\mathbf{A}\,|\,\mathbf{B}]$ is a necessary condition for controllability of the FDLCP system (1).

4. Controllability decomposition of FDLCP systems. From the results of the previous sections, one can see that FDLCP systems are essentially LTI when controllability is concerned. It is well known [11], [21], [39] that an LTI state-space model can always be decomposed according to controllable/uncontrollable modes through a state coordinate transform. It is natural to consider whether or not such controllability decompositions exist in FDLCP systems. Indeed, as a first answer to such a decomposition question, the canonical structure theorems of [21] ensure that controllability canonical decompositions for general dynamic systems are possible and the state coordinate transform can be determined in a pointwise fashion in time (i.e., for each fixed instant t of time, the dynamic system concerned is viewed as an LTI one and thus decomposed accordingly). For FDLCP systems, it is claimed that the state coordinate transform needed can be continuous in time by Theorem 6 of [21] (without proof).

However, in many FDLCP systems merely keeping continuity during the state coordinate transform is not enough to validate harmonic analysis due to convergence issues. There is another decomposition method given in [37] for general linear continuous-time time-varying systems by means of the so-called Doležal theorem with a continuous and differentiable state coordinate transform, under the assumption that the rank of the controllability Gramian is independent of t. This rank assumption is true in FDLCP systems and brings us with periodic transformed systems by applying the Doležal theorem to suitable sequences of Gramians as shown by [4]. In general, the Doležal theorem results in a transformed FDLCP system that has a periodic state matrix, which may be hard to handle when such a controllability decomposition is utilized.

In this section, we show that controllability decompositions for FDLCP systems are also available via state coordinate transforms that possess fairly strong analytic properties, which are highly expected in system analysis and synthesis if the harmonic approach is adopted, while a constant state matrix is attained in the decomposed FDLCP system. How to exploit these analytic features of the controllability canonical decompositions for poles/zeros analysis, positive realness of the so-called harmonic frequency response operators [41] and the harmonic state operators [43], and H_2/H_∞ performance synthesis in FDLCP systems is left for subsequent papers.

4.1. Decomposition algorithm. First, we write $\Phi(t,0) = P(t,0)e^{Qt}$ to be a Floquet factorization for the FDLCP system (1), and then we observe that the FDLCP system given by

$$\Sigma : \dot{z} = Qz + \hat{B}(t)u \tag{13}$$

with $\hat{B}(t) = P^{-1}(t,0)B(t)$ is equivalent to the system (1) in the sense of controllability, after introducing the Floquet state coordinate transform $z = P(t,0)x$ to the system (1).

Second, we form the following approximate FDLCP model for the system (13) by piecewise constant treatment upon all the entries of $\hat{B}(t)$. That is, we have

$$\Sigma_\kappa : \dot{\zeta} = Q\zeta + \hat{B}(\kappa,t)u \tag{14}$$

with

$$\hat{B}(\kappa,t) := \hat{B}(t_{k-1}) \quad \forall t \in [t_{k-1}, t_k),\ k = 1,2,\ldots,\kappa,$$

with $0 = t_0 < t_1 < t_2 < \cdots < t_\kappa = h$. That is, the union of all the subintervals $[t_0,t_1),\ [t_1,t_2),\cdots,[t_{\kappa-1},t_\kappa)$ forms $[0,h)$. Obviously, if $\hat{B}(t)$ is continuous in t, the piecewise constant treatment on $\hat{B}(t)$ is well defined for any subinterval sequences since the set of piecewise constant functions is dense in the set of piecewise continuous functions in the norm sense of $\sup_{t\in[0,h)} ||\cdot||$. For simplicity, the subintervals sequence $\{[t_0,t_1),[t_1,t_2),\cdots,[t_{\kappa-1},t_\kappa)\}$ is called a segmentation of $[0,h)$ and denoted by $\mathcal{S}(\kappa)$. With a bit of abuse of notation we use $\kappa \to \infty$ to mean $\max_{k=1,2,\ldots,\kappa} |t_{k-1} - t_k| \to 0$ in the following. Hence, we obtain by the definition of $\hat{B}(\kappa,t)$ that $\lim_{\kappa\to\infty} \hat{B}(\kappa,t) = \hat{B}(t)$.

Third, let us define the following matrix:

$$\hat{B}_\kappa = [\, e^{-Qt_0}\hat{B}(t_0), e^{-Qt_1}\hat{B}(t_1), \cdots, e^{-Qt_{\kappa-1}}\hat{B}(t_{\kappa-1})\,].$$

Now let $\gamma(\kappa) := \operatorname{rank}[\hat{B}_\kappa, Q\hat{B}_\kappa, \cdots, Q^{n-1}\hat{B}_\kappa]$ and choose $\gamma(\kappa)$ column vectors that are linearly independent from $[\hat{B}_\kappa, Q\hat{B}_\kappa, \cdots, Q^{n-1}\hat{B}_\kappa]$. Based on these $\gamma(\kappa)$ column vectors, one can always construct $\gamma(\kappa)$ orthonormal vectors $\epsilon_\kappa^{(1)}, \epsilon_\kappa^{(2)}, \cdots, \epsilon_\kappa^{(\gamma(\kappa))}$ by Lemma 1 of [8, p. 25]. Then add other $n - \gamma(\kappa)$ more orthonormal vectors $\epsilon_\kappa^{(\gamma(\kappa)+1)}, \cdots, \epsilon_\kappa^{(n)}$ such that $\epsilon_\kappa^{(1)}, \epsilon_\kappa^{(2)}, \cdots, \epsilon_\kappa^{(n)}$ form an orthonormal base for the Euclidean space $\mathcal{C}^n$. For our purpose, let us write

$$T_\kappa := [\epsilon_\kappa^{(1)}, \cdots, \epsilon_\kappa^{(\gamma(\kappa))} |\, \epsilon_\kappa^{(\gamma(\kappa)+1)}, \cdots, \epsilon_\kappa^{(n)}] =: [T_{1\kappa} |\, T_{2\kappa}]. \tag{15}$$

It should be pointed out that $T_\kappa^H = T_\kappa^{-1}$ by the definition of T_κ.

Finally, noncontrollability of the system (1) is connected to that of its approximate models defined in (14). This noncontrollability connection is significant in determining

a controllability canonical decomposition which we will deal with in the next section. A detailed proof of the results in Proposition 4.1 is given in Appendix B.

PROPOSITION 4.1. *In the* FDLCP *system* (1)*, assume that the entries of* $A(t)$ *are piecewise continuous in* t*, while the entries of* $B(t)$ *are continuous in* t*. Then the approximate* FDLCP *system* (14) *is well defined for any segmentation* $\mathcal{S}(\kappa)$*. Moreover, the system* (1) *is not completely controllable if and only if for any* $\mathcal{S}(\kappa)$ *the approximate system* (14) *is not completely controllable.*

4.2. Controllability canonical decomposition theorem. Based on the algorithm for constructing T_κ and the Floquet factorization of the state transition matrix, we can claim the following controllability canonical decomposition theorem for the FDLCP system (1), whose proof is a bit lengthy and thus found in Appendix C.

THEOREM 4.2. *In the* FDLCP *system* (1)*, suppose that the entries of* $A(t)$ *are piecewise continuous in* t*, while the entries of* $B(t)$ *are continuous in* t*. Assume that the pair* $(A(\cdot), B(\cdot))$ *is not completely controllable. If the segmentation* $\mathcal{S}(\kappa)$ *on* $[0, h)$ *is fine enough in the sense of* $\kappa \to \infty$*, then the state coordinate transform* $\tilde{x} = T_\kappa P(t, 0)x$ *transforms the* FDLCP *system* (1) *into a controllability canonical form given by*

$$\dot{\tilde{x}} = \begin{bmatrix} Q_c & Q_{12} \\ 0 & Q_{\bar{c}} \end{bmatrix} \tilde{x} + \begin{bmatrix} B_c(t) \\ 0 \end{bmatrix} u, \tag{16}$$

where Q_c*,* Q_{12}*, and* $Q_{\bar{c}}$ *are constant matrices of appropriate dimensions, while* $B_c(t)$ *is* h*-periodic and continuous in* t *and the pair* $(Q_c, B_c(\cdot))$ *is completely controllable. Moreover,* T_κ *defined in* (15) *can be chosen independently of the segmentation* $\mathcal{S}(\kappa)$ *as long as* κ *is large enough. Hence,* Q_c*,* Q_{12}*,* $Q_{\bar{c}}$*, and* $B_c(t)$ *are also independent of the segmentation* $\mathcal{S}(\kappa)$*.*

When the assumptions on $A(t)$ and $B(t)$ are strengthened mildly, some excellent analytic properties about $P(t, 0)$, $P^{-1}(t, 0)$, and $\hat{B}(t)$ follow from the results in [41]. This, together with the fact that T_κ is a constant matrix, implies the assertions of Corollary 4.3.

COROLLARY 4.3. *In the* FDLCP *system* (1)*, suppose that* $A(t) \in L_{\mathrm{PCD}}[0, h]$ *and* $B(t) \in L_{\mathrm{CAC}}[0, h]$*, and that the pair* $(A(\cdot), B(\cdot))$ *is not completely controllable. If the segmentation* $\mathcal{S}(\kappa)$ *on* $[0, h)$ *is fine enough, i.e.,* $\kappa \to \infty$*, then the state coordinate transform* $x = T_\kappa P(t, 0)\tilde{x}$ *transforms the* FDLCP *system* (1) *into a controllability canonical form given by* (16)*. Moreover, it holds that*

(a) $B_c(t) \in L_{\mathrm{CAC}}[0, h]$*;*

(b) $T_\kappa P(t, 0)$ *is absolutely continuous with respect to* t *and invertible uniformly over* $t \in [0, h)$*;*

(c) *The first-order derivatives of* $T_\kappa P(t, 0)$ *and* $P^{-1}(t, 0)T_\kappa^{-1}$ *are piecewise continuous in* t*.*

5. Controllability of numeric examples. To illustrate the main results obtained in the paper, we investigate controllability of the following 5-periodic differential state-space equation [10] as the first example, in which different constant input matrices $B(t)$ (i.e., b_1, b_2, b_3, and b_4 are constants) are considered in two cases:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{13\pi}{30} & \frac{2\pi}{5}\cos^2(\frac{2\pi t}{5}) & \frac{\pi}{5}\sin(\frac{4\pi t}{5}) \\ \frac{13\pi}{30} & 0 & -\frac{\pi}{5}\sin(\frac{4\pi t}{5}) & \frac{2\pi}{5}\cos^2(\frac{2\pi t}{5}) \\ 0 & 0 & 0 & -\frac{13\pi}{30} \\ 0 & 0 & \frac{13\pi}{30} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} u.$$

For the state matrix $A(t)$ we have the following Floquet factorization:

$$P(t,0) = \begin{bmatrix} \cos(\frac{2\pi t}{5}) & -\sin(\frac{2\pi t}{5}) & \sin(\frac{2\pi t}{5}) & 0 \\ \sin(\frac{2\pi t}{5}) & \cos(\frac{2\pi t}{5}) & 0 & \sin(\frac{2\pi t}{5}) \\ 0 & 0 & \cos(\frac{2\pi t}{5}) & -\sin(\frac{2\pi t}{5}) \\ 0 & 0 & \sin(\frac{2\pi t}{5}) & \cos(\frac{2\pi t}{5}) \end{bmatrix},$$

$$Q = \begin{bmatrix} 0 & -\frac{\pi}{30} & 0 & 0 \\ \frac{\pi}{30} & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{13\pi}{30} \\ 0 & 0 & \frac{13\pi}{30} & 0 \end{bmatrix}.$$

Case A. Let us investigate controllability when $B(t) = [0,0,0,1]^T$. By means of the sufficient rank criterion suggested in [33], the so-called controllability matrix is

$$\begin{bmatrix} 0 & -\frac{\pi}{5}\sin(\frac{4\pi t}{5}) & \begin{pmatrix} -\frac{26\pi^2}{75}\cos^2(\frac{2\pi t}{5}) \\ -\frac{4\pi^2}{25}\cos(\frac{4\pi t}{5}) \end{pmatrix} & \frac{2019\pi^3}{4500}\sin(\frac{4\pi t}{5}) \\ 0 & -\frac{2\pi}{5}\cos^2(\frac{2\pi t}{5}) & \frac{\pi^2}{3}\sin(\frac{4\pi t}{5}) & \begin{pmatrix} \frac{338\pi^2}{1500}\cos^2(\frac{2\pi t}{5}) \\ +\frac{252\pi^3}{750}\cos(\frac{4\pi t}{5}) \end{pmatrix} \\ 0 & \frac{13\pi}{30} & 0 & -\frac{2197\pi^3}{27000} \\ 1 & 0 & -\frac{169\pi^2}{900} & 0 \end{bmatrix},$$

whose rank is 4 for $t = 1.25$, for instance, after tedious computations. Then one can assert by Theorem 3 of [33] that the given FDLCP system is completely controllable.

Since the Floquet factorization is available, we can also test controllability for Case A through Theorem 3.1. Note that $\hat{B}(t) = P^{-1}(t,0)B(t)$ possesses only nonzero harmonic waves within the third order. Then, dropping all zero harmonic blocks we obtain

$$[sI - Q | \hat{B}_{-3}, \hat{B}_{-2}, \hat{B}_{-1}, \hat{B}_0, \hat{B}_1, \hat{B}_2, \hat{B}_3]$$
$$= \left[\begin{array}{cccc|ccccccc} s & \frac{\pi}{30} & 0 & 0 & \frac{1}{4} & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & \frac{1}{4} \\ -\frac{\pi}{30} & s & 0 & 0 & -\frac{j}{4} & 0 & \frac{j}{4} & 0 & -\frac{j}{4} & 0 & \frac{j}{4} \\ 0 & 0 & s & \frac{13\pi}{30} & 0 & 0 & \frac{j}{2} & 0 & -\frac{j}{2} & 0 & 0 \\ 0 & 0 & -\frac{13\pi}{30} & s & 0 & 0 & \frac{1}{2} & 0 & -\frac{1}{2} & 0 & 0 \end{array}\right],$$

which has rank 4 for all $s \in \mathcal{C}$. Hence, Theorem 3.1 says also that the FDLCP system considered is completely controllable. It should be pointed out that the criterion in [33] is merely sufficient but the controllability conditions in Theorem 3.1 are necessary and sufficient.

Case B. Let us investigate controllability when $B(t) = [0,1,0,0]^T$. By means of the rank criterion of [33], the corresponding controllability matrix is

$$\begin{bmatrix} 0 & \frac{13\pi}{30} & 0 & -(\frac{13\pi}{30})^3 \\ 1 & 0 & -(\frac{13\pi}{30})^2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

whose rank is $2 < 4$. Then Theorem 3 of [33] cannot tell us whether the given FDLCP system is completely controllable. Fortunately, however, we test controllability through Theorem 3.1. The corresponding controllability matrix is

$$[sI - Q|\hat{B}_{-1}, \hat{B}_0, \hat{B}_1] = \left[\begin{array}{cccc|ccc} s & \frac{\pi}{30} & 0 & 0 & \frac{j}{2} & 0 & -\frac{j}{2} \\ -\frac{\pi}{30} & s & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & s & \frac{13\pi}{30} & 0 & 0 & 0 \\ 0 & 0 & -\frac{13\pi}{30} & s & 0 & 0 & 0 \end{array}\right].$$

It follows readily that $\operatorname{rank}[sI - Q|\hat{B}_{-1}, \hat{B}_0, \hat{B}_1] < 4$ for $s = \pm\frac{13\pi}{30}j$. This implies by Theorem 3.1 that the system is not completely controllable.

We can also draw the noncontrollability conclusion by Corollary 3.6 for Case B, which does not depend on the Floquet factorization. We notice that the Fourier coefficients of the state matrix $A(t)$ possess only harmonic waves up to the second order and the input matrix $B(t)$ is constant. Then, we can construct the matrix $[(s + j\mu\omega_h)\mathbf{I} - \mathbf{A}|\mathbf{B}]$ as defined in Corollary 3.6 as follows, where all zero harmonic blocks are dropped for brevity:

$$\left[\begin{array}{cccccccccc|c} 0 & 0 & -\frac{\pi}{10} & -\frac{\pi j}{10} & s + j\mu\omega_h & \frac{13\pi}{30} & -\frac{\pi}{5} & 0 & 0 & 0 & -\frac{\pi}{10} & \frac{\pi j}{10} & b_1 \\ 0 & 0 & -\frac{\pi j}{10} & -\frac{\pi}{10} & -\frac{13\pi}{30} & s + j\mu\omega_h & 0 & -\frac{\pi}{5} & 0 & 0 & \frac{\pi j}{10} & -\frac{\pi}{10} & b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & s + j\mu\omega_h & \frac{13\pi}{30} & 0 & 0 & 0 & 0 & b_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{13\pi}{30} & s + j\mu\omega_h & 0 & 0 & 0 & 0 & b_4 \end{array}\right],$$

from which it is straightforward to see that $\operatorname{rank}[(s + j\mu\omega_h)\mathbf{I} - \mathbf{A}|\mathbf{B}] < 4$ for some $s \in \mathcal{C}$ and $\mu \in \mathcal{Z}$ whenever $b_3 = b_4 = 0$. Hence, it is possible to construct a number sequence such that Corollary 3.6 is satisfied. In other words, in this case the system is not completely controllable.

We stress that regarding the two cases of the first numeric example, one can apply the necessary and sufficient controllability conditions in Corollary 3.5, in which only the Fourier coefficient matrices of $A(t)$ and $B(t)$ are involved. However, since the state matrix $A(t)$, which is 4×4 in dimension, possesses harmonic waves up to the second order and thus the conditions of Corollary 3.5 must be examined by working on some matrices of big size, it seems inappropriate to give the detailed arguments in this paper due to the limited space.

To show how to apply the results of Corollary 3.4, we consider controllability of the π-periodic lossy Mathieu differential equation [29], [38] given by

$$\ddot{y}(t) + 2\zeta\dot{y}(t) + (1 - 2\beta\cos(2t))y(t) = u(t),$$

where ζ is the damping ratio. The lossy Mathieu differential equation can be equivalently expressed by the following FDLCP state-space modeling:

$$A(t) = \left[\begin{array}{cc} 0 & 1 \\ -1 + 2\beta\cos(2t) & -2\zeta \end{array}\right], \quad B(t) = \left[\begin{array}{c} 0 \\ 1 \end{array}\right].$$

We notice that the Fourier coefficients of the state matrix $A(t)$ do not possess any harmonic waves higher than the first order and that the input matrix $B(t)$ is constant. Then, we can construct $[s\underline{I} + \underline{E}(j0) - \underline{A}]_{N=0}$ as follows, where all zero harmonic blocks are dropped:

$$\left[\begin{array}{cccccc|c} 0 & 0 & s & -1 & 0 & 0 & 0 \\ -\beta & 0 & 1 & s - 2\zeta & -\beta & 0 & 1 \end{array}\right].$$

Forming a 2×2 matrix by the 4th and 7th columns in $[s\underline{I} + \underline{E}(j0) - \underline{A}]_{N=0}$, we can see that its determinant is -1 no matter how the variables s, β, and ζ are taken. It follows that $\operatorname{rank}[s\underline{I} + \underline{E}(j0) - \underline{A}]_{N=0} = 2$ for all $s \in \mathcal{C}$.

Similarly, we have $[s\underline{I} + \underline{E}(j0) - \underline{A}]_{N=1}$ given by

$$\left[\begin{array}{cccccccccc|ccc}
0 & 0 & s+j\omega_h & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\beta & 0 & 1 & s+j\omega_h-2\zeta & -\beta & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & s & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -\beta & 0 & 1 & s-2\zeta & -\beta & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & s-j\omega_h & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -\beta & 0 & 1 & s-j\omega_h-2\zeta & -\beta & 0 & 0 & 0 & 1
\end{array}\right].$$

Forming a 6×6 matrix by the 4th, 6th, 8th, 11th, 12th, and 13th columns in $[s\underline{I} + \underline{E}(j0) - \underline{A}]_{N=1}$, we obtain a matrix whose determinant is $(-1)^k$ ($k > 0$ is an integer) no matter what s, β, and ζ are taken. It follows that $\operatorname{rank}[s\underline{I} + \underline{E}(j0) - \underline{A}]_{N=1} = 6$ for all $s \in \mathcal{C}$.

We can repeat the above arguments for all $N = 2, 3, \ldots$. In particular, it is always possible to fix a corresponding square matrix whose determinant is 1 or -1. Hence, one can conclude that the conditions of Corollary 3.4 hold for each $N = 0, 1, 2, \ldots$ over $s \in \mathcal{C}$, β, and ζ. It follows that the lossy Mathieu equation is completely controllable.

One can also investigate controllability of the lossy Mathieu equation by means of the rank criterion suggested in [33], which also leads to the lossy Mathieu equation being completely controllable. In both numeric examples, the system matrices $A(t)$ and $B(t)$ contain only finitely many harmonic waves so that the system matrices are analytic. Note 9.4 of [31] indicates that under the analytic assumption the controllability criterion of [33] can also be necessary.

We must add that the numeric examples are relatively simple, for example, compared with those given in [26]. Performance of the theoretical results in practical and complicated systems still needs to be evaluated but is left as one of our future works.

6. Conclusion. In this paper, controllability of a large class of FDLCP systems is considered from a harmonic analysis point of view for the first time, to the best knowledge of the author. Major contributions of this study include an operator-valued PBH controllability criterion claimed in the FDLCP setting and its derivations; that is, Theorems 3.1–3.3. Numeric implementable algorithms for these controllability criteria are also worked out in Corollaries 3.4, 3.5, and 3.6, which are directly applicable via Floquet factorizations of the FDLCP systems and/or Fourier coefficients of the systems matrices. Compared with controllability criteria existing in the literature, the methodology adopted here is highly intuitive and heuristic and keeps some mathematical convenience of their LTI counterparts in the FDLCP setting. Another achievement of this study is a novel controllability canonical decomposition algorithm summarized in Theorem 4.2, which provides us with decomposed FDLCP modeling with a constant state matrix through state coordinate transforms of strong analytic properties; see Corollary 4.3. The results are significant since the paper has succeeded in establishing a harmonic framework in the FDLCP field to exploit the well-developed LTI analysis and synthesis tools relevant to controllability/observability characteristics.

Needless to say, one can simply assert similar results about observability of FDLCP systems through the duality principle [19, pp. 79–103]. However, to avoid any redundancy in the paper, the author paid no attention to observability directly or indirectly. We hope this treatment is allowable for the sake of simplicity.

Appendix A. Proof of lemmas.

Proof of Lemma 2.1. The existence assertion of Floquet simplices can follow from the main branch formula and relevant properties about matrix logarithm theory [40, pp. 55–58]. Here we provide an alternative proof for (a) that is constructive and given by modifying the proof of Theorem 8.1.3 of [23]. This alternative proof also paves a way for showing assertion (b).

To see assertion (a), let $\Xi(\cdot)$ be a fundamental matrix of $A(\cdot)$. That is, $\dot{\Xi}(t) = A(t)\Xi(t)$, a.e. $t \geq 0$. By the properties of the fundamental matrix, $\det(\Xi(t)) \neq 0$ for all $t \geq 0$. Now let $\Theta(t) = \Xi(t+h)$. Then $\dot{\Theta}(t) = A(t+h)\Xi(t+h) = A(t)\Theta(t)$, a.e. for $t \geq 0$, which means that $\Theta(\cdot)$ is also a fundamental matrix of $A(\cdot)$. Also note that $\det(\Theta(t)) \neq 0$ for all t. These facts imply by Theorem 6.7.2 of [23] that $\Xi(t+h) = \Xi(t)M$ for all $t \geq 0$ with some nonsingular matrix M. This implies by Lemma 8.1.1 of [23] that there exists a complex matrix $\tilde{Q}$ such that $M = e^{\tilde{Q}h}$. Then the arguments in the later part of the proof of Theorem 8.1.3 of [23] say that $\tilde{P}(t,0)e^{\tilde{Q}t}$, with $\tilde{P}(t,0) =: \Xi(t)e^{-\tilde{Q}t}$ being h-periodic, is a Floquet factorization for $\Phi(t,0)$.

Next, we express $\tilde{Q}$ through its Jordan canonical form $\tilde{Q} = S\tilde{J}S^{-1}$ with S being a nonsingular matrix, and $\tilde{J} = \operatorname{diag}[\tilde{J}_1, \tilde{J}_2, \cdots, \tilde{J}_\alpha]$, where $\tilde{J}_i$ $(i = 1, 2, \ldots, \alpha)$ is an $n_i \times n_i$ Jordan block defined in the obvious fashion corresponding to an eigenvalue $\tilde{\lambda}_i$ of $\tilde{Q}$. It is straightforward to see that

$$e^{\tilde{Q}t} = Se^{\tilde{J}t}S^{-1}, \tag{17}$$

where $e^{\tilde{J}t} = \operatorname{diag}[e^{\tilde{J}_1 t}, e^{\tilde{J}_2 t}, \cdots, e^{\tilde{J}_\alpha t}]$ and

$$e^{\tilde{J}_i t} = e^{\tilde{\lambda}_i t} \begin{bmatrix} 1 & t & \frac{t}{2} & \cdots & \frac{t^{n_i-1}}{(n_i-1)!} \\ 0 & 1 & t & \cdots & \frac{t^{n_i-2}}{(n_i-2)!} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{18}$$

Write $\tilde{\lambda}_i = \lambda_i + jk_i\omega_h$, $k_i \in \mathcal{Z}, i = 1, 2, \ldots, \alpha$, with $\lambda_i \in \mathcal{C}_f$, which are always possible. In particular, (18) means that $e^{\tilde{J}_i h} = e^{J_i h}$ for each i, where J_i is a Jordan block in terms of λ_i. From (17), we have that $e^{\tilde{Q}h} = Se^{Jh}S^{-1} = e^{SJS^{-1}h} = e^{Qh}$, in which $Q := SJS^{-1}$ and $J := \operatorname{diag}[J_1, J_2, \cdots, J_\alpha]$. Clearly, Q has eigenvalues λ_1, $\lambda_2, \cdots, \lambda_\alpha$ located in $\mathcal{C}_f$.

Now, let us repeat the arguments in the later part of the proof of Theorem 8.1.3 of [23] but in terms of $M = e^{Qh}$. Then we can draw the conclusion that there exists an h-periodic matrix $P(t,0) =: \Xi(t)e^{-Qt}$ such that $P(t,0)e^{Qt}$ is a Floquet factorization of $\Phi(t,0)$ with $\lambda(Q) \subset \mathcal{C}_f$; in other words, $\Phi(t,0) = P(t,0)e^{Qt}$ is a Floquet simplex.

To see assertion (b) under the assumption of (2), let us observe that

$$P(t,0) = \Xi(t)e^{-Qt} = \Xi(t)e^{-\tilde{Q}t}e^{\tilde{Q}t}e^{-Qt} = \tilde{P}(t,0)e^{\tilde{Q}t}e^{-Qt}. \tag{19}$$

On the other hand, (17) yields that

$$\begin{aligned} e^{\tilde{Q}t} &= Se^{\tilde{J}t}S^{-1} = S\operatorname{diag}[e^{\tilde{J}_1 t}, e^{\tilde{J}_2 t}, \cdots, e^{\tilde{J}_\alpha t}]S^{-1} \\ &= S\ \operatorname{diag}[e^{jk_1\omega_h t}I_{n_1}, e^{jk_2\omega_h t}I_{n_2}, \cdots, e^{jk_\alpha\omega_h t}I_{n_\alpha}]\operatorname{diag}[e^{J_1 t}, e^{J_2 t}, \cdots, e^{J_\alpha t}]S^{-1} \\ &= S\ \operatorname{diag}[e^{jk_1\omega_h t}I_{n_1}, e^{jk_2\omega_h t}I_{n_2}, \cdots, e^{jk_\alpha\omega_h t}I_{n_\alpha}]\ S^{-1}S\operatorname{diag}[e^{J_1 t}, e^{J_2 t}, \cdots, e^{J_\alpha t}]S^{-1} \\ &= S\ \operatorname{diag}[e^{jk_1\omega_h t}I_{n_1}, e^{jk_2\omega_h t}I_{n_2}, \cdots, e^{jk_\alpha\omega_h t}I_{n_\alpha}]\ S^{-1}S\,e^{Jt}S^{-1} \\ &= S\ \operatorname{diag}[e^{jk_1\omega_h t}I_{n_1}, e^{jk_2\omega_h t}I_{n_2}, \cdots, e^{jk_\alpha\omega_h t}I_{n_\alpha}]\ S^{-1}e^{Qt}, \end{aligned}$$

where I_{n_i} denotes the $n_i \times n_i$ identity matrix.

Using the last equation in the above back to (19), we have

$$\begin{aligned} P(t,0) &= \tilde{P}(t,0) S \operatorname{diag}[e^{jk_1\omega_h t} I_{n_1}, e^{jk_2\omega_h t} I_{n_2}, \cdots, e^{jk_\alpha\omega_h t} I_{n_\alpha}]\, S^{-1} \\ &= \tilde{P}(t,0) S \sum_{i=1}^{\alpha} \operatorname{diag}[0, \cdots, 0, e^{jk_i\omega_h t} I_{n_i}, 0, \cdots, 0]\, S^{-1} \\ &= \tilde{P}(t,0) \sum_{i=1}^{\alpha} e^{jk_i\omega_h t} \{S \operatorname{diag}[0, \cdots, 0, I_{n_i}, 0, \cdots, 0]\, S^{-1}\} \\ &=: \tilde{P}(t,0) \sum_{i=1}^{\alpha} e^{jk_i\omega_h t} S_i, \end{aligned}$$

where the definition of S_i is obvious. Clearly S_i is a constant matrix for each i. The last equation, together with the assumption on $\tilde{P}(t,0)$, yields that

$$P(t,0) = \sum_{|k| \le N_p} \tilde{P}_k e^{jk\omega_h t} \sum_{i=1}^{\alpha} e^{jk_i\omega_h t} S_i,$$

which implies the desired assertion if we denote $N_q = \max\{k_1, k_2, \cdots, k_\alpha\}$. □

Proof of Lemma 2.5. Without loss of generality, we prove only the case of $t_0 = 0$. The proof is accomplished by modifying the arguments in the proof for Lemma 3 of [2]. First, we observe by Lemma 1 of [2] that

$$W_c[t_0, t_0 + ih] = W_c[t_0, t_0 + (i-1)h] + e^{-Q(i-1)h} W_c[t_0, t_0 + h] e^{-Q^T(i-1)h},$$

where $i = 1, 2, \ldots, k$. Using this, it is straightforward to see that

$$W_c[t_0, t_0 + kh] = \sum_{i=1}^{k} e^{-Q(i-1)h} W_c[t_0, t_0 + h] e^{-Q^T(i-1)h}. \tag{20}$$

Clearly, $W_c[t_0, t_0 + kh] \ge 0$ implies that there exists at least one nonzero vector $\alpha \in \mathcal{C}^n$ satisfying $\alpha^H W_c[t_0, t_0 + kh]\alpha = 0$. Interpreting this term by term in the right-hand side of (20), we obtain

$$0 = \alpha^H W_c[t_0, t_0 + kh]\alpha = \sum_{i=0}^{k-1} \alpha^H e^{-Qih} W_c[t_0, t_0 + h] e^{-Q^T ih} \alpha.$$

Note that each term in the summation of the above equation is nonnegative. Then it follows that $\alpha^H e^{-Qih} W_c[t_0, t_0 + h] e^{-Q^T ih} \alpha = 0$ for any $i = 1, 2, \ldots, k-1$. Also, $W_c[t_0, t_0 + h]$ is symmetric and nonnegative, and its square root $W_c^{1/2}[t_0, t_0 + h]$ exists [12]. We obtain that $W_c^{1/2}[t_0, t_0 + h] e^{-Q^T ih} \alpha = 0$ for any $i = 1, 2, \ldots, k-1$ and thus

$$W_c[t_0, t_0 + h] e^{-Q^T ih} \alpha = 0 \quad \forall i = 1, 2, \ldots, k-1. \tag{21}$$

Let $a(z)$ be a polynomial of minimal degree for $e^{-Q^T h}$ at α; that is, $a(e^{-Q^T h})\alpha = 0$. It is easy to see by the Cayley–Hamilton theorem about the characteristic equations of matrices that $a(z)$ always exists and the degree of $a(z)$ is less than or equal to the

dimension of $e^{-Q^T h}$. Factorizing $a(z)$ in the form of $a(z) = (z - \lambda)b(z)$ with λ being a root of $a(z) = 0$ and the degree of $b(z)$ being strictly less than that of $a(z)$, we can assert that

$$0 = (e^{-Q^T h} - \lambda I)b(e^{-Q^T h})\alpha := (e^{-Q^T h} - \lambda I)\mathbf{a}, \tag{22}$$

where $\mathbf{a} = b(e^{-Q^T h})\alpha$. Then $b(e^{-Q^T h})\alpha \neq 0$ (otherwise $b(z)$ is also a polynomial of minimal degree for $e^{-Q^T h}$ at α but of degree strictly less than that of $a(z)$, which is impossible). The fact of (22) means that λ is actually an eigenvalue of $e^{-Q^T h}$ and $\mathbf{a} \neq 0$ is an associated eigenvector; or, equivalently, $\mathbf{a}^T \neq 0$ is a left eigenvector of Q.

Now we show that $W_c[t_0, t_0 + h]\mathbf{a} = 0$. Note that k is larger than or equal to the degree of the minimal polynomial and that the degree of $b(z)$ must be strictly less than k. Based on these facts, it follows by (21) that $W_c[t_0, t_0 + h]\mathbf{a} = \sum_i b_i W_c[t_0, t_0 + h]e^{-Q^T ih}\alpha = 0$, where b_i are the coefficients of $b(z)$. Recalling that $e^{-Q^T h}\mathbf{a} = \lambda\mathbf{a}$, simple manipulations on the right-hand side of (20), in which k is replaced with v, lead to the desired assertion. □

Appendix B. Proofs for propositions.

Proof of Proposition 2.6. The proof is completed in two steps.

Step 1. It is shown that the system (1) is completely controllable at t_0 if and only if $W_c[t_0, t_0 + kh] > 0$. The sufficiency is obvious. To see the necessity, assume that the system (1) is completely controllable at t_0; or, equivalently, there is $t_1 > t_0$ such that $W_c[t_0, t_1] > 0$.

Case (i) $t_1 = t_0 + kh$. The necessity assertion follows readily.

Case (ii) $t_1 < t_0 + kh$. Note by segmenting the integral interval $[t_0, t_0 + kh]$ that

$$W_c[t_0, t_0 + kh] = W_c[t_0, t_1] + \int_{t_1}^{t_0+kh} \Phi(t_0, \tau)B(\tau)B^T(\tau)\Phi^T(t_0, \tau)d\tau > 0$$

since $W_c[t_0, t_1] > 0$ and the integral in the above equation is at least positive semi-definite.

Case (iii) $t_1 > t_0 + kh$. Note that $W_c[t_0, t_1] > 0$ always implies $W_c[t_0, t_2] > 0$ as long as $t_2 \geq t_1$. Hence, in this case we lose no generality by assuming that $t_1 = t_0 + vh$ for some integer $v > k$. Now we show that $W_c[t_0, t_0 + vh] > 0$ entails $W_c[t_0, t_0 + kh] > 0$. To see this, suppose that $W_c[t_0, t_0 + kh] \geq 0$ under $W_c[t_0, t_0 + vh] > 0$. However, Lemma 2.5 says that if $W_c[t_0, t_0 + kh] \geq 0$, then there exists a left eigenvector $\mathbf{a}^H \neq 0$ of Q such that $\mathbf{a}^H W_c[t_0, t_0 + vh]\mathbf{a} = 0$ for any integer $v \geq 1$. This is contradictory to $W_c[t_0, t_0 + vh] > 0$.

Step 2. It is shown that the system (1) is completely controllable if and only if $W_c[0, kh] > 0$.

(Necessity) Assume that the system (1) is completely controllable. Then, the system (1) is completely controllable at each $t_0 \in [0, \infty)$. In particular, the system is completely controllable at $t_0 = 0$. This, together with the results in Step 1, implies that $W_c[0, kh] > 0$.

(Sufficiency) Assume that $W_c[0, kh] > 0$. It is shown that the system (1) is completely controllable; that is, $W_c[t_0, t_0 + kh] > 0$ for any $t_0 \in [0, \infty)$. To this end, we observe that

$$W_c[\mu h, \mu h + kh] = \int_{\mu h}^{\mu h+kh} \Phi(\mu h, \tau)B(\tau)B^T(\tau)\Phi^T(\mu h, \tau)d\tau$$

$$
\begin{aligned}
&= \int_0^{kh} \Phi(\mu h, \tau' + \mu h) B(\tau' + \mu h) B^T(\tau' + \mu h) \Phi^T(\mu h, \tau' + \mu h) d\tau' \\
&= \int_0^{kh} \Phi(\mu h, 0) \Phi^{-1}(\tau' + \mu h, 0) B(\tau' + \mu h) \\
&\qquad \cdot B^T(\tau' + \mu h) \Phi^{-T}(\tau' + \mu h, 0) \Phi^T(\mu h, 0) d\tau' \qquad (23) \\
&= \int_0^{kh} P(\mu h, 0) e^{-Q\tau'} P^{-1}(\tau' + \mu h, 0) B(\tau' + \mu h) \\
&\qquad \cdot B^T(\tau' + \mu h) P^{-T}(\tau' + \mu h, 0) e^{-Q^T\tau'} P^T(\mu h, 0) d\tau' \\
&= \int_0^{kh} e^{-Q\tau'} P^{-1}(\tau', 0) B(\tau') B^T(\tau') P^{-T}(\tau', 0) e^{-Q^T\tau'} d\tau' = W_c[0, kh] > 0,
\end{aligned}
$$

where we used the fact that $P(\mu h, 0) = P(0, 0) = I$. The above equation implies that the system is completely controllable at $t_0 = \mu h$ for each integer $\mu \geq 0$. Note by the Gramian criterion that if a system is completely controllable at time t_0, then it is completely controllable at any $t \leq t_0$. Since μ is arbitrary, the desired assertion follows. □

Proof of Proposition 4.1. By the assumptions on $A(t)$ and $B(t)$ and the Floquet theorem, it can be shown that $\hat{B}(t) = P^{-1}(t, 0)B(t)$ is continuous in t. Hence, the approximate FDLCP system (14) is well defined for each segmentation $\mathcal{S}(\kappa)$ in the sense described in the paragraph below (14).

Now let us observe the following arguments, by means of Proposition 2.7.

$$
\begin{aligned}
&\text{The FDLCP system (1) is not completely controllable} \\
&\Leftrightarrow \beta^H \int_0^h e^{-Q\tau} \hat{B}(\tau) \hat{B}^T(\tau) e^{-Q^T\tau} d\tau \beta = 0, \quad \exists \beta \neq 0 \in \mathcal{C}^n \\
&\Leftrightarrow \beta^H e^{-Qt} \hat{B}(t) = 0 \quad \forall t \in [0, h), \ \exists \beta \neq 0 \in \mathcal{C}^n \\
&\Rightarrow \beta^H e^{-Qt} \hat{B}(t_{k-1}) = 0 \quad \forall t \in [t_{k-1}, t_k), \ \forall \mathcal{S}(\kappa), \ \exists \beta \neq 0 \in \mathcal{C}^n \\
&\Leftrightarrow \beta^H \sum_{k=1}^{\kappa} \int_{t_{k-1}}^{t_k} e^{-Q\tau} \hat{B}(t_{k-1}) \hat{B}^T(t_{k-1}) e^{-Q^T\tau} d\tau \beta = 0 \quad \forall \mathcal{S}(\kappa), \ \exists \beta \neq 0 \in \mathcal{C}^n \\
&\Leftrightarrow \beta^H \sum_{k=1}^{\kappa} \int_{t_{k-1}}^{t_k} e^{-Q\tau} \hat{B}(\kappa, \tau) \hat{B}^T(\kappa, \tau) e^{-Q^T\tau} d\tau \beta = 0 \quad \forall \mathcal{S}(\kappa), \ \exists \beta \neq 0 \in \mathcal{C}^n \\
&\Leftrightarrow \text{The system (14) is not completely controllable for any } \mathcal{S}(\kappa). \qquad (24)
\end{aligned}
$$

A sufficiency assertion, that is, $\Leftarrow$, in the third relationship in (24) can also be derived. This is because noncontrollability of the approximate system (14) tells us that there exists $\beta \neq 0 \in \mathcal{C}^n$ so that $\beta^H e^{-Qt} \hat{B}(t)$ vanishes almost everywhere in $[0, h)$ as long as $\kappa \to \infty$. □

Appendix C. Proof of Theorem 4.2. The proof will be completed in 4 steps.

Step 1. It is shown that T_κ of (15) can transform the state matrix Q of the system (14) into a state matrix in the form of (16) with a specific segmentation $\mathcal{S}(\kappa)$ when $\kappa \to \infty$.

Note that the pair $(A(\cdot), B(\cdot))$ is not completely controllable. Then, Proposition 4.1 says that the system (14) is not completely controllable, no matter how large κ is taken. From this noncontrollability and Proposition 2.7, the approximate FDLCP system (14) defined via $\mathcal{S}(\kappa)$ is not completely controllable if and only if there exists

$\beta \neq 0 \in \mathcal{C}^n$ such that

$$\begin{aligned}
&\beta^H \int_0^h e^{-Q\tau}\hat{B}(\kappa,\tau)\hat{B}^H(\kappa,\tau)e^{-Q^T\tau}d\tau\beta = 0\\
&\Leftrightarrow \beta^H e^{-Qt}\hat{B}(\kappa,t) = 0 \quad \forall t \in [0,h)\\
&\Leftrightarrow \beta^H e^{-Qt}\hat{B}(t_{k-1}) = 0 \quad \forall t \in [t_{k-1},t_k),\ k = 1,2,\ldots,\kappa,\\
&\Leftrightarrow \beta^H e^{-Q(t+t_{k-1})}\hat{B}(t_{k-1}) = 0 \quad \forall t \in [0,t_k - t_{k-1}),\ k = 1,2,\ldots,\kappa,\\
&\Leftrightarrow \beta^H e^{-Qt}e^{-Qt_{k-1}}\hat{B}(t_{k-1}) = 0 \quad \forall t \in \Big[0, \min_{k=1,2,\ldots,\kappa} |t_k - t_{k-1}|\Big)\\
&\Leftrightarrow \beta^H e^{-Qt}\hat{B}_\kappa = 0 \quad \forall t \in \Big[0, \min_{k=1,2,\ldots,\kappa} |t_k - t_{k-1}|\Big)\\
&\Leftrightarrow \beta^H e^{-Qt}\hat{B}_\kappa = 0 \quad \forall t \geq 0, \qquad (25)
\end{aligned}$$

which implies that the constant pair $(Q, \hat{B}_\kappa)$ is not completely controllable. This tells us that there exists a left eigenvector β_κ^H of Q such that

$$\beta_\kappa^H[\lambda I - Q, \hat{B}_\kappa] = 0. \qquad (26)$$

Now we show that such a left eigenvector β_κ^H is related to some uncontrollable modes of Q, which can be "extracted" by applying the linear transformation T_κ on Q. To see this, we define

$$\mathcal{E}_\kappa := \{\eta \in \mathcal{C}^n : \eta^H[\hat{B}_\kappa, Q\hat{B}_\kappa, \cdots, Q^{n-1}\hat{B}_\kappa] = 0, \eta \neq 0\}. \qquad (27)$$

Clearly, $\mathcal{E}_\kappa$ is nonempty since at least $\beta \in \mathcal{E}_\kappa$ by (25).

Note in (15) that each column vector in $T_{1\kappa}$ is orthogonal to any column vectors in $T_{2\kappa}$. This implies that column vectors of $T_{2\kappa}$ form a base for $\mathcal{E}_\kappa$. Furthermore, we notice that $Q^H\eta \in \mathcal{E}_\kappa$ for any $\eta \in \mathcal{E}_\kappa$, which follows from the definition of $\mathcal{E}_\kappa$ and the Cayley–Hamilton theorem about characteristic polynomials of square matrices. These facts say that each vector in $Q^H T_{2\kappa}$ must belong to $\mathcal{E}_\kappa$. Therefore, we have $T_{2\kappa}^H Q T_{1\kappa} = 0$ and then it follows that

$$T_\kappa^H Q T_\kappa = \begin{bmatrix} X_\kappa & Y_\kappa \\ 0 & Z_\kappa \end{bmatrix} \qquad (28)$$

for some matrices X_κ, Y_κ, and Z_κ of compatible dimensions. Note that Z_κ is square. For an eigenvalue $\lambda(Z_\kappa)$ and a corresponding left eigenvector $\hat{\eta}_\kappa^H$, we construct an augmented vector

$$\eta_\kappa := \begin{bmatrix} 0 \\ \hat{\eta}_\kappa \end{bmatrix} \in \mathcal{C}^n.$$

Then, it follows readily that $\lambda(Z_\kappa)$ is also an eigenvalue of Q with a left eigenvalue $T_\kappa\eta_\kappa$ since

$$\eta_\kappa^H T_\kappa^{-1} Q\, T_\kappa = \eta_\kappa^H \begin{bmatrix} X_\kappa & Y_\kappa \\ 0 & Z_\kappa \end{bmatrix} = \begin{bmatrix} 0 & \hat{\eta}_\kappa^H Z_\kappa \end{bmatrix} = \begin{bmatrix} 0 & \hat{\eta}_\kappa^H \lambda(Z_\kappa) \end{bmatrix} = \lambda(Z_\kappa)\eta_\kappa^H.$$

Let $\beta_\kappa := T_\kappa\eta_\kappa$ and we must show that such a β_κ satisfies (26). To this end, we observe that

$$\beta_\kappa^H \hat{B}_\kappa = \eta_\kappa^H T_\kappa^H \hat{B}_\kappa = \eta_\kappa^H \begin{bmatrix} T_{1\kappa}^H \\ T_{2\kappa}^H \end{bmatrix} \hat{B}_\kappa = \eta_\kappa^H \begin{bmatrix} *(\gamma(\kappa), m\kappa) \\ 0(n-\gamma(\kappa), m\kappa) \end{bmatrix} = 0, \qquad (29)$$

where $0(n-\gamma(\kappa), m\kappa)$ follows from the definition of $T_{2\kappa}$ and the fact that the column vectors in $\hat{B}_\kappa$ are a linear combination of $\epsilon_\kappa^{(1)}, \cdots, \epsilon_\kappa^{(\gamma(\kappa))}$. Based on (29) and the fact that β_κ is a left eigenvector of Q, we obtain (26). In other words, we can conclude from (26) and (29) that eigenvalues in $\lambda(Z_\kappa)$ are uncontrollable modes of the approximate system (14). Note by (28) that $\lambda(Z_\kappa) \subset \lambda(Q)$. It follows that eigenvalues in $\lambda(Z_\kappa)$ are uncontrollable modes of the system (1).

Step 2. We show that T_κ defined in (15) can be taken independently of $\mathcal{S}(\kappa)$ that is taken appropriately as long as $\kappa \to \infty$.

Let us construct consecutive segmentations as follows. First, segment $[0, h)$ into κ_1 subintervals $[t_0, t_1)$, $[t_1, t_2), \cdots, [t_{\kappa_1-1}, t_{\kappa_1})$, which form $\mathcal{S}(\kappa_1)$. In the ensuing segmentation $\mathcal{S}(\kappa_2)$, we partition $[0, h)$ into κ_2 subintervals, each of which is constructed by segmenting one of the subintervals in $\mathcal{S}(\kappa_1)$. Thus, each ending point of a subinterval in $\mathcal{S}(\kappa_1)$, i.e., $t_0, t_1, \cdots, t_{\kappa_1}$, also appears as an ending point of some subinterval contained in $\mathcal{S}(\kappa_2)$. In such a way, we can obtain a sequence of consecutive segmentations $\mathcal{S}(\kappa_1)$, $\mathcal{S}(\kappa_2), \cdots$, satisfying $0 < \kappa_1 < \kappa_2 < \cdots$, and define the approximate systems Σ_{κ_1}, $\Sigma_{\kappa_2}, \cdots$ as in (14) and $\hat{B}_{\kappa_1}$, $\hat{B}_{\kappa_2}, \cdots$ as introduced in (25).

Clearly, $\hat{B}_{\kappa_{i-1}}$ is contained in $\hat{B}_{\kappa_i}$ as a submatrix for any i. Thus

$$\operatorname{rank}[\hat{B}_{\kappa_i}, Q\hat{B}_{\kappa_i}, \cdots, Q^{n-1}\hat{B}_{\kappa_i}] \geq \operatorname{rank}[\hat{B}_{\kappa_{i-1}}, Q\hat{B}_{\kappa_{i-1}}, \cdots, Q^{n-1}\hat{B}_{\kappa_{i-1}}],$$

which means in turn that

$$\cdots \subseteq \mathcal{E}_{\kappa_i} \subseteq \cdots \subseteq \mathcal{E}_{\kappa_2} \subseteq \mathcal{E}_{\kappa_1}, \tag{30}$$

where $\mathcal{E}_{\kappa_i}$ is defined as in (27) but in terms of Q and $\hat{B}_{\kappa_i}$. From (30), it follows that $\cdots \leq \dim(\mathcal{E}_{\kappa_i}) \leq \cdots \leq \dim(\mathcal{E}_{\kappa_2}) \leq \dim(\mathcal{E}_{\kappa_1})$. In view of this, we turn to show that if the system (1) is not completely controllable, there exists an integer $\kappa_1 > 0$ that is large enough such that for all $\kappa_i \geq \kappa_1$ it holds that

$$\dim(\mathcal{E}_{\kappa_i}) = \dim(\mathcal{E}_{\kappa_1}) =: \gamma \geq 1, \tag{31}$$

with γ being an integer.

By contradiction, suppose that (31) is not true; i.e., for some large $\kappa_1 > 0$, one has that $\dim(\mathcal{E}_{\kappa_1}) = 0$. It means by definition of $\mathcal{E}_{\kappa_1}$ that $\operatorname{rank}[\hat{B}_{\kappa_1}, Q\hat{B}_{\kappa_1}, \cdots, Q^{n-1}\hat{B}_{\kappa_1}] = n$, which says that the constant pair $(Q, \hat{B}_{\kappa_1})$ is completely controllable. Thus, for any nonzero vector $b \in \mathcal{C}^n$, $b^T e^{-Qt}\hat{B}_{\kappa_1} \neq 0\ \forall t \geq 0$. If we interpret this along subintervals in $\mathcal{S}(\kappa_1)$ (i.e., we argue similarly to those in (25) but in the sense of being controllable), it follows that at least on one subinterval, say $[t_{k-1}, t_k)$ in $\mathcal{S}(\kappa_1)$, it holds that $b^T e^{-Qt}\hat{B}(\kappa_1, t) \neq 0$. Hence, we obtain that

$$\begin{aligned} & b^T \int_0^h e^{-Q\tau}\hat{B}(\kappa_1, \tau)\hat{B}^T(\kappa_1, \tau)e^{-Q^T\tau}d\tau \\ = & \sum_{k=0,1,\ldots,\kappa_1-1} b^T \int_{t_{k-1}}^{t_k} e^{-Q\tau}\hat{B}(t_{k-1})\hat{B}^T(t_{k-1})e^{-Q^T\tau}d\tau\, b > 0. \end{aligned}$$

Again noticing that in the approximate system Σ_{κ_1} of (14), $W_c[0, vh] = W_c[0, h] + W_c[h, vh]$ with v be any positive integer, we can assert that $W_c[0, vh] > 0$. Hence, Proposition 2.7 says that the approximate system Σ_{κ_1} is completely controllable.

Let us return to (30) and notice that $\dim(\mathcal{E}_{\kappa_i}) \geq 0$ for any κ_i. Then $\dim(\mathcal{E}_{\kappa_1}) = 0$ tells us that for any $\kappa_i > \kappa_1$, $\dim(\mathcal{E}_{\kappa_i}) = 0$. By repeating the arguments in the

previous paragraph but in terms of $\hat{B}_{\kappa_i}$, one can assert that the approximate system Σ_{κ_i} (14) defined on $\mathcal{S}(\kappa_i)$ is completely controllable for each $\kappa_i > \kappa_1$. Bearing this in mind, Proposition 4.1 yields that the system Σ of (13) is also completely controllable or, equivalently, the FDLCP system (1) is completely controllable. However, this is a contradiction.

In summary, (31) indicates that there exist orthonormal vectors $\epsilon_1, \cdots, \epsilon_\gamma$, which are determined during $\mathcal{S}(\kappa_1)$, and available orthonormal vectors for $\mathcal{S}(\kappa_i)$ uniformly over $\kappa_i > \kappa_1$ as well. That is, $\epsilon_1, \cdots, \epsilon_\gamma$ form a base for each $\mathcal{E}_{\kappa_i}$, $\kappa_i > \kappa_1$, while $\epsilon_1, \cdots, \epsilon_n$ form a base for $\mathcal{C}^n$ by including $n-\gamma$ more orthonormal vectors $\epsilon_{\gamma+1}, \cdots, \epsilon_n$. Based on such a base, we have that T_{κ_1} defined in (15) can satisfy (28) for any $\kappa_i \geq \kappa_1$.

Step 3. It is shown that T_{κ_1} defined in Step 2 transforms $\hat{B}(t)$ of the system (13) into the input matrix of the system (16); that is,

$$T_{\kappa_1}^{-1}\hat{B}(t) = \begin{bmatrix} B_*(t) \\ 0 \end{bmatrix}, \tag{32}$$

where $B_*(t)$ is a $\gamma \times m$ h-periodic matrix. Note that $\hat{B}_{\kappa_1}$ is contained in $\hat{B}_{\kappa_i}$ as a submatrix for any $\kappa_i > \kappa_1$. Therefore, for any $\kappa_i > \kappa_1$ we have from (29) that

$$\begin{aligned} T_{\kappa_1}^{-1}\hat{B}_{\kappa_i} = T_{\kappa_1}^H \hat{B}_{\kappa_i} &= \left[\frac{*(\gamma, m\kappa_i)}{0(n-\gamma, m\kappa_i)}\right] \\ &= \left[T_{\kappa_1}^H e^{-Q\tau_0}\hat{B}(\tau_0), \cdots, T_{\kappa_1}^H e^{-Q\tau_{\kappa_i-1}}\hat{B}(\tau_{\kappa_i-1})\right] \\ &= \left[T_{\kappa_1}^H e^{-Q\tau_0} T_{\kappa_1} T_{\kappa_1}^H \hat{B}(\tau_0), \cdots, T_{\kappa_1}^H e^{-Q\tau_{\kappa_i-1}} T_{\kappa_1} T_{\kappa_1}^H \hat{B}(\tau_{\kappa_i-1})\right] \\ &= \left[e^{-T_{\kappa_1}^H Q T_{\kappa_1}\tau_0} T_{\kappa_1}^H \hat{B}(\tau_0), \cdots, e^{-T_{\kappa_1}^H Q T_{\kappa_1}\tau_{\kappa_i-1}} T_{\kappa_1}^H \hat{B}(\tau_{\kappa_i-1})\right] \\ &= \left[e^{-Q_{\kappa_1}\tau_0} T_{\kappa_1}^H \hat{B}(\tau_0), \cdots, e^{-Q_{\kappa_1}\tau_{\kappa_i-1}} T_{\kappa_1}^H \hat{B}(\tau_{\kappa_i-1})\right], \end{aligned} \tag{33}$$

where $T_{\kappa_1}^H Q T_{\kappa_1}$ satisfies (28) in form, and thus we have

$$Q_{\kappa_1} := \begin{bmatrix} X_{\kappa_1} & Y_{\kappa_1} \\ 0 & Z_{\kappa_1} \end{bmatrix}.$$

Furthermore, it is evident by trivial manipulations that

$$e^{-Q_{\kappa_1}\tau_k} = \begin{bmatrix} e^{-X_{\kappa_1}\tau_k} & *(\cdot) \\ 0 & e^{-Z_{\kappa_1}\tau_k} \end{bmatrix}, \quad k = 0, 1, \ldots, \kappa_i - 1.$$

Using $e^{-Q_{\kappa_1}\tau_k}$ $(k = 0, 1, \ldots, \kappa_i - 1)$ back to (33) and denoting $T_{\kappa_1} = [T_{1\kappa_1} \,|\, T_{2\kappa_1}]$ in the same sense as we express T_κ in (15), we obtain

$$\left[\frac{*(\gamma, m\kappa_i)}{0(n-\gamma, m\kappa_i)}\right] = \left[\cdots, \begin{bmatrix} e^{-X_{\kappa_1}\tau_k} & *(\cdot) \\ 0 & e^{-Z_{\kappa_1}\tau_k} \end{bmatrix} \begin{bmatrix} T_{1\kappa_1}^H \\ T_{2\kappa_1}^H \end{bmatrix} \hat{B}(\tau_k), \cdots\right].$$

Comparing the corresponding entries in the lower halves of the left and right sides of the above equation, it is easy to see that

$$e^{-Z_{\kappa_1}\tau_k} T_{2\kappa_1}^H \hat{B}(\tau_k) = 0(n-\gamma, m) \quad \forall k = 0, 1, \ldots, \kappa_i - 1.$$

Note that $e^{-Z_{\kappa_1}\tau_k}$ is always invertible. Then it follows readily that $T_{2\kappa_1}^H \hat{B}(\tau_k) = 0(n-\gamma, m)$ for any $k = 0, 1, \ldots, \kappa_i - 1$. Based on this, it follows readily that

$$T_{\kappa_1}^H \hat{B}(\tau_k) = \left[\frac{*(\gamma, m)}{0(n-\gamma, m)}\right] \quad \forall k = 0, 1, \ldots, \kappa_i - 1.$$

We mention that τ_k $(k = 0, 1, \ldots, \kappa_i - 1)$ can be taken in $[0, h)$ arbitrarily as long as $\kappa_i \geq \kappa_1$ with κ_1 being sufficiently large. This implies nothing but (32).

Step 4. By letting $T_\kappa = T_{\kappa_1}$ and denoting $Q_c = X_{\kappa_1}$, $Q_{12} = Y_{\kappa_1}$, $Q_{\bar{c}} = Z_{\kappa_1}$, and $B_c(t) = B_*(t)$, it follows that the state coordinate transform $z = T_\kappa \tilde{x}$ transforms Σ to the form of (16). The assertion that the pair $(Q_c, B_c(\cdot))$ is completely controllable follows from the fact that γ given in (31) is the biggest one when κ_1 is sufficiently large. This completes the proof. ∎

REFERENCES

[1] S. Bittanti, G. Guardabassi, C. Maffezzoni, and L. Silverman, *Periodic systems: Controllability and the matrix Riccati equation*, SIAM J. Control Optim., 16 (1978), pp. 37–40.

[2] S. Bittanti, P. Colaneri, and G. Guardabassi, *H-controllability and observability of linear periodic systems*, SIAM J. Control Optim., 22 (1984), pp. 889–893.

[3] S. Bittanti, A. J. Laub, and J. C. Willems, eds., *The Riccati Equation*, Springer-Verlag, Berlin, 1991, pp. 131–162.

[4] S. Bittanti and P. Bolzern, *Stabilizability and detectability of linear periodic systems*, Systems Control Lett., 6 (1985), pp. 141–145.

[5] P. Bolzern and P. Colaneri, *The periodic Lyapunov equation*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 499–512.

[6] S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.

[7] P. Brunovsky, *Controllability and linear closed-loop controls in linear periodic systems*, J. Differential Equations, 6 (1969), pp. 296–313.

[8] C. L. Devito, *Functional Analysis and Linear Operator Theory*, Addison-Wesley, Redwood City, CA, 1990.

[9] J. Dugundji and J. H. Wendell, *Some analysis methods for rotating systems with periodic coefficients*, AIAA J., 21 (1983), pp. 890–897.

[10] M. Farkas, *Periodic Motions*, Springer-Verlag, New York, 1994.

[11] E. G. Gilbert, *Controllability and observability in multivariable control systems*, J.S.I.A.M. Control Ser. A, 1 (1963), pp. 128–151.

[12] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators*, Vol. II, Birkhäuser Verlag, Basel, 1993.

[13] M. Green and D. J. N. Limebeer, *Linear Robust Control*, Prentice-Hall, Upper Saddle River, NJ, 1995, pp. 93–96.

[14] M. Grimble, *Industrial Control Systems Design*, Wiley-Interscience, New York, 2001.

[15] G. Guo, J. F. Qiao, and C. Z. Han, *Controllability of periodic systems: Continuous and discrete*, IEEE Proc. Control Theory Appl., 151 (2004), pp. 488–490.

[16] A. Halanay, *Differential Equations: Stability, Oscillations, Time Lags*, Academic Press, New York, 1966.

[17] G. A. Hewer, *Periodicity, detectability and the matrix Riccati equation*, SIAM J. Control, 13 (1975), pp. 1235–1251.

[18] P. T. Kabamba, S. M. Meerkov, and E.-K. Poh, *Pole placement capabilities of vibrational control*, IEEE Trans. Automat. Control, 43 (1998), pp. 1256–1261.

[19] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[20] R. E. Kalman, Y. C. Ho, and K. S. Narendra, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1, (1963), pp. 189–213.

[21] R. E. Kalman, *Mathematical description of linear dynamical systems*, J.S.I.A.M. Control Ser. A, 1 (1963), pp. 152–192.

[22] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, The Clarendon Press, Oxford University Press, New York, 1995.

[23] D. L. Lukes, *Differential Equations: Classical to Controlled*, Academic Press, New York, 1982.

[24] P. Montagnier, C. C. Paige, and R. J. Spiteri, *Real Floquet factors of linear time-periodic systems*, Systems Control Lett., 50 (2003), pp. 251–262.

[25] P. Montagnier and R. J. Spiteri, *A Gramian-based controller for linear periodic systems*, IEEE Trans. Automat. Control, 49 (2004), pp. 1380–1385.

[26] P. Montagnier, R. J. Spiteri, and J. Angeles, *The control of linear time-periodic systems using Floquet-Lyapunov theory*, Internat. J. Control, 77 (2004), pp. 472–490.

[27] G. D. Nicolao, G. Ferrari-Trecate, and S. Pinzoni, *Zeros of continuous-time linear periodic systems*, Automatica J. IFAC, 34 (1998), pp. 1651–1655.

[28] M. Pavella and P. G. Murthy, *Transient Stability of Power Systems: Theory and Practice*, John Wiley, New York, 1994.
[29] J. A. Richards, *Analysis of Periodically Time-Varying Systems*, Springer-Verlag, New York, 1983.
[30] H. H. Rosenbrock, *State-space and Multivariable Theory*, Nelson, London, 1970.
[31] W. J. Rugh, *Linear System Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[32] C. B. Schrader and M. K. Sain, *Research on system zeros: A survey*, Internat. J. Control, 50 (1989), pp. 1407–1433.
[33] L. M. Silverman and H. E. Meadows, *Controllability and observability in time-variable linear systems*, SIAM J. Control, 5 (1967), pp. 64–73.
[34] L. M. Silverman and B. D. O. Anderson, *Controllability, observability and stability of linear systems*, SIAM J. Control, 6 (1968), pp. 121–130.
[35] S. C. Sinha, R. Pandiyan, and J. S. Bibb, *Liapunov-Floquet transformation: Computation and applications to periodic systems*, J. Vib. Acoust., 118 (1996), pp. 209–219.
[36] S. Skogestad and I. Postlethwaite, *Multivariable Feedback Control—Analysis and Design*, Wiley, Chichester, UK, 1996.
[37] L. Weiss, *On the structure theory of linear differential systems*, SIAM J. Control, 6 (1968), pp. 659–680.
[38] N. M. Wereley, *Analysis and Control of Linear Periodically Time Varying Systems*, Ph.D. thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA, 1990.
[39] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer-Verlag, New York, 1979.
[40] V. A. Yakubovich and V. M. Starzhinskii, *Linear Differential Equations with Periodic Coefficients*, Vol. 1, John Wiley, New York, 1975.
[41] J. Zhou and T. Hagiwara, *Existence conditions and properties of the frequency response operators of continuous-time periodic systems*, SIAM J. Control Optim., 40 (2002), pp. 1867–1887.
[42] J. Zhou, T. Hagiwara, and M. Araki, *Stability analysis of continuous-time periodic systems via the harmonic analysis*, IEEE Trans. Automat. Control, 47 (2002), pp. 292–298.
[43] J. Zhou, T. Hagiwara, and M. Araki, *Spectral characteristics and eigenvalues computation of the harmonic state operators in continuous-time periodic systems*, Systems Control Lett., 53 (2004), pp. 141–155.
[44] K. Zhou, *Essentials of Robust Control*, Prentice-Hall, Upper Saddle River, NJ, 1998.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 655–682

LOCAL CONTROLLABILITY FOR A "SWIMMING" MODEL*

A. Y. KHAPALOV†

Abstract. We study the local controllability of a mathematical model of an abstract object which "swims" in the two-dimensional (2D) nonstationary Stokes fluid. We assume that this object consists of finitely many subsequently connected small sets ("thick points"), each of which can act upon any of the adjacent sets in a rotation fashion with the purpose of generating its fish- or snake-like motion. We regard the magnitudes of the respective rotation forces, entering the system's equations as coefficients, as multiplicative (or bilinear) controls. The structural integrity of the object is maintained by the elastic forces acting between the aforementioned adjacent sets according to Hooke's law. Models like this are of a interest in biology and engineering applications dealing with propulsion systems in fluids.

Key words. swimming model, coupled systems, multiplicative control, local controllability, nonstationary Stokes equation

AMS subject classifications. 76, 92, 35

DOI. 10.1137/050638424

1. Introduction: Model and its wellposedness.

1.1. Model description. The subject of our interest in this paper is the study of the swimming phenomenon (see, e.g., the classical works [12], [2]) from the controllability theory viewpoint. We would like to approach this issue by investigating the local controllability properties of an abstract object which applies fish- or snake-like motion to *"swim" in a fluid* (as *opposed to* bodies that are drifting or being pushed/pulled in a fluid by external forces). This object (we also call it an "apparatus" below) can be viewed as a very simplified model of a living organism (see [12], [15], [2], [3], [4], [16], and the references therein) or a "mechanical device (such as a robotic fish or eel, e.g., [5], [13], [14], and the references therein).

Modeling philosophy. It appears that the issue of modeling for the swimming phenomenon should be perceived as a variety of models of different levels of complexity describing various objects that can *propel themselves* in a fluid. Such objects can be modeled as solid bodies or not, can have different geometries, and can employ different "swimming techniques" (such as "snake-like" or "rowing"). Numerous approaches, currently available in the literature in this respect, reflect the interests and preferences of researchers using them, also imposing the respective limitations ("tradeoffs") on the resulting form of the equations involved.

For example, a number of models employ only the finite systems of ODEs to describe the positions of certain points of the swimming object at hand and avoid the use of fluid equations, replacing them with friction forces acting upon the aforementioned points (e.g., [14] and the references therein). On the other end of this spectrum, there are sophisticated infinite dimensional swimming models focusing on detailed study of the interaction between the solid bodies and the surrounding medium (see, e.g., [12], [2], and the references therein). However, in the latter case it can be more difficult to

*Received by the editors August 18, 2005; accepted for publication (in revised form) December 8, 2006; published electronically May 7, 2007.
http://www.siam.org/journals/sicon/46-2/63842.html
†Department of Mathematics, Washington State University, Pullman, WA 99164-3113 (khapala@wsu.edu). The work of this author was supported in part by NSF grant DMS-0504093.

construct a swimming model as a "solvable" system of coupled differential equations containing an equation which describes the *progress of the position of the body (such as, e.g., its center of mass) in the fluid*. This equation is critical if one wants to study the issue of controllability for the swimming phenomenon, which is our goal in this paper.

In this respect, we would like to begin with a "reasonably good starting model" which should, on the one hand, be simple enough from the mathematical viewpoint, while, on the other hand, be adequate enough to represent (at least some of the) principal elements and difficulties arising in the context of swimming processes (so that the developed controllability methodology could later be carried over to more complex models in various fluids). This model, in our opinion, should include (a) a fluid equation (and there are many of these available), (b) an equation describing the motion of the swimming object, and (c) the coupling between them. In this paper we employ the following approach (see Figures 1–3).

1. *We model a swimming object (an "apparatus") as a collection of "small" sets of nonzero measure ("thick" points) in the nonstationary two-dimensional* (2D) *Stokes fluid, which are linked to each other by the set of internal forces, satisfying the third law of Newton*. Some of these forces serve to maintain the structural integrity of the object, while the others excite its swimming motion in the fluid. (Thus, our swimming object and its motion can also be viewed as a version of the classical problem in mechanics about the motion of a system of particles, linked by internal forces, when they are placed in a "resisting" medium.)

2. Since the sum of all internal forces, defining the force acting upon the center of mass of the object at hand, is zero, *no actual motion (of the center of mass) occurs, if the object is not in the fluid*. However, when such an object is placed in the fluid, the interaction with the latter can result in its swimming motion. The variety of these motions is the goal of our study.

3. To further simplify the model *we identify the "thick" points forming the body of the object with the parts of the fluid they "occupy,"* which seems a reasonable assumption if such "points" are "small" and stay away from each other. This assumption allows us to avoid dealing with the mathematics of solid bodies in fluids at this point. (Note that in many theoretical works a solid body is viewed as a limit of a sequence of fluids of increasing density occupying its volume.)

Our modeling approach can be viewed as one derived from the approach developed by Peskin, Fauci, and others (see also the references in [15], [3], [4], [16]) in computational mathematical biology, where an object in a fluid is modeled as *an immaterial curve (immersed boundary), identified with the fluid, further discretized* for computational purposes *on some grid*. In turn, *our model* (1.1)–(1.3) *can be viewed as such an already discretized immaterial curve supported on the respective cells of the aforementioned grid*; see Figures 1–3.

The equations (1.1)–(1.3) below resemble, in particular, the equations (2.9) in [15, p. 223], where an object in a fluid is modeled as a collection of countably many points linked by internal forces instead of our finitely many "thick" points (which allows us to "replace" the δ-functions in the limit description of the forcing term in [15] with the integral terms in (1.2) and a finite sum in (1.3) involving "more analysis-friendly" characteristic functions). In [4, p. 93], e.g., the swimming object is represented as an immaterial curve (immersed boundary), which requires the use of a more sophisticated δ-function.

More precisely, we consider the following model, consisting of two *coupled* systems of equations—one for the nonstationary 2D Stokes fluid and another for the *position*

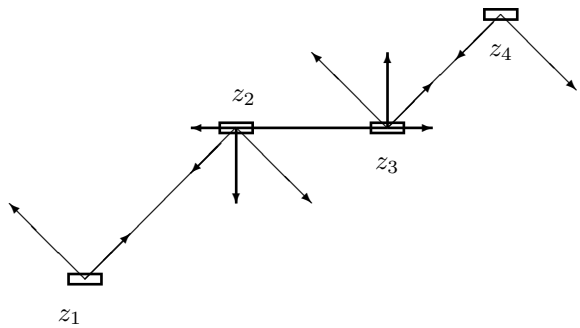


FIG. 1. *The case* $n = 4$.

of the apparatus in it:

$$\frac{\partial y}{\partial t} = \nu \Delta y + F(y, z, v) - \nabla p \quad \text{in} \quad Q_T = \Omega \times (0, T), \tag{1.1}$$

$$\text{div } y = 0 \text{ in } Q_T, \quad y = 0 \quad \text{in } \Sigma_T = \partial\Omega \times (0, T), \quad y\,|_{t=0} = y_0 \text{ in } \Omega,$$

$$\frac{dz_i}{dt} = \frac{1}{\text{mes}\,\{S_r(0)\}} \int_{S_r(z_i(t))} y(x,t)dx, \quad z_i(0) = z_{i0}, \quad i = 1, \dots, n, \quad n > 2, \tag{1.2}$$

where for $t \in [0, T]$

$$\begin{aligned} z(t) = (z_1(t), \dots, z_n(t)),\ z_i(t) \in R^2,\ \ i = 1, \dots, n, \quad v(t) = (v_1(t), \dots, v_{n-1}(t)) \in R^{n-1}, \\ F(y,z,v) = \sum_{i=2}^{n+1} \xi_{i-1}(x,t) \left[k_{i-1} \frac{(\| z_i(t) - z_{i-1}(t) \|_{R^2} - l_{i-1})}{\| z_i(t) - z_{i-1}(t) \|_{R^2}} (z_i(t) - z_{i-1}(t)) \right. \\ \left. + k_{i-2} \frac{(\| z_{i-2}(t) - z_{i-1}(t) \|_{R^2} - l_{i-2})}{\| z_{i-2}(t) - z_{i-1}(t) \|_{R^2}} (z_{i-2}(t) - z_{i-1}(t)) \right] \\ + \sum_{i=2}^{n+1} \xi_{i-1}(x,t) \left(v_{i-1}(t) A \left(z_i(t) - z_{i-1}(t) \right) + v_{i-2}(t) A \left(z_{i-2}(t) - z_{i-1}(t) \right) \right). \end{aligned} \tag{1.3}$$

In the above, Ω is a bounded domain in R^2 with boundary $\partial\Omega$ of class C^2, $y = (y_1(x,t), y_2(x,t))$ and $p(x,t)$ are, respectively, the velocity and the pressure of the fluid at point $x = (x_1, x_2) \in \Omega$ at time t, and ν is a kinematic viscosity constant. Also, to simplify the Σ-notation in (1.3) and below, throughout the paper we use two auxiliary fictitious points z_0 and z_{n+1} as $z_0(t) = z_1(t), z_n(t) = z_{n+1}(t)$, and we set accordingly $v_0 = v_n = k_0 = k_n = l_0 = l_n = 0$ (see below for more details).

Let us explain the terms in (1.1)–(1.3).

Apparatus. The swimming object in (1.1)–(1.3) is modeled as a collection of finitely many points with flexible immaterial internal links (or, say, which have a "negligible affect" on the swimming process), each of which is surrounded by *"very small immaterial"* support (i.e., identified with the fluid it occupies); see Figure 1.

Thus, *as a "mechanical device,"* our apparatus can be viewed as a *sequence of floating platforms connected by flexible links positioned above the surface of fluid*.

At any given moment of time the apparatus is represented by a "broken-line" structure, formed by an ordered sequence of "thick points" $S_r(z_1(t)), \dots, S_r(z_n(t))$, where $z_i(t), i = 1, \dots, n$, are points in Ω (the apparatus's "skeleton"). Accordingly,

in (1.3) the ξ_i's denote the respective characteristic functions of the $S_r(z_i(t))$'s:

$$\xi_i(x,t) = \begin{cases} 1 & \text{if } x \in S_r(z_i(t)), \\ 0 & \text{if } x \in \Omega \backslash S_r(z_i(t)), \end{cases} \qquad i = 1, \dots, n. \tag{1.4a}$$

We assume that (a) $S_r(0)$ is the given open set with its center of mass at the origin (if treated as a plate with uniform mass density) and (b)

$$S_r(0) = \{x \mid -r < x_1 < r,\ \alpha(x_1) < x_2 < \beta(x_1)\}, \tag{1.4b}$$

where α and β are the given continuously differentiable functions. Alternatively, $S_r(0)$ may consist of finitely many nonoverlapping sets similar to (1.4b), namely, of the form

$$\{x \mid -r \le r_* < x_1 < r^* \le r,\ \alpha_*(x_1) < x_2 < \beta_*(x_1)\}$$

or

$$\{x \mid -r \le r_{**} < x_2 < r^{**} \le r,\ \alpha_{**}(x_2) < x_1 < \beta_{**}(x_2)\}.$$

$S_r(a)$ denotes the set $S_r(0)$ shifted to point a.

Forces. The term $F(y,z,v)$ in (1.3) represents the *internal forces* (their sum is zero) generated by the apparatus, acting in turn as *external forces* upon the fluid in the fluid equation (1.1) (see also Remark 1.2 below). We assume that all the apparatus's forces act through the immaterial links attached to the centers of mass of sets $S_r(z_i(t))$, i.e., to the points $z_i(t)$, and then transmitted as such to all points in their respective supports. The latter points will create a pressure upon the surrounding fluid, thus acting as external forces upon it.

Each of the points $z_i(t)$ can force any of the adjacent points to "rotate" about it. In turn, by the third Newton law, the affected point will act back upon $z_i(t)$ with the opposite force. For example, $z_1(t)$ can act upon $z_2(t)$ with the force perpendicular to the vector $z_2(t) - z_1(t)$ and $z_2(t)$ will act back with the opposite force. These two forces, being transmitted to their respective supports, provide two terms in the last line in (1.3), namely,

$$\xi_1(x,t)v_1(t)A\left(z_2(t) - z_1(t)\right) + \xi_2(x,t)v_1(t)A\left(z_1(t) - z_2(t)\right),$$

where

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

The magnitudes and directions of the applied rotation forces (shown in Figure 2) are determined by the coefficients $v_i, i = 1, \dots, n-1$, which we regard as *bilinear or multiplicative* controls (see, e.g., [1], [6], [7], [8]).

The structural integrity of the apparatus is preserved by the elastic forces (shown in Figure 3) which act according to Hooke's law when the distances between any two adjacent points $z_i(t)$ and $z_{i-1}(t), i = 2, \dots, n$, deviate from the respective given values

$$l_{i-1} > 0, \quad i = 2, \dots, n, \tag{1.5}$$

as described in the first two lines in (1.3), where the given parameters $k_i > 0,\ i = 1, \dots, n-1$, characterize the rigidity of the links $z_{i-1}(t)z_i(t),\ i = 2, \dots, n$. (For the auxiliary points/links we set $k_0 = k_n = l_0 = l_n = 0$.)

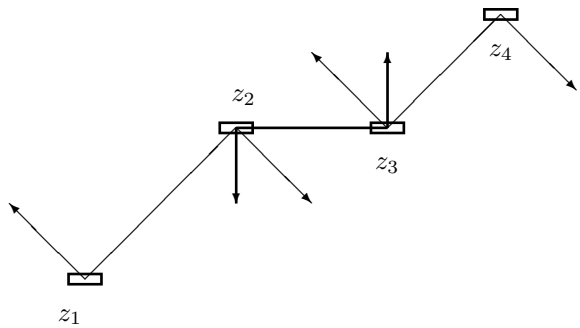


FIG. 2. *Controlling rotation forces,* $n = 4$.

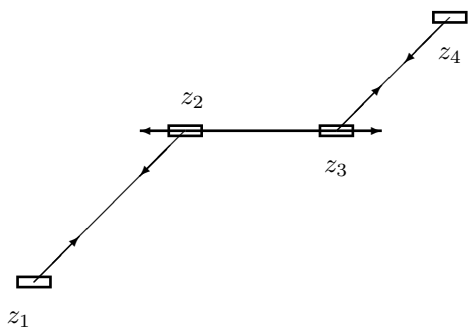


FIG. 3. *Elastic forces,* $n = 4$.

Remark 1.1.

- Note that, when the adjacent points in the apparatus share the same position in space (at some moments of time or on some time-interval), the forcing term F in (1.3) and hence the model (1.1)–(1.3) become undefined. While this situation seems physically plausible, even if, prior to this, the solution to this system exists on some "small" time-interval, it does not necessarily have to happen. The former issue, namely, the existence on some "small" time-interval $(0, T)$, is addressed below in the next section. The latter issue can be viewed as the issue of *controllability*, namely, when one tries to select multiplicative controls v_i with the purpose of ensuring that the apparatus is *"swimming" in the desirable fashion, while avoiding the aforementioned ill-defined situation.*
- Assuming that the set $S_r(0)$ in (1.4a)–(1.4b) is "small," we model all the apparatus's forces in (1.3) as of the same vector-value within their respective supports.

Apparatus's motion. The *dynamics of the "thick points"* $z_i(t)\xi_i(x,t), i = 1, \ldots, n$, are determined by the average motion of the fluid within their respective supports $S_r(z_i(t))$'s as described in (1.2).

Remark 1.2. Internal forces and conservation of momentums.

- We want to emphasize that all forces in (1.3) satisfy the third Newton law and their sum is equal to zero. Thus, they are *internal with respect to the apparatus and cannot move its center of mass without interaction with the fluid.* This is the principal feature of a "swimming-by-itself-device."
- The third Newton law ensures that the linear momentums generated by the apparatus's forces are conserved (see, e.g., [17]). However, the rotation forces

produce, in general, a nonzero torque. This means that the conservation of the angular momentums should hold in a more general framework, which also takes into account some "additional control forces" (from an "engine" such as, e.g., a "watch-and-hand" mechanism with its mutually counterrotating parts), also internal with respect to the apparatus, that generate the corresponding "negating" torques.

1.2. Existence and uniqueness. Let $\dot{J}(\Omega)$ denote the linear space of infinitely differentiable 2D vector functions $\phi(x) \in R^2$ which have compact support in Ω and are *solenoidal* (or *divergence-free*), that is, div $\phi = 0$ in Ω. By $H(\Omega)$ we denote the completion of this space in the norm

$$\| \phi \|_{H(\Omega)} = \left(\int_\Omega (\| \phi_{x_1} \|^2_{R^2} + \| \phi_{x_2} \|^2_{R^2}) dx \right)^{1/2}.$$

We decompose (e.g., [11], [18]) the vector space $(L^2(Q_T))^2$ into two orthogonal subspaces $J_0(Q_T)$ and $G(Q_T)$ assuming that for almost all $t \in (0,T)$ the elements of the former belong to the completion $J_0(\Omega)$ of $\dot{J}(\Omega)$ in the norm of $(L^2(\Omega))^2$ and the elements of the latter to its orthogonal complement $G(\Omega)$.

ASSUMPTION 1.1. *Assume that*

$$\| z_i(0) - z_{i-1}(0) \|_{R^2} = \mu_{i-1} > 0, \quad i = 2, \dots, n; \qquad \bar{S}_r(z_i(0)) \subset \Omega, \quad i = 1, \dots, n, \tag{1.6}$$

where "–" stands for closure, and the set $S_r(0)$ is such that

$$\int_{(S_r(0) \bigcup S_r(h)) \setminus (S_r(0) \bigcap S_r(h))} dx = \int_\Omega | \xi(x) - \xi(x-h) | \, dx \le C_0 \| h \|_{R^2} \qquad \forall h, \| h \|_{R^2} \in (0, h_0) \tag{1.7}$$

for some positive constants h_0 and C_0, where $\xi(x)$ is the characteristic function of $S_r(0)$.

Conditions (1.6) simply mean that our apparatus lies in Ω and that the positions of any two adjacent points forming it are distinct at time $t = 0$. Condition (1.7) is not difficult to satisfy—it holds, e.g., for rectangles and disks.

Throughout the paper we assume that $y_0 \in H(\Omega) \bigcap (H^2(\Omega))^2$. We have the following result from [9].

THEOREM 1.1. *Let $y_0 \in H(\Omega) \bigcap (H^2(\Omega))^2$; $T^* > 0; v_i \in L^\infty(0, T^*), k_i > 0, i = 1, \dots, n-1$; and $z_{i0} \in \Omega, i = 1, \dots, n$ be given, and let Assumption* 1.1 *hold. Then there exists a $T = T(z_{10}, \dots, z_{n0}, \| v_1 \|_{L^\infty(0,T^*)}, \dots, \| v_{n-1} \|_{L^\infty(0,T^*)}, \Omega) \in (0, T^*)$ such that the system* (1.1)–(1.3) *admits a unique solution $\{y, p, z\}$ on $(0,T)$, $\{y, \nabla p, z\} \in J_0(Q_T) \times G(Q_T) \times [C([0,T]; R^2)]^n$. Moreover, $y \in C([0,T]; H(\Omega))$, $y_t, y_{x_i x_j} \in (L^2(Q_T))^2$, $p_{x_i} \in L^2(Q_T)$, where $i, j = 1, 2$, and equations in* (1.1) *and* (1.2) *are satisfied almost everywhere, while*

$$z_i(t) \ne z_{i+1}(t), \quad i = 1, \dots, n-1; \qquad \bar{S}_r(z_i(t) \subset \Omega, \quad t \in [0,T], \quad i = 1, \dots, n. \tag{1.8}$$

Remark 1.3.

- Condition (1.7) is used in the proof of Theorem 1.1 (as a form of the Lipschitz condition).
- Condition (1.8) means that for the solution of (1.1)–(1.7), whose existence is established in Theorem 1.1, we can guarantee that on some "small" time-interval $[0,T]$ any two adjacent points $z_i(t)$ and $z_{i+1}(t)$ in the apparatus do

not share the same point in space, while our swimming device stays "sufficiently away" from the boundary of Ω. The former allows us to maintain the wellposedness (both mathematical and "physical"—the validity of Hooke's law) of the elastic forces in (1.3), while the latter implies that we do not have to deal with any "complications" arising when some of the "thick points" "hit" $\partial\Omega$.

- On the other hand, Theorem 1.1 allows sharing of the same space for some portions of supports of the aforementioned thick points (recall that they are assumed to be immaterial and "parts of the fluid"). At no extra cost, we could equally make the assumption (1.6) more strict to exclude the latter possibility by assuming that μ_i's (and l_i's) strictly exceed $2r$ or even to assume that (1.6) (and then (1.8)) holds with a margin exceeding $2r$ for all $z_i(t)$'s (*not only adjacent*), while modifying the statement of Theorem 1.1 accordingly.

The duration of the time-interval $(0,T)$ in Theorem 1.1 is not quite of local nature. Namely, based on suitable a priori estimates, the value of T is selected small enough to guarantee that condition (1.8) holds on $(0,T)$ for the given choice of data in (1.1)–(1.3). This solution can be extended further in time as long as (1.8) holds.

2. Local controllability: Problem formulation and main results.

2.1. Problem formulation. In this paper we would like to analyze the *"local" swimming capabilities* of model (1.1)–(1.7) in the following sense.

For the given initial datum of this model, namely,

$$\{y_0,\ z_i(0),\ \ i=1,\dots,n\}, \tag{2.1a}$$

denote by

$$\{y^*(x,t),\ z_i^*(t),\ \ i=1,\dots,n\} \tag{2.1b}$$

the solution pair to (1.1)–(1.7) generated by the zero controls $v_i=0,\ i=1,\dots,n-1$ (as long as (1.8) holds).

We also distinguish the following equilibrium initial state for system (1.1)–(1.7):

$$\begin{aligned}\{y_0=0,\ \ z_i(0),\ i=1,\dots,n,\ \ \text{such that}\ \ l_{i-1}=&\parallel z_i(0)-z_{i-1}(0)\parallel_{R^2},\\ &i=2,\dots,n\},\end{aligned} \tag{2.2a}$$

in which case the fluid "rests" and the apparatus does not move for any $t>0$, that is,

$$\{y^*(x,t)\equiv 0,\quad z_i^*(t)\equiv z_i(0),\ \ i=1,\dots,n\},\ \ t\geq 0. \tag{2.2b}$$

We intend to approach the general issue of local controllability for (1.1)–(1.7) by asking first a "rather simple" question.

Given the equilibrium initial datum (2.2a) *in* (1.1)–(1.7)*, can we move at least one point, say, z_i, in the skeleton of the apparatus anywhere within some neighborhood of its initial equilibrium position $z_i(0)$ at some preassigned moment $T>0$?*

We will call this problem the *local controllability problem with respect to z_i near equilibrium* at time T. However, the question of main interest, associated with the actual motion of any object, is the following.

Given the equilibrium initial datum (2.2a) *in* (1.1)–(1.7)*, can we move the "center of mass" of our apparatus, namely, the point*

$$z_c(t)=\frac{1}{n}\sum_{i=1}^{n} z_i(t),$$

anywhere within some neighborhood of its initial equilibrium position

$$z_c(0) = \frac{1}{n}\sum_{i=1}^{n} z_i(0)$$

at some preassigned moment $T > 0$*?*

We will call this problem the *local controllability problem with respect to* $z_c(0)$ *near equilibrium* at time T.

In the case of the general (not necessarily equilibrium) initial conditions (2.1a), the motion of the apparatus, associated with the zero-controls v_i's, is a "drifting" (uncontrolled) motion (2.1b), generated, on the one hand, by the given initial fluid condition y_0 and, on the other hand, by the elastic forces in the first two lines on the right in (1.3) trying to return the apparatus to its natural equilibrium position, that is, when the distances between the adjacent points z_i's are exactly l_i's. In this case, our goal is to investigate the *local controllability of system* (1.1)–(1.7) *near the "drifting" trajectory* $z_i^*(t)$, $i = 1, \dots, n$.

Given the initial datum in (1.1)–(1.7)*, can we move at least one point, say,* z_i*, of the apparatus or its center of mass* z_c *anywhere within some neighborhood, respectively, of* $z_i^*(T)$ *or of*

$$z_c^*(T) = \frac{1}{n}\sum_{i=1}^{n} z_i^*(T),$$

along its uncontrolled drifting trajectory (2.1a)–(2.1b) *for some* $T > 0$*?*

We will call these two problems the *local controllability problems near the drifting positions, respectively, of* $z_i^*(T)$ *and of* $z_c^*(T)$.

Our strategy in this paper is centered around the following propositions.

PROPOSITION 2.1. *Assume that in* (1.1)–(1.7) *only two controls are active, say,* v_j *and* v_l*, where* $j \neq l$ *and* $j, l \in \{1, \dots, n-1\}$*, while* $v_k = 0$ *for* $k = 1, \dots, n-1, k \neq j, l$*. Assume further that* v_j *and* v_l *are independent of time. Then, if for some* $i \in \{1, \dots, n\}$ *there exists a* $T > 0$ *such that the matrix*

$$\left(\frac{dz_i(T)}{dv_j} \mid_{v_m' s = 0}, \frac{dz_i(T)}{dv_l} \mid_{v_m' s = 0} \right) \tag{2.3}$$

is nondegenerate, then the system (1.1)–(1.7) *is locally controllable near its drifting position* $z_i^*(T)$ *in* (2.1b)*. Namely, there is an* $\varepsilon > 0$ *such that*

$$B_\varepsilon(z_i^*(T)) \subset \{z_i(T) \mid v_j, v_l \in R,\ v_k = 0 \text{ for } k = 1, \dots, n-1, k \neq j, l\}. \tag{2.4}$$

In particular, for the initial equilibrium position (2.2a) *condition* (2.3) *implies the local controllability with respect to* z_i *near equilibrium at time* T.

In other words, (2.4) means that the set of all possible positions of $z_i(T)$ when controls v_i run over R will include some ε-neighborhood of $z_i^*(T)$.

In (2.3) and anywhere below the subscript, $v_m' s = 0$ indicates that the corresponding expressions are calculated for $v_m = 0, m = 1, \dots, n-1$.

Proof of Proposition 2.1. This is an immediate consequence of the inverse function theorem, which, in view of (2.3), implies that the mapping

$$R^2 \ni (v_j, v_l) \ \rightarrow \ z_i(T) \in R^2,$$

defined on some (open) neighborhood of the origin, has the inverse mapping, defined on some (open) neighborhood of $z_i^*(T)$; that is, (2.4) holds. □

Clearly, the same argument implies a similar result for the motion of the center of mass $z_c(t)$.

PROPOSITION 2.2. *Assume that in* (1.1)–(1.7) *only two controls are active, say,* v_j *and* v_l*, where* $j \neq l$ *and* $j, l \in \{1, \dots, n-1\}$*, while* $v_k = 0$ *for* $k = 1, \dots, n-1, k \neq j, l$*. Assume further that* v_j *and* v_l *are independent of time. Then, if there exists a* $T > 0$ *such that the matrix*

$$\left(\frac{dz_c(T)}{dv_j} \mid_{v'_m s=0}, \frac{dz_c(T)}{dv_l} \mid_{v'_m s=0} \right) \tag{2.5}$$

is nondegenerate, then the system (1.1)–(1.7) *is locally controllable near its drifting position of the center of mass* $z_c^*(T)$*. Namely, there is an* $\varepsilon > 0$ *such that*

$$B_\varepsilon(z_c^*(T)) \subset \{z_c(T) \mid v_j, v_l \in R,\ v_k = 0 \text{ for } k = 1, \dots, n-1, k \neq j, l\}.$$

In particular, for the initial equilibrium position (2.2a) *we have the local controllability with respect to* z_c *near equilibrium at time* T*.*

Our *main results* below deal with the conditions under which the matrices in (2.3) and (2.5) in Propositions 2.1 and 2.2 are nondegenerate (the general scheme of our proofs is described in the beginning of section 3). They involve various assumptions on the initial position of the apparatus. To formulate them, we will need to introduce some notation first.

Let 2D vector functions $\omega_k \in J_0(\Omega) \bigcap H(\Omega)$ and real numbers $-\lambda_k, k = 1, \dots$ ($\lambda_k > 0, \lambda_k \to \infty$ as $k \to \infty$), denote, respectively, the orthonormalized in $(L^2(\Omega))^2$ eigenfunctions and eigenvalues of the spectral problem associated with (1.1):

$$\begin{aligned} \nu \Delta \omega_k \ - \ \nabla p_k &= -\lambda_k \omega_k \quad \text{in} \quad \Omega, \\ \operatorname{div} \omega_k = 0 \ \text{ in } \ \Omega, \quad \omega_k &= 0 \quad \text{in } \ \partial\Omega. \end{aligned}$$

Then the unique solution to (1.1), described in Theorem 1.1, admits the following *implicit* representation:

$$\begin{aligned} y(x,t) = &\sum_{k=1}^{\infty} e^{-\lambda_k t} \left(\int_\Omega y_0^T \omega_k dq \right) \omega_k(x) \\ &+ \sum_{k=1}^{\infty} \int_0^t e^{-\lambda_k (t-\tau)} \left(\int_\Omega F^T(y, z, v) \omega_k dq d\tau \right) \omega_k(x). \end{aligned} \tag{2.6}$$

(Here and below, where appropriate, we use $q = (q_1, q_2)$ to denote the space variable in the internal integration.)

The series in (2.6) and the series obtained from it by differentiation once with respect to t and twice with respect to the spatial variables converge in $(L^2(\Omega))^2$ uniformly for $t \geq 0$ (e.g., [11], [18]).

Denote the projection of the sum of two rotation forces in the last line of (1.3), generated at the initial moment $t = 0$ by the unit control input $v_j = 1$, on the *divergence-free* space $J_0(\Omega)$ by

$$F_j(x) = F_{j,1}(x) + F_{j,2}(x), \quad j = 1, \dots, n-1, \tag{2.7a}$$

where

$$F_{j,1}(x) = \sum_{k=1}^{\infty} \left[\int_{\Omega} \left(\xi_j(q,0) A(z_{j+1}(0) - z_j(0))\right)^T \omega_k(q) dq \right] \omega_k(x)$$

$$= \sum_{k=1}^{\infty} \left[\left(A(z_{j+1}(0) - z_j(0))\right)^T \int_{S_r(z_j(0))} \omega_k(q) dq \right] \omega_k(x), \tag{2.7b}$$

$$F_{j,2}(x) = \sum_{k=1}^{\infty} \left[\int_{\Omega} \left(-\xi_{j+1}(q,0) A(z_{j+1}(0) - z_j(0))\right)^T \omega_k(q) dq \right] \omega_k(x)$$

$$= -\sum_{k=1}^{\infty} \left[\left(A(z_{j+1}(0) - z_j(0))\right)^T \int_{S_r(z_{j+1}(0))} \omega_k(q) dq \right] \omega_k(x). \tag{2.7c}$$

Here we used the fact that $\{\omega_k\}_{k=1}^{\infty}$ form an orthonormalized basic in $J_0(\Omega) \subset (L^2(\Omega))^2$.

ASSUMPTION 2.1. *Let the* $[2 \times 2]$*-matrix*

$$\left(\int_{S_r(z_i(0))} F_j(x) dx, \ \int_{S_r(z_i(0))} F_l(x) dx \right) \tag{2.8}$$

be nondegenerate for some $i \in \{1, \dots, n\}$ *and* $l, j \in \{1, \dots, n-1\}$.

THEOREM 2.3. *Let* $i \in \{1, \dots, n\}$, $l, j \in \{1, \dots, n-1\}$, *and Assumption* 2.1 *hold. Then there exists a* $T^* > 0$ *such that the matrix* (2.3) *is nondegenerate for any* $T \in (0, T^*]$ *and Proposition* 2.1 *holds. Namely, we have the local controllability of system* (1.1)–(1.7) *near its drifting position* $z_i^*(T)$. *In particular, for the equilibrium position* (2.2a)–(2.2b) *condition* (2.8) *implies the local controllability with respect to* z_i *near equilibrium at time* T.

The argument of Theorem 2.3 establishes that

$$\frac{dz_i(t)}{dv_j} \mid_{v'_m s=0} = \frac{t^2}{2\text{mes}\{S_r(0)\}} \int_{S_r(z_i(0))} F_j(x) dx + t^2 O(t), \ \ j = 1, \dots, n-1, \tag{2.9}$$

which allows us to apply (2.8) to ensure that (2.3) in Proposition 2.1 is nondegenerate.

Condition (2.8) holds for any point $z_i, i = 2, \dots, n-1$, in the original position of the swimming apparatus with controls acting in the adjacent links $A(z_{i+1}(0) - z_i(0))$ and $A(z_i(0) - z_{i-1}(0))$ (i.e., for $i = j, l = i-1 = j-1$), provided (a) that these links are *nonparallel* and (b) that the thick points forming it are *sufficiently small disks*. This conclusion is based on the following lemma.

LEMMA 2.4. *Let* $S_r(0)$ *be a disk of radius* r. *Then*

$$\frac{1}{\text{mes}\{S_r(0)\}} \int_{S_r(z_j(0))} F_j dx = \frac{1}{2} A(z_{j+1}(0) - z_j(0)) \ + \ g(r), \ \ j = 1, \dots, n-1, \tag{2.10a}$$

$$\frac{1}{\text{mes}\{S_r(0)\}} \int_{S_r(z_{j+1}(0))} F_j dx = -\frac{1}{2} A(z_{j+1}(0) - z_j(0)) \ + \ g(r), \ \ j = 1, \dots, n-1, \tag{2.10b}$$

where $\| g(r) \|_{R^2} \leq Cr$ as $r \to 0+$ *for some positive constant* C.

Due to (2.9), at no extra cost, Theorem 2.3 implies the respective statement for the center of mass $z_c(t)$.

THEOREM 2.5. *Let $l, j \in \{1, \dots, n-1\}$, and the matrix*

$$\sum_{i=1}^{n} \left(\int_{S_r(z_i(0))} F_j(x)dx, \ \int_{S_r(z_i(0))} F_l(x)dx \right)$$

is nondegenerate. Then there exists a $T^* > 0$ such that for any $T \in (0, T^*]$ Proposition 2.2 *holds. Namely, we have the local controllability of system* (1.1)–(1.7) *with respect to the position of center of mass $z_c(T)$ near its drifting position $z_c^*(T)$. In particular, for the equilibrium position* (2.2a)–(2.2b) *we have the local controllability with respect to z_c near equilibrium at time T.*

Discussion of Theorems 2.3 and 2.5. Note that the two columns in (2.8) multiplied by $\text{mes}^{-1}\{S_r(0)\}$ describe the "average" forces induced respectively by the forces $F_j(x)$ and $F_l(x)$ over the region $S_r(z_i(0))$. Thus, the sufficient conditions for the local controllability near the drifting position $z_i^*(T)$ in Theorem 2.3 require that these average forces are not colinear. Respectively, for the local controllability of the center of mass z_c Theorem 2.5 requires that the sums of such average forces over all "thick points" in the skeleton of the apparatus generated respectively by the unit controls $v_l = 1$ and $v_j = 1$ are not colinear as well.

Remark 2.1. Supports of $F_j(x)$'s . In spite of the fact that the rotation forces in (1.3) have only local supports this does not have to be so for their projections (as in (2.7a)–(2.7b)) on the solenoidal part $J_0(\Omega)$ of $(L^2(\Omega))^2$, associated with incompressible fluids (see section 6 for more details).

Recall now that the space $(L^2(\Omega))^2$ is the direct sum of the spaces $J_0(\Omega)$ and $G(\Omega)$. In (2.7a)–(2.7c) we denoted the projections of the functions

$$\xi_j(x,0)\,(A(z_{j+1}(0) - z_j(0))) \quad \text{and} \quad -\xi_{j+1}(x,0)\,(A(z_{j+1}(0) - z_j(0)))\,, \quad j = 1, \dots, n-1$$

on $J_0(\Omega)$ by $F_{j,1}(x)$ and $F_{j,2}(x)$. Denote now the projections of the aforementioned functions on the space $G(\Omega)$, respectively, by $F_{j,1}^{\perp}(x)$ and $F_{j,2}^{\perp}(x)$. Since (e.g., [11, p. 28]; [18, p. 15])

$$J_0(\Omega) = \{u \in (L^2(\Omega))^2, \ \text{div}\, u = 0, \ \gamma_\nu u\,|_{\partial\Omega} = 0\}, \tag{2.11}$$

$$G(\Omega) = \{u \in (L^2(\Omega))^2, \ u = \nabla p, \ p \in H^1(\Omega)\}, \tag{2.12}$$

where ν is the unit vector normal to the boundary $\partial\Omega$ (pointing outward) and $\gamma_\nu u\,|_{\partial\Omega}$ is the restriction of $u \cdot \nu$ to $\partial\Omega$, we have

$$F_{j,1}^{\perp}(x) = \nabla w_{j,1}(x), \quad F_{j,2}^{\perp}(x) = \nabla w_{j,2}(x), \quad j = 1, \dots, n-1,$$

for some functions $w_{j,1}, w_{j,2} \in H^1(\Omega), j = 1, \dots, n-1$. Thus,

$$F_{j,1}(x) = \xi_j(x,0)\,(A(z_{j+1}(0) - z_j(0))) - \nabla w_{j,1}(x), \tag{2.13}$$

$$F_{j,2}(x) = -\xi_{j+1}(x,0)\,(A(z_{j+1}(0) - z_j(0))) - \nabla w_{j,2}(x). \tag{2.14}$$

Furthermore, $w_{j,1}$ and $w_{j,2}$ solve the following two generalized Neumann problems:

$$\Delta w_{j,1} = \ \text{div}\ \xi_j(x,0)\,(A(z_{j+1}(0) - z_j(0))) \quad \text{in } \Omega, \quad \frac{\partial w_{j,1}}{\partial \nu}\,|_{\partial\Omega} = 0, \tag{2.15}$$

$$\Delta w_{j,2} = \ -\text{div}\ \xi_{j+1}(x,0)\,(A(z_{j+1}(0) - z_j(0))) \quad \text{in } \Omega, \quad \frac{\partial w_{j,2}}{\partial \nu}\,|_{\partial\Omega} = 0. \tag{2.16}$$

Indeed, (2.15), e.g., can be obtained by applying divergence to (2.13) and recalling that $F_{j,1} \in J_0(\Omega)$, which in particular implies that $\mathrm{div} F_{j,1} = 0$. In turn, the boundary condition in (2.15) follows from (2.13) by recalling that, due to (2.11), $\gamma_\nu F_{j,1} \mid_{\partial\Omega} = 0$ and that $\xi_j(x,0)$ vanishes outside of $S_r(z_j(0))$, which provides $\gamma_\nu \nabla w_{j,1} \mid_{\partial\Omega} = \frac{\partial w_{j,1}}{\partial \nu} \mid_{\partial\Omega} = 0$.

Our qualitative analysis of solutions to (2.15) and (2.16) in section 6 led us to Lemma 2.4. In this section we also (see Remark 6.1) derived certain qualitative estimates and formulas which can be used to verify sufficient conditions for controllability in Theorems 2.3 and 2.5 (i.e., not only for the circular support as in Lemma 2.4).

3. Preliminary results. Our plan to prove Theorem 2.3 is as

1. We intend to use Propositions 2.1 and 2.2 involving derivatives $\frac{dz_i(T)}{dv_j}, i = 1, \dots, n, j = 1, \dots, n-1$. In order to evaluate them, in section 3 we differentiate the implicit solution formula (2.6) with respect to v_j's.
2. In section 4 the results of section 3 are presented as a vector Volterra equation for the aforementioned $\frac{dz_i(T)}{dv_j}$'s and suitable asymptotic analysis is used to qualitatively evaluate them for "small" T's.
3. Making use of all of the above to obtain the qualitative estimates for the terms in (2.8), we complete the proof of Theorem 2.3 in section 5.

In this section we intend to derive a number of auxiliary formulas. Everywhere below, for simplicity of calculations we assume that $S_r(0)$ has the form of (1.4b).

3.1. Solution formula. Let us rewrite (2.6) as

$$
\begin{aligned}
y(x,t) = &\sum_{k=1}^{\infty} e^{-\lambda_k t} \left(\int_\Omega y_0^T \omega_k dx \right) \omega_k(x) \\
&+ \sum_{i=2}^{n+1} \left(P_i(x,t) + v_{i-1}(t) Q_i(x,t) + v_{i-2}(t) R_i(x,t) \right),
\end{aligned} \tag{3.1}
$$

where *here and below we always assume* that (besides all other imposed assumptions, if there are such) T is selected "sufficiently small" as in Theorem 1.1 to ensure the wellposednesss of system (1.1)–(1.7) at hand on $[0,T]$, and

(3.2)

$$
\begin{aligned}
P_i(x,t) = \sum_{k=1}^{\infty} \Bigg(\int_0^t e^{-\lambda_k (t-\tau)} \Bigg[& k_{i-1} \frac{(\| z_i(\tau) - z_{i-1}(\tau) \|_{R^2} - l_{i-1})}{\| z_i(\tau) - z_{i-1}(\tau) \|_{R^2}} (z_i(\tau) - z_{i-1}(\tau))^T \\
& + k_{i-2} \frac{(\| z_{i-2}(\tau) - z_{i-1}(\tau) \|_{R^2} - l_{i-2})}{\| z_{i-2}(\tau) - z_{i-1}(\tau) \|_{R^2}} (z_{i-2}(\tau) - z_{i-1}(\tau))^T \Bigg] \\
& \times \left(\int_\Omega \xi_{i-1}(q,\tau) \omega_k dq \right) d\tau \Bigg) \omega_k(x),
\end{aligned}
$$

(3.3)

$$
Q_i(x,t) = \sum_{k=1}^{\infty} \left(\int_0^t e^{-\lambda_k (t-\tau)} \left(A(z_i(\tau) - z_{i-1}(\tau)) \right)^T \left(\int_\Omega \xi_{i-1}(q,\tau) \omega_k dq \right) d\tau \right) \omega_k(x),
$$

(3.4)

$$
R_i(x,t) = \sum_{k=1}^{\infty} \left(\int_0^t e^{\lambda_k (t-\tau)} \left(A(z_{i-2}(\tau) - z_{i-1}(\tau)) \right)^T \left(\int_\Omega \xi_{i-1}(q,\tau) \omega_k dq \right) d\tau \right) \omega_k(x).
$$

3.2. Differentiation with respect to v_j's. We assume from now on that the functions $v_j, j = 1, \dots, n-1$, are constant in time. Below we formally differentiate various expressions with respect to $v_j, j = 1, \dots, n-1$. The validity of these calculations will be discussed in the next section (see the subsection 4.2).

3.2.1. Derivatives of z_i's. In view of (1.2) and (1.4a)–(1.4b), for any $t \in [0, T]$,

$$\begin{aligned} z_i(t) &= z_{i0} + \frac{1}{\operatorname{mes}\{S_r(0)\}} \int_0^t \int_{S_r(z_i(\tau))} y(x,\tau) dx d\tau \\ &= z_{i0} + \frac{1}{\operatorname{mes}\{S_r(0)\}} \int_0^t \int_{z_{i,1}(\tau)-r}^{z_{i,1}(\tau)+r} \int_{z_{i,2}(\tau)+\alpha(x_1-z_{i,1}(\tau))}^{z_{i,2}(\tau)+\beta(x_1-z_{i,1}(\tau))} y(x,\tau) dx_2 dx_1 d\tau. \end{aligned} \tag{3.5}$$

Denote $z_i(t) = (z_{i,1}(t), z_{i,2}(t))$. Then differentiating (3.5) with respect to v_j, we obtain

$$\begin{aligned} \frac{dz_i(t)}{dv_j} &= \frac{1}{\operatorname{mes}\{S_r(0)\}} \int_0^t \int_{z_{i,1}(\tau)-r}^{z_{i,1}(\tau)+r} \\ &\times \left(y(x_1, z_{i,2}(\tau) + \beta(x_1 - z_{i,1}(\tau)), \tau) \left(\frac{dz_{i,2}(\tau)}{dv_j} - \beta'(x_1 - z_{i,1}(\tau)) \frac{dz_{i,1}(\tau)}{dv_j} \right) \right. \\ &\left. - y(x_1, z_{i,2}(\tau) + \alpha(x_1 - z_{i,1}(\tau)), \tau) \left(\frac{dz_{i,2}(\tau)}{dv_j} - \alpha'(x_1 - z_{i,1}(\tau)) \frac{dz_{i,1}(\tau)}{dv_j} \right) \right) dx_1 d\tau \\ &+ \frac{1}{\operatorname{mes}\{S_r(0)\}} \int_0^t \frac{dz_{i,1}(\tau)}{dv_j} \int_{z_{i,2}(\tau)+\alpha(r)}^{z_{i,2}(\tau)+\beta(r)} y(z_{i,1}(\tau) + r, x_2, \tau) dx_2 d\tau \\ &- \frac{1}{\operatorname{mes}\{S_r(0)\}} \int_0^t \frac{dz_{i,1}(\tau)}{dv_j} \int_{z_{i,2}(\tau)+\alpha(-r)}^{z_{i,2}(\tau)+\beta(-r)} y(z_{i,1}(\tau) - r, x_2, \tau) dx_2 d\tau \\ &+ \frac{1}{\operatorname{mes}\{S_r(0)\}} \int_0^t \int_{S_r(z_i(\tau))} \frac{dy(x,\tau)}{dv_j} dx d\tau. \end{aligned} \tag{3.6}$$

3.2.2. Derivatives of y. Making use of (3.1), we obtain

$$\begin{aligned} \frac{dy}{dv_j} |_{v'_m s=0} &= \frac{d}{dv_j} \left(\sum_{i=2}^{n+1} (P_i + v_{i-1} Q_i + v_{i-2} R_i) \right) |_{v'_m s=0} \\ &= Q_{j+1} |_{v'_m s=0} + R_{j+2} |_{v'_m s=0} + \sum_{i=2}^{n+1} \left(\frac{d}{dv_j} P_i + v_{i-1} \frac{d}{dv_j} Q_i + v_{i-2} \frac{d}{dv_j} R_i \right) |_{v'_m s=0} \\ &= Q_{j+1} |_{v'_m s=0} + R_{j+2} |_{v'_m s=0} + \sum_{i=2}^{n+1} \left(\frac{d}{dv_j} P_i \right) |_{v'_m s=0} . \end{aligned} \tag{3.7}$$

3.2.3. Derivatives of P_i's. Making use of (3.2), while noticing that

$$\int_\Omega \xi_{i-1}(x,\tau) \omega_k dx = \int_{z_{i-1,1}(\tau)-r}^{z_{i-1,1}(\tau)+r} \int_{z_{i-1,2}(\tau)+\alpha(x_1-z_{i-1,1}(\tau))}^{z_{i-1,2}(\tau)+\beta(x_1-z_{i-1,1}(\tau))} \omega_k(x) dx_2 dx_1,$$

we obtain

$$\frac{d}{dv_j}P_i(x,t)\mid_{v'_m s=0}$$

$$=\sum_{k=1}^{\infty}\left[\int_0^t e^{-\lambda_k(t-\tau)}\left\{k_{i-1}\Theta_{1,i,j}(\tau)+k_{i-2}\Theta_{2,i,j}(\tau)\right\}\left(\int_\Omega \xi_{i-1}(q,\tau)\omega_k dq\right)d\tau\right]\omega_k(x)\mid_{v'_m s=0}$$

$$+\sum_{k=1}^{\infty}\left(\int_0^t e^{-\lambda_k(t-\tau)}\Theta_{3,i}(\tau)\;\Theta_{4,i,j}(\tau)d\tau\right)\omega_k(x)\mid_{v'_m s=0},$$

(3.8)

where

$$\Theta_{1,i,j}(\tau)=\left(\frac{dz_i(\tau)}{dv_j}-\frac{dz_{i-1}(\tau)}{dv_j}\right)^T\left(1\;-\;\frac{l_{i-1}}{\parallel z_i(\tau)-z_{i-1}(\tau)\parallel_{R^2}}\right)$$

$$+\,l_{i-1}\left(z_i(\tau)-z_{i-1}(\tau)\right)^T\frac{1}{\parallel z_i(\tau)-z_{i-1}(\tau)\parallel^3_{R^2}}$$

$$<\frac{dz_i(\tau)}{dv_j}\;-\;\frac{dz_{i-1}(\tau)}{dv_j},z_i(\tau)-z_{i-1}(\tau)>_{R^2},$$

$$\Theta_{2,i,j}(\tau)=\left(\frac{dz_{i-2}(\tau)}{dv_j}-\frac{dz_{i-1}(\tau)}{dv_j}\right)^T\left(1-\frac{l_{i-2}}{\parallel z_{i-2}(\tau)-z_{i-1}(\tau)\parallel_{R^2}}\right)$$

$$+\,l_{i-2}(z_{i-2}(\tau)-z_{i-1}(\tau))^T\frac{1}{\parallel z_{i-2}(\tau)-z_{i-1}(\tau)\parallel^3_{R^2}}$$

$$<\frac{dz_{i-2}(\tau)}{dv_j}\;-\;\frac{dz_{i-1}(\tau)}{dv_j},z_{i-2}(\tau)-z_{i-1}(\tau)>_{R^2},$$

$$\Theta_{3,i}(\tau)=k_{i-1}\left(\frac{(\parallel z_i(\tau)-z_{i-1}(\tau)\parallel_{R^2}\,-l_{i-1})}{\parallel z_i(\tau)-z_{i-1}(\tau)\parallel_{R^2}}(z_i(\tau)-z_{i-1}(\tau))^T\right)$$

$$+\,k_{i-2}\left(\frac{(\parallel z_{i-2}(\tau)-z_{i-1}(\tau)\parallel_{R^2}\,-l_{i-2})}{\parallel z_{i-2}(\tau)-z_{i-1}(\tau)\parallel_{R^2}}(z_{i-2}(\tau)-z_{i-1}(\tau))^T\right),$$

$$\Theta_{4,i,j}(\tau)=\int_{z_{i-1,1}(\tau)-r}^{z_{i-1,1}(\tau)+r}\omega_k\left(x_1,z_{i-1,2}(\tau)+\beta(x_1-z_{i-1,1}(\tau))\right)$$

$$\times\left(\frac{dz_{i-1,2}(\tau)}{dv_j}-\beta'(x_1-z_{i-1,1}(\tau))\frac{dz_{i-1,1}(\tau)}{dv_j}\right)dx_1$$

$$-\int_{z_{i-1,1}(\tau)-r}^{z_{i-1,1}(\tau)+r}\omega_k\left(x_1,z_{i-1,2}(\tau)+\alpha(x_1-z_{i-1,1}(\tau))\right)$$

$$\times\left(\frac{dz_{i-1,2}(\tau)}{dv_j}-\alpha'(x_1-z_{i-1,1}(\tau))\frac{dz_{i-1,1}(\tau)}{dv_j}\right)dx_1$$

$$+\,\frac{dz_{i-1,1}(\tau)}{dv_j}\int_{z_{i-1,2}(\tau)+\alpha(r)}^{z_{i-1,2}(\tau)+\beta(r)}\omega_k(z_{i-1,1}(\tau)+r,x_2)dx_2$$

$$-\,\frac{dz_{i-1,1}(\tau)}{dv_j}\int_{z_{i-1,2}(\tau)+\alpha(-r)}^{z_{i-1,2}(\tau)+\beta(-r)}\omega_k(z_{i-1,1}(\tau)-r,x_2)dx_2.$$

To better understand the terms in the second sum in (3.8), denote

$$V_i(\tau)=\Theta_{3,i}(\tau)\mid_{v'_m s=0}.$$

Then we can rewrite, e.g., the term in the second sum in (3.8) associated with the factor $\frac{dz_{i-1,2}(\tau)}{dv_j}$ in the first line in the expression for $\Theta_{4,i,j}(\tau)$ as

$$\int_0^t \frac{dz_{i-1,2}(\tau)}{dv_j}\left[\sum_{k=1}^{\infty} e^{-\lambda_k(t-\tau)}\{(\Phi_1(\tau))(V_i(\tau)\omega_k)\}\right] d\tau\omega_k(x) \mid_{v'_m s=0}, \tag{3.9}$$

where we denoted

$$(\Phi_1(t))(\psi) = \int_{z_{i-1,1}(t)-r}^{z_{i-1,1}(t)+r} \psi(x_1, z_{i-1,2}(t) + \beta(x_1 - z_{i-1,1}(t)))dx_1.$$

Note that $\Phi_1(t) \in H^{-1}(\Omega)$ for any $t \in [0,T]$, where the space $H^{-1}(\Omega)$ is dual of $H_0^1(\Omega)$; namely, it is the space of all linear bounded functionals on $H_0^1(\Omega)$. (As usual, we endow the latter space with the norm $\|\psi\|_{H_0^1(\Omega)} = \{\int_\Omega(\psi_{x_1}^2 + \psi_{x_2}^2)dx\}^{1/2}$.)

Indeed, regardless of t, for any $\psi \in H_0^1(\Omega)$, due to the continuous embedding of $H_0^1(a,b)$ into $C[a,b]$ for any finite interval $[a,b]$, for any $t \in [0,T]$ we have

$$\mid (\Phi_1(t))(\psi) \mid \leq \int_{z_{i-1,1}(t)-r}^{z_{i-1,1}(t)+r} \mid \psi(x_1, z_{i-1,2}(t) + \beta(x_1 - z_{i-1,1}(t))) \mid dx_1$$

$$\leq C_1 \int_{z_{i-1,1}(t)-r}^{z_{i-1,1}(t)+r} \| \psi(x_1,\cdot) \|_{H_0^1(\{\xi|(x_1,\xi)\in\Omega\}} dx_1 \leq \sqrt{2r}C_1 \| \psi \|_{H_0^1(\Omega)},$$

where C_1 is a positive constant. Thus, for any $t \in [0,T]$,

$$\| \Phi_1(t) \|_{H^{-1}(\Omega)} = \sup_{\|\psi\|_{H_0^1(\Omega)}=1, \psi\in H_0^1(\Omega)} \mid (\Phi_1(t))(\psi) \mid \leq \sqrt{2r}C_1. \tag{3.10}$$

In the next subsection we will need the following observation.

Remark 3.1. Consider any vector $\kappa \in R^2$. Then, similar to the derivation of (3.10), we can show that for any $t \in [0,T]$ the expression

$$(\Phi_1(t)\kappa)(\phi) = (\Phi_1(t))(\kappa^T\phi), \quad \text{where } \phi \in (H_0^1(\Omega))^2,$$

defines a linear bounded functional on $(H_0^1(\Omega))^2$ and

$$\sup_{\|\phi\|_{(H_0^1(\Omega))^2}=1, \phi\in(H_0^1(\Omega))^2} \mid (\Phi_1(t)\kappa)(\phi) \mid \leq \| \kappa \|_{R^2} \| \Phi_1(t) \|_{H^{-1}(\Omega)} \leq \sqrt{2r}C_1 \| \kappa \|_{R^2}.$$

On the other hand, if $\phi \in H(\Omega) \subset (H_0^1(\Omega))^2$, then it admits the following representation:

$$\phi = \sum_{k=1}^{\infty} a_k\omega_k, \quad \| \phi \|_{H(\Omega)} = \| \phi \|_{(H_0^1(\Omega))^2} = \left(\sum_{k=1}^{\infty} \frac{\lambda_k}{\nu} a_k^2\right)^{1/2}.$$

In this case, we can introduce the space $H'(\Omega)$ of linear bounded functionals $\mathbf{\Phi}$ on $H(\Omega)$ making use of the duality product

$$\mathbf{\Phi}(\phi) = \sum_{k=1}^{\infty} a_k b_k, \quad \text{where } b_k = \mathbf{\Phi}(\omega_k), \quad k = 1, \ldots,$$

with the norm

$$\| \mathbf{\Phi} \|_{H'(\Omega)} = \left(\sum_{k=1}^{\infty} \frac{\nu}{\lambda_k} b_k^2 \right)^{1/2},$$

equivalent to the regular one on this space (i.e., analogous to that in (3.10)). Thus, in particular, for any $\kappa \in R^2$ we have for any $t \in [0,T]$

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{\nu}{\lambda_k} \{(\Phi_1(t)\kappa)(\omega_k)\}^2 \;&\le\; C_* \left(\sup_{\|\phi\|_{H(\Omega)}=1, \phi \in H(\Omega)} | \, (\Phi_1(t)\kappa)(\phi) \, | \right)^2 \\ \le\; C_* \left(\sup_{\|\phi\|_{(H_0^1(\Omega))^2}=1, \phi \in (H_0^1(\Omega))^2} | \, (\Phi_1(t)\kappa)(\phi) \, | \right)^2 \;&\le\; C_* \left(\sqrt{2r} C_1 \, \| \, \kappa \, \|_{R^2} \right)^2, \end{aligned} \tag{3.11}$$

where $C_* > 0$ is some (generic) constant.

3.3. Kernels. Let us now consider in detail the contribution of (3.9) to the expression in (3.6), multiplied by $\operatorname{mes}\{S_r(0)\}$, when $v_m' s = 0$, namely,

$$\begin{aligned} &\int_0^t \int_0^\tau \frac{dz_{i-1,2}(s)}{dv_j} \int_{S_r(z_i(\tau))} \left[\sum_{k=1}^{\infty} e^{-\lambda_k(\tau - s)} \{(\Phi_1(s) V_i(s))(\omega_k)\} \omega_k(x) \right] dx ds d\tau \\ &= \int_0^t \frac{dz_{i-1,2}(s)}{dv_j} \left\{ \int_s^t \int_{S_r(z_i(\tau))} \left[\sum_{k=1}^{\infty} e^{-\lambda_k(\tau - s)} \{(\Phi_1(s) V_i(s))(\omega_k)\} \omega_k(x) \right] dx d\tau \right\} ds. \end{aligned}$$

In the above, the factor at $\frac{dz_{i-1,2}(s)}{dv_j}$ can be regarded as a 2D kernel $K(t,s)$, $(t,s) \in (0,T) \times (0,T)$, vanishing for $s > t$.

LEMMA 3.1. *The kernel K is an element of $(L^\infty((0,T) \times (0,T)))^2$.*

Proof of Lemma 3.1. Indeed, for almost all $(s,\tau) \in (0,T) \times (0,T)$,

$$\begin{aligned} \| \, K(t,s) \, \|_{R^2}^2 &= \| \int_s^t \int_{S_r(z_i(\tau))} \left[\sum_{k=1}^{\infty} e^{-\lambda_k(\tau - s)} \{(\Phi_1(s) V_i(s))(\omega_k)\} \omega_k(x) \right] dx d\tau \, \|_{R^2}^2 \\ &\le\; T \operatorname{mes}\{S_r(0)\} \sum_{k=1}^{\infty} \left(\int_s^t e^{-2\lambda_k(\tau - s)} d\tau \right) \{(\Phi_1(s) V_i(s))(\omega_k)\}^2 \\ &\le\; CT \sum_{k=1}^{\infty} \frac{1}{\lambda_k} \{(\Phi_1(s) V_i(s))(\omega_k)\}^2 \;\le\; \hat{C} T \left(\sqrt{2r} C_1 \, \| \, V_i \, \|_{C([0,T];R^2)} \right)^2 \end{aligned} \tag{3.12}$$

for some (generic) positive constants C and $\hat{C}$, while C_1 is from (3.11) in Remark 3.1. In (3.12) we also used the following type of estimate, employing Bessel's inequality:

$$\begin{aligned} \left\| \int_\omega \left(\sum_{k=1}^{\infty} a_k \omega_k(x) \right) dx \, \right\|_{R^2}^2 &= \| \int_\Omega \left(\sum_{k=1}^{\infty} a_k \omega_k(x) \right) \xi_\omega(x) dx \, \|_{R^2}^2 \\ \le\; \left\| \sum_{k=1}^{\infty} a_k \omega_k \right\|_{(L^2(\Omega))^2}^2 &\| \, \xi_\omega \, \|_{L^2(\Omega)}^2 = \left(\sum_{k=1}^{\infty} a_k^2 \right) \operatorname{mes}\{\omega\}, \end{aligned}$$

where $\xi_\omega(x)$ is the characteristic function of a set $\omega \subset \Omega$. This ends the proof of Lemma 3.1. □

The assertion of Lemma 3.1 can be established for all other kernels associated with $\frac{dz_i(s)}{dv_j}$'s in (3.8) and (3.6).

4. Volterra equations for $\frac{dz_i(\tau)}{dv_j}$'s. Below we will deal only with the terms $\frac{dz_i(t)}{dv_j}\,|_{v'_m s=0}$. Therefore, to simplify further notation, we will *omit the subscript* $|_{v'_m s=0}$ *from now on.*

4.1. Volterra equations. Equation (3.6) can be rewritten as the following vector Volterra equation:

$$\begin{pmatrix} \frac{dz_1(t)}{dv_j} \\ \vdots \\ \frac{dz_n(t)}{dv_j} \end{pmatrix} + \int_0^t \mathbf{B}_j(t,s) \begin{pmatrix} \frac{dz_1(s)}{dv_j} \\ \vdots \\ \frac{dz_n(s)}{dv_j} \end{pmatrix} ds = \begin{pmatrix} \Xi_{1j}(t) \\ \vdots \\ \Xi_{nj}(t) \end{pmatrix}, \quad j = 1, \dots, n-1, \quad t \in [0,T], \tag{4.1}$$

where

$$\Xi_{ij}(t) = \frac{1}{\operatorname{mes}\{S_r(0)\}} \int_0^t \int_{S_r(z_i(\tau))} (Q_{j+1} + R_{j+2})\, dx d\tau \,|_{v'_m s=0}$$

and $\mathbf{B}_j(t,s)$, $j = 1, \dots, n-1$, are, respectively, the vector and matrix functions defined by (3.6) along (3.7)–(3.12). We will need the following asymptotic result.

LEMMA 4.1.

$$\Xi_{ij}(t) = \frac{t^2}{2\operatorname{mes}\{S_r(0)\}} \left[\sum_{k=1}^{\infty} \left(\int_{S_r(z_i(0))} \omega_k dx\right)\left(\int_{S_r(z_j(0))} \omega_k^T(x) dx\right)\right] A(z_{j+1}(0) - z_j(0))$$

$$+ \frac{t^2}{2\operatorname{mes}\{S_r(0)\}} \left[\sum_{k=1}^{\infty} \left(\int_{S_r(z_i(0))} \omega_k dx\right)\left(\int_{S_r(z_{j+1}(0))} \omega_k^T(x) dx\right)\right] A(z_j(0) - z_{j+1}(0)) + t^2 O(t) \tag{4.2}$$

as $t \to 0+$, where $i = 1, \dots, n$, $j = 1, \dots, n-1$, and $O(t)$ stands for a vector function whose R^2-norm tends to zero as $t \to 0+$.

Proof of Lemma 4.1. *Step* 1. In view of (3.3) and (3.4), we have

$$\operatorname{mes}\{S_r(0)\}\Xi_{ij}(t)$$

$$= \int_0^t \int_0^\tau \int_\Omega \left[\sum_{k=1}^{\infty} e^{-\lambda_k(\tau - s)} \left(\int_\Omega (A(z_{j+1}(s) - z_j(s)))^T \xi_j(q,s)\omega_k dq\right)\omega_k(x)\right] \xi_i(x,\tau) dx ds d\tau$$

$$+ \int_0^t \int_0^\tau \int_\Omega \left[\sum_{k=1}^{\infty} e^{\lambda_k(\tau - s)} \left(\int_\Omega (A(z_j(s) - z_{j+1}(s)))^T \xi_{j+1}(q,s)\omega_k dq\right)\omega_k(x)\right] \xi_i(x,\tau) dx ds d\tau. \tag{4.3}$$

Step 2. We will further deal with the first term in (4.3). Let us show that the following representation holds for the expression in the square brackets in this term:

$$\mathbf{W}(s,\tau) = \left\| \sum_{k=1}^{\infty} e^{-\lambda_k(\tau - s)} \left(\int_\Omega (A(z_{j+1}(s) - z_j(s)))^T \xi_j(q,s)\omega_k dq\right)\omega_k \right.$$

$$\left. - \sum_{k=1}^{\infty} \left(\int_\Omega (A(z_{j+1}(0) - z_j(0)))^T \xi_j(q,0)\omega_k dq\right)\omega_k \right\|^2_{(L^2(\Omega))^2} \le O(t) \tag{4.4}$$

as $t \to 0+$ uniformly over $0 \le s \le \tau \le t$.

Indeed,

$$\begin{aligned}\mathbf{W}(s,\tau) &\le 2\sum_{k=1}^{\infty} e^{-2\lambda_k(\tau-s)} \\ &\times \left[\int_\Omega \left((A(z_{j+1}(s)-z_j(s)))^T\xi_j(q,s) - (A(z_{j+1}(0)-z_j(0)))^T\xi_j(q,0)\right)\omega_k dq\right]^2 \\ &+ 2\sum_{k=1}^{N}(e^{-\lambda_k(\tau-s)}-1)^2\left(\int_\Omega (A(z_{j+1}(0)-z_j(0)))^T\,\xi_j(q,0)\omega_k dq\right)^2 \\ &+ 2\sum_{k=N+1}^{\infty}(e^{-2\lambda_k(\tau-s)}-1)^2\left(\int_\Omega (A(z_{j+1}(0)-z_j(0)))^T\,\xi_j(q,0)\omega_k dq\right)^2.\end{aligned} \tag{4.5}$$

Step 3. Due to Bessel's inequality,

$$\begin{aligned}&\left\|\sum_{k=1}^{\infty}\left(\int_\Omega (A(z_{j+1}(0)-z_j(0)))^T\,\xi_j(q,0)\omega_k dq\right)\omega_k\right\|^2_{(L^2(\Omega))^2} \\ &= \sum_{k=1}^{\infty}\left(\int_\Omega (A(z_{j+1}(0)-z_j(0)))^T\,\xi_j(q,0)\omega_k dq\right)^2 \le \|\,\xi_j(\cdot,0)\,(A(z_{j+1}(0)-z_j(0)))\,\|^2_{(L^2(\Omega))^2}.\end{aligned}$$

Therefore for every $\varepsilon>0$ there is an $N=N(\varepsilon)$ such that the last sum on the right in (4.5) can be made smaller than $\varepsilon/3$ regardless of $0\le s\le\tau\le t$.

In turn, for this $N(\varepsilon)$, determined by ε, there is a $t_*=t_*(\varepsilon)>0$ such that, by continuity of the exponential function, the second sum on the right in (4.5) can be made smaller than $\varepsilon/3$ as well for any $0\le s\le\tau\le t\le t_*$.

Step 4. Now recall that continuity of solutions in time in Theorem 1.1 yields that

$$\|\,A(z_{j+1}(s)-z_j(s)) - A(z_{j+1}(0)-z_j(0))\,\|_{R^2} = O(s)$$

as $s\to 0+$. In turn, (1.7) implies that for any $k=1,\dots,n$

$$\begin{aligned}&\|\,\xi_k(\cdot,s)-\xi_k(\cdot,0)\,\|^2_{L^2(\Omega)} \\ &= \int_{(S_r(z_k(s)))\bigcup(S_r(z_k(0)))\setminus(S_r(z_k(s))\bigcap S_r(z_k(0)))} dx = O(\|\,z_k(s)-z_k(0)\,\|_{R^2}) = O(s)\end{aligned} \tag{4.6}$$

as $s\to 0+$. Therefore, for any selected-above ε there is a $t_{**}=t_{**}(\varepsilon)>0$ such that the first term on the right in (4.5) can be made smaller than $\varepsilon/3$ for any $0\le s\le\tau\le t\le t_{**}$. Indeed,

$$\begin{aligned}&\sum_{k=1}^{\infty} e^{-2\lambda_k(\tau-s)}\left[\int_\Omega \left((A(z_{j+1}(s)-z_j(s)))^T\xi_j(q,s) - (A(z_{j+1}(0)-z_j(0)))^T\xi_j(q,0)\right)\omega_k dq\right]^2 \\ &\le \sum_{k=1}^{\infty}\left[\int_\Omega \left((A(z_{j+1}(s)-z_j(s)))^T\xi_j(q,s) - (A(z_{j+1}(0)-z_j(0)))^T\xi_j(q,0)\right)\omega_k dq\right]^2 \\ &\le \|\,\xi_j(\cdot,s)(A(z_{j+1}(s)-z_j(s))) - \xi_j(\cdot,0)(A(z_{j+1}(0)-z_j(0)))\,\|_{(L^2(\Omega))^2} = O(s)\end{aligned}$$

as $s\to 0+$ uniformly over $0\le s\le\tau\le t$.

Combining all of the above, we obtain that $\mathbf{W}(s,\tau)$ in (4.4), (4.5) can be made smaller than $\varepsilon > 0$ for any $t \in [0, \min\{t_*, t_{**}\}]$, which yields (4.4).

Step 5. Applying (4.6) to $\xi_i(x,\tau)$ in the first term in (4.3) and making use of (4.4) yields the assertion of Lemma 4.1 for the first term in (4.2). The proof for the second term is similar. This ends the proof of Lemma 4.1. □

4.2. Auxiliary estimates. It is well known (e.g., [10]) that (as a special form of the Fredholm equation) (4.1) admits a unique solution in $(L^2(0,T))^{2n}$. This will allow us to prove, making use of the classical methods, that dz_i/dv_j's indeed exist in $(L^2(0,T))^2$, and all the above calculations leading to (4.1) are valid. (Namely, one needs, based on (3.5), to write the Volterra equations for the expressions $\Delta z_i/\Delta v_j$'s and then pass to the limit as Δv_j tend to zero to obtain (4.1).)

Furthermore, it is easy to see that there exists a ("small") $T_1 > 0$ such that for any $T \in (0, T_1]$

$$\left\| \left(\frac{dz_1}{dv_j}, \dots, \frac{dz_n}{dv_j} \right) \right\|_{(L^2(0,T))^{2n}} \le C \parallel (\Xi_{1j}, \dots, \Xi_{nj}) \parallel_{(L^2(0,T))^{2n}}, \tag{4.7}$$

where $C > 0$ is a (generic) positive constant independent of $T \in [0, T_1]$.

Moreover, since in (4.1) the integral terms, with $(L^2(0,T))^2$-derivatives dz_i/dv_j's in them and the right-hand sides, are actually continuous functions (which can be shown, in particular, making use of Lemma 3.1), we have

$$\left(\frac{dz_1}{dv_j}, \dots, \frac{dz_n}{dv_j} \right) \in C([0,T]; R^{2n}),$$

and, similar to (4.7), there exists a ("small") $T_2 \in [0, T_1]$ such that for any $T \in (0, T_2]$

$$\left\| \left(\frac{dz_1}{dv_j}, \dots, \frac{dz_n}{dv_j} \right) \right\|_{C([0,T];R^{2n})} \le C \parallel (\Xi_1, \dots, \Xi_n) \parallel_{C([0,T];R^{2n})}, \tag{4.8}$$

where, to simplify notation, we again used the generic notation C for the constant.

Applying (4.2) and (4.8) to (4.1), we obtain that

$$\left\| \left(\frac{dz_1}{dv_j}, \dots, \frac{dz_n}{dv_j} \right) \right\|_{C([0,T];R^{2n})} \le Lt^2 \tag{4.9}$$

for some constant $L > 0$ as $t \to 0+$.

5. Proof of Theorem 2.3: The inverse function theorem. From (4.1), making use of (4.2), (4.9), and (2.7a)–(2.7b), we derive that

$$\frac{dz_i(t)}{dv_j} = \frac{t^2}{2\mathrm{mes}\,\{S_r(0)\}} \int_{S_r(z_i(0))} F_j(x)dx + t^2 O(t) \tag{5.1}$$

for $j = 1, \dots, n-1$, $t \in [0,T]$.

Thus, the matrix

$$\left(\frac{dz_i(t)}{dv_j}, \frac{dz_i(t)}{dv_l} \right)$$

is not degenerate under Assumption 2.1 of Theorem 2.3 for sufficiently small t. We can now select any such "small" number t as a $T > 0$ in Proposition 2.1 and obtain the statement of Theorem 2.3 from this proposition. This ends the proof of Theorem 2.3. □

6. Proof of Lemma 2.4 and supports of F_j's. Let us show first (see Remark 2.1) that the projections $F_{j,1}(x)$ and $F_{j,2}(x)$ in (2.7a)–(2.7c) of the respective rotation forces in (1.3) with supports in $S_r(z_j(0))$ and $S_r(z_{j+1}(0))$ can have support beyond these two sets in Ω.

For example, let $F_{j,1}(x)$ be supported only in $S_r(z_j(0))$. Then, as a divergence-free function on Ω, it is orthogonal to any function of type ∇p in $(L^2(\Omega))^2$. Since the latter functions include all constant vectors, this yields that

$$\int_{S_r(z_j(0))} F_{j,1}(x)dx = 0.$$

In turn, since

$$F_{j,1}(x) \ + \ \nabla w_{j,1} = (A(z_{j+1}(0) - z_j(0)))\, \xi_j(x,0),$$

and $\nabla w_{j,1}$ is orthogonal to $F_{j,1}$ in $(L^2(\Omega))^2$,

$$\int_\Omega \parallel F_{j,1}(x) \parallel^2_{R^2} dx = (A(z_{j+1}(0) - z_j(0)))^T \int_{S_r(z_j(0))} F_{j,1}(x)dx = 0$$

or $F_{j,1}(x) \equiv 0$, and hence

$$\nabla w_{j,1} = \xi_j(x,0)\,(A(z_{j+1}(0) - z_j(0)))\,,$$

which means that $w_{j,1} \notin H^1(\Omega)$, contradicting (2.12).

Proof of Lemma 2.4. We will prove (2.10a) (the proof of (2.10b) is similar). Without loss of generality, we can assume that some neighborhoods of $S_r(z_j(0))$ and $S_r(z_{j+1}(0))$ do not overlap. We use below the traditional notation $y = (y_1, y_2)$ for a space variable to complement $x = (x_1, x_2)$, where it is necessary. (Since the proof below does not use (1.1)–(1.3), this should not cause confusion.)

Step 1: Green's formula. Our plan is to evaluate the vector columns in (2.8)/(2.9) by making use of (2.13)–(2.16) and the generalized version of the classical Green's formula representing solutions of the boundary problems (2.15), (2.16).

Consider the spectral problem

$$\Delta\theta = \alpha\theta, \quad \frac{\partial\theta}{\partial\nu} \mid_{\partial\Omega} = 0.$$

Denote by $\{\alpha_k\}_{k=1}^\infty$ its *negative* eigenvalues and by $\{\theta_k\}_{k=1}^\infty$ denote their respective orthonormalized in $L^2(\Omega)$ eigenfunctions. Then any solution to (2.15) admits the following representation:

$$w_{j,1}(x) = \sum_{k=1}^\infty \frac{1}{\sqrt{-\alpha_k}} \left(\int_\Omega \xi_j(q,0)\,(A(z_{j+1}(0) - z_j(0)))^T \frac{\nabla\theta_k}{\sqrt{-\alpha_k}} dq \right) \theta_k(x) \ + \ K,$$

where K is any number. (Note that $\{\frac{\nabla\theta_k}{\sqrt{-\alpha_k}}\}_{k=1}^\infty$ is an orthonormal sequence in $(L^2(\Omega))^2$.) Since in (2.13) we deal only with $\nabla w_{j,1}$, without loss of generality we can further assume that $K = 0$.

Let $g_{j,n}(x), n = 1, \ldots$, be a sequence of uniformly bounded infinitely many times differentiable functions vanishing on $\partial\Omega$ which converge to $\xi_j(x,0)$ in $L^2(\Omega)$. Consider now the boundary problem (2.15) with div $g_{j,n}\,(A(z_{j+1}(0) - z_j(0)))$'s in place

of div $\xi_j(x,0)\,(A(z_{j+1}(0)-z_j(0)))$ on the right. Denote by $w_{j,1,n}$ the following particular sequence of solutions to this new boundary problem:

$$w_{j,1,n}(x)=\sum_{k=1}^{\infty}\frac{1}{\sqrt{-\alpha_k}}\left(\int_\Omega g_{j,n}(q)\,(A(z_{j+1}(0)-z_j(0)))^T\,\frac{\nabla\theta_k}{\sqrt{-\alpha_k}}dq\right)\theta_k(x),\quad n=1,\dots.$$

Note that $w_{j,1,n}\to w_{j,1}$ as $n\to\infty$ in $H^1(\Omega)$ and hence in $L^2(\partial\Omega)$, due to the continuous embedding of the former space into the latter (see (6.3) below and also recall that we picked $K=0$).

Recall now that the classical Green's formula yields for $w_{j,1,n}$'s

$$\begin{aligned}
2\pi w_{j,1,n}(x)&=-\int_{\partial\Omega}w_{j,1,n}(\eta)\frac{\partial}{\partial\nu}\left(\ln\frac{1}{\sqrt{(x_1-\eta_1)^2+(x_2-\eta_2)^2}}\right)d\eta\\
&\quad-\int_\Omega\Delta w_{j,1,n}(y)\left(\ln\frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}}\right)dy\\
&=-\int_{\partial\Omega}w_{j,1,n}(\eta)\frac{\partial}{\partial\nu}\left(\ln\frac{1}{\sqrt{(x_1-\eta_1)^2+(x_2-\eta_2)^2}}\right)d\eta\\
&\quad-\int_\Omega\operatorname{div}g_{j,n}(y)\,(A(z_{j+1}(0)-z_j(0)))\left(\ln\frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}}\right)dy\\
&=-\int_{\partial\Omega}w_{j,1,n}(\eta)\frac{\partial}{\partial\nu}\left(\ln\frac{1}{\sqrt{(x_1-\eta_1)^2+(x_2-\eta_2)^2)}}\right)d\eta\\
&\quad+\int_\Omega g_{j,n}(y)\,(A(z_{j+1}(0)-z_j(0)))^T\nabla\left(\ln\frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}}\right)dy,
\end{aligned}$$

where in the last step we used integration by parts. Here and below, when we write ∇ within some integral we mean that the corresponding differentiation is conducted with respect to the integration variables.

Everywhere in this section we understand the improper integral over the given domain E for a function with a discontinuity at x as the limit of the integrals over $E\backslash B_s(x)$ as $s\to 0+$, where $B_s(x)$ is a disk of radius s with center at x. In particular, by switching to the polar coordinates near the "bad point" (x_1,x_2), one can show the last integral in the above is well defined, even if $g_{j,n}$ is replaced by any measurable bounded function.

Making use of the above, we can pass to the limit in the space $L^2(\Omega)$ as $n\to\infty$ in the last two lines in the above expression for $w_{j,1,n}$'s , which gives the following formula:

$$\begin{aligned}
2\pi w_{j,1}(x)&=-\int_{\partial\Omega}w_{j,1}(\eta)\frac{\partial}{\partial\nu}\left(\ln\frac{1}{\sqrt{(x_1-\eta_1)^2+(x_2-\eta_2)^2)}}\right)d\eta\\
&\quad+\int_{S_r(z_j(0))}(A(z_{j+1}(0)-z_j(0)))^T\nabla\left(\ln\frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}}\right)dy.
\end{aligned}\tag{6.1a}$$

Analogously,

$$
\begin{aligned}
2\pi w_{j,2}(x) = &- \int_{\partial\Omega} w_{j,2}(\eta) \frac{\partial}{\partial \nu} \left(\ln \frac{1}{\sqrt{(x_1-\eta_1)^2+(x_2-\eta_2)^2)}} \right) d\eta \\
&- \int_{S_r(z_{j+1}(0))} (A(z_{j+1}(0)-z_j(0)))^T \nabla \left(\ln \frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}} \right) dy.
\end{aligned} \tag{6.1b}
$$

For the first column in (2.8) we have for $j = i$ (see (2.7a), (2.14)):

$$
\begin{aligned}
\int_{S_r(z_j(0))} F_j dx &= \int_{S_r(z_j(0))} F_{j,1}(x)dx \; + \; \int_{S_r(z_j(0))} F_{j,2}(x)dx \\
&= \int_{S_r(z_j(0))} F_{j,1}(x)dx \; - \; \int_{S_r(z_j(0))} \nabla w_{j,2}(x)dx.
\end{aligned} \tag{6.2}
$$

To evaluate (6.2), we intend to evaluate the gradients of the terms in (6.1a)–(6.1b) and their integrals over required $S_r(z_j(0))$ and $S_r(z_{j+1}(0))$, and then to use (2.13). Since the integration with respect to x in (2.10a) is taken over $S_r(z_j(0))$, *everywhere below we consider only* $x \in S_r(z_j(0))$.

Step 2: Evaluation of the integrals of the gradients of the first terms in (6.1a)–(6.1b) over $S_r(z_j(0))$.

We begin with the gradient of the first term on the right in (6.1a). Recall first that

$$
\| w_{j,1} \|_{L^2(\partial\Omega)} \leq \; L_0 \, \| w_{j,1} \|_{H^1(\Omega)}, \tag{6.3}
$$

where L_0 depends on the $\partial\Omega$.

Recall that in Step 1, we selected

$$
w_{j,1}(x) = \sum_{k=1}^{\infty} \frac{1}{\sqrt{-\alpha_k}} \left(\int_\Omega \xi_j(q,0) \, (A(z_{j+1}(0)-z_j(0)))^T \, \frac{\nabla\theta_k}{\sqrt{-\alpha_k}} dq \right) \theta_k(x).
$$

Hence, taking into account that $\{\frac{\nabla\theta_k}{\sqrt{-\alpha_k}}\}_{k=1}^{\infty}$ is an orthonormal sequence in $(L^2(\Omega))^2$, we derive from Bessel's inequality that

$$
\| w_{j,1} \|_{H^1(\Omega)} \leq \; C \, \| A(z_{j+1}(0)-z_j(0)) \|_{R^2} \;\; \mathrm{mes}^{1/2} \{S_r(0)\}, \tag{6.4}
$$

where C denotes a (generic) positive constant.

Furthermore, for $i, j = 1, 2, i \neq j$, and $x \neq y$,

$$
\frac{\partial}{\partial y_i} \left(\ln \frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}} \right) = \frac{x_i - y_i}{(x_1-y_1)^2+(x_2-y_2)^2}, \tag{6.5}
$$

$$
\frac{\partial^2}{\partial y_i \partial x_i} \left(\ln \frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}} \right) = \frac{-(x_i-y_i)^2+(x_j-y_j)^2}{((x_1-y_1)^2+(x_2-y_2)^2)^2}, \tag{6.6}
$$

$$
\frac{\partial^2}{\partial y_i \partial x_j} \left(\ln \frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}} \right) = \frac{-2(x_i-y_i)(x_j-y_j)}{((x_1-y_1)^2+(x_2-y_2)^2)^2}. \tag{6.7}
$$

Denote next by d_0 the shortest distance between the set $S_r(z_j(0)) \bigcup S_r(z_{j+1}(0))$ and $\partial\Omega$:

$$
d_0 = \inf_{x \in S_r(z_j(0)) \bigcup S_r(z_{j+1}(0)), y \in \partial\Omega} \sqrt{(x_1-y_1)^2+(x_2-y_2)^2}. \tag{6.8}
$$

Note first that, since $S_r(z_j(0))$ is strictly separated from $\partial\Omega$, all the denominators in (6.5)–(6.7) are well defined for $x \in S_r(z_j(0))$ and $y \in \partial\Omega$. Therefore, in view of (6.5)–(6.7), we have the following estimate:

$$\left\| \nabla_x \frac{\partial}{\partial \nu} \left(\ln \frac{1}{\sqrt{(x_1-\eta_1)^2+(x_2-\eta_2)^2}} \right) \right\|_{R^2} \le \frac{C_0}{d_0^4}, \quad x \in S_r(z_j(0)), \ \ y \in \partial\Omega,$$

where C_0 is a positive constant and ∇_x means that the gradient is taken with respect to x (while $\frac{\partial}{\partial \nu}$ is taken with respect to y). Denote

$$\nabla_x \frac{\partial}{\partial \nu} \left(\ln \frac{1}{\sqrt{(x_1-\eta_1)^2+(x_2-\eta_2)^2}} \right) = (h_1(x,y), h_2(x,y)).$$

Therefore, for the gradient of the first term in (6.1a), we have

$$\begin{aligned}
&\left\| \int_{S_r(z_j(0))} \nabla \int_{\partial\Omega} w_{j,1}(\eta) \frac{\partial}{\partial \nu} \left(\ln \frac{1}{\sqrt{(x_1-\eta_1)^2+(x_2-\eta_2)^2}} \right) d\eta dx \right\|_{R^2} \\
&= \left[\left(\int_{S_r(z_j(0))} \int_{\partial\Omega} w_{j,1}(\eta) h_1(x,\eta) d\eta dx \right)^2 + \left(\int_{S_r(z_j(0))} \int_{\partial\Omega} w_{j,1}(\eta) h_2(x,\eta) d\eta dx \right)^2 \right]^{1/2} \\
&\le \left(\left(\max_{x \in S_r(z_j(0)),\ y \in \partial\Omega} \mid h_1(x,y) \mid \right)^2 \right. \\
&\qquad \left. + \left(\max_{x \in S_r(z_j(0)),\ y \in \partial\Omega} \mid h_2(x,y) \mid \right)^2 \right)^{1/2} \int_{S_r(z_j(0))} \int_{\partial\Omega} \mid w_{j,1}(\eta) \mid d\eta dx \\
&\le \frac{2C_0}{d_0^4} \int_{S_r(z_j(0))} \| w_{j,1} \|_{L^2(\partial\Omega)} \operatorname{mes}^{1/2} \{\partial\Omega\} dx \\
&\le 2L_0 \frac{C_0}{d_0^4} \| w_{j,1} \|_{H^1(\Omega)} \operatorname{mes}^{1/2} \{\partial\Omega\} \int_{S_r(z_j(0))} dx \\
&\le \frac{2L_0 C C_0}{d_0^4} \| A(z_{j+1}(0) - z_j(0)) \|_{R^2} \operatorname{mes}^{3/2} \{S_r(0)\} \operatorname{mes}^{1/2} \{\partial\Omega\},
\end{aligned} \tag{6.9}$$

where we used (6.3) and (6.4) in the last two steps. Analogously,

$$\begin{aligned}
&\left\| \int_{S_r(z_j(0))} \nabla \int_{\partial\Omega} w_{j,2}(\eta) \frac{\partial}{\partial \nu} \left(\ln \frac{1}{\sqrt{(x_1-\eta_1)^2+(x_2-\eta_2)^2}} \right) d\eta dx \right\|_{R^2} \\
&\qquad \le \frac{C}{d_0^4} \| A(z_{j+1}(0) - z_j(0)) \|_{R^2} \operatorname{mes}^{3/2} \{S_r(0)\} \operatorname{mes}^{1/2} \{\partial\Omega\},
\end{aligned} \tag{6.10}$$

where C is a (generic) positive constant.

Step 3: Evaluation of the integral of the gradient of the second term in (6.1b) over $S_r(z_j(0))$. Denote next by d_1 the distance between the set $S_r(z_j(0))$ and $S_r(z_{j+1}(0))$:

$$d_1 = \inf_{x \in S_r(z_j(0)), y \in S_r(z_{j+1}(0))} \sqrt{(x_1-y_1)^2+(x_2-y_2)^2}.$$

Then for $d_1 > 0$, similar to (6.9), (6.10) from (6.1b), it follows that

$$\left\| \int_{S_r(z_j(0))} \nabla \int_{S_r(z_{j+1}(0))} (A(z_{j+1}(0) - z_j(0)))^T \nabla \left(\ln \frac{1}{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}} \right) dy dx \right\|_{R^2}$$
$$\leq \frac{C}{d_1^4} \parallel A(z_{j+1}(0) - z_j(0)) \parallel_{R^2} \ \mathrm{mes}^2 \{S_r(0)\},$$

(6.11)

where C is a (generic) positive constant.

Step 4: Evaluation of the integral of the gradient of the second term in (6.1a) over $S_r(z_j(0))$. Without loss of generality, we can assume that the origin of our space coordinate system is located at $z_j(0)$, i.e., $S_r(z_j(0)) = S_r(0)$.

Consider any point $x = (x_1, x_2) \in S_r(0)$ and introduce the following function of x (or, more precisely, of $\parallel x \parallel_{R^2}$):

$$\rho(x) = \frac{1}{2} \min\{\parallel x \parallel_{R^2}, r - \parallel x \parallel_{R^2}\}.$$

Let $B_{\rho(x)}(x)$ be the disk of radius $\rho(x)$ with center at x:

$$B_{\rho(x)}(x) = \{(y_1, y_2) \mid (x_1 - y_1)^2 + (x_2 - y_2)^2 < \rho^2(x)\}.$$

Note that $B_{\rho(x)}(x) \subset S_r(0)$ for any $x \in S_r(0)$.

To simplify further notation, denote

$$A(z_{j+1}(0) - z_j(0)) = b = (b_1, b_2).$$

Then, from (6.1a),

$$\int_{S_r(0)} b^T \nabla \left(\ln \frac{1}{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}} \right) dy$$
$$= \int_{S_r(0) \setminus B_{\rho(x)}(x)} b^T \nabla \left(\ln \frac{1}{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}} \right) dy$$
$$+ \int_{B_{\rho(x)}(x)} b^T \nabla \left(\ln \frac{1}{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}} \right) dy.$$

Now note that, in view of (6.5),

$$\int_{B_{\rho(x)}(x)} \left(\ln \frac{1}{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}} \right)_{y_1} dy$$
$$= \lim_{s \to 0+} \int_{B_{\rho(x)}(x) \setminus B_s(x)} \frac{x_1 - y_1}{(x_1 - y_1)^2 + (x_2 - y_2)^2} dy \tag{6.12}$$
$$= - \lim_{s \to 0+} \int_0^{2\pi} \int_s^{\rho(x)} \cos\zeta \, d\rho \, d\zeta = 0.$$

These and similar calculations for the integration with respect to y_2 within the disk $B_\rho(x)$ yield that

$$\int_{B_{\rho(x)}(x)} b^T \nabla \left(\ln \frac{1}{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}} \right) dx = 0.$$

Thus,

$$\begin{aligned}&\int_{S_r(0)} b^T\nabla\left(\ln\frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}}\right)dy\\ &=\int_{S_r(0)\setminus B_{\rho(x)}(x)}\left(b_1\frac{x_1-y_1}{(x_1-y_1)^2+(x_2-y_2)^2}+b_2\frac{x_2-y_2}{(x_1-y_1)^2+(x_2-y_2)^2}\right)dy.\end{aligned}\tag{6.13}$$

We intend now to calculate the gradient of the second term in (6.1a) represented as in (6.13), assuming that $S_r(z_j(0))=S_r(0)$ is as in (1.4b) (recall that we can assume that $z_j(0)$ is the origin).

Fix any $x=(x_1,x_2)\in S_r(0)$.

Due to our selection of h for the given x, for "small" Δx_1,

$$B_{\rho((x_1+\Delta x_1,x_2))}((x_1+\Delta x_1,x_2))\subset S_r(0).$$

Furthermore, notice that, as in the second equality in (6.12),

$$\int_{B_{\rho((x_1+\Delta x_1,x_2))}((x_1+\Delta x_1,x_2))\setminus B_{\rho(x)}((x+\Delta x_1,x_2))}\left(\ln\frac{1}{\sqrt{(x_1+\Delta x_1-y_1)^2+(x_2-y_2)^2}}\right)_{y_1}dy=0.$$

Taking this into account, we obtain from (6.13)

$$\begin{aligned}&\frac{\partial}{\partial x_1}\int_{S_r(0)} b^T\nabla\left(\ln\frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}}\right)dy\\ &=\lim_{\Delta x_1\to 0}\frac{1}{\Delta x_1}\left(\int_{S_r(0)\setminus B_{\rho((x_1+\Delta x_1,x_2))}((x_1+\Delta x_1,x_2))} b^T\nabla\right.\\ &\qquad\times\left(\ln\frac{1}{\sqrt{(x_1+\Delta x_1-y_1)^2+(x_2-y_2)^2}}\right)dy\\ &\left.-\int_{S_r(0)\setminus B_{\rho(x)}(x)} b^T\nabla\left(\ln\frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}}\right)dy\right)\\ &=\lim_{\Delta x_1\to 0}\frac{1}{\Delta x_1}\left(\int_{S_r(0)\setminus B_{\rho(x)}((x_1+\Delta x_1,x_2))} b^T\nabla\left(\ln\frac{1}{\sqrt{(x_1+\Delta x_1-y_1)^2+(x_2-y_2)^2}}\right)dy\right.\\ &\left.-\int_{S_r(0)\setminus B_{\rho(x)}(x)} b^T\nabla\left(\ln\frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}}\right)dy\right)\\ &=b_1\frac{\partial}{\partial x_1}\left(\int_{S_r(0)\setminus B_\rho(x)}\frac{x_1-y_1}{(x_1-y_1)^2+(x_2-y_2)^2}dy\right)\\ &+b_2\frac{\partial}{\partial x_1}\left(\int_{S_r(0)\setminus B_\rho(x)}\frac{x_2-y_2}{(x_1-y_1)^2+(x_2-y_2)^2}dy\right),\end{aligned}$$

where $\rho=\rho(x)$ *in the last line is now treated as independent of* x *when calculating the derivatives.*

Let us calculate the derivative in the first term in the last expression. To simplify notation, we will also further write ρ instead of $\rho(x)$.

$$\begin{aligned}
&\frac{\partial}{\partial x_1}\int_{S_r(0)\setminus B_\rho(x)}\frac{x_1-y_1}{(x_1-y_1)^2+(x_2-y_2)^2}dy\\
&=\frac{\partial}{\partial x_1}\left\{\int_{-r}^{x_1-\rho}\int_{\alpha(y_1)}^{\beta(y_1)}\frac{x_1-y_1}{(x_1-y_1)^2+(x_2-y_2)^2}dy_2dy_1\right.\\
&\quad+\int_{x_1+\rho}^{r}\int_{\alpha(y_1)}^{\beta(y_1)}\frac{x_1-y_1}{(x_1-y_1)^2+(x_2-y_2)^2}dy_2dy_1\\
&\quad+\int_{x_1-\rho}^{x_1+\rho}\left[\int_{\alpha(y_1)}^{x_2-\sqrt{\rho^2-(x_1-y_1)^2}}\frac{x_1-y_1}{(x_1-y_1)^2+(x_2-y_2)^2}dy_2\right.\\
&\quad\left.\left.+\int_{x_2+\sqrt{\rho^2-(x_1-y_1)^2}}^{\beta(y_1)}\frac{x_1-y_1}{(x_1-y_1)^2+(x_2-y_2)^2}dy_2\right]dy_1\right\}\qquad(6.14)\\
&=\int_{S_r(0)\setminus B_\rho(x)}\frac{-(x_1-y_1)^2+(x_2-y_2)^2}{((x_1-y_1)^2+(x_2-y_2)^2)^2}dy\;+\;2\int_{x_1-\rho}^{x_1+\rho}\frac{(x_1-y_1)^2}{\rho^2\sqrt{h^2-(x_1-y_1)^2}}dy_1\\
&=\int_{S_r(0)\setminus B_\rho(x)}\frac{-(x_1-y_1)^2+(x_2-y_2)^2}{((x_1-y_1)^2+(x_2-y_2)^2)^2}dy\;+\;\pi.
\end{aligned}$$

Similar calculations also yield

$$\begin{aligned}
&\frac{\partial}{\partial x_2}\int_{S_r(0)\setminus B_\rho(x)}\frac{x_2-y_2}{(x_1-y_1)^2+(x_2-y_2)^2}dy\\
&=\int_{S_r(0)\setminus B_\rho(x)}\frac{-(x_2-y_2)^2+(x_1-y_1)^2}{((x_1-y_1)^2+(x_2-y_2)^2)^2}dy\;+\;\pi,\qquad(6.15)\\
&\frac{\partial}{\partial x_j}\int_{S_r(0)\setminus B_\rho(x)}\frac{x_i-y_i}{(x_1-y_1)^2+(x_2-y_2)^2}dy\\
&=-\int_{S_r(0)\setminus B_\rho(x)}2\frac{(x_i-y_i)(x_j-y_j)}{((x_1-y_1)^2+(x_2-y_2)^2)^2}dy,\quad i\neq j,i,j=1,2.\qquad(6.16)
\end{aligned}$$

Then, since we assumed that $S_r(0)$ is a disk, due to the symmetry of quadratic function,

$$\int_{S_r(0)}\int_{S_r(0)\setminus B_\rho(x)}\frac{-(x_i-y_i)^2+(x_j-y_j)^2}{((x_1-y_1)^2+(x_2-y_2)^2)^2}dydx=0,\quad i,j=1,2,\;\;i\neq j,\qquad(6.17)$$

while due to the antisymmetry of the linear functions,

$$\int_{S_r(0)}\int_{S_r(0)\setminus B_\rho(x)}\frac{(x_i-y_i)(x_j-y_j)}{((x_1-y_1)^2+(x_2-y_2)^2)^2}dydx=0,\quad i,j=1,2,\;\;i\neq j.\qquad(6.18)$$

Hence, in view of (6.14)–(6.16),

$$\int_{S_r(0)}\nabla\int_{S_r(0)\setminus B_\rho(x)}b^T\nabla\left(\ln\frac{1}{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}}\right)dydx=\pi\ \mathrm{mes}\,\{S_r(0)\}b.\qquad(6.19)$$

Step 5: An approximation formula for (2.8). In view of (6.9)–(6.11) and (6.19), we obtain from (6.1a) that for "small" r's

$$\begin{aligned}(6.20)\qquad &\left\| \int_{S_r(z_j(0))} \nabla w_{j,1} dx \; - \; \frac{\text{mes}\,\{S_r(0)\}}{2} A(z_{j+1}(0) - z_j(0)) \right\|_{R^2} \\ &\leq \; \frac{C}{\min\{d_0^4, d_1^4\}} \parallel A(z_{j+1}(0) - z_j(0)) \parallel_{R^2} \; \text{mes}^{3/2}\,\{S_r(0)\},\end{aligned}$$

where, again, C is a (generic) positive constant. Now, also making use of (6.1a)–(6.1b), (6.9)–(6.11), (6.20), and (2.13)–(2.16), we obtain (2.10a) from (6.2), taking into account that d_0 and d_1 do not decrease as $r \to 0+$. This ends the proof of Lemma 2.4. □

Remark 6.1. Note that we essentially used the assumption that $S_r(0)$ is a disk only to establish (6.17)–(6.18). The qualitative estimates and formulas (6.9)–(6.16) can be used to analyze the sufficient conditions of controllability in Theorems 2.3 and 2.5 along with formulas like (6.2) in other cases as well.

Acknowledgment. The author wishes to thank the referee for numerous helpful comments and suggestions that resulted in substantial changes to the original version of the manuscript.

REFERENCES

[1] J. M. Ball, J. E. Marsden, and M. Slemrod, *Controllability for distributed bilinear systems*, SIAM J. Control Optim., 20 (1982), pp. 575–597.
[2] S. Childress, *Mechanics of Swimming and Flying*, Cambridge University Press, Cambridge, UK, 1981.
[3] L. J. Fauci and C. S. Peskin, *A computational model of aquatic animal locomotion*, J. Comput. Phys., 77 (1988), pp. 85–108.
[4] L. J. Fauci, *Computational modeling of the swimming of biflagellated algal cells*, in Fluid Dynamics in Biology, Contemp. Math. 141, AMS, Providence, RI, 1993, pp. 91–102.
[5] S. Hirose, *Biologically Inspired Robots: Snake-like Locomotors and Manipulators*, Oxford University Press, Oxford, UK, 1993.
[6] A. Y. Khapalov, *Global non-negative controllability of the semilinear parabolic equation governed by bilinear control*, ESAIM: Control., Optim. Calc. Var., 7 (2002), pp. 269–283.
[7] A. Y. Khapalov, *Controllability of the semilinear parabolic equation governed by a multiplicative control in the reaction term: A qualitative approach*, SIAM J. Control Optim., 41 (2003), pp. 1886–1900.
[8] A. Y. Khapalov, *Controllability properties of a vibrating string with variable axial load*, Discrete Contin. Dynam. Systems, 11 (2004), pp. 311–324.
[9] A. Y. Khapalov, *The Well-posedness of a Model of an Apparatus Swimming in the 2-D Stokes Fluid*, preprint (available as Technical Rep. 2005-5, Washington State University, Department of Mathematics, http://www.math.wsu.edu/TRS/2005-5.pdf).
[10] A. N. Kolmogorov and S. V. Fomin, *Elements of the Theory of Functions and Functional Analysis*, Graylock Press, Rochester, NY, 1957.
[11] O. H. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1963.
[12] M. J. Lighthill, *Mathematical Biofluiddynamics*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 17, SIAM, Philadelphia, 1975.
[13] K. A. McIsaac and J. P. Ostrowski, *Motion planning for dynamic eel-like robots*, in Proceedings of the IEEE International Conference on Robotics and Automation, San Francisco, CA, 2000, pp. 1695–1700.
[14] S. Martinez and J. Cortés, *Geometric control of robotic locomotion systems*, in Proceedings of the 10th Fall Workshop on Geometry and Physics, Miraflores de la Sierra, Spain, 2001, Publ. R. Soc. Mat. Esp. 4, R. Soc. Mat. Esp., Madrid, 2003, pp. 183–198.
[15] C. S. Peskin, *Numerical analysis of blood flow in the heart*, J. Comput. Phys., 25 (1977), pp. 220–252.

[16] C. S. Peskin and D. M. McQueen, *A general method for the computer simulation of biological systems interacting with fluids*, in Proceedings of the SEB Symposium on Biological Fluid Dynamics (Leeds 1994), Symp. Soc. Exp. Biol. 49, The Company of Biologists Limited, Cambridge, UK, 1995, pp. 265–276.
[17] K. R. Symon, *Mechanics*, Addison–Wesley, Reading, MA, 1971.
[18] R. Temam, *Navier-Stokes Equations*, North–Holland, Amsterdam, 1984.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 683–693

NULL CONTROLLABILITY WITH VANISHING ENERGY FOR DISCRETE-TIME SYSTEMS IN HILBERT SPACE*

AKIRA ICHIKAWA†

Abstract. In this paper null controllability with vanishing energy is considered for discrete-time systems in Hilbert space. As in the case of continuous time systems, necessary and sufficient conditions in terms of an algebraic Riccati equation are given. Then necessary and sufficient conditions involving the spectrum of the system operator are given. Reachability and controllability with vanishing energy are also considered, and necessary and sufficient conditions for them are given. Finally applications to sampled-data systems, systems with impulse control, and periodic systems are discussed.

Key words. discrete-time systems, null controllability, optimal regulator, Riccati equation

AMS subject classifications. 93C55, 93B05, 93C05, 93C25

DOI. 10.1137/060657637

1. Introduction.

Consider the linear system

$$\dot{x} = Ax + Bu,\ x(0) = x_0 \in H, \tag{1.1}$$

where A is the infinitesimal generator of a strongly continuous semigroup $S(t)$ in a Hilbert space H, u is a control in some Hilbert space U, and $B \in L(U, H)$ is the space of bounded linear operators from U into H. For each locally square integrable function $u : [0, \infty) \to U$, define the solution in the mild sense

$$x(t; x_0, u) = S(t)x_0 + \int_0^t S(t - r)Bu(r)dr,\ t \geq 0.$$

We denote by $|\cdot|$ the norm of vectors and by $\sigma(A)$ the spectrum of the operator A. The following definitions are introduced in [8].

DEFINITION 1.1. (a) *The system* (1.1) *is said to be null controllable with vanishing energy* (*NCVE for short*) *if for each initial* $x(0) = x_0$ *there exists a sequence of pairs* (T_N, u_N), $0 < T_N \uparrow \infty$, $u_N \in L_2(0, T_N; U)$ *such that* $x(T_N; x_0, u_N) = 0$ *and*

$$\lim_{N\to\infty} \int_0^{T_N} |u_N(t)|^2 dt = 0. \tag{1.2}$$

(b) *The system* (1.1) *is said to be exactly controllable with vanishing energy* (*ECVE*) *if for any pair* (x_0, x_1) *of initial and final states there exists a sequence of pairs* (T_N, u_N), $0 < T_N \uparrow \infty$, $u_N \in L_2(0, T_N; U)$ *such that* $x(T_N; x_0, u_N) = x_1$ *and* (1.2) *holds.*

(A, B) is said to be NCVE (ECVE) if the system (1.1) is NCVE (ECVE). The following theorem gives necessary and sufficient conditions.

*Received by the editors April 19, 2006; accepted for publication (in revised form) January 23, 2007, published electronically May 7, 2007.
http://www.siam.org/journals/sicon/46-2/65763.html
†Department of Aeronautics and Astronautics, Kyoto University, Kyoto 606-8501, Japan (ichikawa@kuaero.kyoto-u.ac.jp).

THEOREM 1.2. *(A, B) is NCVE if and only if*
(a) *it is null controllable on some interval $[0, \tau]$, and*
(b) *$X = 0$ is the unique solution of the algebraic Riccati equation (ARE)*

$$A^*X + XA - XBB^*X = 0$$

in the class of nonnegative operators.

Priola and Zabczyk [8] showed that condition (b) is necessary and sufficient for NCVE when (A, B) is null controllable on some interval $[0, \tau]$. The necessity of (a) was then shown by van Neerven [7].

Under the following two assumptions Priola and Zabczyk [8] obtained more explicit necessary and sufficient conditions.

Hypothesis 1. There exists a sequence $\{\lambda_n\} \subset \sigma(A)$ such that λ_n is isolated in $\sigma(A)$ and

$$\lim_{n\to\infty} \mathrm{Re}(\lambda_n) = s(A) = \sup\{\mathrm{Re}(\lambda) : \lambda \in \sigma(A)\}.$$

Hypothesis 2. There exist $S(t)$-invariant subspaces H_s and H_u such that
(a) $H = H_s \oplus H_u$,
(b) A on H_s is exponentially stable, and
(c) the set of all generalized eigenvectors of A contained in H_u is linearly dense in H_u.

THEOREM 1.3. *Suppose that Hypotheses 1 and 2 hold. Then (A, B) is NCVE if and only if*
(a) *(A, B) is null controllable on some interval $[0, \tau]$, and*
(b) *$\mathrm{Re}(\lambda) \le 0$ for any $\lambda \in \sigma(A)$.*

THEOREM 1.4. *Suppose that Hypotheses 1 and 2 hold. Suppose further that $S(t)$ is a strongly continuous group on H. Then (A, B) is ECVE if and only if*
(a) *(A, B) is exactly controllable on some interval $[0, \tau]$, and*
(b) *$\mathrm{Re}(\lambda) = 0$ for any $\lambda \in \sigma(A)$.*

The proof of Theorem 1.2 is based on the theory of optimal quadratic control. For the proof of necessity of Theorem 1.3 the relation between the Riccati equation and the controllability Gramian of the pair $(-A, B)$ is used, while for sufficiency the Riccati equation is directly used. Theorem 1.4 is a consequence of Theorem 1.3 and the fact that $(-A, -B)$ is also NCVE.

If we fix $x_0 = 0$ in (b) of Definition 1.1, (A, B) is said to be reachable with vanishing energy (RVE). It is easy to see that (A, B) is ECVE if and only if it is NCVE and RVE. Suppose that $S(t)$ is a strongly continuous group, and let P_T be the controllability operator defined by

$$P_T x = \int_0^T S(t)BB^*S^*(t)x dt.$$

It is coercive (positive and boundedly invertible) for $T \ge \tau$, if (A, B) is exactly controllable on $[0, \tau]$. The control with minimum norm in $L_2(0, T; U)$ such that $x(T; 0, u) = x_1$ is given by

$$\hat{u}_T = B^*S^*(T - t)P_T^{-1}x_1$$

[8] and its norm by

$$\| \hat{u}_T \|_2 = \langle x_1, P_T^{-1}x_1 \rangle^{\frac{1}{2}}. \tag{1.3}$$

LEMMA 1.5. *(A, B) is RVE if and only if*

(a) *(A, B) is exactly controllable on some interval $[0, \tau]$, and*

(b) *$P_T^{-1} \to 0$ strongly as $T \to \infty$.*

Proof. The proof of necessity of (a) is based on the Baire category theorem and is similar to that of Theorem 1.2(a) given in [7]. The rest follows from (1.3). □

In this paper we shall establish the discrete-time versions of the theorems above. It is important in its own right but also useful when we consider sampled-data systems with zero-order hold, systems with impulse control, and periodic systems. In the discrete-time case the proof of necessity of Theorem 1.3 is more involved, since the Riccati equation is more complicated for discrete-time systems. Lemma 2.5 in section 2 fills this gap and enables us to extend Theorem 1.3. The extension of Theorem 1.4 requires the invertibility of A. It is also useful to introduce RVE. In section 2 we give preliminaries concerning necessary notions of discrete-time systems. In section 3 we consider necessary and sufficient conditions for NCVE and extend Theorems 1.2 and 1.3. In section 4 we introduce RVE and extend Theorem 1.4. Finally in section 5 we apply NCVE and ECVE results to sampled-data systems, systems with impulse control, and periodic systems.

2. Preliminaries. Consider the discrete-time system

$$x(k+1) = Ax(k) + Bu(k),\ x(0) = x_0, \tag{2.1}$$

where $A \in L(H)$, $B \in L(U, H)$, $x \in H$, and $u \in U$. We collect basic definitions and some useful results for (2.1) as in the finite dimensional case [1].

DEFINITION 2.1. (a) *(A, B) is null controllable on $[0, K]$ if for any x_0 there is a sequence of control inputs $u = \{u(0), u(1), \dots, u(K-1)\}$ such that $x(K; x_0, u) = 0$.*

(b) *(A, B) is reachable on $[0, K]$ if for every state x_1 there is a sequence of control inputs $u = \{u(0), u(1), \dots, u(K-1)\}$ such that $x(K; 0, u) = x_1$.*

(c) *(A, B) is exactly controllable on $[0, K]$ if for every pair (x_0, x_1) there is a sequence of control inputs $u = \{u(0), u(1), \dots, u(K-1)\}$ such that $x(K; x_0, u) = x_1$.*

LEMMA 2.2. (a) *(A, B) is reachable on $[0, K]$ if and only if it is exactly controllable on $[0, K]$. In this case it is null controllable on $[0, K]$.*

(b) *If A is invertible and (A, B) is null controllable on $[0, K]$, then (A, B) is exactly controllable on $[0, K]$.*

LEMMA 2.3. *The following statements are equivalent:*

(a) *(A, B) is null controllable on $[0, K]$.*

(b) *$R(A^K) \subset R(M_K)$, where $M_K = [B, AB, \dots, A^{K-1}B]$ is the reachability operator.*

(c) *$|M_K^* x| \geq a|(A^*)^K x|$ for some $a > 0$.*

If these conditions hold, the operator $\begin{bmatrix} B^* \\ \lambda I - A^* \end{bmatrix}$ is 1 to 1 for any nonzero λ.

Proof. Consider the response of the system (2.1) with initial condition x_0 and control $u = \{u(0), u(1), \dots, u(K-1)\}$. Then

$$x(K; x_0, u) = A^K x_0 + \sum_{j=0}^{k-1} A^{k-j-1} B u(j),$$

and the second term on the right-hand side lies in $R(M_K)$, the range of M_K. Hence (a) is equivalent to (b). The equivalence of (b) and (c) follows from Corollary 3.5 of [3]. If there exists a nonzero q such that $B^* q = 0$ and $\lambda q = A^* q$, it contradicts to (c) with $x = q$. □

LEMMA 2.4. *Suppose A is exponentially stable* [5], i.e., $|A^k| \le M\rho^k$, $0 < \rho < 1$, and that (A, B) is exactly controllable on $[0, K]$. Then there exists a coercive operator Y such that

$$Y = AYA^* + BB^*.$$

Y is called the controllability Gramian of (A, B).

Proof. By Lemma 2.2 (A, B) is reachable on $[0, K]$. Hence $M_K = [B, AB, \dots, A^{K-1}B]$ is onto and $M_K M_K^* \ge aI$ for some $a > 0$. Define

$$Y = \lim_{k\to\infty} M_k M_k^* = \lim_{k\to\infty} \sum_{j=0}^{k-1} A^j BB^* (A^*)^j.$$

The right-hand side converges in the uniform operator topology, and $Y \ge M_K M_K^* \ge aI$. Hence Y is coercive. Moreover

$$Y = BB^* + A\sum_{j=0}^{\infty} A^j BB^* (A^*)^j A^* = BB^* + AYA^*. \qquad \square$$

LEMMA 2.5. *Suppose A is invertible and (A, B) is exactly controllable on $[0, K]$. Then $(A^{-1}, A^{-1}B)$ is exactly controllable on $[0, K]$. If A^{-1} is exponentially stable, then the inverse of its controllability Gramian Y exists and satisfies the following ARE:*

$$X = A^*XA - A^*XB(I + B^*XB)^{-1}B^*XA. \tag{2.2}$$

Proof. Since (A, B) is exactly controllable on $[0, K]$, so is $(A^{-1}, A^{-1}B)$. In fact

$$\begin{aligned}&[A^{-1}B, A^{-1}(A^{-1}B), \dots, (A^{-1})^{K-1}(A^{-1}B)]\\ &= (A^{-1})^K [A^{K-1}B, \dots, AB, B].\end{aligned}$$

Now by definition

$$Y = A^{-1}Y(A^{-1})^* + A^{-1}BB^*(A^{-1})^*,$$

which implies

$$AYA^* = Y + BB^*.$$

By Lemma 2.4, Y is coercive and hence invertible. As in Lemma 3.18 in [4], we obtain

$$\begin{aligned}(A^{-1})^* Y^{-1} A^{-1} &= (Y + BB^*)^{-1}\\ &= Y^{-1}(I + BB^*Y^{-1})^{-1}\\ &= Y^{-1}[I - (I + BB^*Y^{-1})^{-1}BB^*Y^{-1}]\\ &= Y^{-1}[I - B(I + B^*Y^{-1}B)^{-1}B^*Y^{-1}]\\ &= Y^{-1} - Y^{-1}B(I + B^*Y^{-1}B)^{-1}B^*Y^{-1},\end{aligned}$$

where for the second equality we have used the equality $Y + BB^* = (I + BB^*Y^{-1})Y$, and for the fourth equality the familiar identity $M(I + NM)^{-1} = (I + MN)^{-1}M$ is used. Hence we obtain

$$Y^{-1} = A^*Y^{-1}A - A^*Y^{-1}B(I + B^*Y^{-1}B)^{-1}B^*Y^{-1}A,$$

and Y^{-1} is a coercive solution of the ARE (2.2). $\square$

3. Null controllability with vanishing energy. Consider the system (2.1)

$$x(k+1) = Ax(k) + Bu(k),\ x(0) = x_0.$$

We shall define NCVE for this system.

DEFINITION 3.1. *(A, B) is NCVE if for each x_0 there exists a sequence of pairs (k_N, u_N), k_N a positive integer $\uparrow \infty$, $u_N \in l_2(0, k_N - 1; U)$, such that $x(k_N; x_0, u_N) = 0$ and*

$$\lim_{N\to\infty} \| u_N \|_2 = 0,$$

where $l_2(0, k_N - 1; U)$ is the set of vectors $u = \{u(0), u(1), \dots, u(k_N - 1)\}$, $u(k) \in U$, with norm

$$\| u \|_2 = \left(\sum_{k=0}^{k_N - 1} |u(k)|^2 \right)^{\frac{1}{2}}.$$

LEMMA 3.2. *If (A, B) is NCVE, then (A, B) is null controllable on some interval $[0, K]$.*

Proof. The proof is based on the Baire category theorem and similar to the proof of Theorem 3.1 in [7]. □

First we shall prove the following.

THEOREM 3.3. *(A, B) is NCVE if and only if*

(a) *(A, B) is null controllable on some interval $[0, K]$, and*

(b) *$X = 0$ is the unique solution of the ARE* (2.2)

$$X = A^*XA - A^*XB(I + B^*XB)^{-1}B^*XA$$

in the class of nonnegative operators.

We modify Hypotheses 1 and 2 as follows.

Hypothesis 3. There exists a sequence $\{\lambda_n\} \subset \sigma(A)$ such that λ_n is isolated in $\sigma(A)$ and

$$\lim_{n\to\infty} |\lambda_n| = s(A) = \sup\{|\lambda| : \lambda \in \sigma(A)\}.$$

Hypothesis 4. There exist A-invariant subspaces H_s and H_u such that

(a) $H = H_s \oplus H_u$,

(b) (b) A on H_s is exponentially stable, and

(c) the set of all generalized eigenvectors of A contained in H_u is linearly dense in H_u.

Under Hypotheses 3 and 4 we shall prove the following.

THEOREM 3.4. *(A, B) is NCVE if and only if*

(a) *(A, B) is null controllable on some interval $[0, K]$, and*

(b) *$|\lambda| \le 1$ for any $\lambda \in \sigma(A)$.*

Proof of Theorem 3.3. We shall follow the proof of Theorem 1.2 in [8]. We first show necessity. Consider the quadratic cost associated with (2.1) on $[0, k_N - 1]$:

$$J(u; x_0, k_N, Q) = \sum_{k=0}^{k_N - 1} |u(k)|^2 + \langle x(k_N), Qx(k_N) \rangle,$$

where $Q \geq 0$. It is known [6], [9], [10] that the optimal control minimizing the cost function is given by the feedback law

$$\bar{u}(k) = -[I + B^*X(k+1)B]^{-1}B^*X(k+1)Ax(k),$$

where $X(k) = X(k; k_N, Q)$ is the sequence of nonnegative operators defined by the Riccati equation

$$\begin{aligned} X(k) &= A^*X(k+1)A - A^*X(k+1)B[I + B^*X(k+1)B]^{-1}B^*X(k+1)A, \\ X(k_N) &= Q. \end{aligned} \tag{3.1}$$

Moreover,

$$J(\bar{u}; x_0, k_N, Q) = \langle x_0, X(0; k_N, Q)x_0 \rangle.$$

Now we consider the case $Q = qI$, $q > 0$, and let $q \to \infty$. Since (A, B) is null controllable on $[0, K]$, for each x_0 and $k_N \geq K$ there exists a control $u \in l_2(0, k_N - 1; U)$ such that $x(k_N; x_0, u) = 0$. Let u_N be the control with minimum norm among them. Then it is given by $u_N = -\bar{M}_N^*(\bar{M}_N\bar{M}_N^*)^{-1}A^{k_N}x_0$, where $\bar{M}_N = [A^{k_N-1}B, \ldots, AB, B]$. Since (A, B) is NCVE, $\lim_{N\to\infty} \| u_N \|_2^2 = 0$ for each x_0, and hence there exists a constant $a > 0$ such that $\| u_N \|_2^2 \leq a|x_0|^2$. Notice that

$$J(\bar{u}; x_0, k_N, qI) = \langle x_0, X(0; k_N, qI)x_0 \rangle \leq J(u_N; x_0, k_N, qI) = \| u_N \|_2^2 \leq a|x_0|^2,$$

which yields $X(0; k_N, qI) \leq aI$. Since $X(0; k_N, qI)$ is monotone increasing in q, there exists a limit as $q \to \infty$, denoted by $X(0; k_N)$, i.e., $X(0; k_N) = \lim_{q\to\infty} X(0; k_N, qI)$. Let $\bar{u}_q$ be the optimal control for $J(u; x_0, k_N, qI)$. Then it is uniformly bounded in q. Hence there exists a subsequence q_j such that $\bar{u}_{q_j}$ converges weakly to some limit $\bar{u}_\infty$. Then $x(k_N; x_0, \bar{u}_\infty) = 0$ and $\| \bar{u}_\infty \|_2^2 \leq \langle x_0, X(0; k_N)x_0 \rangle \leq \| u_N \|_2^2$. But u_N is the control with minimum norm, and hence $\| \bar{u}_\infty \|_2^2 = \langle x_0, X(0; k_N)x_0 \rangle = \| u_N \|_2^2$. Now suppose that (A, B) is null controllable on $[0, K]$, $K \leq k_N$. Since $X(k; k_N, qI) = X(0; k_N - k, qI)$, the following limit exists:

$$\lim_{q\to\infty} X(k; k_N, qI) = \lim_{q\to\infty} X(0; k_N - k, qI) \equiv X(k; k_N) \text{ for } k \leq k_N - K.$$

Moreover, from (3.1), $X(k; k_N)$, $k \leq k_N - K$, satisfies the Riccati equation

$$\begin{aligned} X(k) &= A^*X(k+1)A - A^*X(k+1)B[I + B^*X(k+1)B]^{-1}B^*X(k+1)A, \\ X(k_N - K) &= X(k_N - K; k_N). \end{aligned}$$

Since $\langle x_0, X(0; k_N)x_0 \rangle = \| u_N \|_2^2$, $X(0; k_N)$ is decreasing in N and has a nonnegative limit

$$X_\infty = \lim_{N\to\infty} X(0; k_N).$$

For $k \leq N - K$ we know $X(k; k_N) = X(0; k_N - k)$, and hence $\lim_{N\to\infty} X(k; k_N) = X_\infty$. Letting $N \to \infty$ in the Riccati equation above we see that X_∞ satisfies the ARE (2.2). Recall that (A, B) is NCVE, and hence

$$\langle x_0, X_\infty x_0 \rangle \leq \langle x_0, X(0; k_N)x_0 \rangle = \| u_N \|_2^2 \to 0$$

and $X_\infty = 0$. Now let X be any nonnegative solution of the ARE (2.2). We shall show that $X \le X_\infty$ to conclude $X = 0$. For this purpose consider the Riccati difference equation (3.1) with $Q = X$. Then $X(k) = X$ is a solution. Thus

$$J(\bar{u}_X; x_0, k_N, X) = \langle x_0, Xx_0\rangle \le J(\bar{u}_q; x_0, k_N, qI) = \langle x_0, X(0; k_N, qI)x_0\rangle$$

for $q \ge \| X \|$, where $\bar{u}_X$ and $\bar{u}_q$ denote the optimal controls for the corresponding cost functions. Now passing to the limit $q \to \infty$ and to the limit $N \to \infty$ we obtain $\langle x_0, Xx_0\rangle \le \langle x_0, X(0; k_N)x_0\rangle$ and $\langle x_0, Xx_0\rangle \le \langle x_0, X_\infty x_0\rangle$, respectively. Thus we have shown $X = 0$, which completes the proof of necessity. □

To show sufficiency we recall that $\| u_N \|_2^2 = \langle x_0, X(0; k_N)x_0\rangle \to \langle x_0, X_\infty x_0\rangle$. But by condition (b) $X_\infty = 0$, and hence $\| u_N \|_2 \to 0$ and (A, B) is NCVE.

Proof of Theorem 3.4. We shall follow the proof of Theorem 1.3 in [8]. To show necessity we suppose that $|\lambda| > 1$ for some $\lambda \in \sigma(A)$. Then by Hypothesis 3 there exists an isolated element $\mu \in \sigma(A)$, with $|\mu| > 1$. Consider the spectral Riesz projection P_1 associated with μ

$$P_1 x = \frac{1}{2\pi i}\int_\gamma (\lambda I - A)^{-1} x d\lambda,\ x \in H,$$

where γ is a circle containing μ in its interior and $\sigma(A)/\{\mu\}$ in its exterior. Using projections P_1 and $P_2 = I - P_1$, we can split (2.1) into two subsystems in E_1 and E_2, respectively:

$$x_1(k+1) = A_1 x_1(k) + B_1 u(k),$$
$$x_2(k+1) = A_2 x_2(k) + B_2 u(k),$$

where $E_i = P_i H$, A_i is the restriction of A to E_i, and $B_i = P_i B$. The subspaces E_i are A-invariant and $H = E_1 \oplus E_2$. Since (2.1) is null controllable, (A_1, B_1) and (A_2, B_2) are null controllable. Since $\sigma(A_1) = \{\mu\}$, it is invertible, and (A_1, B_1) is exactly controllable by Lemma 2.2. Hence $(A_1^{-1}, A_1^{-1}B_1)$ is exactly controllable. Since A_1^{-1} is exponentialy stable, by Lemma 2.4 it possesses a coercive controllability Gramian Y

$$Y = A_1^{-1} Y (A_1^{-1})^* + A_1^{-1} B_1 B_1^* (A_1^{-1})^*.$$

By Lemma 2.5, $X_1 = Y^{-1}$ is a coercive solution of the ARE

$$X = A_1^* X A_1 - A_1^* X B_1 (I + B_1^* X B_1)^{-1} B_1^* X A_1.$$

Then $X = I_H X_1 P_1$ is a nontrivial nonnegative solution of the ARE for (2.2), where I_H is the injection of E_1 into H. This contradicts Theorem 3.3 and hence $|\lambda| \le 1$ for any $\lambda \in \sigma(A)$.

To show sufficiency let X be any nonnegative solution of the ARE (2.2). Since $H = H_s \oplus H_u$, it is suficient to show $X = 0$ on both H_s and H_u. As in the proof of Theorem 3.3 consider (3.1) with $Q = X$, and recall the inequality

$$\langle x_0, Xx_0\rangle = J(\bar{u}_X; x_0, k_N, X) \le J(0; x_0, k_N, X) = \langle A^{k_N} x_0, XA^{k_N} x_0\rangle \to 0,\ x_0 \in H_s.$$

Hence $Xx_0 = 0$ for any $x_0 \in H_s$. To show $Xx_0 = 0$ for any $x_0 \in H_u$, let $\lambda \in \sigma(A)$ with $|\lambda| \le 1$, which corresponds to an eigenvector p, i.e., $Ap = \lambda p$. Then

$$\begin{aligned}\langle p, Xp\rangle &= \langle p, A^*XAp\rangle - \langle p, A^*XB(I + B^*XB)^{-1}B^*XAp\rangle \\ &= |\lambda|^2[\langle p, Xp\rangle - \langle p, XB(I + B^*XB)^{-1}B^*Xp\rangle]. \qquad (3.2)\end{aligned}$$

If $|\lambda| < 1$, then (3.2) yields $Xp = 0$. If $|\lambda| = 1$, then it yields $B^*Xp = 0$. In this case we obtain $Xp = \lambda A^*Xp$ from the ARE (2.2). Hence

$$\begin{bmatrix} B^* \\ \frac{1}{\lambda}I - A^* \end{bmatrix} Xp = 0.$$

By Lemma 2.3 the operator above is one to one and hence $Xp = 0$. Thus for any eigenvector of A we have shown $Xp = 0$. We shall show that $Xq = 0$ for any generalized eigenvector of A, which would then conclude $X = 0$. Now let $q \in N((\lambda I - A)^2)$, i.e., $(\lambda I - A)^2 q = 0$. Then $q_1 = (\lambda I - A)q$ satisfies $(\lambda I - A)q_1 = 0$. Repeating the arguments above we conclude $Xq_1 = 0$. Hence $XAq = \lambda Xq$, and from the ARE (2.2) we obtain

$$\begin{aligned} \langle q, Xq\rangle &= \langle q, A^*XAq\rangle - \langle q, A^*XB(I + B^*XB)^{-1}B^*XAq\rangle \\ &= |\lambda|^2[\langle q, Xq\rangle - \langle q, XB(I + B^*XB)^{-1}B^*Xq\rangle]. \end{aligned}$$

This is the same with (3.2) and hence $Xq = 0$. Repeating this process we conclude $Xq = 0$ for any generalized eigenvector of A satisfying $(\lambda I - A)^k q = 0$. Hence $X = 0$ on H_u. Thus $X = 0$ on H, and by Theorem 3.3 (A, B) is NCVE. □

In [7] the reproducing kernel Hilbert space associated with the controllability operator was introduced, and Theorem 1.2 was extended to the case where H is a Banach space. The extension of Theorem 3.3 to a Banach space is also possible using the Riccati equation directly.

4. Exact controllability with vanishing energy. First we introduce RVE, which is useful to consider ECVE.

DEFINITION 4.1. *(A, B) is RVE if for each x_1 there exists a sequence of pairs (k_N, u_N), $k_N \uparrow \infty$, $u_N \in l_2(0, k_N - 1; U)$, such that $x(k_N; 0, u_N) = x_1$ and*

$$\lim_{N\to\infty} \| u_N \|_2 = 0.$$

LEMMA 4.2. *Suppose (A, B) is RVE. Then $0 \notin \sigma_p(A^*)$. If A is invertible, then (A, B) is RVE if and only if $(A^{-1}, A^{-1}B)$ is NCVE.*

Proof. Suppose $0 \in \sigma_p(A^*)$ and $A^*h = 0$, with $|h| = 1$. If (A, B) is reachable on $[0, K]$, then for some sequence $u = (u_j)$

$$\sum_{j=0}^{K-1} A^{K-j-1}Bu_j = h.$$

Then

$$\begin{aligned} 1 = \langle h, h\rangle &= \left\langle h, \sum_{j=0}^{K-1} A^{K-j-1}Bu_j \right\rangle \\ &= \sum_{j=0}^{K-1} \langle B^*(A^*)^{K-j-1}h, u_j\rangle \\ &= \langle B^*h, u_{K-1}\rangle \le |B^*h||u_{K-1}|. \end{aligned}$$

Hence $|u_{K-1}| \ge \frac{1}{|B^*h|}$, and (A, B) cannot be RVE. Now assume that A is invertible. Then the system (2.1) can be written as

$$x(k) = A^{-1}x(k + 1) - A^{-1}Bu(k).$$

Thus if (A, B) is RVE, then by redefining u and x we can easily see that

$$\tilde{x}(k+1) = A^{-1}\tilde{x}(k) + A^{-1}B\tilde{u}(k)$$

is NCVE. The converse is also true, since we can reverse the arguments. □

From Lemma 4.2 we immediately obtain the following.

THEOREM 4.3. *Suppose A is invertible and A^{-1} satisfies Hypotheses* 3 *and* 4. *Then (A, B) is RVE if and only if*

(a) *(A, B) is exactly controllable on some interval $[0, K]$, and*

(b) *$|\lambda| \geq 1$ for any $\lambda \in \sigma(A)$.*

Now we are ready to extend Theorem 1.4.

THEOREM 4.4. *Suppose A and A^{-1} satisfy Hypotheses* 3 *and* 4. *Then (A, B) is ECVE if and only if*

(a) *(A, B) is exactly controllable on some interval $[0, K]$, and*

(b) *$|\lambda| = 1$ for any $\lambda \in \sigma(A)$.*

Proof. Note that (A, B) is ECVE if and only if it is NCVE and RVE. Hence the proof follows from Theorems 3.4 and 4.3. □

5. Applications. In this section we apply our theorems to sampled-data systems, systems with impulse control, and periodic systems. First we consider a sampled-data system with zero-order hold [2]

$$\dot{x} = Ax + Bu,$$

where A is the infinitesimal generator of a strongly continuous semigroup $S(t) \in L(H)$, $B \in L(U, H)$, and u is a control given by

$$u(t) = u(k\tau), k\tau \leq t < (k+1)\tau.$$

Then at times $k\tau$ we have the following:

$$\begin{aligned} x((k+1)\tau) &= S(\tau)x(k\tau) + \int_0^\tau S(r)Bdru(k\tau) \\ &\equiv A_d x(k\tau) + B_d u(k\tau). \end{aligned}$$

The sampled-data system is said to be NCVE (ECVE) if it is NCVE (ECVE) in the sense of Definition 1.1 with $T_N = N\tau$. Note that the sampled-data system is NCVE (ECVE) if and only if (A_d, B_d) is NCVE (ECVE). Hence, if A_d satisfies Hypotheses 3 and 4, then by Theorem 3.4 the sampled-data system is NCVE if and only if

(a) (A_d, B_d) is null controllable on some interval $[0, K]$, and

(b) $|\lambda| \leq 1$ for any $\lambda \in \sigma(A_d)$.

If $S(t)$ is a group and $S(\tau)^{-1}$ satisfies Hypotheses 3 and 4, then the sampled-data system is ECVE if and only if

(a) (A_d, B_d) is exactly controllable on some interval $[0, K]$, and

(b) $|\lambda| = 1$ for any $\lambda \in \sigma(A_d)$.

Next we consider the system (1.1) with impulse control $u(k-1)\delta(t-k\tau)$ at time $k\tau$, $k \geq 1$. Then the state $x(k\tau)$ after the impulse $u(k-1)\delta(t-k\tau)$ satisfies

$$x((k+1)\tau) = S(\tau)x(k\tau) + Bu(k).$$

LEMMA 5.1. *The system* (1.2) *with impulse control is NCVE if and only if $(S(\tau), B)$ is NCVE.*

LEMMA 5.2. *Suppose $S(t)$ is a group and $S(\tau)^{-1}$ satisfies Hypotheses* 3 *and* 4. *Then the system* (1.2) *with impulse control is ECVE if and only if $(S(\tau), B)$ is ECVE.*

Proof. Note that the system (1.2) with impulse control is ECVE if and only if it is NCVE and RVE. Let $K\tau \le T < (K+1)\tau$, and consider the controllability operator

$$\begin{aligned} P_T^{imp} &= S(T-K\tau)\left(\sum_{j=1}^{K} S(\tau)^{K-j}BB^*(S(\tau)^*)^{K-j}\right)S^*(T-K\tau) \\ &= S(T-K\tau)P_{K\tau}^{imp}S^*(T-K\tau). \end{aligned}$$

Hence $(P_T^{imp})^{-1} \to 0$ strongly if and only if $(P_{K\tau}^{imp})^{-1} \to 0$ strongly, and as in Lemma 1.5 the assertion follows. □

Now we have the following.

THEOREM 5.3. (1) *The system* (1.2) *with impulse control is NCVE if and only if*

(a) *$(S(\tau), B)$ is null controllable on some interval $[0, K]$, and*

(b) *$|\lambda| \le 1$ for any $\lambda \in \sigma(S(\tau))$.*

(2) *Suppose $S(t)$ is a group and $S(\tau)^{-1}$ satisfies Hypotheses* 3 *and* 4. *Then the system* (1.2) *with impulse control is ECVE if and only if*

(a) *$(S(\tau), B)$ is exactly controllable on some interval $[0, K]$, and*

(b) *$|\lambda| = 1$ for any $\lambda \in \sigma(S(\tau))$.*

Finally consider the T-periodic system

$$\dot{x} = A(t)x + B(t)u,\ x(t_0) = x_0, 0 \le t_0 < T, \tag{5.1}$$

where $A(t)$ is T-periodic and generates an evolution operator $S(t,s)$, and $B(t)$ is T-periodic and strongly continuous. Then

$$\begin{aligned} x((k+1)T+t_0) &= S(T+t_0,t_0)x(kT+t_0) + \int_{t_0}^{T+t_0} S(T+t_0,r)B(r)u(k,r)dr \\ &\equiv S(T+t_0,t_0)x(kT+t_0) + B_d u(k), \end{aligned} \tag{5.2}$$

where we have used the property $S((k+1)T+t_0, kT+r) = S(T+t_0, r)$, and $u(k,r) = u(kT+r)$, for $t_0 \le r < t_0+T$, $u(k) = u(k,\cdot) \in L_2(t_0, t_0+T; U)$, and B_d is a bounded linear operator in $\mathcal{L}(L_2(t_0, t_0+T; U), H)$. Notice that the periodic system is NCVE if and only if $(S(T+t_0,t_0), B_d)$ is NCVE. Then by Theorem 3.4 the periodic system is NCVE if and only if

(a) it is null controllable on some interval $[t_0, \tau]$, and

(b) $|\lambda| \le 1$ for any $\lambda \in \sigma(S(T+t_0,t_0))$.

Suppose $S(t,s)$ is a two-parameter group so that $S(T+t_0,t_0)$ is boundedly invertible.

LEMMA 5.4. *The periodic system* (5.1) *is ECVE if and only if the discrete-time system* (5.2) *is ECVE.*

Proof. Consider the controllability operator

$$P_L x = \int_{t_0}^{L} S(L,r)B(r)B(r)^*S^*(L,r)x\,dr.$$

Let $KT+t_0 \le L < (K+1)T+t_0$. Then $\alpha P_{(K+1)T} \ge P_L \ge \beta P_{KT}$ for some $\alpha > 0$ and $\beta > 0$. Hence $P_L^{-1} \to 0$ strongly if and only if $(P_{KT})^{-1} \to 0$ strongly. By a periodic version of Lemma 1.5 the assertion follows. □

LEMMA 5.5.

$$S(T+t_0,t_0) = S(t_0,0)S(T,0)S(t_0,0)^{-1}$$

and $\sigma(S(T+t_0,t_0)) = \sigma(S(T,0))$.

Suppose further $S(T+t_0,t_0)^{-1}$ satisfies Hypotheses 3 and 4. Then from Theorem 4.4 we obtain the following.

THEOREM 5.6. *The periodic system* (5.1) *is ECVE if and only if*

(a) *it is exactly controllable on some interval* $[t_0,\tau]$, *and*

(b) $|\lambda| = 1$ *for any* $\lambda \in \sigma(S(T,0))$.

Acknowledgment. The author thanks the anonymous referees for careful reading of the mauscript and many helpful comments on the paper.

REFERENCES

[1] F. M. CALLIER AND C. A. DESOER, *Linear System Theory*, Springer-Verlag, Berlin, 1991.

[2] T. CHEN AND B. A. FRANCIS, *Optimal Sampled-Data Control Systems*, Springer-Verlag, London, 1995.

[3] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Informa. Sci. 8, Springer-Verlag, Berlin, 1978.

[4] A. ICHIKAWA AND H. KATAYAMA, *Linear Time Varying Systems and Sampled-Data Systems*, Lecture Notes in Control and Inform. Sci. 265, Springer-Verlag, London, 2001.

[5] C. S. KUBRUSLY, *Mean square stability for discrete bounded linear systems in Hilbert space*, SIAM J. Control Optim., 23 (1985), pp. 19–29.

[6] K. Y. LEE, S.-N. CHOW, AND R. O. BARR, *On the control of discrete-time distributed parameter systems*, SIAM J. Control Optim., 10 (1972), pp. 361–376.

[7] J. M. A. M. VAN NEERVEN, *Null controllability and the algebraic Riccati equation in Banach spaces*, SIAM J. Control Optim., 43 (2005), pp. 1313–1327.

[8] E. PRIOLA AND J. ZABCZYK, *Null controllability with vanishing energy*, SIAM J. Control Optim., 42 (2003), pp. 1013–1032.

[9] J. ZABCZYK, *Remarks on the control of discrete-time distributed parameter systems*, SIAM J. Control Optim., 12 (1974), pp. 721–735.

[10] J. ZABCZYK, *On optimal stochastic control of discrete-time systems in Hilbert space*, SIAM J. Control Optim., 13 (1975), pp. 1217–1234.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 694–713

IDENTIFIABILITY OF PIECEWISE CONSTANT CONDUCTIVITY IN A HEAT CONDUCTION PROCESS*

SEMION GUTMAN† AND JUNHONG HA‡

Abstract. We study the identification and identifiability problems for heat conduction in a nonhomogeneous rod. The identifiability results are established for two different sets of observations. Given a sequence of distributed type observations, the identifiability is proved for conductivities in a piecewise smooth class of functions. In the case of observations taken at finitely many points the identifiability is established for piecewise constant conductivities. Such conductivities can be uniquely identified using the proposed marching algorithm.

Key words. identification, identifiability, piecewise constant conductivity

AMS subject classifications. 35R30, 93B30

DOI. 10.1137/060657364

1. Introduction.

Consider the heat conduction in a nonhomogeneous insulated rod of a unit length, with the ends kept at zero temperature at all times. Our main interest is in the identification and identifiability of the discontinuous conductivity (thermal diffusivity) coefficient $a(x)$, $0 \le x \le 1$. The identification problem consists of finding a conductivity $a(x)$ in an admissible set K for which the temperature $u(x,t)$ fits given observations in a prescribed sense. Under a wide range of conditions one can establish the continuity of the objective function $J(a)$ representing the best fit to the observations. Then the existence of the best fit to data conductivity follows if the admissible set K is compact in the appropriate topology. However, such an approach usually does not guarantee the uniqueness of the found conductivity $a(x)$. Establishing such a uniqueness is referred to as the identifiability problem. If the conductivity is identifiable and one can design an algorithm for its reconstruction, then we say that a is constructively identifiable.

From physical considerations the conductivity coefficients $a(x)$ are assumed to be in

$$A_{\rm ad} = \{a \in L^\infty(0,1) : 0 < \nu \le a(x) \le \mu\}. \tag{1.1}$$

The temperature $u(a) = u(x,t;a)$ inside the rod satisfies

$$\begin{array}{ll} u_t - (a(x)u_x)_x = 0, & Q = (0,1)\times(0,T), \\ u(0,t) = u(1,t) = 0, & t \in (0,T), \\ u(x,0) = g(x), & x \in (0,1), \end{array} \tag{1.2}$$

where $g \in L^2(0,1)$. In general, the solution of (1.2) is understood in the weak sense. According to [11] for any $a \in A_{\rm ad}$ there exists a unique weak solution $u(a) \in L^2(0,1;H_0^1(0,1)) \cap C([0,1];L^2(0,1))$, and so the map $a \to u(a)$ is well defined. Moreover, this map is continuous from $A_{\rm ad}$ equipped with the $L^2(0,1)$ topology into

*Received by the editors April 15, 2006; accepted for publication (in revised form) January 25, 2007; published electronically May 7, 2007.
http://www.siam.org/journals/sicon/46-2/65736.html

†Department of Mathematics, The University of Oklahoma, Norman, Oklahoma 73019 (sgutman@ou.edu).

‡School of Liberal Arts, Korea University of Technology and Education, Cheonan 330-708, South Korea (hjh@kut.ac.kr).

$C([0,T];L^2(0,1))$. In fact, in [11] these results are established for multidimensional parabolic problems.

The identification (parameter estimation) problem for (1.2) is as follows: Find a conductivity $a \in A_{\rm ad}$ such that the solution $u(a)$ of (1.2) fits a given observation z of the heat conduction process. For example, given $z \in L^2(0,1)$ one defines

$$J(a) = \|u(x,T;a) - z(x)\|_{L^2(0,1)}. \tag{1.3}$$

Then the parameter estimation problem for (1.2) is reduced to the minimization of the objective function J over the admissible set $A_{\rm ad}$ or its subset $K_{\rm ad}$: Find $\overline{a} \in K_{\rm ad} \subset A_{\rm ad}$ such that

$$J(\bar{a}) = \inf\{J(a) : a \in K_{\rm ad}\}. \tag{1.4}$$

The above-mentioned properties of the solutions $u(a)$ imply that the objective function $J(a)$ is continuous on $A_{\rm ad} \cap L^2(0,1)$. Therefore the identification problem (1.4) has a solution if $A_{\rm ad}$ is compact in $L^2(0,1)$. One such choice is $K_{\rm ad} = \{a \in A_{\rm ad} \cap H^1(0,1) : \|a\|_{H^1} \le \text{constant}\}$; see [8]. However, $a \in H^1(0,1)$ implies that the conductivity is continuous. Therefore this choice of $K_{\rm ad}$ is not suitable for the study of the identification problems with discontinuous coefficients.

To overcome this difficulty we have shown in [5] (in a multidimensional case) that one can take for $K_{\rm ad}$ the set of functions in $A_{\rm ad}$ which have a uniformly bounded variation. Such a set $K_{\rm ad}$ is compact in $L^2(0,1)$, and the existence of solutions to the identification problem (1.4) follows. See [5] for additional details and numerical experiments for 2D parameter identification problems. A variety of identification problems is studied in [1] under very general assumptions on the problem's parameters.

The identifiability questions for partial differential equations are much more difficult, and there are just a few available results. Suppose that one is given an observation $z(t) = u(p,t;a)$ of the heat conduction process (1.2) for $t_1 < t < t_2$ at some observation point $0 < p < 1$. From the series solution for (1.2) and the uniqueness of the Dirichlet series expansion (see section 2), one can, in principle, recover all of the eigenvalues of the associated Sturm–Liouville problem. If one also knows the eigenvalues for the heat conduction process with the same coefficient a and different boundary conditions, then the classical results of Gelfand and Levitan [4] show that smooth coefficients $a(x)$ can be uniquely identified from the knowledge of the two spectral sequences. Also, if the entire spectral function is known (i.e., the eigenvalues and the values of the derivatives of the normalized eigenfunctions at $x = 0$), then the conductivity is identifiable as well. However, such results have little practical value, since the observation data $z(t)$ always contain some noise, and therefore one cannot hope to adequately identify more than just the few first eigenvalues of the problem.

A different approach is taken in [6, 12, 13, 14]. These works show that one can identify a constant conductivity a in (1.2) from the measurement $z(t)$ taken at one point $p \in (0,1)$. These works also discuss problems more general than (1.2), including problems with a broad range of boundary conditions, nonzero forcing functions, as well as elliptic and hyperbolic problems. In [7, 3] and references therein identifiability results are obtained for elliptic and parabolic equations with discontinuous parameters in a multidimensional setting. A typical assumption there is that one knows the normal derivative of the solution at the boundary of the region for every Dirichlet boundary input.

The main result of this paper is contained in Theorem 4.6. This theorem describes and justifies the marching algorithm for the unique identification of piecewise constant conductivities from observations of (1.2) given at finitely many points $p_k \in (0,1)$. We

start by recalling some basic properties of (1.2) in section 2. Identifiability results for countably many distributed observations are given in section 3. Identifiability of piecewise constant conductivities a is discussed in section 4. Numerical results for the identifiability algorithms described in this paper require an extensive exposition, and they will be presented elsewhere.

2. Auxiliary results. In this section we collect some well-known results for the solutions $u(x,t;a)$ of (1.2), as well as for its associated Sturm–Liouville problem. Since such results are scattered in the literature, some brief proof outlines are included as well. See [2, 9, 10, 11] for a detailed discussion.

DEFINITION 2.1. *Function $a(x)$ is said to belong to the class $\mathcal{PS}_N$ if*

(i) *$a \in A_{\rm ad} = \{a \in L^\infty(0,1) : 0 < \nu \le a(x) \le \mu\}$ for some positive constants ν and μ;*

(ii) *function a is piecewise smooth; that is, there exists a finite sequence of points $0 = x_0 < x_1 < \cdots < x_{N-1} < x_N = 1$ such that both $a(x)$ and $a'(x)$ are continuous on every open subinterval (x_i, x_{i+1}), $i = 0, \dots, N-1$, and both can be continuously extended to the closed intervals $[x_i, x_{i+1}]$, $i = 0, \dots, N-1$. For definiteness, we assume that a and a' are continuous from the right, i.e., $a(x) = a(x+)$ and $a'(x) = a'(x+)$ for all $x \in [0,1)$. Also let $a(1) = a(1-)$.*

DEFINITION 2.2. $\mathcal{PS} = \cup_{N=1}^\infty \mathcal{PS}_N$.

Everywhere in the following the conductivities a are assumed to be in $\mathcal{PS}$. If $a \in \mathcal{PS}_N$, then the regularity conditions on a and the uniqueness of the weak solutions imply that for any $t > 0$ the weak solution $u(x,t;a)$ of (1.2) satisfies the equation in the classical sense on any subinterval (x_i, x_{i+1}), $i = 0, \dots, N-1$. Also u satisfies the matching conditions for the continuity of the solution and its conormal derivative at $x_i \in (0,1)$, $i = 1, 2, \dots, N-1$:

$$
\begin{aligned}
&u_t - (a(x)u_x)_x = 0, \quad x \ne x_i, \quad t \in (0,T),\\
&u(0,t) = u(1,t) = 0, \quad t \in (0,T),\\
&u(x_i+,t) = u(x_i-,t),\\
&a(x_i+)u_x(x_i+,t) = a(x_i-)u_x(x_i-,t),\\
&u(x,0) = g(x), \quad x \in (0,1),
\end{aligned}
\tag{2.1}
$$

where $g \in L^2(0,1)$; see [11, 16].

Denote by $\|\cdot\|$, $\langle \cdot, \cdot \rangle$ the norm and the inner product, respectively, in $H = L^2(0,1)$.

THEOREM 2.3. *Let $a \in \mathcal{PS}$. Then*

(i) *the associated Sturm–Liouville problem*

$$
\begin{aligned}
&(a(x)v(x)')' = -\lambda v(x), \quad x \ne x_i,\\
&v(0) = v(1) = 0,\\
&v(x_i+) = v(x_i-),\\
&a(x_i+)v_x(x_i+) = a(x_i-)v_x(x_i-)
\end{aligned}
\tag{2.2}
$$

has infinitely many eigenvalues

$$0 < \lambda_1 < \lambda_2 < \cdots \to \infty.$$

The eigenvalues $\{\lambda_k\}_{k=1}^\infty$ and the corresponding orthonormal set of eigenfunctions $\{v_k\}_{k=1}^\infty$ satisfy

(2.3)

$$\lambda_k = \inf \left\{ \frac{\int_0^1 a(x)[v'(x)]^2 dx}{\int_0^1 [v(x)]^2 dx} \ : \ v \in H_0^1(0,1), \ \langle < v, v_j \rangle = 0, \ j = 1, 2, \dots, k-1 \right\},$$

$$\lambda_k = \int_0^1 a(x)[v_k'(x)]^2 dx. \tag{2.4}$$

The normalized eigenfunctions $\{v_k\}_{k=1}^{\infty}$ *form a basis in* $L^2[0,1]$.

(ii) *Each eigenvalue is simple. For each eigenvalue* λ_k *there exists a unique continuous, piecewise smooth normalized eigenfunction* $v_k(x)$ *such that* $v_k'(0+) > 0$, *and the function* $a(x)v_k'(x)$ *is continuous on* $[0,1]$.

(iii) *Eigenvalues* $\{\lambda_k\}_{k=1}^{\infty}$ *satisfy the inequality*

$$\nu\pi^2 k^2 \le \lambda_k \le \mu\pi^2 k^2.$$

(iv) *The first eigenfunction* v_1 *satisfies* $v_1(x) > 0$ *for any* $x \in (0,1)$.

(v) *The first eigenfunction* v_1 *has a unique point of maximum* $q \in (0,1)$: $v_1(x) < v_1(q)$ *for any* $x \ne q$.

(vi) *For any fixed* $t > 0$ *the solution* u *of* (2.1) *is given by*

$$u(x,t;a) = \sum_{k=1}^{\infty} \langle g, v_k\rangle e^{-\lambda_k t} v_k(x),$$

and the series converges uniformly and absolutely on $[0,1]$.

(vii) *For any* $p \in (0,1)$ *the function*

$$z(t) = u(p,t;a), \quad t > 0,$$

is real analytic on $(0,\infty)$.

Proof. (i) The proof is standard; see, e.g., [10].

(ii) On any subinterval (x_i, x_{i+1}) the coefficient $a(x)$ has a bounded continuous derivative. Therefore, on any such interval the initial value problem $(a(x)v'(x))'+\lambda v = 0,\ v(x_i) = A,\ v'(x_i) = B$ has a unique solution. Suppose that two eigenfunctions $w_1(x)$ and $w_2(x)$ correspond to the same eigenvalue λ_k. Then they both satisfy the condition $w_1(0) = w_2(0) = 0$. Therefore their Wronskian is equal to zero at $x = 0$. Consequently, the Wronskian is zero throughout the interval (x_0, x_1), and the solutions are linearly dependent there. Thus $w_2(x) = Cw_1(x)$ on (x_0, x_1), $w_2(x_1-) = Cw_1(x_1-)$, and $w_2'(x_1-) = Cw_1'(x_1-)$. The linear matching conditions imply that $w_2(x_1+) = Cw_1(x_1+)$ and $w_2'(x_1+) = Cw_1'(x_1+)$. The uniqueness of solutions implies that $w_2(x) = Cw_1(x)$ on (x_1, x_2), etc. Thus $w_2(x) = Cw_1(x)$ on $(0,1)$, and each eigenvalue λ_k is simple. In particular λ_1 is a simple eigenvalue. The uniqueness and the matching conditions also imply that any solution of $(a(x)v'(x))' + \lambda v = 0,\ v(0) = 0,\ v'(0) = 0$ must be identically equal to zero on the entire interval $(0,1)$. Thus no eigenfunction $v_k(x)$ satisfies $v_k'(0) = 0$. Assuming that the eigenfunction v_k is normalized in $L^2(0,1)$, it leaves us with the choice of its sign for $v_k'(0)$. Letting $v_k'(0) > 0$ makes the eigenfunction unique.

(iii) The eigenvalues of (2.3) satisfy the min-max principle

$$\lambda_k = \min_{V_k} \max \left\{ \frac{\int_0^1 a(x)[v'(x)]^2 dx}{\int_0^1 [v(x)]^2 dx} \ : \ v \in V_k \right\},$$

where V_k varies over all subspaces of $H_0^1(0,1)$ of finite dimension k; see [10]. Therefore $a(x) \le b(x),\ x \in [0,1]$ implies $\lambda_k^{(a)} \le \lambda_k^{(b)}$. Since the eigenvalues of (2.3) with $a(x) = 1$ are $\pi^2 k^2$, the required inequality follows.

(iv) Recall that $v_1(x)$ is a continuous function on $[0,1]$. Suppose that there exists $p \in (0,1)$ such that $v_1(p) = 0$. Let $w_l(x) = v_1(x)$ for $0 \le x < p$ and $w_l(x) = 0$ for $p \le x \le 1$. Let $w_r(x) = v_1(x) - w_l(x),\ x \in [0,1]$. Then w_l, w_r are continuous, and, moreover, $w_l, w_r \in H_0^1(0,1)$. Also

$$\int_0^1 w_l(x)w_r(x)dx = 0 \quad \text{and} \quad \int_0^1 a(x)w_l'(x)w_r'(x)dx = 0.$$

Suppose that w_l is not an eigenfunction for λ_1. Then

$$\int_0^1 a(x)[w_l'(x)]^2dx > \lambda_1 \int_0^1 [w_l(x)]^2dx.$$

Since

$$\int_0^1 a(x)[w_r'(x)]^2dx \ge \lambda_1 \int_0^1 [w_r(x)]^2dx,$$

we have

$$\lambda_1 = \frac{\int_0^1 a(x)[v_1'(x)]^2dx}{\int_0^1 [v_1(x)]^2dx} = \frac{\int_0^1 a(x)([w_l'(x)]^2 + [w_r'(x)]^2)dx}{\int_0^1 ([w_l(x)]^2 + [w_r(x)]^2)dx}$$

$$> \frac{\int_0^1 (\lambda_1[w_l(x)]^2 + \lambda_1[w_r(x)]^2)dx}{\int_0^1 ([w_l(x)]^2 + [w_r(x)]^2)dx} = \lambda_1.$$

This contradiction implies that w_l (and w_r) must be an eigenfunction for λ_1. However, $w_l(x) = 0$ for $p \le x \le 1$, and as in (ii) it implies that $w_l(x) = 0$ for all $x \in [0,1]$, which is impossible. Since $v_1'(0) > 0$ the conclusion is that $v_1(x) > 0$ for $x \in (0,1)$.

(v) From part (i), any eigenfunction v_k is continuous and satisfies

$$(a(x)v_k'(x))' = -\lambda_k v_k(x)$$

for $x \ne x_i$. Also the function $a(x)v_k'(x)$ is continuous on $[0,1]$ because of the matching conditions at the points of discontinuity $x_i,\ i = 1, 2, \dots, N-1$ of a. The integration gives

$$a(x)v_k'(x) = a(p)v_k'(p) - \lambda_k \int_p^x v_k(s)ds$$

for any $x, p \in (0,1)$.

Let $p \in (0,1)$ be a point of maximum of v_k. If $p \ne x_i$, then $v_k'(p) = 0$. If $p = x_i$, then $v_k'(x_i-) \ge 0$ and $v_k'(x_i+) \le 0$. Therefore $\lim_{x\to p} a(x)v_k'(x) = 0$ and $v_k'(p+) = v_k'(p-) = 0$ since $a(x) \ge \nu > 0$. In any case for such a point p we have

$$a(x)v_k'(x) = -\lambda_k \int_p^x v_k(s)ds, \quad x \in (0,1). \tag{2.5}$$

Since $v_1(x) > 0,\ a(x) > 0$ on $(0,1)$, (2.5) implies that $v_1'(x) > 0$ for any $0 \le x < p$ and $v_1'(x) < 0$ for any $p < x \le 1$. Since the derivative of v_1 is zero at any point of maximum, we have to conclude that such a maximum p is unique.

(vi) We prove only the convergence part. Note that

$$\nu\|v_k'\|^2 \le \int_0^1 a(x)[v_k'(x)]^2 dx = \lambda_k \|v_k\|^2 = \lambda_k.$$

Thus

$$\|v_k'\| \le \frac{\sqrt{\lambda_k}}{\sqrt{\nu}}$$

and

$$|v_k(x)| \le \int_0^x |v_k'(s)| ds \le \|v_k'\| \le \frac{\sqrt{\lambda_k}}{\sqrt{\nu}}.$$

Bessel's inequality implies that the sequence of Fourier coefficients $\langle g, v_k\rangle$ is bounded. Therefore, denoting by C various constants and using the fact that the function $s \to \sqrt{s}e^{-\sigma s}$ is bounded on $[0,\infty)$ for any $\sigma > 0$, one gets

$$|\langle g, v_k\rangle e^{-\lambda_k t} v_k(x)| \le C\frac{\sqrt{\lambda_k}}{\sqrt{\nu}} e^{-\frac{\lambda_k t}{2}} e^{-\frac{\lambda_k t}{2}} \le C e^{-\frac{\lambda_k t}{2}}.$$

From (iii) of this theorem $\lambda_k \ge \nu\pi^2 k^2$. Thus

$$\sum_{k=1}^{\infty} |\langle g, v_k\rangle e^{-\lambda_k t} v_k(x)| \le C\sum_{k=1}^{\infty} e^{-\frac{\nu\pi^2 k^2 t}{2}} \le C\sum_{k=1}^{\infty} \left(e^{-\frac{\nu\pi^2 t}{2}}\right)^k < \infty.$$

(vii) Let $t_0 > 0$ and $p \in (0,1)$. From (vi), the series $\sum_{k=1}^{\infty}\langle g, v_k\rangle e^{-\lambda_k t_0} v_k(p)$ converges absolutely. Therefore $\sum_{k=1}^{\infty}\langle g, v_k\rangle e^{-\lambda_k s} v_k(p)$ is analytic in the part of the complex plane $\{s \in \mathbf{C} : \text{Re } s > t_0\}$, and the result follows. □

Series of the form $\sum_{k=1}^{\infty} C_k e^{-\lambda_k t}$ are known as the Dirichlet series. The following lemma shows that the Dirichlet series representation of a function is unique.

LEMMA 2.4. *Let $\mu_k > 0$, $k = 1, 2, \ldots$, be a strictly increasing sequence. Suppose that $T_1 \ge 0$ and $\sum_{k=1}^{\infty} |C_k| < \infty$. If*

$$\sum_{k=1}^{\infty} C_k e^{-\mu_k t} = 0 \quad \textit{for all} \quad t \in (T_1, T_2),$$

then $C_k = 0$ for $k = 1, 2, \ldots$.

The result follows at once from the observation that the series $\sum_{k=1}^{\infty} C_k e^{-\mu_k z}$ converges uniformly in the Re $z > 0$ region of the complex plane, implying that it is an analytic function there. See Chapter 9 of [15] for additional results on Dirichlet series.

Remark. According to Theorem 2.3(vi) for each fixed $p \in (0,1)$ the solution $z(t) = u(p,t;a)$ of (2.1) is given by a Dirichlet series. However, Lemma 2.4 is not directly applicable since the coefficients $C_k = \langle g, v_k\rangle v_k(p)$ are only square summable. Nevertheless, the conclusion of Lemma 2.4 remains valid, since the exponents μ_k in the Dirichlet series are the eigenvalues λ_k which satisfy the growth condition stated in Theorem 2.3(iii). This allows one to conclude (Theorem 2.3(vii)) that the solution $z(t)$ is a real analytic function on $(0,\infty)$, and the uniqueness of such a representation follows. Thus it would be a mistake to simply refer to the standard results such as Lemma 2.4 for the uniqueness of the Dirichlet series representation to justify the paper's conclusions.

3. Identifiability by distributed measurements. Suppose that one is given some observations of the heat conduction process (2.1) with an unknown conductivity $a(x)$ and that they coincide with the observations of the model process

$$\begin{aligned} &u_t^m - (a^m(x)u_x^m)_x = 0, \quad x \neq x_i^m, \quad t \in (0,T),\\ &u^m(0,t) = u^m(1,t) = 0, \quad t \in (0,T),\\ &u^m(x_i^m+,t) = u^m(x_i^m-,t),\\ &a^m(x_i^m+)u_x^m(x_i^m+,t) = a^m(x_i^m-)u_x^m(x_i^m-,t),\\ &u^m(x,0) = g(x), \quad x \in (0,1), \end{aligned} \tag{3.1}$$

where g is the same as in (2.1). The conductivity a is said to be identifiable in some class of functions $\mathcal{M}$ if the coincidence of the measurements of the observed and the model processes implies that $a = a^m$, provided $a, a^m \in \mathcal{M}$.

THEOREM 3.1. *Let $\{\psi_n\}_{n=1}^\infty$ be a complete orthonormal set in $H = L^2(0,1)$. Suppose that nonzero initial data $g \in H$ and the observations $z_n(t) = \langle u(x,t;a), \psi_n\rangle$ for $n = 1,2,\dots$ and $0 \le T_1 < t < T_2$ of the heat conduction process (2.1) are given. Then the conductivity $a(x) \in A_{\rm ad}$ is constructively identifiable in the class of piecewise smooth functions $\mathcal{PS}$.*

Proof. To show the identifiability of a we give an algorithm for its reconstruction from the data $z_n(t)$, $n = 1,2,\dots$, guaranteeing the uniqueness in each step. Using Theorem 2.3(vi) we have

$$z_n(t) = \sum_{k=1}^{\infty} \langle g, v_k\rangle e^{-\lambda_k t}\langle v_k, \psi_n\rangle \tag{3.2}$$

for each $n = 1,2,\dots$ and $0 \le T_1 < t < T_2$.

Fix an $n > 0$. Since $\psi_n \in H$ and $\{v_k\}_{k=1}^\infty$ form a basis in H, the Bessel inequality implies that the sequence of the Fourier coefficients $\{\langle v_k, \psi_n\rangle\}_{k=1}^\infty$ is bounded. From Theorem 2.3(vi) one concludes that the above series converges absolutely. Note that some products $\langle g, v_k\rangle\langle v_k, \psi_n\rangle$ may be equal to zero. The zero value products present a difficulty, since we would not know how to associate the sequence of exponents recovered from (3.2) with the (unknown) eigenvalues λ_k: Some eigenvalues may be missing from the sequence. Define (possibly empty) subsets $Q_n \subset \mathbf{N}$ by

$$Q_n = \{k \in \mathbf{N} \ : \ \langle g, v_k\rangle\langle v_k, \psi_n\rangle \neq 0\}, \quad n = 1,2,\dots.$$

For $Q_n \neq \emptyset$ reindex (3.2) so that there would be no vanishing coefficients:

$$z_n(t) = \sum_{l=1}^{\infty} C_{l,n} e^{-\mu_{l,n} t}, \quad t \in (T_1, T_2). \tag{3.3}$$

By Theorem 2.3(vii) the solutions $z_n(t)$ are real analytic. Therefore, since all of the coefficients $C_{l,n} \neq 0$, one can uniquely determine them and the sequences $\mu_{l,n}$, $l = 1,2,\dots$. Recall that $\{\mu_{l,n}\}_{l=1}^\infty \subset \{\lambda_k\}_{k=1}^\infty$ for any n with a nonempty Q_n so each $\mu_{l,n} \ge \lambda_1 > 0$. For each n such that $Q_n \neq \emptyset$ let

$$\gamma_n = \min\{\mu_{l,n} \ : \ l \in \mathbf{N}\}.$$

Define

$$\gamma = \min\{\gamma_n \ : \ Q_n \neq \emptyset\}$$

and

$$A_n = \begin{cases} C_{1,n} & \text{if} \quad \gamma_n = \gamma, \quad Q_n \neq \emptyset, \\ 0 & \text{if} \quad \gamma_n > \gamma, \quad Q_n \neq \emptyset, \\ 0 & \text{if} \quad Q_n = \emptyset. \end{cases} \tag{3.4}$$

Let

$$w(x) = \sum_{n=1}^{\infty} A_n \psi_n(x). \tag{3.5}$$

We claim that w is a nonzero multiple of some eigenfunction v_J of (2.1). Indeed, let J be the smallest index for which $\langle g, v_J \rangle \neq 0$. Such an index exists since $g \neq 0$, and the eigenfunctions form a basis in H. Now, since $v_J \neq 0$ and $\{\psi_n\}_{n=1}^{\infty}$ also form a basis in H, there exists $n \in \mathbf{N}$ such that $\langle v_J, \psi_n \rangle \neq 0$. Thus $\langle g, v_J \rangle \langle v_J, \psi_n \rangle \neq 0$ and $\gamma \leq \lambda_J$. The choice of J implies that $\gamma = \lambda_J$. By the definition $A_n = \langle g, v_J \rangle \langle v_J, \psi_n \rangle$ for nonzero products. Therefore

$$\begin{aligned} w(x) &= \sum_{n=1}^{\infty} \langle g, v_J \rangle \langle v_J, \psi_n \rangle \psi_n(x) \\ &= \langle g, v_J \rangle \sum_{n=1}^{\infty} \langle v_J, \psi_n \rangle \psi_n(x) = \langle g, v_J \rangle v_J(x) \end{aligned}$$

as claimed.

Now we show that the set of points $y \in (0,1)$ where $w'(y) = 0$ is finite. Assuming the opposite, and since w is a nonzero multiple of v_J, there exists a sequence $y_j \in (0,1)$ such that $v_J'(y_j) = 0$ and $\lim_{j \to \infty} y_j = c \in [0,1]$. The continuity of $a(x)v_J'(x)$ implies that $v_J'(c) = 0$. From $(a(x)v_J'(x))' = -\gamma v_J(x)$ one gets

$$0 = a(y_{j+1})v_J'(y_{j+1}) - a(y_j)v_J'(y_j) = -\gamma \int_{y_j}^{y_{j+1}} v_J(s)ds \tag{3.6}$$

and concludes that v_J cannot be strictly positive or strictly negative on (y_j, y_{j+1}). Let $\zeta_j \in (y_j, y_{j+1})$ be such that $v_J(\zeta_j) = 0$. Then $\lim_{j \to \infty} \zeta_j = c \in [0,1]$ and $v_J(c) = 0$. Now we have both $v_J(c) = 0$ and $v_J'(c) = 0$. But then the uniqueness of the Cauchy problem for the second order linear equations, and the matching conditions (see the proof of Theorem 2.3(ii)) imply that $v_J(x) = 0$ for all $x \in [0,1]$, which is impossible.

Let q be a point of maximum of w. Note that w may have several maxima, unless it is a multiple of v_1. In any case, (2.5) implies

$$a(x)w'(x) = -\gamma \int_q^x w(s)ds, \quad x \in (0,1). \tag{3.7}$$

Then, outside of the finite sets $\{x_i\}$ and $\{y_j\}$, the conductivity $a(x)$ is uniquely determined from (3.7) by

$$a(x) = -\gamma \frac{\int_q^x w(s)ds}{w'(x)}.$$

Because a is assumed to be in $\mathcal{PS}$, it can be uniquely extended to the entire interval $[0,1]$. □

Remark 1. In an application one can choose $\psi_n = \sqrt{2}\sin \pi n x$ and the initial condition $g(x) > 0$ on $(0,1)$. Then $\langle g, v_1\rangle\langle v_1, \psi_1\rangle \neq 0$, since $v_1(x) > 0$ on $(0,1)$ by Theorem 2.3(iv). Thus $J = 1$ in this case, and $w(x) = \langle g, v_1\rangle v_1(x)$ in the above algorithm. Also, one can see from (3.6) that there is only one point $y_1 = q \in (0,1)$ where $v_1'(q) = 0$, and it is the point of maximum of $v_1(x)$ on $(0,1)$. Indeed, if there were two such points, then by (3.6) $v_1(x)$ would have to become negative between them, which would contradict $v_1(x) > 0$ on $(0,1)$.

Remark 2. Since the system $\{\psi_n\}_{n=1}^{\infty}$ is complete, the conditions of Theorem 3.1 imply that for any $t > 0$ one knows $u(x,t;a)$ almost everywhere on $[0,1]$. Since $u(x,t;a)$ is continuous in x and analytic in t, Theorem 3.1, in fact, assumes that the solution is known in $[0,1]\times(0,T)$ or (equivalently) in $[0,1]\times(0,\infty)$. Thus, Theorem 3.1 can be stated under any of these conditions. However, a practical reconstruction of the conductivity a directly from the equation $u_t = (au_x)_x$ is extremely unstable. The algorithm presented in the above theorem does not reconstruct the entire solution u but just the first eigenfunction of the associated Sturm–Liouville problem. Its stability properties will be studied elsewhere.

4. Identifiability of piecewise constant conductivities from finitely many observations. The identifiability is sought within the following set.

DEFINITION 4.1. *Let $\mathcal{PC} \subset \mathcal{PS}$ be the class of piecewise constant conductivities, and let $\mathcal{PC}_N = \mathcal{PC} \cap \mathcal{PS}_N$.*

Functions $a \in \mathcal{PC}_N$ have the form $a(x) = a_i$ for $x \in [x_{i-1}, x_i)$, $i = 1, 2, \dots, N$. In this case the governing system (2.1) is

$$\begin{aligned} &u_t - a_i u_{xx} = 0, \quad x \in (x_{i-1}, x_i), \quad t \in (0,T),\\ &u(0,t) = u(1,t) = 0, \quad t \in (0,T),\\ &u(x_i+,t) = u(x_i-,t),\\ &a_{i+1}u_x(x_i+,t) = a_i u_x(x_i-,t),\\ &u(x,0) = g(x), \quad x \in (0,1), \end{aligned} \tag{4.1}$$

where $g \in L^2(0,1)$ and $i = 1, 2, \dots, N-1$. The associated Sturm–Liouville problem is

$$\begin{aligned} &a_i v''(x) = -\lambda v(x), \quad x \in (x_{i-1}, x_i),\\ &v(0) = v(1) = 0,\\ &v(x_i+) = v(x_i-),\\ &a_{i+1}v'(x_i+) = a_i v'(x_i-) \end{aligned} \tag{4.2}$$

for $i = 1, 2, \dots, N-1$.

We are interested only in the first eigenfunction v_1 of (4.2). Let λ_1 be the first eigenvalue. Suppose that $p^* \in (x_{i-1}, x_i)$. Then v_1 can be expressed on (x_{i-1}, x_i) as

$$v_1(x) = A\cos\left(\sqrt{\frac{\lambda_1}{a_i}}(x - p^*) + \gamma\right), \quad A > 0, \quad -\frac{\pi}{2} < \gamma < \frac{\pi}{2}.$$

The range for γ in the above representation follows from the fact that $v_1(p^*) = A\cos\gamma > 0$ by Theorem 2.3(iv).

The identifiability of piecewise constant conductivities is based on the following three lemmas.

LEMMA 4.2. *Suppose that $\delta > 0$. Assume $Q_1, Q_3 \geq 0$, $Q_2 > 0$, and $0 < Q_1 + Q_3 < 2Q_2$. Let*

$$\Gamma = \left\{(A,\omega,\gamma) \ : \ A > 0, \quad 0 < \omega < \frac{\pi}{2\delta}, \quad -\frac{\pi}{2} < \gamma < \frac{\pi}{2}\right\}.$$

Then the system of equations

$$A\cos(\omega\delta-\gamma)=Q_1,\quad A\cos\gamma=Q_2,\quad A\cos(\omega\delta+\gamma)=Q_3$$

has a unique solution $(A,\omega,\gamma)\in\Gamma$ *given by*

$$\omega=\frac{1}{\delta}\arccos\frac{Q_1+Q_3}{2Q_2},\quad \gamma=\arctan\left(\frac{Q_1-Q_3}{2Q_2\sin\omega\delta}\right),\quad A=\frac{Q_2}{\cos\gamma}.$$

Proof. For $(A,\omega,\gamma)\in\Gamma$ one has $A>0$ and $\cos\gamma>0$. Therefore

$$\frac{\cos(\omega\delta-\gamma)}{\cos\gamma}=\cos(\omega\delta)+\sin(\omega\delta)\frac{\sin\gamma}{\cos\gamma}=\frac{Q_1}{Q_2}, \tag{4.3}$$

$$\frac{\cos(\omega\delta+\gamma)}{\cos\gamma}=\cos(\omega\delta)-\sin(\omega\delta)\frac{\sin\gamma}{\cos\gamma}=\frac{Q_3}{Q_2}. \tag{4.4}$$

Adding (4.3) and (4.4) yields

$$\cos\omega\delta=\frac{Q_1+Q_3}{2Q_2}.$$

Since $0<\omega\delta<\frac{\pi}{2}$ and $0<(Q_1+Q_3)/2Q_2<1$ the above equation is uniquely solvable. Now subtracting (4.4) from (4.3) yields

$$\tan\gamma=\frac{Q_1-Q_3}{2Q_2\sin\omega\delta},$$

which is also uniquely solvable, since $-\pi/2<\gamma<\pi/2$. Finally, we have $A=Q_2/\cos\gamma$. □

LEMMA 4.3. *Suppose that* $\delta>0,\ 0<p\le x_1<p+\delta<1,\ 0<\omega_1,\omega_2<\pi/2\delta$. *Let* $w(x),\ v(x),\ x\in[p,p+\delta]$ *be such that*

$$w(x)=A_1\cos\omega_1x+B_1\sin\omega_1x,$$
$$v(x)=A_2\cos\omega_2x+B_2\sin\omega_2x.$$

Suppose that

$$v(x_1)=w(x_1),\quad \omega_1^2v'(x_1)=\omega_2^2w'(x_1),$$
$$v'(x_1)>0,\quad v(x_1)>0.$$

Then

(i) *conditions* $v(p+\delta)=w(p+\delta),\ v'(p+\delta)\ge0$, *and* $\omega_1\le\omega_2$ *imply* $\omega_1=\omega_2$;

(ii) *conditions* $v(p+\delta)=w(p+\delta),\ w'(p+\delta)\ge0$, *and* $\omega_1\ge\omega_2$ *imply* $\omega_1=\omega_2$.

Proof. Since $v(x_1)>0,\ v'(x_1)>0$ we have

$$v(x)=A\sin[\omega_2(x-x_1)+\gamma],\quad 0<\gamma<\frac{\pi}{2},$$

where $A>0$. The matching conditions for $w(x)$ at x_1 imply

$$\begin{aligned}w(x)&=A\sin\gamma\cos\omega_1(x-x_1)+A\frac{\omega_1}{\omega_2}\cos\gamma\sin\omega_1(x-x_1)\\&=A\sin[\omega_1(x-x_1)+\gamma]+A\left[\frac{\omega_1}{\omega_2}-1\right]\cos\gamma\sin\omega_1(x-x_1).\end{aligned}$$

Thus

$$\begin{aligned} v(p+\delta) - w(p+\delta) &= A\left[1 - \frac{\omega_1}{\omega_2}\right]\cos\gamma \sin\omega_1(p+\delta-x_1) \\ &+ A\sin[\omega_2(p+\delta-x_1)+\gamma] - A\sin[\omega_1(p+\delta-x_1)+\gamma] \\ &= A\frac{\omega_2-\omega_1}{\omega_2}\cos\gamma\sin\omega_1(p+\delta-x_1) \\ &+ 2A\sin\frac{\omega_2-\omega_1}{2}(p+\delta-x_1)\cos\left[\frac{\omega_2+\omega_1}{2}(p+\delta-x_1)+\gamma\right]. \end{aligned}$$

Observe that $0 < p+\delta-x_1 \le \delta$. Thus $\sin\omega_1(p+\delta-x_1) > 0$.

For $\omega_2 > \omega_1$ and $v'(p+\delta) \ge 0$ one has

$$\cos[\omega_2(p+\delta-x_1)+\gamma] = \frac{1}{\omega_2}v'(p+\delta) \ge 0.$$

Therefore

$$\frac{\omega_2+\omega_1}{2}(p+\delta-x_1)+\gamma < \omega_2(p+\delta-x_1)+\gamma \le \frac{\pi}{2}$$

and

$$\cos\left[\frac{\omega_2+\omega_1}{2}(p+\delta-x_1)+\gamma\right] > \cos[\omega_2(p+\delta-x_1)+\gamma] \ge 0.$$

Thus $v(p+\delta) - w(p+\delta) > 0$, and the conclusion (i) of the lemma follows.

The case $\omega_2 < \omega_1$ and $w'(p+\delta) \ge 0$ is reduced to the already established one by interchanging ω_1 with ω_2 and w with v. □

LEMMA 4.4. *Let $\delta > 0$, $0 < \eta \le 2\delta$, $\omega_1 \ne \omega_2$, with $0 < \omega_1\delta, \omega_2\delta < \pi/2$. Also let $A, B > 0$, $0 \le p < p+\eta \le 1$, and*

$$\begin{aligned} w(x) &= A\cos[\omega_1(x-p)+\gamma_1], \\ v(x) &= B\cos[\omega_2(x-p-\eta)+\gamma_2], \end{aligned}$$

with $|\gamma_1|, |\gamma_2| < \pi/2$.

Then the system

$$w(q) = v(q), \tag{4.5}$$

$$\omega_2^2 w'(q) = \omega_1^2 v'(q), \tag{4.6}$$

$$w(q) > 0, \quad v(q) > 0 \tag{4.7}$$

admits at most one solution q on $[p, p+\eta]$. This unique solution q can be computed as follows:

If $\gamma_1 \ge 0$, then

$$q = p + \frac{1}{\omega_1}\left[\arctan\left(\omega_1\sqrt{\left|\frac{B^2-A^2}{A^2\omega_2^2 - B^2\omega_1^2}\right|}\right) - \gamma_1\right]. \tag{4.8}$$

If $\gamma_2 \le 0$, then

$$q = p + \eta + \frac{1}{\omega_2}\left[-\arctan\left(\omega_2\sqrt{\left|\frac{B^2-A^2}{A^2\omega_2^2 - B^2\omega_1^2}\right|}\right) - \gamma_2\right]. \tag{4.9}$$

Otherwise, compute q_1 and q_2 according to (4.8) *and* (4.9) *and discard the one that does not satisfy the conditions of the lemma.*

Proof. Let $\alpha > 0$ and

$$\mathbf{c}(t;\alpha) = \begin{pmatrix} \cos t \\ \alpha \sin t \end{pmatrix}, \quad t \in \mathbf{R}.$$

Vector function $\mathbf{c}(t,\alpha)$ traverses the ellipse $\mathcal{E}(1,\alpha)$ centered in the origin with the x semiaxis equal to 1 and the y semiaxis equal to α. This function can be rewritten as

$$\mathbf{c}(t;\alpha) = \mathbf{P}(\alpha)\mathbf{M}(t)\mathbf{e}_1,$$

where

$$\mathbf{M}(t) = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}, \quad \mathbf{P}(\alpha) = \begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix}, \quad \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Note that $\mathbf{M}(t)$ is the counterclockwise rotation in $\mathbf{R}^2$ by the angle t, $\mathbf{P}(\alpha)$ is the α-contraction (expansion) in $\mathbf{R}^2$ along the y axis, and $\mathbf{e}_1$ is the standard basis vector along the x axis. Furthermore

$$\mathbf{c}(t_1 + t_2;\alpha) = \mathbf{P}(\alpha)\mathbf{M}(t_1)\mathbf{M}(t_2)\mathbf{e}_1. \tag{4.10}$$

With this notation system (4.5)–(4.6) is

$$A\mathbf{c}(\omega_1(q-p)+\gamma_1; 1/\omega_1) = B\mathbf{c}(\omega_2(q-p-\eta)+\gamma_2; 1/\omega_2)$$

or

$$\mathbf{P}(\omega_1^{-1})\mathbf{M}[\omega_1(q-p)+\gamma_1]A\mathbf{e}_1 = \mathbf{P}(\omega_2^{-1})\mathbf{M}[\omega_2(q-p-\eta)+\gamma_2]B\mathbf{e}_1. \tag{4.11}$$

If q is a solution of (4.5)–(4.6), then the vectors in the right and the left sides of (4.11) are identical. Thus they belong to the intersection of the ellipses $\mathcal{E}(A, A/\omega_1)$ and $\mathcal{E}(B, B/\omega_2)$, and this intersection is not empty. In general the ellipses intersect in four points: one in each quadrant.

Suppose that $q^* \neq q$ is another solution of (4.5)–(4.6) on $[p, p+\eta]$. We can assume that $q^* = q + \tau$ for some $\tau > 0$, $0 < \omega_1\tau, \omega_2\tau < \pi$. System (4.5)–(4.6) at $x = q^*$ is

$$\mathbf{P}(\omega_1^{-1})\mathbf{M}[\omega_1(q+\tau-p)+\gamma_1]A\mathbf{e}_1 = \mathbf{P}(\omega_2^{-1})\mathbf{M}[\omega_2(q+\tau-p-\eta)+\gamma_2]B\mathbf{e}_1. \tag{4.12}$$

Using (4.10) and $\mathbf{P}(\alpha)\mathbf{P}(\alpha^{-1}) = \mathbf{I}$ the right side of (4.12) can be written as

$$\begin{aligned}
&\mathbf{P}(\omega_2^{-1})\mathbf{M}[\omega_2(q+\tau-p-\eta)+\gamma_2]B\mathbf{e}_1 = \mathbf{P}(\omega_2^{-1})\mathbf{M}[\omega_2\tau]\mathbf{M}[\omega_2(q-p-\eta)+\gamma_2]B\mathbf{e}_1 \\
&= \mathbf{P}(\omega_2^{-1})\mathbf{M}[\omega_2\tau]\mathbf{P}(\omega_2)\mathbf{P}(\omega_2^{-1})\mathbf{M}[\omega_2(q-p-\eta)+\gamma_2]B\mathbf{e}_1 \\
&= \mathbf{P}(\omega_2^{-1})\mathbf{M}[\omega_2\tau]\mathbf{P}(\omega_2)\mathbf{P}(\omega_1^{-1})\mathbf{M}[\omega_1(q-p)+\gamma_1]A\mathbf{e}_1.
\end{aligned}$$

Similarly the left side of (4.12) can be written as

$$\begin{aligned}
&\mathbf{P}(\omega_1^{-1})\mathbf{M}[\omega_1(q+\tau-p)+\gamma_1]A\mathbf{e}_1 = \mathbf{P}(\omega_1^{-1})\mathbf{M}[\omega_1\tau]\mathbf{M}[\omega_1(q-p)+\gamma_1]A\mathbf{e}_1 \\
&= \mathbf{P}(\omega_1^{-1})\mathbf{M}[\omega_1\tau]\mathbf{P}(\omega_1)\mathbf{P}(\omega_1^{-1})\mathbf{M}[\omega_1(q-p)+\gamma_1]A\mathbf{e}_1.
\end{aligned}$$

Let $\mathbf{v} = \mathbf{P}(\omega_1^{-1})\mathbf{M}[\omega_1(q-p)+\gamma_1]A\mathbf{e}_1$ and

$$\mathbf{D} = \mathbf{P}(\omega_2^{-1})\mathbf{M}[\omega_2\tau]\mathbf{P}(\omega_2) - \mathbf{P}(\omega_1^{-1})\mathbf{M}[\omega_1\tau]\mathbf{P}(\omega_1).$$

Then (4.12) is $\mathbf{D}\mathbf{v} = \mathbf{0}$. Since $\mathbf{v} \neq \mathbf{0}$ we must have $\det(\mathbf{D}) = 0$. Note that

$$\begin{aligned}\det(\mathbf{D}) &= \frac{1}{\omega_1\omega_2}\left[2\omega_1\omega_2 - (\omega_1^2+\omega_2^2)\sin\omega_1\tau\sin\omega_2\tau - 2\omega_1\omega_2\cos\omega_1\tau\cos\omega_2\tau\right]\\ &= \frac{1}{\omega_1\omega_2}\left[2\omega_1\omega_2 - \frac{1}{2}(\omega_1+\omega_2)^2\cos(\omega_1-\omega_2)\tau + \frac{1}{2}(\omega_1-\omega_2)^2\cos(\omega_1+\omega_2)\tau\right].\end{aligned}$$

Let us define $f(\tau)$ on $[0,\pi/\omega_1)\cap[0,\pi/\omega_2)$ as

$$f(\tau) = 2\omega_1\omega_2 + \frac{1}{2}(\omega_1-\omega_2)^2\cos(\omega_1+\omega_2)\tau - \frac{1}{2}(\omega_1+\omega_2)^2\cos(\omega_1-\omega_2)\tau.$$

Function f is smooth on $(0,\pi/\omega_1)\cap(0,\pi/\omega_2)$, and its first and second derivatives are

$$\begin{aligned}f'(\tau) &= -\frac{1}{2}(\omega_1-\omega_2)^2(\omega_1+\omega_2)\sin(\omega_1+\omega_2)\tau + \frac{1}{2}(\omega_1+\omega_2)^2(\omega_1-\omega_2)\sin(\omega_1-\omega_2)\tau,\\ f''(\tau) &= (\omega_1-\omega_2)^2(\omega_1+\omega_2)^2\sin\omega_1\tau\sin\omega_2\tau.\end{aligned}$$

Since $f(0) = f'(0) = 0$, $f''(\tau) > 0$, and $f'(\tau) > 0$ on $(0,\pi/\omega_1)\cap(0,\pi/\omega_2)$, we conclude that $f(\tau) > 0$ for all $\tau \in (0,\pi/\omega_1)\cap(0,\pi/\omega_2)$. Thus $\det(\mathbf{D}) = 0$ if and only if $\tau = 0$. This contradicts the assumption $\tau > 0$. Therefore the solution q of (4.5)–(4.7) is unique on $[p, p+\eta]$.

To obtain formulas (4.8) and (4.9) notice that the ellipses $\mathcal{E}(A, A/\omega_1)$ and $\mathcal{E}(B, B/\omega_2)$ are given by

$$x^2 + \omega_1^2 y^2 = A^2 \quad \text{and} \quad x^2 + \omega_2^2 y^2 = B^2.$$

At the intersection points we have

$$y^2 = \frac{B^2-A^2}{\omega_2^2-\omega_1^2} \quad \text{and} \quad x^2 = \frac{A^2\omega_2^2 - B^2\omega_1^2}{\omega_2^2-\omega_1^2}.$$

The polar angle of the intersection point in the first quadrant is

$$\zeta = \arctan\sqrt{\left|\frac{B^2-A^2}{A^2\omega_2^2 - B^2\omega_1^2}\right|}.$$

Since $w(q) = v(q) > 0$ the intersection points corresponding to the solution q are in either the first or the fourth quadrants, and $0 \le \zeta < \pi/2$.

If $w'(p) \le 0$, then $\gamma_1 \ge 0$. Therefore $0 \le \gamma_1 \le \omega_1(q-p)+\gamma_1$. In this case the intersection point is in the first quadrant. Accordingly $\tan[\omega_1(q-p)+\gamma_1] = \omega_1\tan\zeta$. Thus

$$q = p + \frac{1}{\omega_1}\left[\arctan\left(\omega_1\sqrt{\left|\frac{B^2-A^2}{A^2\omega_2^2 - B^2\omega_1^2}\right|}\right) - \gamma_1\right].$$

If $v'(p+\eta) \ge 0$, then $\gamma_2 \le 0$ and $\omega_2(q-p-\eta)+\gamma_2 \le \gamma_2 \le 0$. In this case the intersection point is in the fourth quadrant. Accordingly $\tan[\omega_2(q-p-\eta)+\gamma_2] = -\omega_2\tan\zeta$ and one gets (4.9). ∎

Now we would like to define a class of piecewise constant conductivities with sufficiently separated points of discontinuity.

DEFINITION 4.5. *By the definition of $a \in \mathcal{PC}$ there exists $N \in \mathbf{N}$ and a finite sequence $0 = x_0 < x_1 < \cdots < x_{N-1} < x_N = 1$ such that a is a constant on each subinterval (x_{n-1}, x_n), $n = 1, \ldots, N$. Let $\sigma > 0$. Define*

$$\mathcal{PC}(\sigma) = \{a \in \mathcal{PC} \ : \ x_n - x_{n-1} \geq \sigma, \quad n = 1, 2, \ldots, N\}.$$

Note that $a \in \mathcal{PC}(\sigma)$ attains at most $N = [[1/\sigma]]$ distinct values $a_i, 0 < \nu \leq a_i \leq \mu$.

The following theorem is our main result. It describes and justifies the marching algorithm for the unique identification of piecewise constant conductivities in the class $\mathcal{PC}(\sigma)$.

THEOREM 4.6. *Given $\sigma > 0$ let an integer M be such that*

$$M \geq \frac{3}{\sigma} \quad \textit{and} \quad M > 2\sqrt{\frac{\mu}{\nu}}.$$

Suppose that the initial data $g(x) > 0$, $0 < x < 1$, and the observations $z_m(t) = u(p_m, t; a), p_m = m/M$ for $m = 1, 2, \ldots, M-1$ and $0 \leq T_1 < t < T_2$ of the heat conduction process (4.1) *are given. Then the conductivity $a \in A_{\mathrm{ad}}$ is constructively identifiable in the class of piecewise constant functions $\mathcal{PC}(\sigma)$.*

First, we present the marching algorithm for the unique identification of the conductivity a and then justify it. The algorithm marches from the left end $x = 0$ to a certain observation point $p_{l-1} \in (0, 1)$ and identifies the values a_n and the discontinuity points x_n of the conductivity a on $[0, p_{l-1}]$. Then the algorithm marches from the right end point $x = 1$ to the left until it reaches the observation point $p_{l+1} \in (0, 1)$ identifying the values and the discontinuity points of a on $[p_{l+1}, 1]$. Finally, the values of a and its discontinuity are identified on the interval $[p_{l-1}, p_{l+1}]$. The overall goal of the algorithm is to determine the number $N - 1$ of the discontinuities of a on $[0, 1]$, the discontinuity points x_n, $n = 1, 2, \ldots, N-1$, and the values a_n of a on $[x_{n-1}, x_n]$, $n = 1, 2, \ldots, N$ ($x_0 = 0$, $x_N = 1$). As a part of the process the algorithm determines certain functions $H_n(x)$ defined on intervals $[x_{n-1}, x_n]$, $n = 1, 2, \ldots, N$. The resulting function $H(x)$ defined on $[0, 1]$ is a multiple of the first eigenfunction v_1.

MARCHING ALGORITHM.

(i) Represent the data $z_m(t)$ as

$$z_m(t) = \sum_{k=1}^{\infty} c_{k,m} e^{-\lambda_k t}, \quad m = 1, 2, \ldots, M-1, \quad 0 \leq T_1 < t < T_2, \tag{4.13}$$

and use it to uniquely identify the first eigenvalue λ_1 and the coefficients $G_m = c_{1,m}$, $m = 1, 2, \ldots, M-1$. Let $G_0 = G_M = 0$.

(ii) Find l, $0 < l < M$, such that $G_l = \max\{G_m : m = 1, 2, \ldots, M-1\}$ and $G_m < G_l$ for any $0 \leq m < l$.

(iii) Let $i = 1$, $m = 0$.

(iv) Use Lemma 4.2 to find A_i, ω_i, and γ_i from the system

$$\begin{cases} A_i \cos(\omega_i \delta - \gamma_i) = G_m, \\ A_i \cos \gamma_i = G_{m+1}, \\ A_i \cos(\omega_i \delta + \gamma_i) = G_{m+2}. \end{cases} \tag{4.14}$$

Let

$$H_i(x) = A_i \cos(\omega_i (x - p_{m+1}) + \gamma_i).$$

(v) If $m+3 \geq l$, then go to step (viii). If $H_i(p_{m+3}) \neq G_{m+3}$, or $H_i(p_{m+3}) = G_{m+3}$ and $H_i'(p_{m+3}) \leq 0$, then a has a discontinuity x_i on interval $[p_{m+2}, p_{m+3})$. Proceed to the next step (vi). If $H_i(p_{m+3}) = G_{m+3}$ and $H_i'(p_{m+3}) > 0$, then let $m := m+1$ and repeat this step (v).

(vi) Use Lemma 4.2 to find A_{i+1}, ω_{i+1}, and γ_{i+1} from the system

$$\begin{cases} A_{i+1}\cos(\omega_{i+1}\delta - \gamma_{i+1}) = G_{m+3}, \\ A_{i+1}\cos\gamma_{i+1} = G_{m+4}, \\ A_{i+1}\cos(\omega_{i+1}\delta + \gamma_{i+1}) = G_{m+5}. \end{cases} \tag{4.15}$$

Let

$$H_{i+1}(x) = A_{i+1}\cos(\omega_{i+1}(x - p_{m+4}) + \gamma_{i+1}).$$

(vii) Use the formulas in Lemma 4.4 to find the unique discontinuity point $x_i \in [p_{m+2}, p_{m+3})$. The parameters and functions used in Lemma 4.4 are defined as follows. Let $p = p_{m+2}$, $\eta = \delta$. To avoid confusion we are going to use the notation Ω_1, Ω_2, Γ_1, Γ_2 for the corresponding parameters ω_1, ω_2, γ_1, γ_2 required in Lemma 4.4. Let $\Omega_1 = \omega_i$, $\Omega_2 = \omega_{i+1}$. For $w(x)$ use function $H_i(x)$ recentered at $p = p_{m+2}$; i.e., rewrite $H_i(x)$ in the form

$$w(x) = H_i(x) = A\cos(\Omega_1(x - p_{m+2}) + \Gamma_1), \quad |\Gamma_1| < \pi/2.$$

For $v(x)$ use function H_{i+1} recentered at $p + \eta = p_{m+3}$; i.e.,

$$v(x) = H_{i+1}(x) = B\cos(\Omega_2(x - p_{m+3}) + \Gamma_2), \quad |\Gamma_2| < \pi/2.$$

Let $i := i+1$, $m := m+3$. If $m < l$, then return to step (v). If $m \geq l$, then go to the next step (viii).

(viii) Do steps (iii)–(vii) in the reverse direction of x, advancing from $x = 1$ to $x = p_{l+1}$. Identify the values and the discontinuity points of a on $[p_{l+1}, 1]$, and determine the corresponding functions $H_i(x)$.

(ix) Using the notation introduced in (vii) let $H_j(x)$ be the previously determined function H on interval $[p_{l-2}, p_{l-1}]$. Recenter it at $p = p_{l-1}$; i.e., $w(x) = H_j(x) = A\cos(\Omega_1(x - p_{l-1}) + \Gamma_1)$. Let $H_{j+1}(x)$ be the previously determined function H on interval $[p_{l+1}, p_{l+2}]$. Recenter it at p_{l+1}: $v(x) = H_{j+1}(x) = B\cos(\Omega_2(x - p_{l+1}) + \Gamma_2)$. If $\Omega_1 = \Omega_2$, then stop; otherwise, use Lemma 4.4 with $\eta = 2\delta$ and the above parameters to find the discontinuity $x_j \in [p_{l-1}, p_{l+1}]$. Stop.

Proof. To prove Theorem 4.6 we need to justify the marching algorithm and to show the uniqueness of the identification in each step.

(i) Using Theorem 2.3(vi) we get

$$z_m(t) = \sum_{k=1}^{\infty} g_k e^{-\lambda_k t} v_k(p_m), \quad m = 1, 2, \ldots, M-1, \quad 0 \leq T_1 < t < T_2, \tag{4.16}$$

where $g_k = \langle g, v_k \rangle$ for $k = 1, 2, \ldots$. By Theorem 2.3(iv) $v_1(x) > 0$ on interval $(0, 1)$. Since g is positive on $(0, 1)$ we conclude that $g_1 v_1(p_m) > 0$. According to Theorem 2.3(vii) each observation $z_m(t)$ is a real analytic function. Thus one can uniquely determine the nonzero coefficients in (4.16) and the corresponding exponents. In particular, one determines the first eigenvalue λ_1 and the values of

$$G_m = g_1 v_1(p_m) > 0, \quad p_m = m/M, \ m = 1, 2, \ldots, M-1. \tag{4.17}$$

Because of the zero boundary conditions we can let $G_0 = G_M = 0$. The crucial point is that the numbers $\{G_m\}_{m=1}^{M-1}$ are not arbitrary but are the values (up to a nonzero multiplicative constant g_1) of the still-undetermined eigenfunction v_1.

(ii) Let index l be defined as in (ii) of the marching algorithm. By Theorem 2.3(v) there exists a unique point q^* of maximum of v_1 on $(0,1)$. Note that $q^* \in (p_{l-1}, p_{l+1})$. Thus $G_{l+1} \le G_l$ and $G_m < G_l$ for $m > l+1$. Also $v_1'(p_m-) > 0$ for $m = 1, 2, \dots, l-1$ and $v_1'(p_m-) < 0$ for $m = l+1, l+2, \dots, M-1$.

(iii) Start at the left end point $p_0 = 0$ and work on interval $[0, x_1]$, where x_1 is the first discontinuity point of a.

(iv) Let $\delta = 1/M$. Since $\sigma \ge 3\delta$ and $a \in \mathcal{PC}(\sigma)$ we conclude that $[0, p_2) \subset [0, x_1)$ and $a = a_1$ on $[0, x_1)$. To apply Lemma 4.2 we just need to check the conditions for Q_1, Q_2, Q_3 required there.

We have $Q_1 = G_0 = 0,\ Q_2 = G_1 = g_1 v_1(p_1) > 0\ , Q_3 = G_2 = g_1 v_1(p_2) > 0$. Let

$$\omega_1 = \sqrt{\frac{\lambda_1}{a_1}}.$$

By Theorem 2.3(iii) $0 < \lambda_1 \le \mu\pi^2$. Since $0 < \nu \le a_1$ we have

$$0 < \omega_1\delta < \sqrt{\frac{\mu\pi^2}{\nu}}\frac{1}{2}\sqrt{\frac{\nu}{\mu}} = \frac{\pi}{2}.$$

This inequality and $v_1(x) > 0$ on $(0,1)$ imply that the first eigenfunction v_1 of (4.2) can be represented on $(0, x_1)$ as

$$v_1(x) = C_1 \cos(\omega_1(x - p_1) + \gamma_1)$$

for some $(C_1, \omega_1, \gamma_1) \in \Gamma$, where Γ was defined in Lemma 4.2.

Also $Q_1 + Q_3 = g_1 C_1(\cos(\omega_1\delta + \gamma_1) + \cos(\omega_1\delta - \gamma_1)) = 2g_1 C_1 \cos(\omega_1\delta)\cos\gamma_1 < 2G_1 = 2Q_2$ since $0 < \omega_1\delta < \pi/2$; hence $0 < \cos(\omega_1\delta) < 1$. Now Lemma 4.2 guarantees a unique solution of the system

$$\begin{cases} g_1 C_1 \cos(\omega_1\delta - \gamma_1) = G_0, \\ g_1 C_1 \cos\gamma_1 = G_1, \\ g_1 C_1 \cos(\omega_1\delta + \gamma_1) = G_2. \end{cases} \tag{4.18}$$

It also gives formulas for the computation of $A_1 = g_1 C_1$, γ_1, ω_1 from the known values of G_0, G_1, and G_2. Thus one can determine $a_1 = \lambda_1/\omega_1^2$ and obtain

$$H_1(x) = g_1 C_1 \cos(\omega_1(x - p_1) + \gamma_1) = g_1 v_1(x) \tag{4.19}$$

for $x \in [0, x_1)$.

(v) Let a_i be the value of a on the part of the interval $[p_{m+1}, p_{m+2})$ adjacent to $[p_{m+2}, p_{m+3})$. By construction this value and the associated function $H_i(x) = g_1 v_1(x)$ are already determined by the algorithm. If there is no discontinuity of a on interval $[p_{m+2}, p_{m+3})$, then a has the same value a_i on interval $[p_{m+2}, p_{m+3})$ as well. Therefore $H_i(x) = g_1 v_1(x)$ on this interval, and we must have $G_{m+3} = H_i(p_{m+3})$ by (4.17). If one has $G_{m+3} \ne H_i(p_{m+3})$, then the implication is that there is a discontinuity of a on $[p_{m+2}, p_{m+3})$, and one proceeds to step (vi).

On the other hand, if $G_{m+3} = H_i(p_{m+3})$, then one cannot, in general, conclude that there is no discontinuity of a on $[p_{m+2}, p_{m+3})$. However, since we have $m + 3 < l$

then (ii) of the proof implies that $v_1'(p_{m+3}-) > 0$. Then the assumption $a = a_i$ on $[p_{m+2}, p_{m+3})$ implies $H_i'(p_{m+3}) = g_1 v_1'(p_{m+3}-)$. Therefore the equality $G_{m+3} = H_i(p_{m+3})$ together with $H_i'(p_{m+3}) \le 0$ lead to a contradiction. The conclusion is that $G_{m+3} = H_i(p_{m+3})$ and $H_i'(p_{m+3}) \le 0$ imply a discontinuity of a on $[p_{m+2}, p_{m+3})$, and one proceeds to step (vi).

Finally, one uses Lemma 4.3 to conclude that $m+3 < l$, $G_{m+3} = H_i(p_{m+3})$, and $H_i'(p_{m+3}) > 0$ imply that there is no discontinuity of a on $[p_{m+2}, p_{m+3})$. Indeed, suppose that there is a discontinuity point x_i of a on interval $[p_{m+2}, p_{m+3})$. Then $a = a_i$ on $[p_{m+1}, x_i)$ and $a = a_{i+1}$ on $[x_i, p_{m+3}]$. We are going to use the notation x_i, Ω_1, Ω_2 for the corresponding variables x_1, ω_1, and ω_2 used in Lemma 4.3. Let $p = p_{m+2}$, $p + \delta = p_{m+3}$, $\Omega_1 = \sqrt{\lambda_1/a_i}$, $\Omega_2 = \sqrt{\lambda_1/a_{i+1}}$, and

$$w(x) = H_i(x) = g_1 v_1(x) = A_1 \cos \Omega_1 x + B_1 \sin \Omega_1 x, \quad x \in [p, x_i),$$
$$v(x) = g_1 v_1(x) = A_2 \cos \Omega_2 x + B_2 \sin \Omega_2 x, \quad x \in [x_i, p+\delta].$$

Note that the condition $\Omega_1^2 v'(x_i) = \Omega_2^2 w'(x_i)$ is just the matching condition (4.2) at $x = x_i$. Since $m+3 < l$, the maximum q^* of v_1 satisfies $q^* > p_{m+3}$. Because w is a positive multiple of v_1, it implies $w(x_i) > 0$ and $w'(x_i) > 0$. Therefore $v(x_i) > 0$ and $v'(x_i) > 0$. Because v is a positive multiple of v_1, we have $v'(p+\delta) > 0$. The condition $v(p+\delta) = w(p+\delta)$ means $v(p_{m+3}) = g_1 v_1(p_{m+3}) = G_m = H_i(p_{m+3}) = w(p_{m+3})$.

Suppose that $\Omega_1 < \Omega_2$. We have $v(p+\delta) = w(p+\delta)$ and $v'(p+\delta) > 0$. According to Lemma 4.3(i), this is impossible.

Suppose that $\Omega_1 > \Omega_2$. We have $v(p+\delta) = w(p+\delta)$ and $w'(p+\delta) = H_i'(p_{m+3}) > 0$. According to Lemma 4.3(ii), this is also impossible.

Thus the conclusion is that there is no point of discontinuity of a on $[p_{m+2}, p_{m+3})$ in this case. By assigning $m := m+1$ one advances to the next observation interval $[p_{m+3}, p_{m+4})$ and repeats the analysis of (v).

(vi) Since it is already determined that there is a discontinuity point on interval $[p_{m+2}, p_{m+3})$, the assumption $a \in \mathcal{PC}(\sigma)$ implies that a is constant on $[p_{m+3}, p_{m+5}]$. This value a_{i+1} of a can be uniquely determined from the system in (vi) similarly to the argument presented in (iv). Note that $H_{i+1}(x) = g_1 v_1(x)$ on $[p_{m+3}, p_{m+5}]$.

(vii) One knows that the discontinuity $x_i \in [p_{m+2}, p_{m+3})$ as well as the values a_i and a_{i+1} of a on the adjacent intervals $[p_{m+1}, p_{m+2}]$ and $[p_{m+3}, p_{m+4}]$ together with the corresponding functions $H_i(x)$ and $H_{i+1}(x)$. According to Lemma 4.4 one can determine the unique location of the discontinuity x_i by the formulas given there.

(viii) The advance of the algorithm from $x = 1$ to $x = p_{l+1}$ is justified by reducing it to (iii)–(vii) using the change of variables $z = 1 - x$.

(ix) Lemma 4.4 is applicable with $\eta = 2\delta$. Note that there can be only one discontinuity of a on $[p_{l-1}, p_{l+1}]$, since $2\delta < \sigma$. The values of a as well as the corresponding functions H_j and H_{j+1} are already known on the adjacent intervals. The discontinuity of a exists on $[p_{l-1}, p_{l+1})$ if $\omega_j \ne \omega_{j+1}$. □

The marching algorithm of Theorem 4.6 requires measurements of the system at a possibly large number of observation points. Our next theorem shows that if a piecewise constant conductivity a is known to have just one point of discontinuity x_1, and its values a_1 and a_2 are known beforehand, then the discontinuity point x_1 can be determined from just one measurement of the heat conduction process.

THEOREM 4.7. *Let $p \in (0,1)$ be an observation point, $g(x) > 0$ on $(0,1)$, and the observation $z_p(t) = u(x_p, t; a)$, $t \in (T_1, T_2)$, of the heat conduction process (4.1) be given. Suppose that the conductivity $a \in A_{\rm ad}$ is piecewise constant and has only one (unknown) point of discontinuity $x_1 \in (0,1)$. Given positive values $a_1 \ne a_2$ such that*

$a(x) = a_1$ for $0 \le x < x_1$ and $a(x) = a_2$ for $x_1 \le x < 1$, the point of discontinuity x_1 is constructively identifiable.

Proof. Arguing as in the previous theorem,

$$z_p(t) = \sum_{k=1}^{\infty} g_k e^{-\lambda_k t} v_k(p), \quad 0 \le T_1 < t < T_2,$$

where $g_k = \langle g, v_k \rangle$ for $k = 1, 2, \ldots$. Since $g_1 v_1(p) > 0$, the uniqueness of the Dirichlet series representation implies that one can uniquely determine the first eigenvalue λ_1 and the value of $G_p = g_1 v_1(p)$.

Without loss of generality one can assume that $a_1 > a_2$. In this case we show that the first eigenvalue λ_1 is strictly increasing as a function of $x_1 \in [0, 1]$. Indeed, suppose that

$$0 \le x_1^a < x_1^b \le 1;$$

that is,

$$a(x) = \begin{cases} a_1, & 0 < x < x_1^a \\ a_2, & x_1^a < x < 1 \end{cases} \quad \text{and} \quad b(x) = \begin{cases} a_1, & 0 < x < x_1^b \\ a_2, & x_1^b < x < 1. \end{cases}$$

By Theorem 2.3(i)

$$\lambda_1^b = \frac{\int_0^1 b(x)[v'_{1,b}(x)]^2 dx}{\int_0^1 [v_{1,b}(x)]^2 dx} > \frac{\int_0^1 a(x)[v'_{1,b}(x)]^2 dx}{\int_0^1 [v_{1,b}(x)]^2 dx} \ge \inf_{v \in H_0^1(0,1)} \frac{\int_0^1 a(x)[v'(x)]^2 dx}{\int_0^1 [v(x)]^2 dx} = \lambda_1^a$$

provided that the derivative $v'_{1,b}(x)$ of the first eigenfunction $v_{1,b}(x)$ is not identically zero on (x_1^a, x_1^b). But, from $(b(x)v'_{1,b}(x))' = -\lambda_1^b v_{1,b}(x)$, the assumption $v'_{1,b}(x) = 0$ on (x_1^a, x_1^b) implies $v_{1,b}(x) = 0$ on (x_1^a, x_1^b), and this is impossible, since $v_{1,b}(x) > 0$ on $(0, 1)$. Thus there exists a unique conductivity of the type sought in the theorem for which its first eigenvalue is equal to λ_1; i.e., a is identifiable.

Now the unique discontinuity point x_1 of a can be determined as follows. Let

$$\omega_1 = \sqrt{\frac{\lambda_1}{a_1}}, \quad \omega_2 = \sqrt{\frac{\lambda_1}{a_2}}.$$

Then the first eigenfunction v_1 is given by

$$v_1(x) = \begin{cases} A \sin \omega_1 x, & 0 < x < x_1, \\ B \sin \omega_2 (1 - x), & x_1 < x < 1, \end{cases} \tag{4.20}$$

for some $A, B > 0$. The matching conditions at x_1 give

$$A \sin \omega_1 x_1 = B \sin \omega_2 (1 - x_1) \quad \text{and} \quad \frac{A}{\omega_1} \cos \omega_1 x_1 = \frac{B}{\omega_2} \cos \omega_2 (1 - x_1).$$

Since $v_1(x_1) > 0$ we have $0 < \omega_1 x_1 < \pi$ and $0 < \omega_2(1 - x_1) < \pi$. Therefore x_1 satisfies

$$\frac{1}{\omega_1} \cot \omega_1 x = \frac{1}{\omega_2} \cot \omega_2 (1 - x).$$

The existence and the uniqueness of the solution x_1 of the above nonlinear equation follows from the monotonicity and the continuity of the cotangent functions. Practically, the value of x_1 can be found by a numerical method. ∎

5. Conclusions. The prevalent approach to parameter identification (estimation) problems is to find such parameters from the best fit to data minimization. However, such an approach usually does not guarantee the uniqueness of the identified parameters. The identifiability problem consists of finding sufficient conditions assuring such a uniqueness, and there have been just a few results for the identifiability in distributed parameter systems.

In this paper we have shown that in some cases a variable conductivity in a 1D heat conduction process can be uniquely identified from observations of this process. The identifiability has been established for two sets of observations. In one case it is assumed that the conductivity is piecewise smooth, and we are given a sequence of distributed observations of the form $z_n(t) = \langle u(x,t;a), \psi_n \rangle$ for $n = 1, 2, \ldots$ on a finite time interval, where functions $\{\psi_N\}_{n=1}^{\infty}$ form a basis in $H = L^2(0,1)$. An algorithm for the conductivity identification is proposed. Its numerical study will be reported elsewhere.

In the second case it is assumed that the conductivity is piecewise constant with sufficiently separated points of discontinuity. The observations of the process are taken at equidistant points $p_m \in (0,1)$. The total number of points needed for the unique conductivity identification can be computed from a priori known parameters of the process. A marching algorithm for the conductivity identification is presented and justified.

In both cases the plant does not require a special external input for its identifiability; i.e., it is modeled by $u_t = (au_x)_x$ rather than by $u_t = (au_x)_x + f(x,t)$. It will be of interest to extend the developed methods to vibration and steady-state processes.

Our current research shows that the methods described in this paper can be extended to identifiability problems for heat conduction processes admitting various boundary (e.g., periodic) inputs and to other cases. A numerical implementation shows that the marching algorithm achieves a perfect identification for observations with low noise levels. These results will be presented elsewhere.

REFERENCES

[1] N. U. Ahmed, *Optimization and identification of systems governed by evolution equations on Banach space*, Pitman Res. Notes Math. Ser. 184, Longman Scientific and Technical, Harlow, UK, 1988.

[2] G. Birkhoff and G. Rota, *Ordinary Differential Equations*, Wiley & Sons, New York, 1989.

[3] A. Elayyan and V. Isakov, *On uniqueness of recovery of the discontinuous coefficient of a parabolic equation*, SIAM J. Math. Anal., 28 (1997), pp. 49–59.

[4] I. M. Gelfand and B. M. Levitan, *On the determination of a differential equation from its spectral function*, Amer. Math. Soc. Transl. Ser. 2, 2 (1955), pp. 253–304.

[5] S. Gutman, *Identification of discontinuous parameters in flow equations*, SIAM J. Control Optim., 28 (1990), pp. 1049–1060.

[6] S. Kitamura and S. Nakagiri, *Identifiability of spatially-varying and constant parameters in distributed systems of parabolic type*, SIAM J. Control Optim., 15 (1977), pp. 785–802.

[7] R. Kohn and M. Vogelius, *Determining conductivity by boundary measurements* II: *Interior results*, Comm. Pure Appl. Math., 41 (1988), pp. 865–877.

[8] C. Kravaris and J. H. Seinfeld, *Identification of parameters in distributed parameter systems by regularization*, SIAM J. Control Optim., 23 (1985), pp. 217–241.

[9] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and quasi-linear equations of parabolic type*, Trans. Math. Monogr., 23, AMS, Providence, RI, 1968.

[10] S. Larsson and V. Thomée, *Partial Differential Equations with Numerical Methods*, Springer-Verlag, New York, 2003.

[11] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[12] S. Nakagiri, *Review of Japanese work of the last* 10 *years on identifiability in distributed parameter systems*, Inverse Problems, 9 (1993), pp. 143–191.
[13] Y. Orlov and J. Bentman, *Adaptive distributed parameter systems identification with enforceable identifiability conditions and reduced-order spatial differentiation*, IEEE Trans. Automat. Control, 45 (2000), pp. 203–216.
[14] A. Pierce, *Unique identification of eigenvalues and coefficients in a parabolic problem*, SIAM J. Control Optim., 17 (1979), pp. 494–499.
[15] S. Saks and A. Zygmund, *Analytic Functions*, Monografie Matematyczne, Warsaw, 1965.
[16] E. Zauderer, *Partial Differential Equations of Applied Mathematics*, Wiley & Sons, New York, 1983.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 714–737

AN ABELIAN LIMIT APPROACH TO A SINGULAR ERGODIC CONTROL PROBLEM*

ANANDA WEERASINGHE†

Abstract. We consider an ergodic stochastic control problem for a class of one-dimensional Itô processes where the available control is an added bounded variation process. The corresponding infinite horizon discounted control problem was solved in [A. Weerasinghe, *SIAM J. Control Optim.*, 44 (2005), pp. 389–417]. Here, we show that as the discount factor approaches zero, the optimal strategies derived in [A. Weerasinghe, *SIAM J. Control Optim.*, 44 (2005), pp. 389–417] "converge" to an optimal strategy for the ergodic control problem. Under different assumptions, two types of optimal strategies were derived. Also, the Abelian limit relationships among the ergodic control problem, the infinite horizon discounted control problem, and the finite time horizon control problem are established here. A solution to a constrained optimization problem is obtained as an application.

Key words. ergodic control, local-time processes, diffusions with reflecting boundaries

AMS subject classifications. 93E20, 60H30

DOI. 10.1137/050646998

1. Introduction.

Consider a weak solution to the one-dimensional stochastic differential equation

$$X_x(t) = x + \int_0^t \mu(X_x(s-))ds + \int_0^t \sigma(X_x(s-))dW(s) + A(t), \tag{1.1}$$

where x is a real number, and $\{W(t) : t \geq 0\}$ is a standard Brownian motion adapted to a right continuous filtration $\{\mathfrak{F}_t : t \geq 0\}$ on a probability space $(\Omega, \mathfrak{F}, P)$. The σ-algebra $\mathfrak{F}_0$ contains all the null sets in $\mathfrak{F}$ and the Brownian increments $W(t+s)-W(t)$ are independent of $\mathfrak{F}_t$ for all $t \geq 0$ and $s \geq 0$. The control process $A(\cdot)$ is $\{\mathfrak{F}_t\}$-adapted, right continuous with left limits, and of bounded variation on finite time intervals. Also, $A(0) = 0$.

We further assume that there is a $\delta_0 > 0$ so that for each $X_x(\cdot)$ and $0 < \alpha < \delta_0$, there exists a sequence of stopping times (τ_n) satisfying $\lim_{n\to\infty} \tau_n = \infty$ and

$$\begin{aligned} &\text{(i)} \quad E_x \int_0^{T\wedge\tau_n} [|\mu(X_x(s-)| + \sigma^2(X_x(s-))]ds < \infty \text{ for each } T > 0, \text{ and} \\ &\text{(ii)} \quad \lim_{n\to\infty} E_x[|X_x(\tau_n)|e^{-\alpha\tau_n} I_{[\tau_n<\infty]}] = 0. \end{aligned} \tag{1.2}$$

The first condition helps us to make sense of (1.1) and the second condition will be used in verifying the Abelian limits described below. Throughout this article, we closely rely on several results obtained in [28], and the above two conditions imply the assumption (1.2) in [28].

The quintuple $((\Omega, \mathfrak{F}, P), (\mathfrak{F}_t), W, A, X_x)$ is called an admissible control system if the corresponding state process X_x satisfies (1.1) and (1.2). Let $\mathcal{U}$ denote the collection of all such admissible systems and let $C : \mathbf{R} \to \mathbf{R}$ be a running cost

*Received by the editors December 7, 2005; accepted for publication (in revised form) November 11, 2006; published electronically May 14, 2007. This work was partially supported by Army Research Office grant W 911NF0510032.

http://www.siam.org/journals/sicon/46-2/64699.html

†Department of Mathematics, Iowa State University, Ames, IA 50011 (ananda@iastate.edu).

function. We shall study the ergodic stochastic control problem with optimal value λ_0 defined by

$$\lambda_0 \triangleq \inf_{\mathcal{U}} \limsup_{T\to\infty} \frac{1}{T} E_x \int_0^T [C(X_x(s))ds + d|A|(s)]. \tag{1.3}$$

Notice that λ_0 is a constant which is independent of the initial value x, since an initial jump does not alter the above lim sup value for a given state process X_x.

Our goal here is to characterize an optimal control with a Markovian state process X that achieves the value λ_0 and to relate it to the value functions of the family of discounted control problems defined by

$$V_\alpha(x) \equiv \inf_{\mathcal{U}} E_x \int_0^\infty e^{-\alpha s}[C(X_x(s))ds + d|A|(s)], \tag{1.4}$$

as well as to the value functions of the family of finite horizon control problems

$$V_0(x,T) \equiv \inf_{\mathcal{U}} E_x \int_0^T [C(X_x(s))ds + d|A|(s)]. \tag{1.5}$$

It turns out that under the assumptions (1.6)–(1.8) below, $V_0(x,T)$ remains the same even when the infimum is taken over all processes $X_x(\cdot)$ which satisfy (1.1) together with the condition $E|X_x(t)| < \infty$ for each t in $[0,T]$. This is because we can extend such a process $X_x(\cdot)$ to $[0,\infty)$ as an admissible process by taking $A(t) \equiv 0$ for all $t > T$, and then using the results outlined in section 3 below.

Throughout this article, we make the basic assumptions (1.6), (1.7), and (1.8) below. Here μ', σ', and C' denote the derivatives of μ, σ, and C, respectively.

(1.6) (i) The functions μ and σ are continuously differentiable on $\mathbf{R}$, $\mu'(x) \le 0$ for all x, $\inf_{\mathbf{R}} \sigma(x) > 0$, and $x\mu(x) < 0$ for all $x \neq 0$.

(1.7) (ii) $$\int_{-\infty}^0 \frac{\mu(x)-x}{\sigma^2(x)}dx = \int_0^\infty \frac{x-\mu(x)}{\sigma^2(x)}dx = \infty.$$

(iii) The cost function C is continuously differentiable on $\mathbf{R}$, decreasing on $(-\infty,0)$, increasing on $(0,\infty)$, $C(0) = 0$, and satisfies one of the following conditions:

(1.8) either (a) $\liminf_{|x|\to\infty} \dfrac{C(x)}{|x|} > 0$ or (b) $\limsup_{|x|\to\infty} \dfrac{C(x)}{|x|} < \infty$.

The condition $C(0) = 0$ is made for convenience. If $C(0)$ is any other value, then it only shifts the value functions of (1.3)–(1.5) by an appropriate constant. The diffusion coefficient $\sigma(\cdot)$ is allowed to be an unbounded function subject to the above conditions (1.6) and (1.7).

Under assumption (1.6), the ordinary differential equation $\dot{x} = \mu(x)$ has a unique global, asymptotically stable equilibrium point at the origin. The cost function $C(\cdot)$ also has its unique minimum at the origin, and it increases as x moves away from the origin. Therefore, our study concerns the long term stability of a randomly perturbed stable dynamical system with a minimal control effort.

The qualitative nature of the optimal policies depends on the growth rates of $|\mu'(x)|$ and $|C'(x)|$. Therefore, we introduce a function H defined by

$$H(x) = \mu'(x) + |C'(x)| \qquad \text{for all } x \text{ in } \mathbf{R}. \tag{1.9}$$

The basic relationships among $\lambda_0, V_\alpha(x)$, and $V_0(x,T)$ are known as the "Abelian limit relations" (see [13]). These relations (which hold uniformly over the compact sets) are described by

$$\text{(a)} \quad \lim_{\alpha\to 0} \sup_{|x|\le K} |\alpha V_\alpha(x) - \lambda_0| = 0 \quad \text{and} \quad \text{(b)} \quad \lim_{T\to\infty} \sup_{|x|\le K} \left| \frac{V_0(x,T)}{T} - \lambda_0 \right| = 0 \tag{1.10}$$

for each $K > 0$.

We will establish these limits in this article. In an interesting article, Karatzas [13] derived optimal strategies for (1.3), (1.4), (1.5) and established the Abelian relations (1.10) when the drift coefficient μ is identically zero and the diffusion coefficient σ is a constant. In the symmetric case, where μ is an odd function and σ and C are even functions, an optimal strategy for (1.3) was obtained in [27]. In both of these articles, first a smooth solution to the corresponding Hamilton–Jacobi–Bellman (HJB) equation was obtained and then a verification lemma was used to prove the optimality of the chosen candidate for optimal control.

Here we develop a different approach. Using the properties of the value function $V_\alpha(x)$ of (1.4) derived in [28], we first establish that $\lim_{\alpha\to 0} \alpha V_\alpha(x)$ exists and is equal to a constant Λ_0, and we also show that $\Lambda_0 \le \lambda_0$, where λ_0 is the value of (1.3). Finally, to describe the derivation of an optimal strategy, let the initial point be at the origin. In this case, the optimal state process for $V_\alpha(0)$ derived in [28] induces a probability measure ν_α on $C[0,\infty)$. Under our assumptions, ν_α converges weakly (through a subsequence) to a probability measure ν_0 on $C[0,\infty)$ as α tends to zero. Thereafter, we derive an admissible strategy $((\Omega,\mathfrak{F},P),(\mathfrak{F}_t),W,A^*,X_0^*)$ in $\mathcal{U}$ so that the corresponding state process X_0^* induces the measure ν_0 on $C[0,\infty)$. It turned out that the corresponding value for the ergodic cost criteria is indeed Λ_0. Hence, Λ_0 is equal to the optimal value λ_0, and the above strategy $((\Omega,\mathfrak{F},P),(\mathfrak{F}_t),W,A^*,X_0^*)$ is optimal for (1.3). A complete solution to a constrained optimization problem is derived in section 6 by applying the results in section 4.

In [3], Arisawa and Lions considered an ergodic stochastic control problem in a compact state space. Hence, their cost function is bounded. In their model, there is no added bounded variation process, but the drift and diffusion coefficients are controlled. Their aim was to analyze the solution to the HJB equation and to establish the uniform convergence of the Abelian limits described in (1.10). But they did not derive any optimal strategies. This work complements their results and shows that uniform convergence on the compact sets for the Abelian limits in (1.10) is the best possible when the state space is noncompact.

In [16], the authors used a controlled martingale formulation for the cost structures in (1.3), (1.4), and (1.5) when there is no added bounded variation process. They related these problems with linear programming problems over a space of measures and proved the existence of optimal Markovian controls. Problem (1.4) is considered in [17] when the drift term is linear and the cost function is convex. There, the optimal control is a local-time process which is similar to our results in section 4. In a series of articles [1], [2], Alvarez considered singular control problems for diffusions with an absorbing barrier at the origin and with an added increasing control process.

He used the connection between stochastic control and optimal stopping to derive optimal strategies. In [8], the existence theorems for optimal policies for the ergodic control problem of multidimensional diffusions were developed. A higher-dimensional singular control problem for the standard Brownian motion was treated in [24]. For a discussion on singular control problems and related references, we refer to [11].

This article is organized as follows: Section 2 gathers the preliminary results regarding the value functions defined in (1.4) and (1.5). Our main theorem of section 3 is Theorem 3.1. It shows that under certain assumptions, zero control policy is optimal for the ergodic control problem (1.3). It also establishes the Abelian relation (1.10a). In section 4, under a different set of assumptions, we derive an optimal control policy which can be described in terms of local-time processes. The corresponding optimal state process is a reflecting diffusion on a finite interval. We also establish the limit (1.10a) for this case. Section 5 is devoted to the proof of the Abelian limit (1.10b). We apply the results obtained in section 4 to find an optimal policy for a constrained minimization problem in section 6. The main results in this article are Theorems 3.1, 4.1, 5.1, and 6.3. Next, we describe a motivating example from finance.

Example 1 (foreign exchange rates). Consider the currency exchange rate that governs the transactions between two countries (which we label as "domestic" and "foreign"). We assume that the economies of both countries are stable. Hence, in the absence of interventions, the currency exchange rate resembles a dynamical system which fluctuates around a stable equilibrium point. In the presence of uncertainty, it is common practice to model currency exchange rates using stochastic differential equations (see Chapter 7 of [21] and also [10], [12], [15], [19], [20], [25], [26]). Here we consider the problem of a central bank which would like to keep the exchange rate as close as possible to a target value through minimal intervention.

In a pioneering work [15], Krugman introduced a model where the exchange rate takes values in an exogenously given target interval, which is commonly called the "target zone" or the "target band." In recent years, exchange rate target zones have been an area of intense research activity in finance [10], [12], [19], [20], [25], [26]. When the exchange rate is high, the central bank may intervene by reducing the money supply by selling the foreign currency reserves or by adjusting the domestic interest rates. The central bank may also intervene appropriately when the exchange rate is low. To keep the exchange rate within the target band, the central bank may intervene while the exchange rate is still within the target band and there is empirical evidence to support this claim [7]. Such an optimal intervention policy with jumps within the target band is derived for a target zone exchange rate model using impulse controls in [12].

For a detailed discussion of Krugman's model, its underlying assumptions, its implications and drawbacks, and for modifications to accompany empirical data, we refer to [26]. One important prediction of Krugman's original model is that the long term distribution of the logarithm of the exchange rate within the target band must be U-shaped. This implies that in the long run, the logarithm of the exchange rate must spend most of its time near the end points of the target band. However, empirical data rejected this fact, and, as described in [26], there is a "hump"-shaped distribution within the band. In a detailed discussion about an extension of Krugman's model to agree with data, Svensson [26] pointed out that the logarithm of the exchange rate within the target band displays a "mean reversion behavior," and this is an important property of target zone exchange rates. This property also supports the "hump"-shaped long run distribution of the logarithm of the exchange rate.

In [20], a discrete impulse control intervention coupled with a continuous domestic interest rate control policy is used to derive an optimal target band. In [10], a similar problem for a geometric Brownian motion with a cost function $C(x) = x^2$ is considered. The authors also allow a fixed cost and a cost proportional to the size of the intervention. They derive an optimal intervention policy and the explicit form of the value function.

In this model, we consider a target value or a benchmark for the exchange rate which we simply assume to be at 1. The controlled state process $X_x(\cdot)$ represents the logarithm of the exchange rate. The central bank would like to keep the $X_x(\cdot)$ process near the origin. Here, there is no a priori assigned target band. We assume that $X_x(\cdot)$ satisfies (1.1) and the drift and diffusion coefficients μ and σ satisfy (1.6) and (1.7). Hence, $X_x(\cdot)$ clearly shows the mean reversion behavior around the origin as observed in [26]. The bounded variation control process $A(\cdot)$ represents the changes in the exchange rate due to central bank interventions. The corresponding total variation process $|A|(\cdot)$ represents the cumulative cost incurred by the central bank interventions. There is also a running cost associated with the deviation of the exchange rate from its benchmark, and this running cost function satisfies the assumption (1.8).

With our model, the central bank would like to know the answers to the following two important questions:

1. What optimal intervention policy will minimize the long term average cost criteria? Furthermore, to what extent will such an optimal intervention policy verify the validity of a target band for the exchange rate?

2. Suppose that the central bank insists on the intervention policy not to exceed the long term average intervention cost above a given target value $m > 0$ (i.e., $\limsup_{T\to\infty} \frac{E|A|(T)}{T} \le m$). What would be an optimal intervention policy which minimizes the long term average running cost $\limsup_{T\to\infty} \frac{1}{T} E \int_0^T C(X_x(s))ds$? Under what conditions on μ, σ, and $C(\cdot)$ does this constrained problem lead to a target zone model?

Sections 3 and 4 provide answers to question 1. Question 2 will be analyzed in section 6. In section 3, we derive a set of sufficient conditions for the optimality of zero intervention policy. Hence, there is no optimal target band under these conditions. With a different set of assumptions on μ, σ, and C, we prove the existence of an optimal target band in section 4. We also derive an optimal intervention policy. In section 6, under the given constraint on the long term average intervention rate, we provide an optimal target band and also an optimal intervention policy.

2. Preliminaries. Here we develop some results related to the value functions of the control problems in (1.3), (1.4), and (1.5). They will be important in establishing Abelian limits in (1.10). Throughout this section, we assume the conditions (1.6), (1.7), and (1.8).

PROPOSITION 2.1. *Let $\lambda_0, V_\alpha(x)$, and $V_0(x,T)$ be as in* (1.3), (1.4), *and* (1.5). *Then the following hold:*

(i) *For each $T > 0$ and $0 < \alpha < \delta_0$, the quantities $\lambda_0, V_\alpha(x)$, and $V_0(x,T)$ are finite. Also, for each $K > 0$, there is a constant M_K so that*

$$\sup_{0<\alpha<\delta_0} \sup_{|x|\le K} \alpha V_\alpha(x) \le M_K.$$

(ii) *For each $T > 0$, $0 < \alpha < \delta_0$, and x, y in $\mathbf{R}$, $|V_\alpha(x,T) - V_\alpha(y,T)| \le |x-y|$ and $|V_\alpha(x) - V_\alpha(y)| \le |x-y|$.*

(iii) $\limsup_{\alpha\to 0} \alpha V_\alpha(x) \le \lambda_0$ *and* $\limsup_{T\to\infty} \frac{1}{T} V_0(x,T) \le \lambda_0$, *where* λ_0 *is the value of the ergodic control problem* (1.3).

Proof. Given $K > 0$ and $|x| < K$, we pick an interval $[a,b]$ so that $a < -K < K < b$. Consider the reflected diffusion process on $[a,b]$, which is given by

$$X_x(t) = x + \int_0^t \mu(X_x(s))ds + \int_0^t \sigma(X_x(s))dW(s) + L_a(t) - L_b(t). \tag{2.1}$$

Here L_a and L_b are local-time processes of X_x at a and b, respectively. In comparison with (1.1), $A(t) = L_a(t) - L_b(t)$ and $|A|(t) = L_a(t) + L_b(t)$ for all $t \ge 0$. Consider the solution to the differential equation

$$\begin{aligned} \frac{\sigma^2(x)}{2} Q''(x) + \mu(x)Q'(x) &= \gamma \quad \text{for all } x \text{ in } (a,b), \\ Q'(a) = -1 \quad &\text{and} \quad Q'(b) = 1, \end{aligned} \tag{2.2}$$

where $\gamma > 0$ is a constant which will be chosen appropriately. Notice that (2.2) is a first order equation in $Q'(\cdot)$, and it can be solved using the boundary condition $Q'(a) = -1$. Then, for each x in $[a,b]$, we obtain

$$Q'(x)e^{2\int_0^x \rho(u)du} + e^{-2\int_a^0 \rho(u)du} = \gamma \int_a^x \frac{2}{\sigma^2(y)} e^{2\int_0^y \rho(u)du} dy.$$

In the above equation, $\rho(x) = \frac{\mu(x)}{\sigma^2(x)}$ for all x in $[a,b]$. Now we choose the positive constant γ so that it satisfies

$$\gamma \int_a^b \frac{2}{\sigma^2(y)} e^{2\int_0^y \rho(u)du} dy = e^{2\int_0^b \rho(u)du} + e^{-2\int_a^0 \rho(u)du}; \tag{2.3}$$

then it enforces $Q'(\cdot)$ to satisfy the other boundary condition $Q'(b) = 1$ in (2.2). The solution to (2.2) is unique up to a constant and we consider the solution

$$Q(x) = \int_a^x u(y)dy,$$

where

$$u(x) = e^{-2\int_0^x \rho(u)du} \left[\gamma \int_a^x \frac{2}{\sigma^2(y)} e^{2\int_0^y \rho(u)du} dy - e^{-2\int_a^0 \rho(u)du} \right].$$

Here the constant γ satisfies (2.3). Next, we apply Itô's lemma to $Q(X_x(t))$ and obtain

$$E|A|(T) = E[L_a(T) + L_b(T)] = \gamma T + Q(x) - E[Q(X_x(T))]. \tag{2.4}$$

At this point, using (2.3) together with (2.4), we can derive the following limit (independent of the initial point x):

$$\lim_{T\to\infty} \frac{E|A|(T)}{T} = \gamma = \frac{e^{2\int_0^b \rho(u)du} + e^{-2\int_a^0 \rho(u)du}}{\int_a^b \frac{2}{\sigma^2(y)} e^{2\int_0^y \rho(u)du} dy}, \tag{2.5}$$

which will be used in section 6.

Let $M_1 > 0$ be a constant so that

$$\sup_{[a.b]} [|C(x)| + |Q(x)|] < M_1. \tag{2.6}$$

This combined with (2.4) yields $E|A|(T) < \gamma T + 2M_1$. Therefore, we can conclude $V_0(x,T) \leq (M_1+\gamma)T + 2M_1$ and $\lambda_0 \leq \limsup_{T\to\infty} \frac{1}{T} E\int_0^T [C(X_x(s))ds + d|A|(s)] \leq (M_1+\gamma)$. Consequently, $V_0(x,T)$ and λ_0 are finite. Next, by (1.4), we have the inequality

$$V_\alpha(x) \leq E\int_0^\infty e^{-\alpha t}[C(X_x(s))ds + d|A|(s)] \leq \frac{M_1}{\alpha} + E\int_0^\infty e^{-\alpha t} d|A|(t).$$

Hence, using integration by parts and the estimate for $E|A|(T)$, we obtain

$$V_\alpha(x) \leq \frac{M_1}{\alpha} + \alpha\int_0^\infty e^{-\alpha t}(\gamma t + 2M_1)dt \leq \frac{M_1}{\alpha} + \frac{\gamma}{\alpha} + 2M_1.$$

Consequently, $V_\alpha(x)$ is also finite and the following uniform estimate holds:

$$\sup_{0<\alpha<\delta_0} \sup_{[a,b]} \alpha V_\alpha(x) \leq M_0 \equiv (M_1 + \gamma + 2M_1\delta_0).$$

Hence, part (i) follows.

For any admissible process X_x, introduce the cost functional

$$J(x, X_x, T) = E\int_0^T [C(X_x(t))dt + d|A|(t)].$$

For a given $\epsilon > 0$ and any $T > 0$, we pick a process X_x so that $V_0(x,T) + \epsilon > J(x, X_x, T)$. Then for any y, consider the process $\tilde{X}_y(0-) = y$ and with an initial jump to the point x so that $\tilde{X}_y(0) = x$. Thereafter it satisfies $\tilde{X}_y(t) \equiv X_x(t)$ for all $t > 0$. Hence we observe that $J(y, \tilde{X}_y, T) = |x-y| + J(x, X_x, T)$ and consequently $V_0(y,T) < |x-y| + J(x, X_x, T) < |x-y| + V_0(x,T) + \epsilon$.

Since ϵ is arbitrary and x and y are arbitrary points, we obtain

$$|V_0(x,T) - V_0(y,T)| \leq |x-y|. \tag{2.7}$$

By Theorems 4.3 and 5.5 of [28], V_α satisfies $|V'_\alpha(x)| \leq 1$ for all x and hence $|V_\alpha(x) - V_\alpha(y)| \leq |x-y|$. Thus, part (ii) follows.

To prove part (iii), we can alter an argument from classical analysis (see [23, p. 107]). We pick any constant K_1 so that $K_1 > \lambda_0$. Then there is an admissible process $X_x(.)$ so that

$$\limsup_{T\to\infty} \frac{J(x, X_x, T)}{T} < K_1. \tag{2.8}$$

Introduce a Borel measure ν on $[0,\infty)$, induced by the distribution function F, where F is defined by $F(T) = J(x, X_x, T)$ for all $T > 0$. Hence, $\nu([0,T]) = F(T)$ for all $T > 0$. Also, introduce the function G on $[0,\infty)$ by $G(T) = \frac{F(T)}{T+1}$ for $T > 0$. Therefore,

$$\begin{aligned} \alpha\int_0^\infty e^{-\alpha t} d\nu(t) &= \alpha^2\int_0^\infty e^{-\alpha t} F(t)dt \qquad \text{(by integration by parts)} \\ &= \alpha^2\int_0^\infty e^{-\alpha t}(t+1)G(t)dt = \int_0^\infty e^{-y}(y+\alpha)G\left(\frac{y}{\alpha}\right)dy. \end{aligned} \tag{2.9}$$

By (2.8), $G(t) < K_1$ for all $t > T_0 > 1$ and $||G||_\infty = \sup_{[0,\infty)} |G(t)| < \infty$. Hence, by (2.9),

$$\begin{aligned}
\alpha V_\alpha(x) &= \alpha \int_0^\infty e^{-\alpha t} d\nu(t) \\
&= \int_0^{\alpha T_0} e^{-y}(y+\alpha) G\left(\frac{y}{\alpha}\right) dy + \int_{\alpha T_0}^\infty e^{-y}(y+\alpha) G\left(\frac{y}{\alpha}\right) dy \\
&< \alpha^2 (T_0^2 + T_0)||G||_\infty + K_1 \int_0^\infty e^{-y}(y+\alpha) dy \\
&< \alpha^2 (T_0^2 + T_0)||G||_\infty + K_1(1+\alpha).
\end{aligned}$$

Consequently, $\limsup_{\alpha \to 0} \alpha V_\alpha(x) \le K_1$. Since $K_1 > \lambda_0$ is arbitrary this implies that $\limsup_{\alpha \to 0} \alpha V_\alpha(x) \le \lambda_0$. Next $V_0(x,T) \le F(T)$ for all $T > 0$ and by (2.8), we obtain $\limsup_{T\to\infty} \frac{V_0(x,T)}{T} \le \limsup_{T\to\infty} \frac{F(T)}{T} < K_1$. But $K_1 > \lambda_0$ is arbitrary. Hence $\limsup_{T\to\infty} \frac{V_0(x,T)}{T} \le \lambda_0$ and the proof of part (iii) is complete. □

3. Optimality of the zero control. First we introduce the state process $Z_x(\cdot)$ which corresponds to zero control policy, namely, $A(t) \equiv 0$ for all t in (1.1). Let $Z_x(\cdot)$ be a weak solution (see [14]) to

$$Z_x(t) = x + \int_0^t \mu(Z_x(s)) ds + \int_0^t \sigma(Z_x(s)) dW(s), \tag{3.1}$$

where $W(t)$ is a one-dimensional Brownian motion. The existence of $Z_x(t)$ for all $t \ge 0$ and the finiteness of the first moment $E|Z_x(t)|$ for each $t \ge 0$ are obtained in section 4 of [28] (see also Chapter 5, Theorem 5.15 in [14]). The main theorem in this section is the following.

THEOREM 3.1. *Assume* (1.6), (1.7), (1.8) *and that* $H(x) \le 0$ *for all* x, *where* H *is given in* (1.9). *Then the following hold:*

(i) $\lim_{\alpha \to 0} \alpha V_\alpha(x) = \Lambda_0$ *exists, where* Λ_0 *is a constant. Moreover, this limit converges uniformly over compact sets.*

(ii) *The process* $Z_x(.)$ *of* (3.1) *is an optimal process which corresponds to the zero control policy for the ergodic control problem in* (1.3). *Its value* $\lambda_0 = \Lambda_0$, *where* Λ_0 *is the limit in part* (i).

This theorem implies that under the above set of assumptions, there is no optimal target band for the exchange rates related to question 1 of our example on foreign exchange rates in section 1. Furthermore, an optimal policy for the central bank is not to intervene at all.

To prove this theorem, first we describe some results related to the value function $V_\alpha(x)$, which were developed in sections 3 and 4 of [28]. Let $Y(\cdot)$ be the weak solution to

$$Y(T) = x + \int_0^T [\sigma(Y(t))\sigma'(Y(t)) + \mu(Y(t))] dt + \int_0^T \sigma(Y(t)) dB(t), \tag{3.2}$$

where $\{B(t) : t \ge 0\}$ is a Brownian motion. The existence and uniqueness of a weak solution to (3.2) follows from Theorem 5.15 of Chapter 5 in [14]. This process was also introduced in (3.1) of [28]. Next we consider the function W_∞ introduced in

Lemma 4.2 of [28] and relabel it as W_α to specify its dependence on α. Then, by Lemma 4.2 of [28], $W_\alpha(x)$ has the stochastic representation

$$W_\alpha(x) = E_x \int_0^{\tau_\infty} e^{-\int_0^t (\alpha - \mu'(Y(s)))ds} C'(Y(t))dt, \tag{3.3}$$

where τ_∞ is the explosion time of the $Y(\cdot)$ process. Next, the assumption $H(x) \leq 0$ for all x implies that $|C'(x)| < (\alpha - \mu'(x))$ for all x. Using this estimate in (3.3), we obtain $|W_\alpha(x)| < 1$ for all x. Furthermore, as in Lemma 4.2 of [28], W_α satisfies

$$\frac{\sigma^2(x)}{2} W''_\alpha(x) + (\sigma(x)\sigma'(x) + \mu(x))W'_\alpha(x) - (\alpha - \mu'(x))W_\alpha(x) + C'(x) = 0 \tag{3.4}$$

for all x and $|W_\alpha(x)| < 1$ for all x. Theorem 4.3 of [28] also implies the following representation for the value function V_α:

$$V_\alpha(x) = \frac{\sigma^2(0)}{2\alpha} W'_\alpha(0) + \int_0^x W_\alpha(u)du. \tag{3.5}$$

Next, we prove a technical lemma.

LEMMA 3.2. *Let W_α be as in* (3.3) *above. Then the following results hold.*

(i) $\lim_{\alpha \to 0} W_\alpha(x)$ *exists for all x. Let $W_0(x) \triangleq \lim_{\alpha \to 0} W_\alpha(x)$; then $W_0(x)$ has the stochastic representation*

$$W_0(x) = E_x \int_0^{\tau_\infty} e^{\int_0^t \mu'(Y(s))ds} C'(Y(t))dt. \tag{3.6}$$

(ii) $W_0(\cdot)$ *also satisfies*

$$\frac{\sigma^2(x)}{2} W''_0(x) + (\sigma(x)\sigma'(x) + \mu(x))W'_0(x) + \mu'(x)W_0(x) + C'(x) = 0 \tag{3.7}$$

and $|W_0(x)| \leq 1$ for all x.

(iii)

$$\lim_{\alpha \to 0} W'_\alpha(0) = W'_0(0). \tag{3.8}$$

Proof. Consider the representation (3.3) for W_α. Since $H(x) \leq 0$ for all x, we obtain

$$|C'(Y(t))|e^{-\int_0^t (\alpha - \mu'(Y(s)))ds} \leq -\mu'(Y(t)) \cdot e^{-\int_0^t (\alpha - \mu'(Y(s)))ds}.$$

Notice that $-\int_0^{\tau_\infty} e^{\int_0^t \mu'(Y(s))ds} \cdot \mu'(Y(t))dt = 1 - e^{\int_0^{\tau_\infty} \mu'(Y(s))ds} \leq 1$. Now using (3.3) and the dominated convergence theorem, it follows that $\lim_{\alpha \to 0} W_\alpha(x) \equiv W_0(x)$ exists and W_0 has the representation (3.6). Hence, part (i) follows.

Next, we integrate (3.4) twice and obtain

$$\frac{\sigma^2(0)}{2} \cdot W'_\alpha(0) \cdot \int_0^x \frac{2}{\sigma^2(r)} dr = W_\alpha(x) - W_\alpha(0) + 2\int_0^x \frac{\mu(r)}{\sigma^2(r)} \cdot W_\alpha(r)dr + 2\int_0^x \frac{C(r)}{\sigma^2(r)} dr - 2\alpha \int_0^x \int_0^r \frac{W_\alpha(u)}{\sigma^2(u)} dudr.$$

Since $\lim_{\alpha\to 0} W_\alpha(x) \equiv W_0(x)$ exists and $|W_\alpha(x)| < 1$ for all x, we obtain that the right-hand side of the above equation converges as α tends to zero and $|W_0(x)| \le 1$ for all x. Therefore, $\lim_{\alpha\to 0} W'_\alpha(0)$ exists and we label it β. Hence,

$$\frac{\sigma^2(0)}{2} \cdot \beta \cdot \int_0^x \frac{2}{\sigma^2(r)} dr = W_0(x) - W_0(0) + \int_0^x \frac{2\mu(r)}{\sigma^2(r)} \cdot W_0(r) dr + \int_0^x \frac{2C(r)}{\sigma^2(r)} dr.$$

By differentiating this equation, we obtain $W'_0(0) = \beta$ and W_0 satisfies the differential equation (3.7). Hence, the proofs of parts (ii) and (iii) are complete. □

Now we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. By (3.5), we can write

$$\alpha V_\alpha(x) = \Lambda_\alpha + \alpha \int_0^x W_\alpha(u) du \tag{3.9}$$

for all x, where

$$\Lambda_\alpha = \frac{\sigma^2(0)}{2} W'_\alpha(0). \tag{3.10}$$

Next, we let

$$\Lambda_0 = \frac{\sigma^2(0)}{2} W'_0(0). \tag{3.11}$$

Using part (iii) of Lemma 3.2, we have $\lim_{\alpha\to 0} \Lambda_\alpha = \Lambda_0$. Since $|W_\alpha(x)| < 1$ for all x, using (3.9) we obtain $|\alpha V_\alpha(x) - \Lambda_0| \le |\Lambda_\alpha - \Lambda_0| + \alpha|x|$. Then $\lim_{\alpha\to 0} \alpha V_\alpha(x) = \Lambda_0$ and the convergence is uniform on compact sets. Thus, the proof of part (i) is complete.

To prove part (ii), we consider the process $Z_x(\cdot)$ in (3.1) and first show that $\limsup_{T\to\infty} \frac{1}{T} E \int_0^T C(Z_x(t)) dt = \Lambda_0$. Under the assumptions of this section, the $Z_x(\cdot)$ process is also optimal for the discounted control problem (1.4) as shown in [28]. Hence,

$$V_\alpha(x) = E \int_0^\infty e^{-\alpha t} C(Z_x(t)) dt = \int_0^\infty e^{-\alpha t} E[C(Z_x(t))] dt.$$

Now we can apply the classical Abelian limit theorem [23, p. 117] to obtain

$$\lim_{\alpha\to 0} \alpha V_\alpha(x) = \lim_{T\to\infty} \frac{1}{T} E \int_0^T C(Z_x(t)) dt.$$

This theorem also guarantees the existence of the ergodic limit on the right-hand side of the above equation. Consequently, $\lim_{T\to\infty} \frac{1}{T} E \int_0^T C(Z_x(t)) dt = \Lambda_0$. By part (iii) of Proposition 2.1, we know that $\Lambda_0 \le \lambda_0$, where λ_0 is the value of (1.3). Therefore, we can conclude $Z_x(\cdot)$ is an optimal process for (1.3) and $\lim_{\alpha\to 0} \alpha V_\alpha(x) = \Lambda_0 = \lambda_0$. This completes the proof. □

Remarks.

1. It should be noted that if the process $Z_x(\cdot)$ has a stationary distribution π, then the limit Λ_0 is equal to $\int_{-\infty}^{\infty} C(y)\pi(dy)$.

2. Under the assumptions of Theorem 3.1, one can show that the process $Z_x(\cdot)$ is also optimal for the finite-horizon problem with the value function $V_0(x, T)$ in (1.5). Here we sketch the proof.

Let $Q(x,T) = E\int_0^T C(Z_x(t))dt$ be the payoff from $Z_x(\cdot)$ in (3.1), which corresponds to zero control. Then Q satisfies $\frac{\sigma^2(x)}{2}Q_{xx} + \mu(x)Q_x + C(x) = Q_t$ for all x and $t > 0$. Also, $Q(x,0) = 0$. Now let $U(x,T) = Q_x(x,T)$; then U satisfies

$$\frac{\sigma^2(x)}{2}U_{xx} + (\sigma(x)\sigma'(x) + \mu(x))U_x + \mu'(x)U + C'(x) = U_t$$

for all x and $t > 0$, and $U(x,0) = 0$. By Itô's lemma (with the same notation as in (3.6)), U has the stochastic representation

$$U(x,T) = E\int_0^{\tau_\infty \wedge T} e^{\int_0^t \mu'(Y(s))ds} C'(Y(t))dt.$$

Thus, as in the proof of Theorem 3.1, $|U(x,T)| \le 1$ since $H(x) \le 0$. Now, for any given process $X_x(\cdot)$, by applying Itô's lemma to $Q(X_x(t), T-t)$, we obtain

$$Q(x,T) \le E\int_0^T [C(X_x(s))ds + d|A|(s)].$$

Hence, $Q(x,T) = V_0(X,T)$ and $Z_x(\cdot)$ is optimal for the finite time horizon problem (1.5).

4. Optimality of a reflected diffusion. In this section, we assume the following conditions (4.1) and (4.2) in addition to the basic assumptions (1.6)–(1.8). Let the function H be as in (1.9). We assume the existence of a constant $\alpha_0 > 0$ which satisfies the following:

(4.1) (i) For each $0 \le \alpha < \alpha_0$, there exist two points $\theta_\alpha < 0 < \beta_\alpha$ so that $H(x) > \alpha$ outside $[\theta_\alpha, \beta_\alpha]$. Furthermore, if $\alpha > 0$, then $H(x) < \alpha$ in $(\theta_\alpha, \beta_\alpha)$. Finally, if $\alpha = 0$, then $\{x : H(x) \le 0\} = [\theta_0, \beta_0]$.

(4.2) (ii) For each $0 < \alpha < \alpha_0$, there are two constants $\epsilon_\alpha > 0$ and $M_\alpha > 0$ so that $H(x) + \epsilon_\alpha\mu'(x) > (1+\epsilon_\alpha)\alpha$ for all $|x| > M_\alpha$.

Remarks.

1. Without loss of generality, we assume that $\alpha_0 < \delta_0$, where δ_0 is given in (1.2).
2. By (4.1), $H(0) \le 0$ and $\theta_\alpha \le \theta_0 < 0 < \beta_0 \le \beta_\alpha$ for each $0 \le \alpha < \alpha_0$.
3. By (4.2), it follows that $\lim_{x\to\infty} C(x) + \mu(x) = +\infty$ and $\lim_{x\to-\infty} C(x) - \mu(x) = +\infty$.
4. For each α in $(0, \alpha_0)$, the above assumptions imply those of section 5 in [28], and hence we can use the results related to $V_\alpha(x)$ in there.

Our main theorem in this section is the following.

THEOREM 4.1. *Assume* (4.1) *and* (4.2), *in addition to the basic assumptions* (1.6)–(1.8). *Then the following hold:*

(i) $\lim_{\alpha\to 0} \sup_{|x|\le K} |\alpha V_\alpha(x) - \lambda_0| = 0$ *for each* $K > 0$.

(ii) *There exist two points* a^* *and* b^* *so that the corresponding reflected diffusion process on the state space* $[a^*, b^*]$ *(if the initial point is outside this interval, then there will be an initial jump to the nearest point of the interval) is an optimal state process for the ergodic control problem* (1.3). *Hence the optimal control policy here is given by the difference of two local-time processes at* a^* *and* b^*.

This theorem verifies the existence of an optimal target band for our example on exchange rates in section 1 under the above set of assumptions. In this case, the optimal intervention policy of the central bank involves local-time processes of the exchange rate as described in the above theorem.

First we gather the necessary technical results in Lemma 4.2.

LEMMA 4.2. *Assume the same assumptions as in Theorem* 4.1. *Let* l_0 *be any limit point of the set* $\{\alpha V_\alpha(0)\}$ *as* α *tends to zero. Then there exist two points* a^*, b^* *and a continuously differentiable function* W_0 *defined on* $\mathbf{R}$ *satisfying the following conditions:*

(i) $-\infty < a^* \le \theta_0 < \beta_0 \le b^* < +\infty$, *where* θ_0 *and* β_0 *are given in* (4.1).

(ii) W_0 *satisfies* $\frac{\sigma^2(x)}{2} W_0'(x) + \mu(x) W_0(x) + C(x) = l_0$ *for* $a^* < x < b^*$.

(iii) $W_0(x) = -1$ *for all* $x \le a^*$, $W_0(x) = +1$ *for all* $x \ge b^*$, *and* $|W_0(x)| \le 1$ *for all* x.

(iv) *The value of* l_0 *can be identified by the formula*

$$l_0 = \frac{e^{2\int_0^{b^*} \rho(u)du} + e^{-2\int_{a^*}^0 \rho(u)du}}{2D} + \int_{a^*}^{b^*} C(u)\phi(u)du,$$

where $D = \int_{a^*}^{b^*} \frac{1}{\sigma^2(x)} e^{2\int_0^x \rho(u)du} dx$, $\rho(x) = \frac{\mu(x)}{\sigma^2(x)}$ *on* $[a^*, b^*]$, *and the density function* ϕ *is given by* $\phi(x) = \frac{1}{D}\frac{1}{\sigma^2(x)} e^{2\int_0^x \rho(u)du}$ *on* $[a^*, b^*]$.

Proof. Using [28, Proposition 5.4, (5.28), and Theorem 5.5], we obtain the following representation for the value function V_α: $V_\alpha(x) = V_\alpha(0) + \int_0^x W_\alpha(u)du$. Here, we write W_α for the function W in [28, Proposition 5.4] to represent the dependence on α. We also observe that

$$V_\alpha(x) = \frac{\sigma^2(0)}{2\alpha} W_\alpha'(0) + \int_0^x W_\alpha(u)du. \tag{4.3}$$

Furthermore, by [28, Proposition 5.4 and Theorem 5.5], for each α, there exist two points a_α^*, b_α^* and a C^1 function W_α so that $a_\alpha^* < \theta_0 < \beta_0 < b_\alpha^*$ and W_α satisfies

$$\frac{\sigma^2(x)}{2} W_\alpha''(x) + (\sigma(x)\sigma'(x) + \mu(x)) W_\alpha'(x) - (\alpha - \mu'(x)) W_\alpha(x) + C'(x) = 0 \tag{4.4}$$

for $a_\alpha^* < x < b_\alpha^*$. Also,

$$W_\alpha(a_\alpha^*) = -1, \quad W_\alpha(b_\alpha^*) = 1, \quad |W_\alpha(x)| < 1 \text{ on } (a_\alpha^*, b_\alpha^*), \tag{4.5}$$

and

$$\begin{aligned} W_\alpha'(a_\alpha^*) = W_\alpha'(b_\alpha^*) = 0, \quad W_\alpha(x) = -1 \text{ for } x \le a_\alpha^*, \\ W_\alpha(x) = 1 \text{ for } x \ge b_\alpha^*. \end{aligned} \tag{4.6}$$

This solution W_α was obtained in [28] by deriving a solution to an optimal stopping problem. For details, we refer to [28]. Now consider the limit point l_0 of the set $\{\alpha V_\alpha(0)\}$ as α tends to zero. Then there is a decreasing sequence $\{\alpha_n\}$ so that $\lim_{n\to\infty} \alpha_n = 0$ and $\lim_{n\to\infty} \alpha_n V_{\alpha_n}(0) = l_0$. By part (iii) of Proposition 2.1, it follows that $l_0 \le \lambda_0$. Hence, there is a constant $C_0 > 0$ so that $0 < \alpha_n V_{\alpha_n}(0) < C_0$ for all n. Notice that $[\theta_0, \beta_0] \subseteq [a_{\alpha_n}^*, b_{\alpha_n}^*]$. We intend to show that there is a finite $K > 0$

so that $[a^*_{\alpha_n}, b^*_{\alpha_n}] \subseteq [-K, K]$ for all α_n. Next, by integrating (4.4) over $[0, b^*_{\alpha_n}]$, we obtain

$$C(b^*_{\alpha_n}) + \mu(b^*_{\alpha_n}) - \frac{\sigma^2(0)}{2} W'_{\alpha_n}(0) = \alpha_n \int_0^{b^*_{\alpha_n}} W_{\alpha_n}(x)dx < \alpha_n b^*_{\alpha_n}.$$

Hence,

$$\begin{aligned} \int_0^{b^*_{\alpha_n}} [H(x) - \alpha_n]dx &= C(b^*_{\alpha_n}) + \mu(b^*_{\alpha_n}) - \alpha_n b^*_{\alpha_n} \\ &< \frac{\sigma^2(0)}{2} W'_{\alpha_n}(0) = \alpha_n V_{\alpha_n}(0) < C_0, \end{aligned} \tag{4.7}$$

where C_0 is a constant independent of n. But, from (4.2), $\int_0^\infty [H(x) - \alpha_n]dx = \infty$ for each n. Hence, by picking $n = 1$, we have a constant $K_1 > 0$ so that $\int_0^x [H(x) - \alpha_1]dx > C_0$ for all $x > K_1$. However, using (4.7) we obtain $\int_0^{b^*_{\alpha_n}} [H(x) - \alpha_1]dx < \int_0^{b^*_{\alpha_n}} [H(x) - \alpha_n]dx < C_0$. Consequently,

$$\beta_0 \le b^*_{\alpha_n} < K_1 \qquad \text{for all } n. \tag{4.8}$$

Similarly, by integrating (4.4) over $[a^*_{\alpha_n}, 0]$, we obtain

$$\int_{a^*_{\alpha_n}}^0 [H(x) - \alpha_n]dx = C(a^*_{\alpha_n}) - \mu(a^*_{\alpha_n}) - \alpha_n |a^*_{\alpha_n}| < \alpha_n V_{\alpha_n}(0) < C_0 \tag{4.9}$$

for all n. Then, using (4.2) and following an argument similar to that above, we obtain a constant $K_2 > 0$ so that

$$-K_2 < a^*_{\alpha_n} \le \theta_0 \qquad \text{for all } n. \tag{4.10}$$

Now we let $K = \max\{K_1, K_2\}$. Then by (4.8) and (4.10), we obtain

$$[a^*_{\alpha_n}, b^*_{\alpha_n}] \subseteq [-K, K] \qquad \text{for all } n. \tag{4.11}$$

In the rest of the proof we show that $\{W_{\alpha_n}\}$ and $\{W'_{\alpha_n}\}$ are equicontinuous families and conclude that $\{W_{\alpha_n}\}$ converges (possibly through a subsequence) to the desired function W_0 which satisfies parts (ii) and (iii) of the lemma. For this, we consider the sequence of functions (W_{α_n}) defined on $[-K, K]$. By integrating (4.4), we have

$$\frac{\sigma^2(x)}{2} W'_{\alpha_n}(x) = \alpha_n V_{\alpha_n}(0) + \alpha_n \int_0^x W_{\alpha_n}(u)du - C(x) - \mu(x) \cdot W_{\alpha_n}(x).$$

Using the facts that $\inf_{[-K,K]} \sigma^2(x) > 0$, $|W_{\alpha_n}(x)| \le 1$ for all x, the functions μ, σ, and C are bounded on $[-K, K]$, and $0 < \alpha_n V_{\alpha_n}(0) < C_0$ for all n, we obtain

$$\sup_n \sup_{[-K,K]} |W'_{\alpha_n}(x)| < C_1, \qquad \text{where } C_1 \text{ is a constant.} \tag{4.12}$$

Now, using (4.4) and (4.12) together with the same reasoning as above, we conclude

$$\sup_n \sup_{[a^*_{\alpha_n}, b^*_{\alpha_n}]} |W''_{\alpha_n}(x)| < C_2, \tag{4.13}$$

where $C_2 > 0$ is a constant. Here, at the end points $a^*_{\alpha_n}$ and $b^*_{\alpha_n}$, we considered the one-sided limits $W''_{\alpha_n}(a^*_{\alpha_n}+)$ and $W''_{\alpha_n}(b^*_{\alpha_n}-)$. Next, using (4.12), (4.13), and the fact that $W''_{\alpha_n}(x) = 0$ outside $[a^*_{\alpha_n}, b^*_{\alpha_n}]$, we conclude that $\{W_{\alpha_n}\}$ and $\{W'_{\alpha_n}\}$ are equicontinuous families on $[-K, K]$. By (4.11), we can pick a subsequence of (α_n), so that $(a^*_{\alpha_n})$ and $(b^*_{\alpha_n})$ converge to limit points a^* and b^*, respectively. Furthermore, these limit points are inside $[-K, K]$. Since $\{W_{\alpha_n}\}$ and $\{W'_{\alpha_n}\}$ are equicontinuous families on $[-K, K]$ (through a further subsequence, if necessary), using the Arzelà–Ascoli theorem, we can conclude that there exists a continuously differentiable function W_0 on the interval $[-K, K]$ which satisfies the following:

$$\lim_{\alpha_n \to 0} W_{\alpha_n}(x) = W_0(x) \qquad \text{and} \qquad \lim_{\alpha_n \to 0} W'_{\alpha_n}(x) = W'_0(x).$$

Furthermore, $-K \le a^* \le \theta_0 < \beta_0 \le b^* \le K$. Now by integrating (4.4) and letting α_n tend to zero in the resulting integral equation, we obtain that W_0 satisfies the differential equation in part (ii) of the lemma. To derive part (iii), we only need to extend W_0 to all $\mathbf{R}$, so that $W_0(x) = -1$ for $x \le -K$ and $W_0(x) = 1$ for $x \ge K$. Now, the equicontinuity of $\{W_{\alpha_n}\}$ and $\{W'_{\alpha_n}\}$ implies part (iii).

Next, similar to the argument in (2.2) and (2.3), we can solve the first order differential equation in part (ii) with the boundary condition $W_0(a^*) = -1$ and then use the other boundary condition $W_0(b^*) = 1$ to obtain the formula for l_0. Hence, the proof of part (iv) is complete. □

Remarks. In the following proof of the theorem, we show that $l_0 = \lambda_0$. Hence, we obtain the uniqueness of l_0 as well as the existence of the limit $\lim_{\alpha \to 0} \alpha V_\alpha(0)$.

Proof of Theorem 4.1. Introduce the function Q by

$$Q(x) = \int_0^x W_0(u)du,$$

where W_0 is as in the previous lemma. Let a^* and b^* be also as in the previous lemma. Next, consider the reflected diffusion process X^*_x on the interval $[a^*, b^*]$, given by (4.14). This process is positive recurrent on the interval $[a^*, b^*]$ and its ergodic limit for the cost functional can be derived explicitly as given below (see also [9, Chapter II, sec. 6]). In the following discussion, we simply assume x is in $[a^*, b^*]$, since an initial jump to the set $\{a^*, b^*\}$ does not alter the cost functional in (1.3). Let

$$X^*_x(t) = x + \int_0^t \mu(X^*_x(s))ds + \int_0^t \sigma(X^*_x(s))ds + A^*(t), \tag{4.14}$$

where

$$A^*(t) = L_{a^*}(t) - L_{b^*}(t) \qquad \text{and} \qquad |A^*|(t) = L_{a^*}(t) + L_{b^*}(t). \tag{4.15}$$

The processes L_{a^*} and L_{b^*} are local-time processes of X^*_x at the points a^* and b^*, respectively. Clearly, X^*_x is an admissible process for (1.3). We apply Itô's lemma to $Q(X^*_x(T))$ and use Lemma 4.2 to obtain

$$Q(X^*_x(T)) = Q(x) + l_0 T - E\int_0^T [C(X^*_x(t))dt + d|A^*|(t)].$$

Consequently, $\lim_{T\to\infty} \frac{1}{T} E\int_0^T [C(X^*_x(t))dt + d|A^*|(t)] = l_0$. Therefore, $l_0 \ge \lambda_0$ and the value of l_0 is given in part (iv) of Lemma 4.2. However, $l_0 \le \lambda_0$ from Proposition 2.1. Therefore, $l_0 = \lambda_0$ and the above X^*_x process is an optimal state process. Furthermore, every limit point of $\{\alpha V_\alpha(0)\}$ is equal to λ_0. Consequently, $\lim_{\alpha\to 0} \alpha V_\alpha(0) = \lambda_0$. Using part (ii) of Proposition 2.1, we obtain $\lim_{\alpha\to 0} \sup_{|x|\le K} |\alpha V_\alpha(x) - \lambda_0| = 0$ for each $K > 0$. Hence, the proofs of both parts (i) and (ii) of Theorem 4.1 are complete. □

5. Asymptotics for $V_0(x,T)$. In this section, we intend to prove the following theorem, which describes the long term behavior of $V_0(x,T)$ defined in (1.5).

THEOREM 5.1. *Under the assumptions of Theorem* 3.1 *or* 4.1, *the following Abelian limit holds:*

$$\lim_{T\to\infty} \sup_{|x|\le K} \left| \frac{V_0(x,T)}{T} - \lambda_0 \right| = 0 \quad \textit{for each } K > 0. \tag{5.1}$$

Proof. It suffices to show $\liminf_{T\to\infty} \frac{V_0(0,T)}{T} \ge \lambda_0$, because this together with parts (ii) and (iii) of Proposition 2.1 implies (5.1). First, we observe that a given process $X_x(\cdot)$ which satisfies (1.1) on $[0,T]$ with the condition $E\int_0^T [C(X_x(t))dt + d|A|(t)] < \infty$ can be extended to $[0,\infty)$ as an admissible process for (1.3), simply by using the zero control policy on $[T,\infty)$. Hence

$$X_x(T+s) = X_x(T) + \int_T^{T+s} \mu(X_x(u))du + \int_T^{T+s} \sigma(X_x(u))dW(u),$$

where $\{W(t) : t \ge 0\}$ is a Brownian motion. Since we have observed that the process corresponding to zero control policy is an admissible process in section 3, it easily follows that $X_x(\cdot)$ is also an admissible process. Second, we have observed that $\limsup_{T\to\infty} \frac{V_0(0,T)}{T} \le \lambda_0$ in the Proposition 2.1. Thus, if we take a constant $M_1 > \lambda_0$, then there is a $T_0 > 0$ so that

$$V_0(0,T) < M_1 T \tag{5.2}$$

for all $T > T_0$. Because of this fact, it suffices to consider the collection of admissible processes $X_0(\cdot)$ which satisfy (1.1) together with

$$E\int_0^T [C(X_0(t))dt + d|A|(t)] < M_1 T. \tag{5.3}$$

Next, we establish that the quantity $\frac{E|X_0(T)|}{T}$ is bounded for such an admissible process $X_0(\cdot)$ as $T \to \infty$. For this consider the even C^2 function ϕ defined by

$$\phi(x) = \frac{1}{8}(3 + 6x^2 - x^4)I_{[|x|<1]} + |x|I_{[|x|\ge 1]}.$$

Here, I_A denotes the indicator function of the set A. Then $0 \le \phi'(x) \le 1$ for $x \ge 0$ and $1 + \phi(x) \ge |x|$ for all x. Also, ϕ is nonnegative and $\phi'(x)\mu(x) \le 0$ for all x. Now, we use the generalized Itô lemma [18, p. 285] together with the above facts to obtain

$$E\phi(X_0(T)) \le E\int_0^T \frac{\sigma^2(X_0(s))}{2}\phi''(X_0(s))I_{[-1,1]}(X_0(s))ds + E|A|(T).$$

Since we can find a constant $C_2 > 0$ so that $\sup_{[-1,1]} \left[\sigma^2(x)|\phi''(x)|\right] \le C_2$ and by (5.3), for each $T > \max\{T_0, 1\}$, we obtain

$$E|X_0(T)| \le 1 + E\phi(X_0(T)) \le C_2 T + E|A|(T) + 1 < (M_1 + C_2 + 1)T. \tag{5.4}$$

Notice that the constants M_1 and C_2 are independent of the process $X_0(\cdot)$. Hence, the quantity $\frac{E|X_0(T)|}{T}$ remains bounded as $T \to \infty$. Next, by the dynamic programming

principle (or applying the generalized Itô lemma), we derive

$$\begin{aligned}\alpha V_\alpha(0) \le \alpha E \int_0^{T\wedge\tau_n} e^{-\alpha s}[C(X_0(s))ds + d|A|(s)] \\ + E[e^{-\alpha(T\wedge\tau_n)}\alpha V_\alpha(X_0(T\wedge\tau_n))].\end{aligned} \tag{5.5}$$

Here (τ_n) is a sequence of stopping times which satisfy the assumption (1.2). From the proofs of Theorems 3.1 and 4.1 (see (3.9), (3.10), and (4.3)), $V_\alpha(x)$ has the representation

$$\alpha V_\alpha(x) = \Lambda_\alpha + \alpha\int_0^x W_\alpha(u)du, \tag{5.6}$$

where $\lim_{\alpha\to 0}\Lambda_\alpha = \lambda_0$ and $|W_\alpha(x)| \le 1$ for all x. Hence, $\alpha V_\alpha(x) \le \Lambda_\alpha + \alpha|x|$ for all x. Combining this with (5.5), we obtain

$$\begin{aligned}\Lambda_\alpha[1 - E(e^{-\alpha(T\wedge\tau_n)})] \le \alpha E\int_0^T [C(X_0(s))ds + d|A|(s)] \\ + \alpha E|X_0(T\wedge\tau_n)|e^{-\alpha(T\wedge\tau_n)}.\end{aligned} \tag{5.7}$$

Notice that $E[|X_0(T\wedge\tau_n)|e^{-\alpha(T\wedge\tau_n)}] \le E[|X_0(\tau_n)|e^{-\alpha\tau_n}I_{[\tau_n<T]}] + E[|X_0(T)|e^{-\alpha T}]$. Now keeping $\alpha > 0$ fixed and letting (τ_n) tend to infinity and using (1.2), we derive $\lim_{n\to\infty} E[|X_0(\tau_n)|e^{-\alpha\tau_n}I_{[\tau_n<T]}] = 0$. Therefore,

$$\begin{aligned}\Lambda_\alpha[1 - e^{-\alpha T}] &\le \alpha E\int_0^T [C(X_0(s))ds + d|A|(s)] + \alpha E|X_0(T)|e^{-\alpha T} \\ &\le \alpha E\int_0^T [C(X_0(s))ds + d|A|(s)] + \alpha K e^{-\alpha T}.\end{aligned}$$

We have used (5.4) in the last inequality and here $K = (M_1 + C_2 + 1)$, where the constants M_1 and C_2 are as in (5.4). Thus, K is independent of α as well as the process $X_0(\cdot)$. Therefore, $\Lambda_\alpha[1 - e^{-\alpha T}] \le \alpha V_0(0,T) + \alpha K e^{-\alpha T}$. Consequently,

$$\Lambda_\alpha\frac{1 - e^{-\alpha T}}{\alpha T} \le \frac{V_0(0,T)}{T} + \frac{K}{T}e^{-\alpha T}.$$

We choose $\alpha = \frac{\delta}{T}$, where $0 < \delta < 1$ as in the argument in section (vi) of [3] for large $T > 0$. Thus,

$$\Lambda_\alpha\frac{1 - e^{-\delta}}{\delta} \le \frac{V_0(0,T)}{T} + \frac{K}{T}e^{-\delta}.$$

Since $\lim_{\alpha\to 0}\Lambda_\alpha = \lambda_0$, first we let δ tend to zero and then let T tend to infinity to obtain $\lambda_0 \le \liminf_{T\to\infty}\frac{V_0(0,T)}{T}$ as desired. This completes the proof. □

6. A constrained optimization problem. In this section, we address a constrained optimization problem which can be solved by using our results in section 4. For the purposes of this section, we need to strengthen the assumption (1.7) by the following assumption:

$$\int_{-\infty}^0 \frac{1 + \mu(x)}{\sigma^2(x)}dx = \int_0^\infty \frac{1 - \mu(x)}{\sigma^2(x)}dx = \infty. \tag{6.1}$$

At the end of this section, in Theorem 6.4, we will describe the results we can obtain when we replace (6.1) by the assumption (1.7). We will introduce further assumptions on μ, σ, and C after we describe the constrained optimization problem.

Consider the collection $\mathcal{U}$ of all admissible control systems used in the ergodic control problem (1.3). Let $\mathcal{U}_0$ be the subcollection of $\mathcal{U}$ obtained by enforcing μ and σ to also satisfy (1.6) and (6.1). Let $m > 0$ be any fixed positive real number. Here, we address the following constrained minimization problem:

$$\text{Minimize} \qquad \limsup_{T\to\infty} \frac{1}{T} E \int_0^T C(X_x(s))ds \tag{6.2}$$

subject to

$$\limsup_{T\to\infty} \frac{E|A|(T)}{T} \le m \tag{6.3}$$

over all admissible systems in $\mathcal{U}_0$. Notice that an initial jump does not affect our cost criteria or the constraint. Therefore, throughout this section, we simply consider the initial point x to be the origin and omit the dependence on x in our notation. To be more precise, for each $m > 0$, we define

$$\mathcal{U}_m = \left\{ ((\Omega, \mathfrak{F}, P), (\mathfrak{F}_t), W, A, X) \in \mathcal{U}_0 : \limsup_{T\to\infty} \frac{E|A|(T)}{T} \le m \right\}. \tag{6.4}$$

The collection $\mathcal{U}_m$ is nonempty, since the zero control policy developed in section 3 is there. The constrained minimization problem is to find the value function and characterize an optimal policy for

$$\inf_{\mathcal{U}_m} \limsup_{T\to\infty} \frac{1}{T} E \int_0^T C(X(s))ds. \tag{6.5}$$

We intend to characterize an optimal strategy that achieves the infimum in (6.5). We develop a "Lagrange multiplier"-type method by introducing an unconstrained optimization problem whose cost criteria includes a "penalty rate" $p > 0$. This penalty rate can be considered as the Lagrange multiplier variable. For each $p > 0$, we can obtain an optimal strategy for the unconstrained problem from the results in section 4. Furthermore, we also obtain an explicit formula for the derivative of the value with respect to p. Then we show that there exists a unique value for p, say p^*, where the corresponding optimal control A^* of the unconstrained problem satisfies $\lim_{T\to\infty} \frac{E|A^*|(T)}{T} = m$. This enables us to conclude that the same control policy is also optimal for the constrained minimization problem (6.5). At the end of this section, we also point out that if we assume (1.7) instead of (6.1), we can solve the constrained minimization problem with the same optimal policy only when $m > \gamma_0$, where γ_0 is a constant explicitly given in Theorem 6.4.

In a continuous-time setting, the idea of using both Lagrange multipliers and Kuhn–Tucker characterization of optimal policies for constrained stochastic control problems is considered in [6]. A problem of finite-fuel singular control with dynamic constraints for the control process is addressed in [5]. The Lagrange multiplier method was applied to a stochastic control problem with terminal conditions in [22, p. 241]. When the state space is a finite interval, a similar constrained optimization problem and an application to dynamic power control in wireless communication are addressed in [4] and we are motivated by their work.

Let $p > 0$ be a positive constant which represents the penalty rate. For each $p > 0$, we let

$$\Gamma(p) = \inf_{\mathcal{U}_0} \limsup_{T\to\infty} \frac{1}{T} E\left[\int_0^T C(X(s))ds + p \cdot |A|(T)\right]. \tag{6.6}$$

Introduce the function $C_p : \mathbf{R} \to [0, \infty)$ by

$$C_p(x) = \frac{C(x)}{p} \qquad \text{for all } x. \tag{6.7}$$

Notice that

$$\frac{\Gamma(p)}{p} = \inf_{\mathcal{U}_0} \limsup_{T\to\infty} \frac{1}{T} E\left[\int_0^T \frac{C(X(s))}{p} ds + |A|(T)\right] \tag{6.8}$$

and the function C_p satisfies the assumption (1.8). Similar to the definition of H in (1.9), for each $p > 0$, we define the function $H_p : \mathbf{R} \to \mathbf{R}$ by

$$H_p(x) = \mu'(x) + \frac{|C'(x)|}{p}. \tag{6.9}$$

We make the following additional assumptions throughout this section. They will enable us to use the results in section 4. For each $p > 0$, we assume there exists a constant $\alpha_0(p) > 0$ which satisfies the conditions below.

(6.10) (i) For each $0 \le \alpha < \alpha_0(p)$, there exist two points $\theta_\alpha(p) < 0 < \beta_\alpha(p)$ so that $H_p(x) > \alpha$ outside the interval $[\theta_\alpha(p), \beta_\alpha(p)]$. Furthermore, if $\alpha > 0$, then $H_p(x) < \alpha$ in $(\theta_\alpha(p), \beta_\alpha(p))$. Finally, if $\alpha = 0$, then $\{x : H_p(x) \le 0\} = [\theta_0(p), \beta_0(p)]$.

(6.11) (ii) For each $p > 0$, $\lim_{|x|\to\infty} \frac{\alpha_0(p) - \mu'(x)}{|C'(x)|} = 0$.

(6.12) (iii) $C(x) > 0$ for all $x \neq 0$.

Remarks. Assumption (6.10) is similar to (4.1) in section 4, and (6.11) clearly implies (4.2). Condition (6.12) enables us to simplify the proofs.

Examples. The following examples of μ, σ, and C satisfy all the assumptions in this section:

(i) Let $\mu(x) = -\theta x^3$ for some $\theta > 0$, $\sigma(x) = 1 + x^2$ and $C(x) = x^{2n}$ for any $n \ge 2$.

(ii) Let $\mu(x) = -\theta x$ for some $\theta > 0$, $\sigma(x) = \sigma_0$, where $\sigma_0 > 0$ is a constant. Let $C(\cdot)$ be any C^2 convex function which has a unique minimum at zero, $C(0) = 0$ and $\lim_{|x|\to\infty} |C'(x)| = \infty$.

When μ, σ, and C satisfy all the assumptions in this section, then for each $p > 0$, μ, σ, and C_p clearly satisfy all the assumptions of Theorem 4.1. Therefore, for each $p > 0$, $\frac{\Gamma(p)}{p}$ is finite and there is a finite interval $[a^*(p), b^*(p)]$ so that the corresponding reflecting diffusion process $X_p^*(\cdot)$ which satisfies (2.1) on this interval is an optimal process. The corresponding optimal bounded variation control process $A_p^*(\cdot)$ satisfies

$$A_p^*(t) = L_{a^*(p)}(t) - L_{b^*(p)}(t) \qquad \text{for all } t > 0, \tag{6.13}$$

where $L_{a^*(p)}(\cdot)$ and $L_{b^*(p)}(\cdot)$ are local time processes of $X_p^*(\cdot)$ at the end points $a^*(p)$ and $b^*(p)$, respectively. By Lemma 4.2, we know that $a^*(p)$ and $b^*(p)$ satisfy

$$-\infty < a^*(p) \leq \theta_0(p) < 0 < \beta_0(p) \leq b^*(p) < \infty. \tag{6.14}$$

It is known (see [9]) that the reflected diffusion $X_p^*(\cdot)$ on the finite interval $[a^*(p), b^*(p)]$ has a unique stationary probability distribution with the probability density ϕ given below in (6.16). Therefore,

$$\lim_{T\to\infty} \frac{1}{T} E \int_0^T C(X_p^*(s))ds = \int_{a^*(p)}^{b^*(p)} C(u)\phi(u)du. \tag{6.15}$$

The density function ϕ is given by

$$\phi(x) = \frac{1}{D}\frac{1}{\sigma^2(x)} e^{2\int_0^x \rho(u)du}, \tag{6.16}$$

where $\rho(x) = \frac{\mu(x)}{\sigma^2(x)}$ on the interval $[a^*(p), b^*(p)]$ and the normalization constant $D > 0$ is given by

$$D = \int_{a^*(p)}^{b^*(p)} \frac{1}{\sigma^2(x)} e^{2\int_0^x \rho(u)du} dx. \tag{6.17}$$

Also, we can use the limit in (2.5) for (6.13) to obtain

$$\lim_{T\to\infty} \frac{E|A_p^*|(T)}{T} = \frac{e^{2\int_0^{b^*(p)} \rho(u)du} + e^{-2\int_{a^*(p)}^0 \rho(u)du}}{2D}, \tag{6.18}$$

where the constant $D > 0$ is given in (6.17).

Consequently, $\Gamma(p)$ has the representation

$$\Gamma(p) = \lim_{T\to\infty} \frac{1}{T} E \int_0^T C(X_p^*(s))ds + p \cdot \lim_{T\to\infty} \frac{E|A_p^*|(T)}{T}. \tag{6.19}$$

Our next lemma shows the differentiability of $\Gamma(p)$ and computes the derivative explicitly.

LEMMA 6.1. *For each $p > 0$, consider the value function $\Gamma(p)$ defined in* (6.6). *Then the following statements are true:*

(i) $\Gamma(\cdot)$ *is a differentiable, strictly increasing function and its derivative is given by*

$$\Gamma'(p) = \lim_{T\to\infty} \frac{E|A_p^*|(T)}{T}, \tag{6.20}$$

where $A_p^*(\cdot)$ is the optimal control process described in (6.13).

(ii) $\Gamma(\cdot)$ *satisfies $0 < p \cdot \Gamma'(p) < \Gamma(p)$ for each $p > 0$ and $\lim_{p\to 0} \Gamma(p) = 0$.*

(iii) *The function $\frac{\Gamma(p)}{p}$ is strictly decreasing on $(0, \infty)$.*

Proof. We introduce the function $F(a, b)$ on the set $\{(a, b) \in \mathbf{R}^2 : a < 0 < b\}$ by

$$F(a,b) = \frac{e^{2\int_0^b \rho(u)du} + e^{-2\int_a^0 \rho(u)du}}{\int_a^b \frac{2}{\sigma^2(x)} e^{2\int_0^x \rho(u)du} dx}, \tag{6.21}$$

where ρ is as in (6.16). Notice that $\rho > 0$ on $(-\infty, 0)$ and $\rho < 0$ on $(0, \infty)$. Using this fact, a direct computation yields

$$\frac{\partial F}{\partial a}(a,b) > 0 \quad \text{and} \quad \frac{\partial F}{\partial b}(a,b) < 0 \quad \text{when} \quad a < 0 < b. \tag{6.22}$$

By (6.18), for each $p > 0$, we observe that

$$F(a^*(p), b^*(p)) = \lim_{T\to\infty} \frac{E|A_p^*|(T)}{T}. \tag{6.23}$$

Next, consider $0 < p < q$. By (6.6), it is clear that $\Gamma(p) \le \Gamma(q)$. Since the optimal strategy (X_p^*, A_p^*) for $\Gamma(p)$ is also an admissible strategy for $\Gamma(q)$, using the definitions of $\Gamma(p)$, $\Gamma(q)$ and (6.23), we obtain

$$0 \le \Gamma(q) - \Gamma(p) \le (q-p) \cdot F(a^*(p), b^*(p)).$$

Similarly, (X_q^*, A_q^*) is optimal for $\Gamma(q)$ and is an admissible strategy for $\Gamma(p)$ and hence

$$0 < (q-p) \cdot F(a^*(q), b^*(q)) \le \Gamma(q) - \Gamma(p).$$

Combining these two inequalities, we obtain

$$0 < F(a^*(q), b^*(q)) \le \frac{\Gamma(q) - \Gamma(p)}{q-p} \le F(a^*(p), b^*(p)) \quad \text{when} \ \ 0 < p < q. \tag{6.24}$$

Notice that (6.24) also implies that $F(a^*(p), b^*(p))$ is a decreasing function in the variable p. Now let $p_0 > 0$ be fixed. Let $\delta_1 > 0$ be such that $0 < \delta_1 < p_0$. Then by (6.24), for all $0 < |h| < \delta_1$, we have

$$F(a^*(p_0+\delta_1), b^*(p_0+\delta_1)) \le \frac{\Gamma(p_0+h) - \Gamma(p_0)}{h} \le F(a^*(p_0-\delta_1), b^*(p_0-\delta_1)). \tag{6.25}$$

Clearly, (6.25) implies the continuity of $\Gamma(\cdot)$ at p_0, and it also shows that if $\lim_{p\to p_0} F(a^*(p), b^*(p)) = F(a^*(p_0), b^*(p_0))$, then $\Gamma(\cdot)$ is differentiable at p_0 and $\Gamma'(p_0) = F(a^*(p_0), b^*(p_0))$. Notice that if $a^*(\cdot)$ and $b^*(\cdot)$ are continuous at $p = p_0$, then by (6.21), $\lim_{p\to p_0} F(a^*(p), b^*(p)) = F(a^*(p_0), b^*(p_0))$ holds. Therefore, to prove part (i) of the lemma, it suffices to show the continuity of the functions $a^*(\cdot)$ and $b^*(\cdot)$. Here, we prove the continuity of $b^*(\cdot)$. The proof of the continuity of $a^*(\cdot)$ is very similar, and therefore we omit it.

By parts (ii) and (iii) of Lemma 4.2 (notice that $l_0 = \lambda_0$ there, as shown in the proof of Theorem 4.1), we have

$$\frac{C(b^*(p))}{p} + \mu(b^*(p)) = \frac{C(a^*(p))}{p} - \mu(a^*(p)) = \frac{\Gamma(p)}{p}. \tag{6.26}$$

First we show that $b^*(\cdot)$ is bounded on the interval $[p_0 - \delta_1, p_0 + \delta_0]$ and notice that $b^*(p) > 0$ by (6.14). By the monotonicity of $\Gamma(\cdot)$, (6.9), and by (6.26), we have

$$\frac{\Gamma(p_0+\delta_1)}{p_0-\delta_1} \ge \frac{\Gamma(p)}{p} = \int_0^{b^*(p)} H_p(u)du \ge \int_0^{b^*(p)} H_{p_0+\delta_1}(u)du.$$

But (6.11) implies that $\int_0^\infty H_{p_0+\delta_1}(u)du = \infty$. Therefore, we can conclude that $b^*(\cdot)$ is a bounded function on $[p_0-\delta_1, p_0+\delta_1]$. Now consider a sequence (q_n) in $(p_0-\delta_1, p_0+\delta_1)$

such that $\lim_{n\to\infty} q_n = p_0$. Now let m_0 be any limit point of the bounded sequence $(b^*(q_n))$. Then $m_0 \geq 0$ by using (6.14). By (6.26) and by the continuity of $\Gamma(\cdot)$, we obtain $\frac{C(m_0)}{p_0} + \mu(m_0) = \frac{\Gamma(p_0)}{p_0} > 0$ and consequently $m_0 > 0$. But $b^*(p_0) > 0$ and also satisfies $\frac{C(b^*(p_0))}{p_0} + \mu(b^*(p_0)) = \frac{\Gamma(p_0)}{p_0} > 0$. By (6.10), we have $H_{p_0}(x) \leq 0$ on $[0, \beta_0(p_0)]$ and $H_{p_0}(x) > 0$ on $(\beta_0(p_0), \infty)$. Therefore, $\{x \geq 0 : \frac{C(x)}{p_0} + \mu(x) > 0\} \subseteq (\beta_0(p_0), \infty)$ and $\frac{C(x)}{p_0} + \mu(x)$ is strictly increasing on $(\beta_0(p_0), \infty)$ and hence $m_0 = b^*(p_0)$. This implies that $b^*(\cdot)$ is continuous at $p = p_0$. A similar proof yields the continuity of $a^*(\cdot)$ at $p = p_0$. Consequently, we have $\lim_{p\to p_0} F(a^*(p), b^*(p)) = F(a^*(p_0), b^*(p_0))$. This together with (6.25) implies that $\Gamma(\cdot)$ is differentiable at p_0, and its derivative is given by $\Gamma'(p_0) = F(a^*(p_0), b^*(p_0)) > 0$. Hence, the proof of part (i) is complete.

To prove part (ii), using (6.15)–(6.19), (6.23), and the above result we can write

$$\Gamma(p) = \int_{a^*(p)}^{b^*(p)} C(u)\phi(u)du + p \cdot \Gamma'(p),$$

where ϕ is given in (6.16). By (6.12) and (6.16), it is clear that the above integral is strictly positive. Therefore, $\Gamma(p) > p \cdot \Gamma'(p) > 0$ holds. To show $\lim_{p\to 0} \Gamma(p) = 0$, we consider the reflecting diffusion process described in (2.1) with $a = -\sqrt{p}$ and $b = \sqrt{p}$ on the interval $[-\sqrt{p}, \sqrt{p}]$ and obtain the following upper bound for $\Gamma(p)$:

$$0 < \Gamma(p) \leq \max\{C(\sqrt{p}), C(-\sqrt{p})\} + p \cdot F(-\sqrt{p}, \sqrt{p}).$$

Clearly, $\lim_{p\to 0^+} \max\{C(\sqrt{p}), C(-\sqrt{p})\} = 0$. A direct computation shows that $\lim_{p\to 0^+} \sqrt{p} \cdot F(-\sqrt{p}, \sqrt{p}) = \frac{\sigma^2(0)}{2}$. Therefore, $\lim_{p\to 0^+} p \cdot F(-\sqrt{p}, \sqrt{p}) = 0$ and, consequently, $\lim_{p\to 0} \Gamma(p) = 0$. This completes the proof of part (ii).

For the proof of part (iii), consider $0 < p < q$. Then the optimal strategy (X_p^*, A_p^*) for $\Gamma(p)$ is also an admissible strategy for $\Gamma(q)$, and therefore, by using (6.15)–(6.19), we obtain

$$\begin{aligned}
\frac{\Gamma(p)}{p} &> \int_{a^*(p)}^{b^*(p)} \frac{C(u)}{q}\phi(u)du + F(a^*(p), b^*(p)) \\
&= \lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T C_q(X_p^*(t))dt + |A_p^*|(T)\right] \\
&\geq \frac{\Gamma(q)}{q}.
\end{aligned}$$

This completes the proof of the lemma. □

In our next lemma, we derive the second order properties of the function $\Gamma(\cdot)$. We show that $\Gamma(\cdot)$ is a strictly concave function and its derivative $\Gamma'(\cdot)$ takes all the values in the interval $(0, \infty)$.

LEMMA 6.2. *The functions $a^*(\cdot)$, $b^*(\cdot)$, and $\Gamma(\cdot)$ satisfy the following conditions:*

(i) *The functions $a^*(\cdot)$ and $b^*(\cdot)$ are differentiable and their derivatives satisfy $\frac{da^*}{dp} < 0$ and $\frac{db^*}{dp} > 0$.*

(ii) *$\lim_{p\to 0^+} a^*(p) = \lim_{p\to 0^+} b^*(p) = 0$, $\lim_{p\to\infty} a^*(p) = -\infty$, and $\lim_{p\to\infty} b^*(p) = \infty$.*

(iii) *$\Gamma(\cdot)$ is a twice differentiable function which is strictly concave on $(0, \infty)$. Furthermore, $\lim_{p\to 0^+} \Gamma'(p) = \infty$ and $\lim_{p\to\infty} \Gamma'(p) = 0$.*

Proof. Introduce the function

$$U(p,x) = \frac{C(x)}{p} + \mu(x)$$

for $p > 0$ and $x > 0$. Then $U(p, b^*(p)) = \frac{\Gamma(p)}{p}$ by (6.26) and $\frac{\partial U}{\partial x}(p,x) = H_p(x)$ by (6.9). We notice that $b^*(p) > \beta_0(p) > 0$ by (6.14) and by the argument below (6.26). Hence, $\frac{\partial U}{\partial x}(p, b^*(p)) = H_p(b^*(p)) > 0$. Therefore, we can use the implicit function theorem to conclude that $b^*(\cdot)$ is differentiable at p. Now, by differentiating $\Gamma(p)$ with respect to p using (6.26), we obtain

$$p \cdot H_p(b^*(p)) \cdot \frac{db^*}{dp}(p) + \mu(b^*(p)) = \Gamma'(p). \tag{6.27}$$

By the previous lemma, $\Gamma'(p) > 0$ and $\mu(b^*(p)) < 0$ by (1.6). Also, $H_p(b^*(p)) > 0$ as we noted above. Hence, we conclude that $\frac{db^*}{dp}(p) > 0$. A similar proof yields $\frac{da^*}{dp}(p) < 0$. This completes the proof of part (i).

Consequently, $b^*(\cdot)$ is a strictly increasing function on $(0,\infty)$ and thus the limit $\lim_{p\to 0^+} b^*(p)$ exists. Let $\lim_{p\to 0^+} b^*(p) = b_0$. By (6.26), we obtain $\Gamma(p) = C(b^*(p)) + p \cdot \mu(b^*(p))$. Using Lemma 6.1 and letting p tend to zero, we can conclude that $b_0 \geq 0$ and $C(b_0) = 0$. By (6.12), it follows that $b_0 = 0$. A similar proof shows $\lim_{p\to 0^+} a^*(p) = 0$.

Next, we intend to show $\lim_{p\to\infty} b^*(p) = \infty$. Since $b^*(\cdot)$ is strictly increasing, we let $b_\infty = \lim_{p\to\infty} b^*(p)$. If b_∞ is finite, using (1.6) and (1.8) we obtain

$$0 < \Gamma(p) = C(b^*(p)) + p \cdot \mu(b^*(p)) < C(b_\infty) + p \cdot \mu(b^*(1))$$

for all $p > 1$. If b_∞ is finite, then the right-hand side of the above expression tends to $-\infty$ as p tends to infinity, which is a contradiction. Hence $b_\infty = \infty$. The proof of $\lim_{p\to -\infty} a^*(p) = -\infty$ is similar, and this completes the proof of part (ii).

To prove part (iii), using (6.23) and Lemma 6.1, we obtain the representation

$$\Gamma'(p) = F(a^*(p), b^*(p)) \quad \text{for } p > 0. \tag{6.28}$$

Since $F(\cdot,\cdot)$ is differentiable, using the proof of part (i) of this lemma, we have that $\Gamma(\cdot)$ is twice differentiable and its second derivative is given by

$$\Gamma''(p) = \frac{\partial F}{\partial a}(a^*(p), b^*(p)) \cdot \frac{da^*}{dp}(p) + \frac{\partial F}{\partial b}(a^*(p), b^*(p)) \cdot \frac{db^*}{dp}(p).$$

Now using (6.22) and the above part (i) of the lemma, it is evident that $\Gamma''(p) < 0$. Next, using (6.21), (6.28), and part (ii) of this lemma, we obtain $\lim_{p\to 0^+} \Gamma'(p) = \infty$. To compute $\lim_{p\to\infty} \Gamma'(p)$, again we use (6.21), (6.28), the fact that $F(a^*(p), b^*(p))$ is a decreasing function in the variable p, and the concavity of $\Gamma(\cdot)$. Then we can conclude that

$$\lim_{p\to\infty} \Gamma'(p) = F(-\infty,\infty) = \frac{e^{2\int_0^\infty \rho(u)du} + e^{-2\int_{-\infty}^0 \rho(u)du}}{\int_{-\infty}^\infty \frac{2}{\sigma^2(x)} e^{2\int_0^x \rho(u)du} dx}. \tag{6.29}$$

By (1.6), we obtain $0 < \int_{-\infty}^0 \rho(u)du \leq \infty$ and $-\infty \leq \int_0^\infty \rho(u)du < 0$. If $\int_{-\infty}^0 \rho(u)du = \infty$ and $\int_0^\infty \rho(u)du = -\infty$, then by (6.29) clearly $\lim_{p\to\infty} \Gamma'(p) = 0$. Next consider

the case $\int_0^\infty \rho(u)du$ to be convergent. Let $\int_0^\infty \rho(u)du = -L$, where L is a positive constant. Then, by (6.1), $\int_0^\infty \frac{1}{\sigma^2(r)}dr = \infty$. Therefore, the numerator of the right-hand side of (6.29) is less than 2, while the denominator is greater than or equal to $2e^{-2L}\int_0^\infty \frac{1}{\sigma^2(r)}dr$. Hence, the denominator is infinite and, consequently, $\lim_{p\to\infty}\Gamma'(p) = 0$. If the integral $\int_{-\infty}^0 \rho(u)du$ is convergent, a similar proof shows $\lim_{p\to\infty}\Gamma'(p) = 0$. This completes the proof. □

Remarks. The assumption (6.1) is used only in the proof of $\lim_{p\to\infty}\Gamma'(p) = 0$. If we assume (1.7) instead of (6.1), our conclusion for the limit $\lim_{p\to\infty}\Gamma'(p)$ will be given by the right-hand side of (6.29).

Next, we present the main theorem of this section.

THEOREM 6.3. *Assume* (1.6), (1.8), (6.1), *and* (6.10)–(6.12). *Then for any positive constant* $m > 0$, *the constrained optimization problem* (6.5) *has an optimal strategy of the type described in part* (ii) *of Theorem* 4.1*: Namely, there exist two points* $a_0^* < 0$ *and* $b_0^* > 0$ *so that the reflecting diffusion process described in* (4.14) *and* (4.15) *on the state space* $[a_0^*, b_0^*]$ *is an optimal state process.*

Proof. Let $m > 0$ be a constant. Then by Lemma 6.2, there is a unique constant $p^* > 0$ so that $\Gamma'(p^*) = m$. Consider the optimal strategy (X_p^*, A_p^*) for (6.6) with $p = p^*$. Then, using Lemma 6.1, we can write

$$\Gamma(p^*) = \lim_{T\to\infty}\frac{1}{T}E\left[\int_0^T C(X_{p^*}^*(u))du\right] + p^*\cdot m. \tag{6.30}$$

Now consider any admissible control system in $\mathcal{U}_m$ with the corresponding state process $X(\cdot)$ and the control process $A(\cdot)$, where $\mathcal{U}_m$ is given in (6.4). Then (X, A), which satisfies (1.1), is also an admissible strategy for $\Gamma(p^*)$ and the following inequalities hold:

(6.31)

$$\begin{aligned}\limsup_{T\to\infty}\frac{1}{T}E\int_0^T C(X(u))du + p^*\cdot m &\geq \limsup_{T\to\infty}\frac{1}{T}E\left[\int_0^T C(X(u))du\right]\\ &\quad + p^*\cdot\limsup_{T\to\infty}\frac{E|A|(T)}{T}\\ &\geq \limsup_{T\to\infty}\frac{1}{T}E\left[\int_0^T C(X(u))du + p^*\cdot|A|(T)\right]\\ &\geq \Gamma(p^*).\end{aligned}$$

Therefore, by (6.30) and (6.31), we conclude that

$$\limsup_{T\to\infty}\frac{1}{T}E\int_0^T C(X(u))du \geq \limsup_{T\to\infty}\frac{1}{T}E\int_0^T C(X_{p^*}^*(u))du.$$

Hence, (X_{p^*}, A_{p^*}) is an optimal policy for the constrained optimization problem (6.5). This completes the proof. □

When we replace assumption (6.1) by (1.7), we can obtain a partial solution to the constrained optimization problem. To describe it, first we introduce the nonnegative constant γ_0 by $\gamma_0 \equiv F(-\infty, \infty)$. Here, $F(-\infty, \infty)$ is defined as in the right-hand side of (6.29). Then we have the following result.

THEOREM 6.4. *Assume* (1.6)–(1.8) *and* (6.10)–(6.12). *Let* γ_0 *be the constant defined above. Then for each* $m > \gamma_0$, *the conclusion of Theorem* 6.3 *holds.*

Proof. By Lemma 6.2, $\Gamma'(\cdot)$ is a strictly increasing function whose range is (γ_0, ∞). Then for a given $m > \gamma_0$, there exists a unique $p^* > 0$ so that $\Gamma'(p^*) = m$. Now, the rest of the proof is identical to that of Theorem 6.3. □

REFERENCES

[1] L. R. Alvarez, *A class of solvable singular stochastic control problems*, Stoch. Stoch. Rep., 67 (1999), pp. 83–122.

[2] L. H. R. Alvarez, *Singular stochastic control, linear diffusions, and optimal stopping: A class of solvable problems*, SIAM J. Control Optim., 39 (2001), pp. 1697–1710.

[3] M. Arisawa and P.-L. Lions, *On ergodic stochastic control,* Comm. Partial Differential Equations, 23 (1998), pp. 2187–2217.

[4] B. Ata, J. M. Harrison, and L. A. Shepp, *Drift rate control of a Brownian processing system*, Ann. Appl. Probab., 15 (2005), pp. 1145–1160.

[5] P. Bank, *Optimal control under a dynamic fuel constraint*, SIAM J. Control Optim., 44 (2005), pp. 1529–1541.

[6] P. Bank and F. Riedel, *Optimal consumption with intertemporal substitution*, Ann. Appl. Probab., 11 (2001), pp. 750–788.

[7] G. Bertola and R. Caballero, *Target zones and realignments*, Amer. Econ. Rev., 27 (1992), pp. 520–536.

[8] V. S. Borkar and M. K. Ghosh, *Ergodic control of multidimensional diffusions* I: *The existence results*, SIAM J. Control Optim., 26 (1988), pp. 112–126.

[9] A. N. Borodin and P. Salminen, *Handbook of Brownian Motion: Facts and Formulae*, 2nd ed., Birkhäuser-Verlag, Basel, 2002.

[10] A. Cadenillas and F. Zapatero, *Optimal central bank intervention in the foreign exchange market*, J. Econom. Theory, 87 (1999), pp. 218–242.

[11] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, 2nd ed., Springer-Verlag, New York, 2006.

[12] M. Jeanblanc-Picqué, *Impulse control method and exchange rate*, Math. Finance, 3 (1993), pp. 161–177.

[13] I. Karatzas, *A class of singular stochastic control problems*, Adv. Appl. Prob., 15 (1983), pp. 225–254.

[14] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.

[15] P. R. Krugman, *Target zones and exchange rate dynamics*, Quart. J. Econom., 106 (1991), pp. 669–682.

[16] T. G. Kurtz and R. H. Stockbridge, *Existence of Markov controls and characterization of optimal Markov controls*, SIAM J. Control Optim., 36 (1998), pp. 609–653.

[17] J. Ma, *On the principle of smooth fit for a class of singular stochastic control problems for diffusions*, SIAM J. Control Optim., 30 (1992), pp. 975–999.

[18] P. A. Meyer, *Un cours sur les integrales stochastiques*, in Seminaire de Probabilities X, Lecture Notes in Math. 511, Springer-Verlag, New York, 1974, pp. 245–400.

[19] M. Miller and L. Zhang, *Optimal target zones: How an exchange rate mechanism can improve upon discretion*, J. Econom. Dynam. Control, 20 (1996), pp. 1641–1660.

[20] G. Mundaca and B. Øksendal, *Optimal stochastic intervention control with application to the exchange rate*, J. Math. Econom., 29 (1998), pp. 225–243.

[21] M. Musiela and M. Rutkowski, *Martingale Methods in Financial Modelling*, Springer-Verlag, New York, 1997.

[22] B. Øksendal, *Stochastic Differential Equations*, 5th ed., Springer-Verlag, New York, 1998.

[23] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Academic Press, New York, 1972.

[24] H. M. Soner and S. E. Shreve, *Regularity of the value function for a two-dimensional singular stochastic control problem*, SIAM J. Control Optim., 27 (1989), pp. 876–907.

[25] L. Svensson, *The term structure of interest rate differentials in a target zone*, J. Monet. Econom., 28 (1991), pp. 87–116.

[26] L. Svensson, *An interpretation of recent research on exchange rate target zones*, J. Econom. Persp., 6 (1992), pp. 119–144.

[27] A. Weerasinghe, *Stationary stochastic control for Itô processes*, Adv. Appl. Prob., 34 (2002), pp. 128–140.

[28] A. Weerasinghe, *A bounded variation control problem for diffusion processes*, SIAM J. Control Optim., 44 (2005), pp. 389–417.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 738–749

COMPARISON RESULTS AND ESTIMATES ON THE GRADIENT WITHOUT STRICT CONVEXITY*

A. CELLINA†

Abstract. In this paper we establish a comparison result for solutions to the problem

$$\text{minimize} \int_\Omega f(\nabla u(x))\,dx \quad \text{on } \{u : u - u_0 \in W_0^{1,1}(\Omega)\}.$$

Key words. strict convexity, comparison theorem, bounded slope condition

AMS subject classification. 49K20

DOI. 10.1137/060655869

1. Introduction. This paper deals with properties of solutions to minimization problems of the kind

$$\text{minimize} \int_\Omega f(\nabla u(x))\,dx \quad \text{on } \{u : u - u_0 \in W_0^{1,1}(\Omega)\}. \tag{1}$$

More precisely, we are interested in establishing a comparison result among solutions and an estimate on the gradient of a solution, derived from the boundary datum.

A comparison result is a statement of the kind, "for w and v solutions, $w \le v$ on $\partial\Omega$ implies $w \le v$ on Ω." In this generality, the only possible condition for the validity of this statement is the assumption of strict convexity on the Lagrangian f. However, rarely a comparison result is needed in this full generality: in general, one of the two solutions, w or v, belongs to a special class of solutions, as the affine functions (sometimes one is the function identically zero), and one aims at results for this more restricted class of problems. As an example, the first step in the proof of the existence of solutions for the minimal surface problem in parametric form depends on obtaining a priori bounds on the solution based on comparing the unknown solution with a constant solution [5].

The following example shows, however, that even when one of the solutions is an affine function (in particular, a constant), without the assumption of strict convexity, the comparison result is false.

Example 1. Let Ω be the interval $[-1, 1]$; let

$$f(\xi) = \begin{cases} 0, & |\xi| \le 1, \\ (|\xi| - 1)^2, & |\xi| > 1, \end{cases}$$

and consider the problem

$$\text{minimize} \int_{[-1,1]} f(u'(x))\,dx; \qquad u \in W_0^{1,1}([-1, 1]).$$

*Received by the editors March 31, 2006; accepted for publication (in revised form) December 6, 2006; published electronically May 14, 2007.
http://www.siam.org/journals/sicon/46-2/65586.html

†Dipartimento di Matematica e Applicazioni, Università degli Studi di Milano-Bicocca, Via R. Cozzi 53, 20125 Milano, Italy (arrigo.cellina@unimib.it).

Both functions $v(x) \equiv 0$ and $w(x) = -|x| + 1$ are solutions, and $w \leq v$ at $\partial\Omega$, but it is not true that $w \leq v$ on Ω.

Still, non–strictly convex Lagrangians do appear in the literature, frequently arising from the convexification of nonconvex Lagrangians: well-known examples, originating from problems of optimal design, are the Lagrangians in [1], [6], [7], [8].

This paper aims at dropping the requirement of strict convexity of f: its purpose is to introduce a class of solutions (to be used instead of the affine functions) such that the corresponding comparison theorem holds true without any requirement of strict convexity. The main property of this class of solutions is that it automatically reduces to the affine functions when f is strictly convex.

As a further motivation for the present study, notice that a comparison theorem involving affine functions is the main tool for the estimates on the gradient of a solution w to problem (1) provided by the bounded slope condition. In the situation described by this condition, affine functions of a given slope K bound the boundary datum u^0, and one is able to show that the same K bounds $\|\nabla w(x)\|$. Example 1 shows, again, that this result cannot possibly be true without the assumption of strict convexity of f. In fact, here the affine function identically zero bounds the boundary datum, with $K = 0$, but it is not true that, for the solution $w(x) = -|x| + 1$, one has $\|\nabla w(x)\| = 0$. In section 3 we prove a new result of the bounded slope condition type. When the Lagrangian f is strictly convex, the new condition we introduce reduces to the classical bounded slope condition, and the estimate on the gradient it provides is the classical estimate.

For a discussion of affine functions as solutions to variational problems, see [3]. In [9], Mariconda and Treu present a comparison theorem for variational problems of general form; their results, in particular, generalize those of [2] and will be used in the present paper. The book [5] contains several references to classical results connected with the bounded slope condition.

2. A comparison theorem. For K a subset of $\mathbb{R}^N$, K^c is its complement; m is the N-dimensional Lebesgue measure, while $m_{(N-1)}$ is the $(N-1)$-dimensional Lebesgue measure; the Hausdorff N-dimensional measure is H^N. The unit ball of $\mathbb{R}^N$ is B. A direction is a vector of unit norm; for a vector $v \in \mathbb{R}^N$, we will frequently use the notation $v = (v_1, \hat{v})$, where $\hat{v}$ is the $N-1$ vector consisting of the components from 2 to N of v. We denote by I_K the *indicator function* of the set K. The notation f^* denotes the *polar* or the *Legendre–Fenchel conjugate* of f. For f a convex function, $Dom(f)$ is the effective domain of f and $\partial f(x)$ is the subdifferential computed at x. For the notions of convex analysis we refer to [11]. When u and v are in $W^{1,1}(\Omega)$, by saying that at $\partial\Omega$ we have $v \geq u$ we mean, as usual, that $(u-v)^+ \in W_0^{1,1}(\Omega)$.

For $\theta \in Dom(\partial f^*)$, we will consider the functions

$$\left(I_{\partial f^*(\theta)}\right)^*(x) = \sup_{k \in \partial f^*(\theta)} \{\langle k, x\rangle\}$$

and

$$-\left(I_{\partial f^*(\theta)}\right)^*(-x) = -\sup_{z \in \mathbb{R}^N} \{\langle -x, z\rangle - I_{\partial f^*(\theta)}(z)\} = \inf_{k \in \partial f^*(\theta)} \{\langle k, x\rangle\}.$$

As an example, for the Lagrangian (see [8])

$$f(\xi) = \begin{cases} \sqrt{2}\|\xi\|, & \|\xi\| \leq \sqrt{2}, \\ 1 + \frac{1}{2}\|\xi\|^2, & \|\xi\| \geq \sqrt{2}, \end{cases} \tag{2}$$

we obtain that $\left(I_{\partial f^*(\theta)}\right)^*(x)$ is the family of maps

$$\left(I_{\partial f^*(\theta)}\right)^*(x) = \begin{cases} 0, & \|\theta\| \le \sqrt{2}, \\ \sqrt{2}\langle \frac{\theta}{\|\theta\|}, x\rangle \chi_{\{x:\langle\theta,x\rangle\ge 0\}}, & \|\theta\| = \sqrt{2}, \\ \langle\theta, x\rangle, & \|\theta\| > \sqrt{2}. \end{cases}$$

Next Theorem 1 shows that the functions defined above are solutions to the minimum problem (1), among those functions satisfying the same boundary conditions. In it, we assume the following growth assumption: $Dom(f^*)$ is open. The following proposition discusses how general this assumption is.

PROPOSITION 1. *Let f be an extended valued, convex, lower semicontinuous function with superlinear growth; then $Dom(f^*) = \mathbb{R}^N$.*

However, the Lagrangian $f(t) = |t| - \sqrt{|t|}$, whose polar is $f^*(p) = \frac{1}{4}\frac{1}{1-|p|}$ for $p \in (-1, 1)$, satisfies the condition $Dom(f^*)$ open, without being of superlinear growth.

It is convenient to use the following notation.

DEFINITION 1. *For $\theta \in Dom(f^*)$, $x^0 \in \mathbb{R}^N$, and $r \in \mathbb{R}$, set*

$$h^+_{\theta,x^0,r}(x) = \left(I_{f^*(\theta)}\right)^*(x - x^0) + r \quad \text{and} \quad h^-_{\theta,x^0,r}(x) = -\left(I_{f^*(\theta)}\right)^*(-x - x^0) + r.$$

THEOREM 1. *Let $Dom(f^*)$ be open. For $\theta \in Dom(f^*)$, $x^0 \in \mathbb{R}^N$, and $r \in \mathbb{R}$, the maps $h^+_{\theta,x^0,r}(x)$ and $h^-_{\theta,x^0,r}(x)$ are solutions to the minimum problem (1), among those u in*

$$\mathcal{S}^+_{\theta,x^0,r} = \left\{u \in W^{1,1}(\Omega) : u - h^+_{\theta,x^0,r} \in W^{1,1}_0(\Omega)\right\}$$

and

$$\mathcal{S}^-_{\theta,x^0,r} = \left\{u \in W^{1,1}(\Omega) : u - h^-_{\theta,x^0,r} \in W^{1,1}_0(\Omega)\right\},$$

respectively.

Proof. The proof is presented for $h^+_{\theta,x^0,r}$. For brevity, we set $h_\theta = h^+_{\theta,x^0,r}$.

The function h_θ, a supremum of convex functions, is convex; since $Dom(f^*)$ is open, $\partial f^*(\theta)$ is bounded; this implies that h_θ is finite on $\mathbb{R}^N$, and hence locally Lipschitzian and differentiable a.e. Let x be a point where it is differentiable, and let $\delta(x)$ be its gradient at x. The set $K(x) = \{k \in \partial f^*(\theta) : \langle k, x - x^0\rangle = \sup_{v\in\partial f^*(\theta)}\langle v, x - x^0\rangle\}$ is nonempty, compact, and convex.

(a) We claim that $\delta(x) \in K(x)$.

In fact, fix any h; let $k_t \in K(x + th)$ be such that $h_\theta(x + th) = \langle k_t, x + th - x^0\rangle + r$ and choose a sequence $t_i \to 0$ such that (k_{t_i}) converge to some k^*_h. Since the map $K(\cdot)$ has closed graph, we have $k^*_h \in K(x)$.

From the definition we have

$$\langle k_{t_i} - k^*_h, x - x^0\rangle \le 0 \quad \text{and} \quad \langle k_{t_i} - k^*_h, x - x^0 + t_i h\rangle \ge 0$$

so that, recalling that $k_{t_i} - k^*_h \to 0$, we obtain

$$\left\langle \frac{k_{t_i} - k^*_h}{t_i}, x - x^0 \right\rangle \to 0.$$

Hence,

$$\begin{aligned} \langle \delta(x), t_i h\rangle + t_i\|h\| o(t_i\|h\|) &= h_\theta(x + t_i h) - h_\theta(x) \\ &= t_i \left\langle \frac{k_{t_i} - k^*_h}{t_i}, x - x^0 \right\rangle + t_i\langle k_{t_i}, h\rangle = t_i\langle k^*_h, h\rangle + t_i o(t_i), \end{aligned}$$

so that

$$\langle \delta(x), h\rangle = \langle k_h^*, h\rangle. \tag{3}$$

In particular, choosing $h = x - x^0$, we obtain $\langle \delta(x) - k_0, x - x^0\rangle = 0$.

Then, in the case $\delta(x) \notin K(x)$, there exist $x^\perp \neq 0$ and $\varepsilon > 0$ such that $\langle x^\perp, x - x^0\rangle = 0$ and $\sup_{k\in K(x)}\langle k, x^\perp\rangle = \langle \delta(x), x^\perp\rangle - \varepsilon$. Take $h = x^\perp$ in (3), and notice that $k^*_{x^\perp} \in K(x)$: we have

$$\sup_{k\in K(x)} \langle k, x^\perp\rangle < \langle \delta(x), x^\perp\rangle = \langle k_h^*, x^\perp\rangle \leq \sup_{k\in K(x)} \langle k, x^\perp\rangle.$$

Hence, $\delta(x) \in K(x)$.

(b) Since $\delta(x) \in K(x) \subset \partial f^*(\theta)$, it follows that $\theta \in \partial f(\delta(x))$. Hence

$$f(\nabla u(x)) \geq \langle \theta, \nabla u(x) - \delta(x)\rangle + f(\delta(x)),$$

and hence

$$\int_\Omega f(\nabla u(x))\,dx - \int_\Omega f(\nabla h_\theta(x))\,dx \geq \int_\Omega \langle \theta, \nabla u(x) - \nabla h_\theta(x)\rangle\,dx.$$

Since $u \in \mathcal{S}_\theta$, the right-hand side equals zero, ending the proof. □

Remarks. In the case in which f is strictly convex, $\partial f^*(\theta)$ is single-valued and the maps h_θ^+ and h_θ^- are affine maps.

In general, we have proved that, a.e. in Ω,

$$\langle \nabla h_\theta, x - x^0\rangle = \sup_{v\in\partial f^*(\theta)} \langle v, x - x^0\rangle. \tag{4}$$

For the comparison theorem we shall need the following lemma: it is a modification of the classical statement saying that Sobolev functions are absolutely continuous on a.e. parallel to a given line. A general discussion of these results is presented in [10] and the lemma follows from these results.

LEMMA 1. *Let Ω be open and convex and let η be in $W^{1,1}(\Omega)$; let $B(x^*, \rho) \subset \Omega$ and, for $x \in B(x^*, \rho)$, set $\tilde{x}(t) = x^0 + (x - x^0)e^t$, $t \leq 0$; then, for a.e. $x \in B(x^*, \rho)$, as long as $\tilde{x}(t) \in \Omega$ and $\|\tilde{x}(t) - x^0\| \geq \delta > 0$,*

(i) *the map $t \to \eta(\tilde{x}(t))$ is absolutely continuous, and*

(ii) *for a.e. t, we have $\frac{d}{dt}\eta(\tilde{x}(t)) = \langle \nabla\eta(\tilde{x}(t)), \tilde{x}(t) - x^0\rangle$.*

The following is the comparison result.

THEOREM 2. *Let Ω be convex, let f be a (possibly extended valued) lower semicontinuous, convex function such that $Dom(f^*)$ is open. Let w be a solution to the problem of minimizing the functional*

$$\mathcal{J}(u) = \int_\Omega f\left(\nabla u(x)\right) dx$$

on $\{u : u - u^0 \in W_0^{1,1}(\Omega)\}$.

Assume that for $\theta \in Dom(f^*)$, $x^0 \in (\Omega)^c$, and $r \in \mathbb{R}$, we have $h^+_{\theta,x^0,r} \geq w$ on $\partial\Omega$. Then $h^+_{\theta,x^0,r} \geq w$ on Ω.

Remarks. Notice that any affine function $\ell(x) = \langle a, x\rangle + b$ can be written as $\ell(x) = \langle a, x - x^0\rangle + r$ with $x^0 \notin \Omega$ and $r = b + \langle a, x^0\rangle$.

Notice also that Example 1 shows that the analogous theorem, where we had an affine function ℓ (in particular, $\ell(x) \equiv 0$) instead of the convex function $h^+_{\theta,x^0,r}$, would be false.

Finally notice that the functions $u(x) \equiv 0$ and $h^+_{0,0,-1}(x) = -1 + |x|$ are solutions to the problem of Example 1; still, $h^+_{0,0,-1} \geq u$ on $\partial\Omega$, but it is not true that $h^+_{0,0,-1} \geq u$ on Ω. Here, the point $x^0 = 0 \in \Omega$, opposite to our assumptions.

Example 2. Consider the problem

$$\text{minimize} \int_{[-1,1]} f(u'(x))\,dx; \quad u^0 = \varepsilon(x-1); \quad u - u^0 \in W^{1,1}_0([-1,1]),$$

where f is defined in (2) and $\varepsilon > 0$ is small. The best upper bound for a solution w is

$$w(x) \leq \sup\left\{ \left(I_{\partial f^*(0)}\right)^* (x+1) + 2\varepsilon, \left(I_{\partial f^*(\sqrt{2})}\right)^* (x-1) \right\} = \sup\left\{ 2\varepsilon, \sqrt{2}(x-1) \right\}.$$

The proof of Theorem 2 is partially based on the following general lemma.

LEMMA 2. *Let Ω be open, bounded, and convex, let η be in $W^{1,1}(\Omega)$, and assume that there exists a point $x^* \in \Omega$ and a set of directions Z^+ having $H^{N-1}(Z^+) > 0$, such that, on the intersection of the half lines $\{x^* + \lambda z : z \in Z^+ \text{ and } \lambda \geq 0\}$ with Ω, we have $\eta \geq \delta > 0$. Then $\eta \notin W^{1,1}_0(\Omega)$.*

Proof. (a) Let D be the diameter of Ω. There exists ρ such that $B(x^*, 3\rho) \subset \Omega$. For $\xi^i \in \partial B$, set $t^i(z) = \frac{z - \langle z, \xi^i \rangle \xi^i}{\langle z, \xi^i \rangle}$; set $R = \frac{3\rho^2}{D^2 + \rho D}$ (so that, in particular, $3\rho - RD \geq 0$); set $Z^i = \{z \in \partial B : \|t^i(z)\| \leq R\}$. Choose finitely many ξ^i so that the corresponding sets Z^i cover ∂B: since $H^{N-1}(Z^+) > 0$, for at least one of them (say, $i = \hat{i}$), one has $H^{N-1}(Z^+ \cap Z^{\hat{i}}) > 0$. We will call Z^* the set $Z^+ \cap Z^{\hat{i}}$: we have

$$H^{N-1}(Z^*) > 0. \tag{5}$$

The half line $\{x^* + \lambda \xi^{\hat{i}} : \lambda \geq 0\}$ meets $\partial\Omega$ at one point, where we set the origin, so that $\xi^{\hat{i}} = -\frac{x^*}{\|x^*\|}$; moreover, we set the x_1 axis to be the half line from the origin through x^*: in this notation we obtain that $x^* = (x^*_1, \hat{0})$ and that $z \in Z^*$ implies

$$\frac{\|\hat{z}\|}{|z_1|} \leq R \quad \text{and} \quad |z_1| \geq \frac{1}{\sqrt{1+R^2}}.$$

It is convenient to call Ω^* the intersection of the half lines $\{x^* + \lambda z : \lambda \geq 0; z \in Z^*\}$ with Ω, and $\partial\Omega^*$ the intersection of the same half lines with $\partial\Omega$.

(b) The map $Z(x) = \frac{x - x^*}{\|x - x^*\|}$ is Lipschitzian on $\Omega \setminus B(x^*, \rho)$, since

$$\left| \frac{x^1 - x^*}{\|x^1 - x^*\|} - \frac{x^2 - x^*}{\|x^2 - x^*\|} \right| \leq \|x^1 - x^2\| \left(\frac{1}{\rho} + \frac{D}{\rho^2} \right).$$

Hence, from the equality $Z^* = Z(\partial\Omega^*)$, we obtain that

$$H^{N-1}(Z^*) \leq \left(\frac{1}{\rho} + \frac{D}{\rho^2} \right) H^{N-1}(\partial\Omega^*)$$

and, from (5), we conclude that

$$H^{N-1}(\partial\Omega^*) > 0. \tag{6}$$

On $\Omega^* \setminus B(x^*, \rho)$ we have $|x_1 - x_1^*| \geq \rho |z_1| \geq \frac{\rho}{\sqrt{1+R^2}}$ so that

$$\left\| \frac{\hat{x}^1 - \hat{x}^*}{x_1^1 - x_1^*} - \frac{\hat{x}^2 - \hat{x}^*}{x_1^2 - x_1^*} \right\| \leq \frac{1}{|x_1^1 - x_1^*|} \|\hat{x}^1 - \hat{x}^2\| + \frac{\|\hat{x}^2 - \hat{x}^*\|}{|x_1^1 - x_1^*||x_1^2 - x_1^*|} |x_1^1 - x_1^2|$$
$$\leq \frac{\sqrt{1+R^2}}{\rho} \|\hat{x}^1 - \hat{x}^2\| + D \frac{1+R^2}{\rho^2} |x_1^1 - x_1^2| \leq L \|x^1 - x^2\|.$$

The above shows that the map $\hat{T}(x) = \frac{\hat{x} - \hat{x}^*}{x_1 - x_1^*}$ is Lipschitzian on $\Omega \setminus B(x^*, \rho)$.

(c) The half line $\{x = x^* + \lambda z : \lambda \geq 0\}$ can as well be described by $\{x : \frac{\hat{x} - \hat{x}^*}{x_1 - x_1^*} = \hat{t} = \frac{\hat{z}}{z_1}\}$: there exists a unique point $b(\hat{t})$ on its intersection with $\partial\Omega$. We will prove the Lipschitzianity of this map. Consider two points $b^1 = b^1(\frac{\hat{x}^1 - \hat{x}^*}{x_1^1 - x_1^*})$ and $b^2 = b^2(\frac{\hat{x}^2 - \hat{x}^*}{x_1^2 - x_1^*})$ in $\partial\Omega$, such that $b_1^1 < b_1^2 < x_1^*$. The half line $b^1 + \lambda(b^2 - b^1)$ meets the plane $\{x_1 = x_1^*\}$ at the point $(x_1^*, \hat{b}^1 + \frac{x_1^* - b_1^1}{b_1^2 - b_1^1}(\hat{b}^2 - \hat{b}^1))$: it cannot be that this point is in $B(x^*, 3\rho)$, since, otherwise, a half line issued from a point in the interior of a convex set would meet its boundary in two distinct points. Hence we must have that $\|\hat{b}^1 + \frac{x_1^* - b_1^1}{b_1^2 - b_1^1}\hat{b}^2 - \hat{b}^1\| \geq 3\rho$, i.e., since $\|\hat{b}^1\| \leq RD$, that $\frac{\|\hat{b}^2 - \hat{b}^1\|}{b_1^2 - b_1^1} \geq \frac{3\rho - RD}{D}$. We have obtained that

$$|b_1^2 - b_1^1| = b_1^2 - b_1^1 \leq \frac{D}{3\rho - RD} \|\hat{b}^2 - \hat{b}^1\| = M \|\hat{b}^2 - \hat{b}^1\|. \tag{7}$$

In particular, since $b_1(\hat{0}) = 0$, we have also obtained that $|b_1| \leq M\|\hat{b}\| \leq MRD = \frac{RD^2}{3\rho - RD} = \rho$, so that $|b_1(\hat{t}) - x_1^*| \geq 2\rho$.

The equalities

$$\frac{\hat{z}^1}{z_1^1} = \frac{\hat{b}^1}{b_1^1 - x_1^*} \qquad \text{and} \qquad \frac{\hat{z}^2}{z_1^2} = \frac{\hat{b}^2}{b_1^2 - x_1^*}$$

yield

$$\hat{b}^1 - \hat{b}^2 = \left(\frac{\hat{z}^1}{z_1^1} - \frac{\hat{z}^2}{z_1^2} \right) (b_1^1 - x_1^*) + \frac{\hat{z}^2}{z_1^2} (b_1^1 - b_1^2)$$

so that (7) gives

$$(1 - MR) \|\hat{b}^1 - \hat{b}^2\| \leq \left\| \frac{\hat{z}^1}{z_1^1} - \frac{\hat{z}^2}{z_1^2} \right\| |b_1^1 - x_1^*| \leq \left\| \frac{\hat{z}^1}{z_1^1} - \frac{\hat{z}^2}{z_1^2} \right\| D;$$

hence we have both

$$\|\hat{b}^1 - \hat{b}^2\| \leq \frac{D}{1 - MR} \left\| \frac{\hat{z}^1}{z_1^1} - \frac{\hat{z}^2}{z_1^2} \right\|$$

and

$$|b_1^1 - b_1^2| \leq \frac{MD}{1 - MR} \left\| \frac{\hat{z}^1}{z_1^1} - \frac{\hat{z}^2}{z_1^2} \right\|,$$

which show that the map $\hat{t} \to b(\hat{t})$ is Lipschitzian (call β its Lipschitz constant).

(d) On $\Omega^* \setminus B(x^*, \rho)$, define the map

$$\Lambda(x) = \frac{x_1 - x_1^*}{b_1(\frac{\hat{x} - \hat{x}^*}{x_1 - x_1^*}) - x_1^*}.$$

We have that $\Lambda(x) = 1$ when $x = b(\frac{\hat{x}-\hat{x}^*}{x_1-x_1^*})$, so that the level set $\Lambda(x) = 1$ is $\partial\Omega^*$. We already know that $|b_1(\frac{\hat{x}-\hat{x}^*}{x_1-x_1^*}) - x_1^*| > 2\rho$ and that $b_1(\hat{t})$ is Lipschitzian. From (b) above, $x \to \hat{t}(x)$ is Lipschitzian. Hence, $\Lambda(x)$ is Lipschitzian and its Jacobian, $J\Lambda$, is bounded.

Consider the sets $\Lambda^{-1}(\lambda)$, which we shall call $(\partial\Omega)_\lambda$, and $(\Omega)_{\lambda^*} = \cup_{\lambda^* \le \lambda < 1}(\partial\Omega)_\lambda$. For x in it, from the equality $x_1 - x_1^* = \lambda(b_1 - x_1^*)$, we obtain $\|x - x^*\| \ge \lambda 2\rho$; hence, whenever $\frac{1}{2} \le \lambda^* \le \lambda \le 1$, we have that $\|x - x^*\| \ge \rho$, so that $(\Omega)_{\lambda^*} \subset \Omega^* \setminus B(x^*, \rho)$. For any set $V \subset \Omega_{\lambda^*}$, the coarea theorem yields

$$\int_{\Omega^*_{\lambda^*}} \chi_V(x) J\Lambda(x)\, dx = \int_{\lambda^*}^1 \left(\int_{(\partial\Omega)_\lambda \cap V} dH^{N-1} \right) d\lambda.$$

Hence, whenever $\int_{\Omega_{\lambda^*}} \chi_V(x) J\Lambda(x)\, dx < \varepsilon$ (ε and V to be fixed later), we must have $m(\{\lambda : H^{N-1}((\partial\Omega)_\lambda \cap V) \le \sqrt{\varepsilon}\}) \ge 1 - \lambda^* - \sqrt{\varepsilon}$, so that, for some λ_ε: $\lambda_\varepsilon - \lambda^* \le 2\sqrt{\varepsilon}$, we must have

$$H^{N-1}((\partial\Omega)_{\lambda_\varepsilon} \cap V) < \sqrt{\varepsilon}. \tag{8}$$

(e) Let $\psi_n \in C_c^\infty(\Omega)$, $\psi_n \to \eta$ in $W^{1,1}(\Omega)$. Fix any $\tilde{\lambda}$ with $\lambda^* \le \tilde{\lambda} < 1$. For $x \in \Omega^*_{\lambda^*}$, consider the Lipschitzian map $\hat{T}(x) = \frac{\hat{x}-\hat{x}^*}{x_1-x_1^*}$ and call $\hat{T}^* = \hat{T}(Z^*)$. We have

$$\begin{aligned}\int_{\Omega_{\lambda^*}} \|\nabla\psi_n(x)\| J\hat{T}(x)\, dx &\ge \int_{\Omega_{\tilde{\lambda}}} \left\langle \nabla\psi_n(x), \frac{x - x^*}{\|x - x^*\|} \right\rangle J\hat{T}(x)\, dx \\ &= \int_{\hat{T}^*} \left(\int_{\hat{T}^{-1}(\hat{t}) \cap \Omega_{\tilde{\lambda}}} \left\langle \nabla\psi_n(x), \frac{x - x^*}{\|x - x^*\|} \right\rangle dH^1 \right) d\hat{t}.\end{aligned}$$

The set $\hat{T}^{-1}(\hat{t}) \cap \Omega_{\tilde{\lambda}}$ is described by $\{x = x^* + \lambda(b(\hat{t}) - x^*) : \tilde{\lambda} < \lambda < 1\}$ and can be parametrized setting $\lambda = e^\tau$, so that $\tilde{x}(\tau) = x^* + (b(\hat{t}) - x^*)e^\tau$, $\tilde{x}'(\tau) = \tilde{x}(\tau) - x^*$, and $dH^1 = \|\tilde{x}(\tau) - x^*\| d\tau$. Hence

$$\begin{aligned}\int_{\Omega_{\lambda^*}} \|\nabla\psi_n(x)\| J\hat{T}(x)\, dx &\ge \int_{\hat{T}^*} \left(\int_0^{\ln\tilde{\lambda}} \langle \nabla\psi_n(\tilde{x}(\tau)), \tilde{x}(\tau) - x^* \rangle\, d\tau \right) d\hat{t} \\ &= \int_{\hat{T}^*} \left(\int_0^{\ln\tilde{\lambda}} \frac{d}{d\tau}\psi_n(\tilde{x}(\tau))\, d\tau \right) d\hat{t} = \int_{\hat{T}^*} \psi_n(x^* + \tilde{\lambda}(b(\hat{t}) - x^*))\, d\hat{t}.\end{aligned} \tag{9}$$

We wish to estimate the last integral by suitably choosing $\tilde{\lambda}$. Set $V_n = \{x \in \Omega_{\lambda^*} : |\psi_n(x) - \eta(x)| \ge \frac{\delta}{2}\}$: we know that, in particular, ψ_n converges to η in measure and, since $J\Lambda$ is bounded, we obtain that

$$\int_{\Omega_{\lambda^*}} \chi_{V_n} J\Lambda(x)\, dx = \varepsilon_n \to 0,$$

and hence, by (8), that there exist λ_n, such that at once we have $\lambda_n - \lambda^* \le 2\sqrt{\varepsilon_n}$ and $H^{N-1}((\partial\Omega)_{\lambda_n} \cap V_n) < \sqrt{\varepsilon_n}$. Since $H^{N-1}(\partial\Omega)_{\lambda_n} = \lambda_n^{N-1} H^{N-1}(\partial\Omega^*)$, we obtain

$$H^{N-1}((\partial\Omega)_{\lambda_n} \setminus V_n) \ge \lambda_n^{N-1} H^{N-1}(\partial\Omega^*) - \sqrt{\varepsilon_n}. \tag{10}$$

There exists n^* such that for every $n \ge n^*$ we have

$$H^{N-1}((\partial\Omega)_{\lambda_n} \setminus V_n) \ge \frac{1}{2}(\lambda^*)^{N-1} H^{N-1}(\partial\Omega^*) = C > 0. \tag{11}$$

(f) Recall that, for $x \in (\partial\Omega)_{\lambda_n} \setminus V_n$, we have both $|\psi_n(x)-\eta(x)| \le \frac{\delta}{2}$ and $\eta(x) \ge \delta$, inferring that $\psi_n(x) \ge \frac{\delta}{2}$ there. Call $\hat{T}_n = \{\hat{T}(x) : x \in (\partial\Omega)_{\lambda_n} \setminus V_n\}$: we also obtain that $(\partial\Omega)_{\lambda_n} \setminus V_n = \{b_n(\hat{t}) : \hat{t} \in \hat{T}_n\}$, where $b_n(\hat{t}) = x^* + \lambda_n(b(\hat{t}) - x^*)$. Hence, we have

$$\int_{\hat{T}_n} \psi_n(x^* + \lambda_n(b(\hat{t}) - x^*))\, d\hat{t} \ge \frac{\delta}{2} m_{(N-1)}(\hat{T}_n).$$

On the other hand, the map b_n is Lipschitzian, since, from the conclusion of (c) above, so is b, and its Lipschitz constant is bounded by β, the Lipschitz constant of b, so that $H^{N-1}((\partial\Omega)_{\lambda_n} \setminus V_n) = H^{N-1}(b_n(\hat{T}_n)) \le \beta m_{(N-1)}(\hat{T}_n)$. Hence, for $n \ge n^*$, we have

$$m_{(N-1)}(\hat{T}_n) \ge \frac{1}{\beta} H^{N-1}(\partial\Omega)_{\lambda_n} \setminus V_n \ge \frac{1}{\beta} C$$

and, from (9), we conclude

$$\begin{aligned}
\int_{\Omega_{\lambda_n}} \|\nabla\eta(x)\| J\hat{T}(x)\, dx &\ge \int_{\Omega_{\lambda_n}} \|\nabla\psi_n(x)\| J\hat{T}(x)\, dx - \int_{\Omega_{\lambda_n}} \|\nabla\eta(x) - \nabla\psi_n(x)\| J\hat{T}(x)\, dx \\
&\ge \int_{\hat{T}_n} \psi_n(x^* + \lambda^*(b(\hat{t}) - x^*))\, d\hat{t} - \int_{\Omega} \|\nabla\eta(x) - \nabla\psi_n(x)\| J\hat{T}(x)\, dx \\
&\ge \frac{\delta}{2}\frac{1}{\beta} C - \int_{\Omega} \|\nabla\eta(x) - \nabla\psi_n(x)\| J\hat{T}(x)\, dx,
\end{aligned}$$

so that

$$\int_{\Omega_{\lambda^*}} \|\nabla\eta(x)\| J\hat{T}(x)\, dx = \lim_{n\to\infty} \int_{\Omega_{\lambda_n}} \|\nabla\eta(x)\| J\hat{T}(x)\, dx \ge \frac{\delta}{2}\frac{1}{\beta} C.$$

However, as $\lambda^* \to 1$, we have $m(\Omega_{\lambda^*}) \to 0$, so the previous estimate contradicts the integrability of $\nabla\eta$. □

Proof of Theorem 2. For brevity, set $h_\theta = h^+_{\theta,x^0,r}$. Let $\eta = (w - h_\theta)^+$, so that $\eta \in W^{1,1}_0(\Omega)$. Set $E^+ = \{x : \eta(x) > 0\}$: we wish to prove that $m(E^+) = 0$.

(a) Consider $v = w - \eta$: we have

$$\nabla v(x) = \begin{cases} \nabla u(x), & x \in (E^+)^c, \\ \nabla h_\theta(x), & x \in E^+. \end{cases}$$

Since $w - v \in W^{1,1}_0(\Omega)$, from the above we obtain

$$0 = \int_\Omega \langle\theta, \nabla w(x) - \nabla v(x)\rangle dx = \int_{E^+} \langle\theta, \nabla w(x) - \nabla h_\theta(x)\rangle dx.$$

Moreover, $\nabla h_\theta(x) \in \partial f^*(\theta)$, so that $\theta \in \partial f(\nabla h_\theta(x))$, and the convexity of f implies

$$\begin{aligned}
&\int_\Omega (f(\nabla w(x)) - f(\nabla v(x)))dx \\
&= \int_{E^+} (f(\nabla w(x)) - f(\nabla h_\theta(x))\, dx \ge \int_{E^+} \langle\theta, \nabla w(x) - \nabla h_\theta(x)\rangle dx = 0.
\end{aligned}$$

On the other hand, w is a minimum, so that

$$0 \ge \int_\Omega (f(\nabla w(x)) - f(\nabla v(x)))dx = \int_{E^+} (f(\nabla w(x)) - f(\nabla h_\theta(x)))dx,$$

so that we conclude that

$$\int_{E^+} (f(\nabla w(x)) - f(\nabla h_\theta(x)))dx = 0.$$

Adding the equalities, we have

$$\int_{E^+} (f(\nabla w(x)) - f(\nabla h_\theta(x)) + \langle \theta, \nabla w(x) - \nabla v(x)\rangle)\, dx = 0.$$

Since the integrand is nonnegative, we obtain that, for a.e. $x \in E^+$,

$$f(\nabla w(x)) - f(\nabla h_\theta(x)) + \langle \theta, \nabla w(x) - \nabla v(x)\rangle = 0.$$

(b) The previous equality can be rewritten as

$$\langle \theta, \nabla w\rangle - f(\nabla w) = \langle \theta, \nabla v\rangle - f(\nabla h_\theta).$$

Since $\theta \in \partial f(\nabla h_\theta(x))$, as is well known we have $\langle \theta, \nabla v\rangle - f(\nabla h_\theta) = f^*(\theta)$, so that also $\langle \theta, \nabla w\rangle - f(\nabla w) = f^*(\theta)$; then we obtain that $\theta \in \partial f(\nabla w)$ and finally that, a.e. on E^+,

$$\nabla w(x) \in \partial f^*(\theta),$$

and hence that, recalling (4), a.e. on E^+, we have

$$\langle \nabla w(x), x - x^0\rangle \le \sup_{k \in \partial f^*(\theta)} \{\langle k, x - x^0\rangle\} = \langle \nabla h_\theta(x), x - x^0\rangle.$$

We have obtained that, a.e. on Ω,

$$\langle \nabla \eta(x), x - x^0\rangle \le 0. \tag{12}$$

In addition, from the assumption on f^*, we obtain that there exists K such that, a.e. on E^+, $\|\nabla w(x)\| \le K$. This inequality will be used in (f).

(c) To show that $m(E^+) = 0$, we shall prove that the assumption that there exist $\delta > 0$ and $E^+_\delta \subset E^+$ such that $m(E^+_\delta) > 0$ and $\eta(x) \ge \delta$ on E^+_δ leads to a contradiction.

Let x^* be a point of density of E^+_δ and let ρ^* be such that $B(x^*, \rho^*) \subset \Omega$. Then, on a.e. line connecting $x \in B(x^*, \rho^*)$ and x^0, the map η is absolutely continuous. The estimate (12) implies that on any such segment (x, x^0), we have that $\eta \ge \delta$, and one would like to conclude that $\eta \notin W^{1,1}_0(\Omega)$, the contradiction we seek. The reasoning to show this contradiction is based on some version of the Fubini–Tonelli theorem, and has been used in [2] and [9]. In the present situation, however, x^0 can belong to $\partial\Omega$ and, in this case, knowing that $\eta \ge \delta$ on segments of the form (x, x^0), does not by itself prevent η from being in $W^{1,1}_0(\Omega)$. It is this case we are going to examine in some detail.

Consider $x^0 \in \partial\Omega$; let $H = \{x : \langle h, x - x^0\rangle = 0\}$ be a supporting plane to Ω through x^0, and set $H^- = \{x : \langle h, x - x^0\rangle < 0\}$ with $\Omega \subset H^+$. Let $\tilde{x}^2$ be in H^-; for $\lambda \in (0, 1)$, set $x^*_\lambda = x^0 + \lambda(x^* - x^0)$ and $\tilde{x}^2_\lambda = x^0 + \lambda(\tilde{x}^2 - x^0)$. Let $x^1_\lambda = \alpha x^*_\lambda + (1-\alpha)\tilde{x}^2_\lambda$ be the intersection of the segment $(x^*_\lambda, \tilde{x}^2_\lambda)$ with H, described by

$$\alpha = \frac{\langle h, x^0 - \tilde{x}^2_\lambda\rangle}{\langle h, x^*_\lambda - \tilde{x}^2_\lambda\rangle} = \frac{\langle h, x^1_\lambda - \tilde{x}^2_\lambda\rangle}{\langle h, x^*_\lambda - \tilde{x}^2_\lambda\rangle}.$$

Since

$$\frac{\langle h, x^1_\lambda - \tilde{x}^2_\lambda\rangle}{\langle h, x^*_\lambda - \tilde{x}^2_\lambda\rangle} = \frac{\langle h, x^0 - \tilde{x}^2\rangle}{\langle h, x^* - \tilde{x}^2\rangle},$$

α is independent on λ. Consider the expression

$$\frac{\|x^*_\lambda - \tilde{x}^2_\lambda\|}{\|x^*_\lambda - x^1_\lambda\|}\left(1 - e^{-\frac{\delta}{2K\|\tilde{x}^2_\lambda - x^0\|}}\right).$$

As λ decreases to zero, the ratio

$$\frac{\|x^*_\lambda - \tilde{x}^2_\lambda\|}{\|x^*_\lambda - x^1_\lambda\|} = \frac{1}{1-\alpha} > 1$$

does not change, while $(1 - e^{-\frac{\delta}{2K\|\tilde{x}^2_\lambda - x^0\|}}) \uparrow 1$. Hence, we can assume to have chosen $\lambda > 0$ such that

$$\frac{\|x^*_\lambda - \tilde{x}^2_\lambda\|}{\|x^*_\lambda - x^1_\lambda\|}\left(1 - e^{-\frac{\delta}{2K\|\tilde{x}^2_\lambda - x^0\|}}\right) = \ell > 1.$$

Having fixed λ, in what follows we will call x^* the point x^*_λ and $\tilde{x}^2$ the point $\tilde{x}^2_\lambda$. For x^2 in a small neighborhood of $\tilde{x}^2$, the unique intersection of the segment (x^2, x^*) with H will be denoted by $x^1_{x^2}$. We notice that, since the original point x^* was of density for E^+_δ and a.e. segment (x^0, x) with x in E^+_δ belongs to E^+_δ, we obtain that the point x^* just fixed is again of density for E^+_δ.

Set $x_{x^2}(t) = x^2 + e^t(x^* - x^2)$ and $t(x^2) = -\frac{\delta}{2K\|x^2 - x^0\|}$. In particular, for $x^2 = \tilde{x}^2$, we have

$$\|x_{\tilde{x}^2}(t(\tilde{x}^2)) - x^*\| = \|x^* - x^1_{\tilde{x}^2}\|\frac{\|x^* - \tilde{x}^2\|}{\|x^* - x^1_{\tilde{x}^2}\|}\left(1 - e^{-\frac{\delta}{2K\|\tilde{x}^2 - x^0\|}}\right) = \ell\|x^* - x^1_{\tilde{x}^2}\|.$$

By continuity, there exists $r > 0$ so that $x^2 \in B(\tilde{x}^2, r)$ implies

$$\frac{\|x^* - x^2\|}{\|x^* - x^1_{x^2}\|}\left(1 - e^{-\frac{\delta}{2K\|x^2 - x^0\|}}\right) \geq \frac{1}{2} + \frac{\ell}{2} > 1.$$

In other words, for every $x^2 \in B(\tilde{x}^2, r)$, $\|x_{x^2}(t(x^2)) - x^*\| > \|x^* - x^1(x^2)\|$, i.e., the segment $(x^*, x_{x^2}(t(x^2)))$ intersects H. Equivalently, the map $x_{x^2}(t)$ takes values in Ω on some interval $(\alpha_{x^2}, 0)$ with $\alpha_{x^2} \geq t(x^2)$.

(d) Since x^* is a point of density of E^+_δ, we have that $m(x^* + rB \cap E^+_\delta) = (1 - \varepsilon(r))m(x^* + rB)$, and by the coarea theorem we obtain

$$\int_0^r H^{(N-1)}(x^* + s\partial B \cap E^+_\delta)\, ds = m(x^* + rB \cap E^+_\delta) = (1 - \varepsilon(r))m(x^* + rB)$$
$$= \int_0^r (1 - \varepsilon(r))H^{(N-1)}(x^* + s\partial B)\, ds = \int_0^r (1 - \varepsilon(r))s^{N-1}N\omega_N\, ds,$$

so that there exist $r_\nu \downarrow 0$ such that $H^{(N-1)}(x^* + r_\nu\partial B \cap E^+_\delta) \geq (1 - \varepsilon(r_\nu))r_\nu^{N-1}N\omega_N$.

For brevity, when $x \neq x^*$, we will use the notation $z(x) = \frac{x - x^*}{\|x - x^*\|}$. Consider $Z = z(B(\tilde{x}^2, r))$. We have that

$$\frac{H^{(N-1)}(Z)}{N\omega_N} = L > 0.$$

Since $H^{(N-1)}(x^* + r_\nu Z) = r_\nu^{(N-1)} N\omega_N L$, whenever $\varepsilon(r_\nu) < L$, we have that

$$H^{(N-1)}(x^* + r_\nu Z \cap E_\delta^+) = M_\nu > 0.$$

Fix one such ν; set Z^+ to be the subset of Z defined by $x^* + r_\nu Z^+ = x^* + r_\nu Z \cap E_\delta^+$, so that $H^{(N-1)}(x^* + r_\nu Z^+) = M_\nu$, i.e.,

$$H^{(N-1)}(Z^+) = \frac{M_\nu}{r_\nu^{N-1}} > 0. \tag{13}$$

(e) Define β_{x^2} by the equation $x^* - (x^2 + e^{\beta_{x^2}}(x^* - x^2)) = r_\nu$. For almost every $z \in Z^+$, the map $\eta(x_{x^2}(t))$ is absolutely continuous on $(\alpha_{x^2}, \beta_{x^2})$ and we can write

$$\begin{aligned}
&\eta(x_{x^2}(\beta_{x^2}) = \eta(x_{x^2}(t)) + \int_t^{\beta_{x^2}} \frac{d}{ds}\eta(x_{x^2}(s))\, ds \\
&= \eta(x_{x^2}(t)) + \int_t^{\beta_{x^2}} \langle \nabla\eta(x_{x^2}(s)), x'_{x^2}(s)\rangle\, ds \\
&= \eta(x_{x^2}(t)) + \int_t^{\beta_{x^2}} \langle \nabla\eta(x_{x^2}(s)), [(x_{x^2}(s) - x^0) + (x^0 - x^2)]\rangle\, ds.
\end{aligned}$$

Recalling that $\eta(x_{x^2}(\beta_{x^2}) \geq \delta$; that, for $x \in E^+$, we have $\langle \nabla\eta(x), x - x^0\rangle \leq 0$; that $\|\nabla\eta(x)\| \leq K$; and that $-t \leq -t(x^2) = \frac{\delta}{2K\|x^2 - x^0\|}$, we obtain

$$\delta \leq \eta(x_{x^2}(\beta_{x^2})) \leq \eta(x_{x^2}(t)) + (-t)K\|x^0 - x^2\| \leq \eta(x_{x^2}(t)) + \frac{\delta}{2},$$

so that $\eta(x_{x^2}(t)) \geq \frac{\delta}{2}$.

Hence, we have obtained that $\eta \geq \frac{\delta}{2}$ on the intersection of Ω with the set of half lines $x^* + \lambda Z^+$ issuing from x^*, where $H^{(N-1)}(Z^+) > 0$. Applying Lemma 2 we obtain that $\eta \notin W_0^{1,1}(\Omega)$, a contradiction. □

3. The bounded slope condition. The bounded slope condition is imposed on the boundary datum u_0: classically, under the assumption of strict convexity on f, it demands the existence, for every $x^0 \in \partial\Omega$, of two vectors k^+ and k^- (depending on x^0) such that, for every $x \in \partial\Omega$, one has

$$u_0(x^0) + \langle k^-, x - x^0\rangle \leq u_0(x) \leq u_0(x^0) + \langle k^+, x - x^0\rangle$$

and, in addition, the existence of K such that

$$K \geq \sup_{x^0 \in \partial\Omega} \max\{\|k^-\|, \|k^+\|\}.$$

Its purpose is to infer that, for a solution w to problem (1), one has $\|\nabla w(x)\| \leq K$ for almost every $x \in \Omega$.

Example 1 shows that this result, the proof of which depends on a comparison theorem, cannot possibly be true without the assumption of strict convexity of the Lagrangian f. Our present aim is to be able to provide estimates for the gradient of a solution in those cases, as the examples mentioned in this paper, where the Lagrangian f is non–strictly convex on a bounded set and becomes strictly convex for large values of ξ. In it, the notation $\|A\| = \sup_{a\in A}\{\|a\|\}$ will be used.

THEOREM 3. *Let Ω be an open, bounded, and convex set; let f be an extended valued, lower semicontinuous convex function, such that $Dom(f^*)$ is open; assume that for every $x^0 \in \partial\Omega$ there exist $\theta^+(x^0)$ and $\theta^-(x^0)$ such that, for every $x \in \partial\Omega$,*

$$u_0(x^0) + (I_{\partial f^*(\theta_)})^*(x - x^0) \le u_0(x) \le u_0(x^0) + (I_{\partial f^*(\theta+)})^*(x - x^0).$$

In addition, assume that there is K such that

$$K \ge \sup_{x^0 \in \partial\Omega} \max\{\|\partial f^*(\theta^+(x^0))\|, \|\partial f^*(\theta^-(x^0))\|\}.$$

Furthermore, assume that when $\|\xi\| > K$, f is strictly convex at ξ. If $w \in C(\Omega) \cap W^{1,1}(\Omega)$ is a solution to problem (1), then w is Lipschitzian and, for a.e. $x \in \Omega$, we have

$$\|\nabla w(x)\| \le K.$$

For the problem in Example 2, we obtain, for a solution w, the bound $|w'(x)| \le \sqrt{2}$, independent of ε.

Proof. The proof is a minor modification of the proof of Theorem 4.1 of [2]. ∎

REFERENCES

[1] A. CELLINA, *Minimizing a functional depending on ∇u and on u*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 339–352.

[2] A. CELLINA, *On the bounded slope condition and the validity of the Euler Lagrange equation*, SIAM J. Control Optim., 40 (2001), pp. 1270–1279.

[3] A. CELLINA, *The Euler Lagrange equation and the Pontriagin maximum principle*, Bol. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8), 8 (2005), pp. 323–347.

[4] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992

[5] E. GIUSTI, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Basel, 1984.

[6] J. GOODMAN, R. V. KOHN, AND L. REYNA, *Numerical study of a relaxed variational problem from optimal design*, Comput. Methods Appl. Mech. Engrg., 57 (1986), pp. 107–127.

[7] B. KAWOHL, J. STARÁ, AND G. WITTUM, *Analysis and numerical studies of a problem of shape design*, Arch. Rational Mech. Anal., 114 (1991), pp. 349–363.

[8] R. V. KOHN, AND G. STRANG, *Optimal design and relaxation of variational problems.* I, Comm. Pure Appl. Math., 39 (1986), pp. 113–137.

[9] C. MARICONDA AND G. TREU, *A comparison principle and the Lipschitz continuity for minimizers*, J. Convex Anal., 12 (2005), pp. 197–212.

[10] C. MARICONDA AND G. TREU, *Absolutely continuous representatives on curves for Sobolev functions*, J. Math. Anal. Appl., 281 (2003), pp. 171–185.

[11] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 2, pp. 750–774

LINEAR QUADRATIC DIFFERENTIAL GAMES: SADDLE POINT AND RICCATI DIFFERENTIAL EQUATION*

MICHEL C. DELFOUR†

Abstract. Zhang [*SIAM J. Control Optim.*, 43 (2005), pp. 2157–2165] recently established the equivalence between the finiteness of the *open loop value* of a two-player zero-sum linear quadratic (LQ) game and the finiteness of its open loop lower and upper values. In this paper we complete and sharpen the results of Zhang for the finiteness of the lower value of the game by providing a set of necessary and sufficient conditions that emphasizes the *feasibility condition*: $(0,0)$ is a solution of the open loop lower value of the game for the zero initial state. Then we show that, under the assumption of an open loop saddle point in the time horizon $[0,T]$ for all initial states, there is an open loop saddle point in the time horizon $[s,T]$ for all initial times s, $0 \le s < T$, and all initial states at time s. From this we get an *optimality principle* and adapt the *invariant embedding approach* to construct the decoupling symmetrical matrix function $P(s)$ and show that it is an $H^1(0,T)$ solution of the matrix Riccati differential equation. Thence an open loop saddle point in $[0,T]$ yields closed loop optimal strategies for both players. Furthermore, a necessary and sufficient set of conditions for the existence of an open loop saddle point in $[0,T]$ for all initial states is the convexity-concavity of the utility function and the existence of an $H^1(0,T)$ symmetrical solution to the matrix Riccati differential equation. As an illustration of the cases where the open loop lower/upper value of the game is $-\infty/+\infty$, we work out two informative examples of solutions to the Riccati differential equation with and without blow-up time.

Key words. linear quadratic differential game, saddle point, value of a game, Riccati differential equation, open loop and closed loop strategies, conjugate point, blow-up time

AMS subject classifications. 91A05, 91A23, 49N70, 91A25

DOI. 10.1137/050639089

1. Introduction.

We consider the *two-player zero-sum game* with *linear dynamics* and a *quadratic utility function* over a *finite time horizon*. The min sup problem was studied in 1969 by [5]. The fundamental theory of closed loop linear quadratic (LQ) games was given in 1979 by Bernhard [4] followed by the seminal book of Başar and Bernhard [1] in 1991 and 1995. A very nice paper by Zhang [10] in 2005 established the equivalence between the finiteness of the open loop value of a two-player zero-sum LQ game and the finiteness of its open loop lower and upper values.

In this paper we complete and sharpen the results of [10] for the finiteness of the lower value of the game by providing a set of necessary and sufficient conditions (Theorem 2.2) that emphasizes the *feasibility condition*: $(0,0)$ is a solution of the open loop lower value of the game for the zero initial state. A similar feasibility condition holds for the finiteness of the open loop upper value and value of the game. It also recasts the results in the more intuitive state-adjoint state framework.

Then we show that, under the assumption of an open loop saddle point in the time horizon $[0,T]$ for all initial states, there is a unique open loop saddle point in

*Received by the editors August 29, 2005; accepted for publication (in revised form) October 30, 2006; published electronically May 22, 2007. This research has been supported by National Sciences and Engineering Research Council of Canada discovery grants and by an FQRNT grant from the Ministère de l'Éducation du Québec.

http://www.siam.org/journals/sicon/46-2/63908.html

†Centre de recherches mathématiques et Département de mathématiques et de statistique, Université de Montréal, C. P. 6128, succ. Centre-ville, Montréal H3C 3J7, QC, Canada (delfour@crm.UMontreal.ca).

the time horizon $[s,T]$ for all initial times s, $0 \le s < T$, and all initial states at time s (Theorem 5.2(iii)). From this we get an *optimality principle* and adapt the *invariant embedding approach* of Bellman in the style of Lions [9] to construct the decoupling symmetrical matrix function $P(s)$ (Theorem 2.9) and show that it is an $H^1(0,T)$ solution of the matrix Riccati differential equation. Thence an open loop saddle point in $[0,T]$ yields closed loop optimal strategies for both players who achieve a closed loop-closed loop saddle point in the sense of Bernhard [4]. Furthermore, a necessary and sufficient set of conditions for the existence of an open loop saddle point in $[0,T]$ for all initial states is the convexity-concavity of the utility function and the existence of a symmetrical $H^1(0,T)$ solution to the matrix Riccati differential equation (Theorem 2.10). As an illustration of the cases where the open loop lower/upper value of the game is $-\infty/+\infty$, we work out two informative examples of solutions to the Riccati differential equation with and without blow-up time.

2. Definitions, notation, and main results.

2.1. System, utility function, values of the game. Given a finite dimensional Euclidean space $\mathbf{R}^d$ of dimension $d \ge 1$, the *norm* and *inner product* will be denoted by $|x|$ and $x \cdot y$, respectively, irrespective of the dimension d of the space. Given $T > 0$, the norm and inner product in $L^2(0,T;\mathbf{R}^n)$ will be denoted $\|f\|$ and (f,g). The norm in the Sobolev space $H^1(0,T;\mathbf{R}^n)$ will be written $\|f\|_{H^1}$.

Consider the following two-player zero-sum game over the finite time interval $[0,T]$ characterized by the quadratic *utility function*

$$C_{x_0}(u,v) \stackrel{\text{def}}{=} Fx(T)\cdot x(T) + \int_0^T Q(t)x(t)\cdot x(t) + |u(t)|^2 - |v(t)|^2\,dt, \tag{2.1}$$

where x is the solution of the linear differential system

$$\frac{dx}{dt}(t) = A(t)x(t) + B_1(t)u(t) + B_2(t)v(t) \quad \text{a.e. in } [0,T], \quad x(0) = x_0, \tag{2.2}$$

x_0 is the *initial state* at time $t = 0$, $u \in L^2(0,T;\mathbf{R}^m)$, $m \ge 1$, is the strategy of the first player, and $v \in L^2(0,T;\mathbf{R}^k)$, $k \ge 1$, is the strategy of the second player. We assume that F is an $n \times n$ matrix and that A, B_1, B_2, and Q are matrix functions of appropriate order that are measurable and bounded a.e. in $[0,T]$. Moreover, $Q(t)$ and F are symmetrical. It will be convenient to use the following compact notation and drop the "a.e. in $[0,T]$" wherever no confusion arises:

$$C_{x_0}(u,v) = Fx(T)\cdot x(T) + \int_0^T Qx\cdot x + |u|^2 - |v|^2\,dt, \tag{2.3}$$

$$x' = Ax + B_1u + B_2v \quad \text{in } [0,T], \quad x(0) = x_0. \tag{2.4}$$

The above assumptions on F, A, B_1, B_2, and Q will be used throughout this paper.

Remark 2.1. The more general quadratic structure involving cross terms and different quadratic weights $N_1u\cdot u$ and $N_2v\cdot v$ on u and v (cf., for instance, Bernhard [4, section 2, p. 53]),

$$\int_0^T (x,u,v)\cdot \begin{bmatrix} Q & S & T \\ S^* & N_1 & 0 \\ T^* & 0 & -N_2 \end{bmatrix} \begin{bmatrix} x \\ u \\ v \end{bmatrix} dt,$$

can be brought back to the simpler form (2.1)–(2.2) by the following change of variables:

$$\begin{bmatrix} x \\ u \\ v \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ -N_1^{-1}S^* & N_1^{-1/2} & 0 \\ N_2^{-1}T^* & 0 & N_2^{-1/2} \end{bmatrix} \begin{bmatrix} x \\ \bar{u} \\ \bar{v} \end{bmatrix},$$

where $N_1(t)$ and $N_2(t)$ are symmetrical positive definite matrices such that

$$\begin{aligned} &\exists \nu_1 > 0 \text{ such that } \forall u \in \mathbf{R}^m \text{ and almost all } t, \quad N_1(t)u \cdot u \geq \nu_1 |u|^2, \\ &\exists \nu_2 > 0 \text{ such that } \forall v \in \mathbf{R}^k \text{ and almost all } t, \quad N_2(t)v \cdot v \geq \nu_2 |v|^2. \end{aligned} \tag{2.5}$$

This yields the simpler initial structure with the system and the utility function

$$x' = \bar{A}x + \bar{B}_1\bar{u} + \bar{B}_2\bar{v}, \quad \int_0^T \bar{Q}x \cdot x + |\bar{u}|^2 - |\bar{v}|^2 \, dt$$

by introducing the new matrices

$$\begin{aligned} &\bar{A} = A - B_1N_1^{-1}S^* + B_2N_2^{-1}T^*, \quad \bar{B}_1 = B_1N_1^{-1/2}, \quad \bar{B}_2 = B_2N_2^{-1/2}, \\ &\bar{Q} = Q - SN_1^{-1}S^* + TN_2^{-1}T^*. \end{aligned}$$

The matrix functions N_1, N_2, S, and T are all assumed to be measurable and bounded.

DEFINITION 2.1. *Let x_0 be an initial state in $\mathbf{R}^n$ at time $t = 0$.*

(i) *The game is said to achieve its* open loop lower value *(resp.,* upper value*) if*

$$v^-(x_0) \stackrel{\text{def}}{=} \sup_{v \in L^2(0,T;\mathbf{R}^k)} \inf_{u \in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,v) \tag{2.6}$$

$$\textit{(resp., } v^+(x_0) \stackrel{\text{def}}{=} \inf_{u \in L^2(0,T;\mathbf{R}^m)} \sup_{v \in L^2(0,T;\mathbf{R}^k)} C_{x_0}(u,v)\textit{)} \tag{2.7}$$

is finite. By definition $v^-(x_0) \leq v^+(x_0)$.

(ii) *The game is said to achieve its* open loop value *if its open loop lower value $v^-(x_0)$ and upper value $v^+(x_0)$ are achieved and $v^-(x_0) = v^+(x_0)$. The* open loop value *of the game will be denoted by $v(x_0)$.*

(iii) *A pair $(\bar{u}, \bar{v})$ in $L^2(0,T;\mathbf{R}^m) \times L^2(0,T;\mathbf{R}^k)$ is an* open loop saddle point *of $C_{x_0}(u,v)$ in $L^2(0,T;\mathbf{R}^m) \times L^2(0,T;\mathbf{R}^k)$ if for all u in $L^2(0,T;\mathbf{R}^m)$ and all v in $L^2(0,T;\mathbf{R}^k)$*

$$C_{x_0}(\bar{u},v) \leq C_{x_0}(\bar{u},\bar{v}) \leq C_{x_0}(u,\bar{v}). \tag{2.8}$$

In general, (ii) *does not necessarily imply* (iii)*, but we shall see that it does for LQ games.*

DEFINITION 2.2. *Associate with $x_0 \in \mathbf{R}^n$ the sets and the functions*

$$V(x_0) \stackrel{\text{def}}{=} \left\{ v \in L^2(0,T;\mathbf{R}^k) : \inf_{u \in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,v) > -\infty \right\}, \tag{2.9}$$

$$U(x_0) \stackrel{\text{def}}{=} \left\{ u \in L^2(0,T;\mathbf{R}^m) : \sup_{v \in L^2(0,T;\mathbf{R}^k)} C_{x_0}(u,v) < +\infty \right\}, \tag{2.10}$$

$$J^-_{x_0}(v) \stackrel{\text{def}}{=} \inf_{u \in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,v), \quad J^+_{x_0}(u) \stackrel{\text{def}}{=} \sup_{v \in L^2(0,T;\mathbf{R}^k)} C_{x_0}(u,v). \tag{2.11}$$

By definition, $V(x_0) \neq \varnothing$ if and only if $v^-(x_0) > -\infty$, and $U(x_0) \neq \varnothing$ if and only if $v^+(x_0) < +\infty$.

2.2. Saddle points of the game and solution of the Riccati differential equation. In the literature, an important issue is the connection between the existence of a symmetrical solution to the *matrix Riccati differential equation*

$$P' + PA + A^*P - PRP + Q = 0 \text{ a.e. in } [0,T], \quad P(T) = F, \tag{2.12}$$

where $R = B_1B_1^* - B_2B_2^*$, and the existence of either an open or closed loop[1] lower value, upper value, or saddle point of the game. For instance, in the closed loop case, quoting Bernhard [4] in his introduction,

> "It has long been known that, for the two-person, zero-sum differential game with linear dynamics, quadratic payoff, fixed end-time, and free end-state (*standard LQ game*), the existence of a solution to a Riccati equation is a sufficient condition for the existence of a saddle point within the class of instantaneous state feedback strategies (cf. [8], [7]), and therefore within any wider class (cf. [3])."

Similarly, we quote Zhang [10] in his introduction,

> "(a) if the Riccati differential equation admits a solution, then, the game admits a closed loop-closed loop saddle point,"

where he refers to [4].

In the open loop case, the above statements are incomplete (cf. Example 2.2), even under the assumptions

$$F \geq 0 \text{ and } Q(t) \geq 0 \text{ a.e. in } [0,T]$$

used in [4] that necessarily imply the convexity of $C_{x_0}(u,v)$ with respect to u and $V(x_0) = L^2(0,T;\mathbf{R}^k)$ for all $x_0 \in \mathbf{R}^n$. Even when the solution of the Riccati differential equation (2.12) is $H^1(0,T)$ or bounded (Remark 2.5), it is also *necessary* that the utility function be convex in u and concave in v (Theorem 2.10) to get an open loop-open loop saddle point.

This leaves the cases where either the open loop lower or upper value of the game explodes. In such cases the solution of the Riccati differential equation might have a blow-up time as illustrated in Example 2.1 (cf. Bernhard [4, Example 5.1, p. 67]:

> "The following game has a saddle point that *survives* a conjugate point,"

where he means a closed loop-closed loop saddle point). The conjugate point corresponds to a *blow-up time* of the solution of the Riccati equation (2.12), where the solution is not of the $H^1(0,T)$ type. Finally, an open loop saddle point yields closed loop optimal strategies that achieve a closed loop-closed loop saddle point (Theorem 2.9), but the *converse is not necessarily true*. It is informative to first detail the example of Bernhard.

Example 2.1. Consider the dynamics and utility function in the time interval $[0,2]$:

$$x'(t) = (2-t)\,u(t) + t\,v(t) \text{ a.e. in } [0,2], \quad x(0) = x_0, \tag{2.13}$$

$$C_{x_0}(u,v) = \frac{1}{2}|x(2)|^2 + \int_0^2 |u(t)|^2 - |v(t)|^2\,dt. \tag{2.14}$$

Here $A = 0$, $B_1(t) = 2-t$, $B_2(t) = t$, $F = 1/2$, $Q = 0$, and $R = B_1B_1^* - B_2B_2^* = 4(1-t)$. It is shown in [4] that the Riccati equation reduces to

$$P' - 4(1-t)P^2 = 0, \quad P(2) = 1/2 \quad \Rightarrow P(t) = \frac{1}{2(t-1)^2}.$$

[1] The reader is referred to Bernhard [4] for the *closed loop* definitions.

Its solution is positive and blows up at $t = 1$. It is not an element of $H^1(0,2)$. We now show that there is no open loop saddle point in the time interval $[0,2]$. For the open loop lower value of the game, the minimization with respect to u has a unique solution for all (x_0, v) since the utility function $u \mapsto C_{x_0}(u,v)$ is convex and bounded below by $-\|v\|^2_{L^2}$. The minimizer is completely characterized by the coupled system

$$\begin{cases} x'(t) = (2-t)\,\hat{u}(t) + t\,v(t) \text{ a.e. in } [0,2], & x(0) = x_0, \\ p'(t) = 0 \text{ a.e. in } [0,2], & p(2) = \dfrac{1}{2}x(2), \\ \hat{u}(t) = -(2-t)\,p(t). \end{cases}$$

From this

$$x(2) = \frac{3}{7}\left[x_0 + \int_0^2 s\,v(s)\,ds\right] \quad \text{and} \quad p(t) = \frac{1}{2}x(2)$$

and

$$\begin{aligned} J^-_{x_0}(v) &\stackrel{\text{def}}{=} \inf_{u \in L^2(0,2;\mathbf{R})} C_{x_0}(u,v) \\ &= C_{x_0}(\hat{u},v) = \frac{1}{2}x(2)^2 + \frac{1}{4}\,x(2)^2 \int_0^2 (2-t)^2\,dt - \int_0^2 |v(t)|^2\,dt \\ &= \frac{7}{6}x(2)^2 - \int_0^2 |v(t)|^2\,dt = \frac{3}{14}\left[x_0 + \int_0^2 s\,v(s)\,ds\right]^2 - \int_0^2 |v(t)|^2\,dt. \end{aligned}$$

It is readily seen that $J^-_{x_0}$ is concave in v and that the supremum with respect to v of $J^-_{x_0}(v)$ exists. Indeed, from the first order condition,[2]

$$\forall v,\ \frac{1}{2}dJ^-_{x_0}(\hat{v};v) = \frac{3}{14}\left[x_0 + \int_0^2 s\,\hat{v}(s)\,ds\right]\int_0^2 s\,v(s)\,ds - \int_0^2 \hat{v}(t)\,v(t)\,dt = 0,$$

there is a unique stationary point $\hat{v}(t) = t\,x_0/2$, the expression of the Hessian

$$\begin{aligned} \frac{1}{2}d^2J^-_{x_0}(\hat{v};v;v) &= \frac{3}{14}\left[\int_0^2 s\,v(s)\,ds\right]^2 - \int_0^2 |v(t)|^2\,dt \\ &\le \frac{3}{14}\left[\int_0^2 s^2\,ds\right]\left[\int_0^2 |v(s)|^2\,ds\right] - \int_0^2 |v(t)|^2\,dt \\ &\le \left[\frac{3}{14}\,\frac{2^3}{3} - 1\right]\int_0^2 |v(t)|^2\,dt = -\frac{3}{7}\int_0^2 |v(t)|^2\,dt \le 0 \end{aligned}$$

is negative, and the open loop lower value of the game is $v^-(x_0) = J^-_{x_0}(\hat{v}) = (x_0)^2/2$.

However, the open loop upper value of the game is $v^+(x_0) = +\infty$ for all $x_0 \in \mathbf{R}$. Indeed pick the sequence of controls $\{v_n\}$, $n \ge 1$, $v_n(t) = 0$ in $[0,1]$, and $v_n(t) = n$ in

[2]Given a real function f defined on a Banach space B, the *first directional semiderivative* at x in the direction v (when it exists) is defined as $df(x;v) \stackrel{\text{def}}{=} \lim_{t \searrow 0} \frac{f(x+tv)-f(x)}{t}$. When the map $v \mapsto df(x;v) : B \to \mathbf{R}$ is linear and continuous, it defines the *gradient* $\nabla f(x)$ as an element of the dual B^* of B. The *second order bidirectional derivative* at x in the directions (v,w) (when it exists) is defined as $d^2f(x;v,w) \stackrel{\text{def}}{=} \lim_{t \searrow 0} \frac{df(x+tw;v)-df(x;v)}{t}$. When the map $(v,w) \mapsto d^2f(x;v,w) : B \times B \to \mathbf{R}$ is bilinear and continuous, it defines the *Hessian operator* $Hf(x)$ as a continuous linear operator from B to B^*.

$[1,2]$. The corresponding sequence of states at time $t=2$ is

$$x_n(2) = x_0 + \int_0^2 (2-t)\,u(t)\,dt + n\int_1^2 t\,dt = \left[x_0 + \int_0^2 (2-t)\,u(t)\,dt\right] + \frac{3}{2}n.$$

Denote by X the square bracket that does not depend on n. Then

$$\begin{aligned} C_{x_0}(u,v_n) &= \frac{1}{2}\left|X + \frac{3}{2}n\right|^2 + \int_0^2 |u(t)|^2\,dt - \int_1^2 n^2\,dt \\ &= \frac{1}{8}n^2 + \frac{3}{2}n\,X + \frac{X^2}{2} + \int_0^2 |u(t)|^2\,dt \to +\infty \text{ as } n \to +\infty. \end{aligned}$$

Thus for all $x_0 \in \mathbf{R}$ and $u \in L^2(0,T;\mathbf{R})$

$$\sup_{v\in L^2(0,T;\mathbf{R})} C_{x_0}(u,v) = +\infty \quad \Rightarrow v^+(x_0) = +\infty \quad \text{and} \quad U(x_0) = \varnothing.$$

Therefore, whatever the initial state x_0 is, $C_{x_0}(u,v)$ has no open loop saddle point.

We now consider the example of Zhang [10] of a game without open loop saddle point. We show that the solution of the Riccati differential equation (2.12) is unique, strictly positive, and *infinitely differentiable*.

Example 2.2. Consider the utility function and linear dynamics

$$C_{x_0}(u,v) = \int_0^1 2x^2 + u^2 - v^2\,dt, \quad x' = x + u + v, \quad x(0) = x_0 \tag{2.15}$$

given by Zhang [10]. Here $A = B_1 = B_2 = 1$, $F = 0$, and $Q = 2$. Now $R = B_1B_1^* - B_2B_2^* = 0$, and the associated Riccati differential equation (2.12) reduces to

$$P' + 2P + 2 = 0 \text{ in } [0,1], \quad P(1) = 0.$$

It has a unique infinitely differentiable solution $P(t) = e^{2(1-t)} - 1$ that is strictly positive in $[0,1)$.

We now extend the result of Zhang [10] on the nonexistence of an open loop saddle point from the initial state $x_0 = 0$ to any initial state. For all $x_0 \in \mathbf{R}$ the open loop lower value $v^-(x_0)$ of the game is finite, but the open loop upper value $v^+(x_0)$ is $+\infty$. Indeed for each $v \in L^2(0,T;\mathbf{R})$

$$\begin{aligned} &\inf_{u\in L^2(0,T;\mathbf{R})} C_{x_0}(u,v) \le C_{x_0}(-v,v) = \int_0^1 2\left(x_0\,e^t\right)^2\,dt = \left(e^2-1\right)(x_0)^2 \\ &\quad \Rightarrow v^-(x_0) = \sup_{v\in L^2(0,T;\mathbf{R})} \inf_{u\in L^2(0,T;\mathbf{R})} C_{x_0}(u,v) \le \left(e^2-1\right)(x_0)^2. \end{aligned}$$

By definition of the sup,

$$\begin{aligned} v^-(x_0) &= \sup_{v\in L^2(0,T;\mathbf{R})} \inf_{u\in L^2(0,T;\mathbf{R})} C_{x_0}(u,v) \\ &\ge \inf_{u\in L^2(0,T;\mathbf{R})} C_{x_0}(u,0) = \inf_{u\in L^2(0,T;\mathbf{R})} \int_0^1 2x^2 + u^2\,dt \ge 0 \\ &\quad \Rightarrow \forall x_0 \in \mathbf{R}, \quad 0 \le v^-(x_0) \le \left(e^2-1\right)(x_0)^2. \end{aligned}$$

For the open loop upper value, associate with each $u \in L^2(0,T;\mathbf{R})$ the sequence of functions $v_n(t) = -u(t) + n$, $n \geq 1$. The corresponding sequence of states is

$$
\begin{aligned}
x_n(t) &= e^t x_0 + n \int_0^t e^{t-s}\, ds = e^t x_0 + n(e^t - 1),\\
C_{x_0}(u, v_n) &= n^2 \int_0^1 2(e^t - 1)^2 - 1\, dt + 2n \int_0^1 u(t)\, dt\\
&\quad + \int_0^1 (e^t x_0)^2\, dt + 2n\, x_0 \int_0^1 e^t\, (e^t - 1)\, dt\\
&= n^2 \int_0^1 1 + 2e^{2t} - 4e^t\, dt + 2n \int_0^1 u(t)\, dt\\
&\quad + \int_0^1 (e^t x_0)^2\, dt + 2n\, x_0\, (e-1)^2\\
&\geq (e-2)^2 n^2 + 2n \left[x_0\, (e-1)^2 + \int_0^1 u(t)\, dt \right] + \int_0^1 (e^t x_0)^2\, dt\\
&\Rightarrow \sup_{v \in L^2(0,T;\mathbf{R})} C_{x_0}(u, v) \geq C_{x_0}(u, v_n) \to +\infty
\end{aligned}
$$

as n goes to infinity. Therefore for all $x_0 \in \mathbf{R}^n$ and all $u \in L^2(0,T;\mathbf{R})$,

$$
\sup_{v \in L^2(0,T;\mathbf{R})} C_{x_0}(u, v) = +\infty \quad \Rightarrow v^+(x_0) = +\infty \quad \text{and} \quad U(x_0) = \varnothing,
$$

and there is no open loop saddle point.

2.3. Properties of the utility function, convexity, concavity, and saddle points. We use a state-adjoint state equation approach to characterize the existence of the open loop upper and lower values as well as the open loop saddle point of the quadratic utility function.

The utility function $C_{x_0}(u, v)$ is infinitely differentiable and, since it is quadratic, its Hessian of second order derivatives is independent of the point (u, v). Indeed,

$$
\frac{1}{2} dC_{x_0}(u, v; \bar{u}, \bar{v}) = Fx(T) \cdot \bar{y}(T) + (Qx, \bar{y}) + (u, \bar{u}) - (v, \bar{v}), \tag{2.16}
$$

where x is the solution of (2.4) and $\bar{y}$ is the solution of

$$
\bar{y}' = A\bar{y} + B_1 \bar{u} + B_2 \bar{v}, \quad \bar{y}(0) = 0. \tag{2.17}
$$

It is customary to introduce the adjoint system

$$
p' + A^* p + Qx = 0, \quad p(T) = Fx(T) \tag{2.18}
$$

and rewrite expression (2.16) for the gradient in the following form:

$$
\frac{1}{2} dC_{x_0}(u, v; \bar{u}, \bar{v}) = (B_1^* p + u, \bar{u}) + (B_2^* p - v, \bar{v}). \tag{2.19}
$$

As predicted, the Hessian is independent of (u, v):

$$
\frac{1}{2} d^2 C_{x_0}(u, v; \bar{u}, \bar{v}; \tilde{u}, \tilde{v}) = F\tilde{y}(T) \cdot \bar{y}(T) + (Q\tilde{y}, \bar{y}) + (\tilde{u}, \bar{u}) - (\tilde{v}, \bar{v}), \tag{2.20}
$$

where $\bar{y}$ is the solution of (2.17) and $\tilde{y}$ is the solution of

$$\tilde{y}' = A\tilde{y} + B_1\tilde{u} + B_2\tilde{v}, \quad \tilde{y}(0) = 0. \tag{2.21}$$

In particular, for all x_0, u, v, $\bar{u}$, and $\bar{v}$

$$d^2C_{x_0}(u, v; \bar{u}, \bar{v}; \bar{u}, \bar{v}) = 2\, C_0(\bar{u}, \bar{v}), \tag{2.22}$$

and this yields the following characterizations of the u-convexity, v-concavity, and (u, v)-convexity-concavity under the assumptions of section 2.1 on the matrix F and the matrix functions A, B_1, B_2, and Q. Note that the matrices F and $Q(t)$ are symmetrical, but they are not necessarily positive semidefinite.

LEMMA 2.1. *Let F be an $n \times n$ matrix and A, B_1, B_2, and Q be bounded measurable matrix functions of appropriate dimensions, and assume that F and $Q(t)$ are symmetrical for almost all t. Then the following statements are equivalent.*

(i) *The map $u \mapsto C_0(u, 0) : L^2(0, T; \mathbf{R}^m) \to \mathbf{R}$ is convex.*
(ii) *For all $u \in L^2(0, T; \mathbf{R}^m)$, $C_0(u, 0) \geq 0$.*
(iii) $\inf_{u \in L^2(0,T;\mathbf{R}^m)} C_0(u, 0) = C_0(0, 0)$.
(iv) *For all v and x_0 the map $u \mapsto C_{x_0}(u, v) : L^2(0, T; \mathbf{R}^m) \to \mathbf{R}$ is convex.*

COROLLARY 2.1. *The following statements are equivalent.*

(i) *The map $v \mapsto C_0(0, v) : L^2(0, T; \mathbf{R}^k) \to \mathbf{R}$ is concave.*
(ii) *For all $v \in L^2(0, T; \mathbf{R}^k)$, $C_0(0, v) \leq 0$.*
(iii) $\sup_{v \in L^2(0,T;\mathbf{R}^k)} C_0(0, v) = C_0(0, 0)$.
(iv) *For all u and x_0, the map $v \mapsto C_{x_0}(u, v) : L^2(0, T; \mathbf{R}^k) \to \mathbf{R}$ is concave.*

COROLLARY 2.2. *The following statements are equivalent.*

(i) *The map, $(u, v) \mapsto C_0(u, v) : L^2(0, T; \mathbf{R}^m) \times L^2(0, T; \mathbf{R}^k) \to \mathbf{R}$ is (u, v)-convex-concave; that is,*

$$\begin{aligned} &\forall v \in L^2(0, T; \mathbf{R}^k), \quad u \mapsto C_0(u, v) \text{ is convex, and} \\ &\forall u \in L^2(0, T; \mathbf{R}^m), \quad v \mapsto C_0(u, v) \text{ is concave.} \end{aligned} \tag{2.23}$$

(ii) *The pair $(0, 0)$ is a saddle point of $C_0(u, v)$:*

$$\sup_{v \in L^2(0,T;\mathbf{R}^k)} C_0(0, v) = C_0(0, 0) = \inf_{u \in L^2(0,T;\mathbf{R}^m)} C_0(u, 0). \tag{2.24}$$

(iii) $\sup_{v \in L^2(0,T;\mathbf{R}^k)} C_0(0, v) = C_0(0, 0) = \inf_{u \in L^2(0,T;\mathbf{R}^m)} C_0(u, 0)$.
(iv) *For all x_0 the map $(u, v) \mapsto C_{x_0}(u, v) : L^2(0, T; \mathbf{R}^m) \times L^2(0, T; \mathbf{R}^k) \to \mathbf{R}$ is (u, v)-convex-concave; that is,*

$$\begin{aligned} &\forall v \in L^2(0, T; \mathbf{R}^k), \quad u \mapsto C_{x_0}(u, v) \text{ is convex, and} \\ &\forall u \in L^2(0, T; \mathbf{R}^m), \quad v \mapsto C_{x_0}(u, v) \text{ is concave.} \end{aligned} \tag{2.25}$$

2.4. Saddle point and coupled state-adjoint state system. We first obtain necessary and sufficient conditions for the existence of a saddle point of the game and introduce the *coupled (state-adjoint state) system* (cf. Notation 2.1 on page 758) that will also arise in the characterization of the open loop lower and upper values of the game in section 2.5. Theorem 2.4 in section 2.5 will later complete this theorem with the equivalent condition that the value $v(x_0)$ of the game is finite.

THEOREM 2.1. *The following conditions are equivalent.*

(i) *There exists an open loop* saddle point *of $C_{x_0}(u, v)$.*

(ii) *There exists a solution* $(\hat{u}, \hat{v})$ *in* $L^2(0,T;\mathbf{R}^m) \times L^2(0,T;\mathbf{R}^k)$ *of the system*

$$\forall u \in L^2(0,T;\mathbf{R}^m),\ \forall v \in L^2(0,T;\mathbf{R}^k), \quad dC_{x_0}(\hat{u},\hat{v};u,v) = 0, \tag{2.26}$$

and C_{x_0} *is convex-concave in the sense of* (2.25).

(iii) *There exists a solution* $(x,p) \in H^1(0,T;\mathbf{R}^n)^2$ *of the* coupled system

$$\begin{cases} x' = Ax - B_1B_1^*p + B_2B_2^*p, & x(0) = x_0, \\ p' + A^*p + Qx = 0, & p(T) = Fx(T), \end{cases} \tag{2.27}$$

$$\hat{u} = -B_1^*p, \quad \hat{v} = B_2^*p, \tag{2.28}$$

and

$$\sup_{v \in L^2(0,T;\mathbf{R}^k)} C_0(0,v) = C_0(0,0) = \inf_{u \in L^2(0,T;\mathbf{R}^m)} C_0(u,0). \tag{2.29}$$

Under any one of the above conditions, the value of the game is given by

$$v(x_0) = C_{x_0}(\hat{u},\hat{v}) = p(0) \cdot x_0. \tag{2.30}$$

Proof. (i) $\Rightarrow$ (ii). Let $(\bar{u},\bar{v})$ in $L^2(0,T;\mathbf{R}^m) \times L^2(0,T;\mathbf{R}^k)$ be an open loop saddle point of $C_{x_0}(u,v)$ in $L^2(0,T;\mathbf{R}^m) \times L^2(0,T;\mathbf{R}^k)$. Then by Definition 2.1

$$\sup_{L^2(0,T;\mathbf{R}^k)} C_{x_0}(\bar{u},v) = C_{x_0}(\bar{u},\bar{v}) = \inf_{L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,\bar{v}). \tag{2.31}$$

Since $C_{x_0}(u,v)$ is infinitely differentiable, the minimizing point $\bar{u}$ of $C_{x_0}(u,\bar{v})$ with respect to u is characterized by the first order condition $dC_{x_0}(\bar{u},\bar{v};u,0) = 0$ for all u and the second order condition $d^2C_{x_0}(\bar{u},\bar{v};u,0;u,0) \geq 0$ for all u. Since $d^2C_{x_0}(\bar{u},\bar{v};u,0;u,0)$ is independent of $(\bar{u},\bar{v})$ and x_0, $C_{x_0}(u,v)$ is convex in u for all x_0 and all v. A similar argument for the maximum yields $dC_{x_0}(\bar{u},\bar{v};,w) = 0$ and $d^2C_{x_0}(\bar{u},\bar{v};0,w;0,w) \leq 0$ for all w and the concavity of $C_{x_0}(u,v)$ in v.

(ii) $\Rightarrow$ (i). By assumption $C_{x_0}(\hat{u},\hat{v})$ is convex-concave and infinitely differentiable and there is a solution to the two first order conditions. By [6, Proposition 1.6], there exists a saddle point.

(ii) $\Leftrightarrow$ (iii). This follows from the previous computations of the gradient and Corollary 2.2.

Finally, we compute the value

$$\begin{aligned} C_{x_0}(\hat{u},\hat{v}) &= Fx(T) \cdot x(T) + (Qx,x) + \|B_1^*p\|^2 - \|B_2^*p\|^2 \\ &= p(T) \cdot x(T) - (p' + A^*p, x) + ([B_1B_1^* - B_2B_2^*]p, p) \\ &= p(0) \cdot x(0) + (p, x' - Ax + ([B_1B_1^* - B_2B_2^*]p) = p(0) \cdot x_0. \end{aligned}$$ ∎

NOTATION 2.1. *It will be useful to introduce the set* $\mathcal{N}_{x,p}$ *of all solutions* (y,q) *of the homogeneous coupled system*

$$\begin{cases} y' = Ay - B_1B_1^*q + B_2B_2^*q, & y(0) = 0, \\ q' + A^*q + Qy = 0, & q(T) = Fy(T). \end{cases} \tag{2.32}$$

Thus the coupled system has a solution up to an additive pair of $\mathcal{N}_{x,p}$.

2.5. Necessary and sufficient conditions for games with finite values. The quadratic character of the problem yields surprising equivalences that reduce the complexity of its solution. We start with the open loop lower value of the game.

THEOREM 2.2. *The following conditions are equivalent.*

(i) *There exist* $\hat{u}$ *in* $L^2(0,T;\mathbf{R}^m)$ *and* $\hat{v}$ *in* $L^2(0,T;\mathbf{R}^k)$ *such that*

$$C_{x_0}(\hat{u},\hat{v}) = \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,\hat{v}) = \sup_{v\in L^2(0,T;\mathbf{R}^k)} \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,v). \tag{2.33}$$

(ii) *The open loop lower value* $v^-(x_0)$ *of the game is finite.*

(iii) *There exists a solution* $(x,p)\in H^1(0,T;\mathbf{R}^n)\times H^1(0,T;\mathbf{R}^n)$ *of the* coupled system (2.27) *such that* $B_2^*p\in V(x_0)$*, the solution pairs* $(\hat{u},\hat{v})$ *and the open loop lower value are given by the expressions*

$$\hat{u} = -B_1^*p, \quad \hat{v} = B_2^*p, \quad \textit{and} \quad v^-(x_0) = C_{x_0}(\hat{u},\hat{v}) = p(0)\cdot x_0, \tag{2.34}$$

and

$$\sup_{v\in V(0)} \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_0(u,v) = \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_0(u,0) = C_0(0,0). \tag{2.35}$$

Proof. The proof of this main theorem will be given in sections 3 and 3.3. □

Remark 2.2. The above necessary and sufficient conditions for the finiteness of the open loop value of the game complete the results of Zhang [10] by introducing the new *feasibility condition* (2.35) that is equivalent to saying that the open loop lower value of the game is zero and that $(0,0)$ is a solution for the zero initial state. It also recasts the results in the more intuitive state-adjoint state framework. Condition (2.35) is equivalent to the convexity of $C_{x_0}(u,v)$ with respect to u and the concavity of $J^-_{x_0}(v) = \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,v)$ with respect to $v\in V(x_0)$.

Theorem 2.2 has a counterpart for the upper value $v^+(x_0)$ of the game.

THEOREM 2.3. *The following conditions are equivalent.*

(i) *There exist* $\hat{u}$ *in* $L^2(0,T;\mathbf{R}^m)$ *and* $\hat{v}$ *in* $L^2(0,T;\mathbf{R}^k)$ *such that*

$$C_{x_0}(\hat{u},\hat{v}) = \sup_{v\in L^2(0,T;\mathbf{R}^k)} C_{x_0}(\hat{u},v) = \inf_{u\in L^2(0,T;\mathbf{R}^m)} \sup_{v\in L^2(0,T;\mathbf{R}^k)} C_{x_0}(u,v). \tag{2.36}$$

(ii) *The open loop upper value* $v^+(x_0)$ *of the game is finite.*

(iii) *There exists a solution* $(x,p)\in H^1(0,T;\mathbf{R}^n)\times H^1(0,T;\mathbf{R}^n)$ *of the* coupled system (2.27) *such that* $-B_1^*p\in U(x_0)$*, the solution pairs* $(\hat{u},\hat{v})$ *and the open loop upper value are given by the expressions*

$$\hat{u} = -B_1^*p, \quad \hat{v} = B_2^*p, \quad \textit{and} \quad v^+(x_0) = C_{x_0}(\hat{u},\hat{v}) = p(0)\cdot x_0, \tag{2.37}$$

and

$$\inf_{u\in U(0)} \sup_{v\in L^2(0,T;\mathbf{R}^k)} C_0(u,v) = \sup_{v\in L^2(0,T;\mathbf{R}^k)} C_0(0,v) = C_0(0,0). \tag{2.38}$$

Condition (2.38) says that $C_{x_0}(u,v)$ is concave with respect to v and that $J^+_{x_0}(u) = \sup_{v\in L^2(0,T;\mathbf{R}^k)} C_{x_0}(u,v)$ is convex with respect to $u\in U(x_0)$.

Finally, the necessary and sufficient condition for the finiteness of the value $v(x_0)$ of the game can now be obtained from the above two theorems and Theorem 2.1(iii).

THEOREM 2.4. *The following conditions are equivalent.*

(i) *There exists an open loop* saddle point *of* $C_{x_0}(u,v)$.
(ii) *The open loop value* $v(x_0)$ *of the game is finite.*
(iii) *There exists a solution* $(x,p) \in H^1(0,T;\mathbf{R}^n) \times H^1(0,T;\mathbf{R}^n)$ *of the* coupled system (2.27)*, the solution pair* $(\hat{u},\hat{v})$ *is given by the expressions* (2.28)*, and the convexity-concavity* (2.29) *is verified.*

Under any one of the above conditions, the open loop value is given by expression (2.30).

Proof. (i) ⇒ (ii). Since the utility function has a saddle point, the value of the game is finite. (ii) ⇒ (iii). From Theorems 2.2 and 2.3 there exists a solution to the coupled system (2.27), and the convexity-concavity condition (2.29) readily follows from (2.38) and (2.35). (iii) ⇒ (i). This follows from Theorem 2.1. □

Remark 2.3. The common necessary condition for the finiteness of the lower value $v^-(x_0)$, value $v(x_0)$, and upper value $v^+(x_0)$ of the game is the existence of a solution of the coupled system (2.27). The difference is in the respective *feasibility conditions* (2.35), (2.29), and (2.38): $v^-(0) = 0$, $v(0) = 0$, and $v^+(0) = 0$.

We conclude with the enlightening result proved by Zhang [10, Thm. 4.1] that has shed new light on the characterization of a game with finite value. One of the consequences is that only three cases can occur: (i) $v^+(x_0)$ finite and $v^-(x_0) = -\infty$, (ii) $v^+(x_0) = +\infty$ and $v^-(x_0)$ finite, or (iii) $v(x_0)$ finite.

THEOREM 2.5. *Given* $x_0 \in \mathbf{R}^n$*, the following statements are equivalent.*
(i) *There exists an open loop saddle point of* $C_{x_0}(u,v)$.
(ii) *The open loop value of the game of* $C_{x_0}(u,v)$ *is finite.*
(iii) *Both the open loop lower and upper values of* $C_{x_0}(u,v)$ *are finite.*

Proof. (i) ⇒ (ii) ⇒ (iii) are obvious. It remains to prove that (iii) ⇒ (i). From condition (2.35) of Theorem 2.2 and condition (2.38) of Theorem 2.3 we get condition (2.29) of Theorem 2.4. Finally, both Theorems 2.2 and 2.3 give the existence of a pair $(x,p) \in H^1(0,T;\mathbf{R}^n) \times H^1(0,T;\mathbf{R}^n)$ solution of the coupled system (2.27). Therefore by Theorem 2.4 the utility function has a saddle point. □

2.6. Games with finite values for each initial state. In this section we sharpen the results of the previous section when the open loop lower value, value, or upper value of the game is finite for *all initial states* $x_0 \in \mathbf{R}^n$. In each case this global assumption yields the *uniqueness of solution.*

THEOREM 2.6. *The following conditions are equivalent.*
(i) *For each* $x_0 \in \mathbf{R}^n$*, there exist* $\hat{u}$ *in* $L^2(0,T;\mathbf{R}^m)$ *and* $\hat{v}$ *in* $L^2(0,T;\mathbf{R}^k)$ *such that*

$$C_{x_0}(\hat{u},\hat{v}) = \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,\hat{v}) = \sup_{v\in L^2(0,T;\mathbf{R}^k)} \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,v). \tag{2.39}$$

(ii) *For each* $x_0 \in \mathbf{R}^n$*, the open loop lower value* $v^-(x_0)$ *of the game is finite.*
(iii) *For each* $x_0 \in \mathbf{R}^n$*, there exists a unique pair* $(x,p) \in H^1(0,T;\mathbf{R}^n)^2$ *solution of the* coupled system (2.27) *such that* $B_2^* p \in V(x_0)$*, there exists a unique pair* $(\hat{u},\hat{v})$ *that verifies* (2.28)*, and*

$$\sup_{v\in V(0)} \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_0(u,v) = \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_0(u,0) = C_0(0,0). \tag{2.40}$$

Remark 2.4. The uniqueness under condition (i) was originally given by Zhang et al. in [11] by a different argument. Our short and transparent proof seems to be new. The same proof can readily be used in the context of optimal control [9].

Proof. (i) ⇒ (ii) is obvious. (ii) ⇒ (iii). This follows from Theorem 2.2 where condition (2.40) is condition (2.35). We need only show the uniqueness of the solution to the coupled system (2.27). By linearity, this amounts to proving that the solution (y, q) of the homogeneous system (2.32) such that $B_2^* q \in V(0)$ is $(0, 0)$. Given an arbitrary x_0, consider the expression

$$\begin{aligned} q(0) \cdot x_0 &= q(T) \cdot x(T) - \int_0^T q' \cdot x + q \cdot x' \, dt \\ &= Fx(T) \cdot y(T) + \int_0^T Qx \cdot y + B_1^* p \cdot B_1^* q - B_2^* p \cdot B_2^* q \, dt \\ &= \frac{1}{2} dC_{x_0}(\hat{u}, \hat{v}; -B_1^* q, B_2^* q) = 0 \end{aligned}$$

from (2.16), (2.27), (2.34), and the fact that $B_2^* q \in V(0)$. Since this identity is true for all $x_0 \in \mathbf{R}^n$, $q(0) = 0$. But now we can look at the coupled system (2.32) as a linear differential system in (x, p) with zero initial condition $(y(0), q(0)) = (0, 0)$ whose unique solution is $(y, q) = (0, 0)$. This proves uniqueness. (iii) ⇒ (i). This follows, again from Theorem 2.2, since the conditions are verified for each $x_0 \in \mathbf{R}^n$. □

We readily have the dual result.

THEOREM 2.7. *The following conditions are equivalent.*

(i) *For each $x_0 \in \mathbf{R}^n$, there exist $\hat{u}$ in $L^2(0, T; \mathbf{R}^m)$ and $\hat{v}$ in $L^2(0, T; \mathbf{R}^k)$ such that*

(2.41)
$$C_{x_0}(\hat{u}, \hat{v}) = \sup_{v \in L^2(0,T;\mathbf{R}^k)} C_{x_0}(\hat{u}, v) = \inf_{u \in L^2(0,T;\mathbf{R}^m)} \sup_{v \in L^2(0,T;\mathbf{R}^k)} C_{x_0}(u, v).$$

(ii) *For each $x_0 \in \mathbf{R}^n$, the open loop upper value $v^+(x_0)$ of the game is finite.*

(iii) *For each $x_0 \in \mathbf{R}^n$, there exists a unique pair $(x, p) \in H^1(0, T; \mathbf{R}^n)^2$ solution of the* coupled system (2.27) *such that $-B_1^* p \in U(x_0)$, there exists a unique pair $(\hat{u}, \hat{v})$ that verifies* (2.28)*, and*

$$\inf_{u \in U(0)} \sup_{v \in L^2(0,T;\mathbf{R}^k)} C_0(u, v) = \sup_{v \in L^2(0,T;\mathbf{R}^k)} C_0(0, v) = C_0(0, 0). \tag{2.42}$$

Finally, by combining the last two theorems, we get the saddle point case.

THEOREM 2.8. *The following conditions are equivalent.*

(i) *For each $x_0 \in \mathbf{R}^n$, there exists an open loop* saddle point *of $C_{x_0}(u, v)$.*

(ii) *For each $x_0 \in \mathbf{R}^n$, the open loop value $v(x_0)$ of the game is finite.*

(iii) *For each $x_0 \in \mathbf{R}^n$, there exists a unique pair $(x, p) \in H^1(0, T; \mathbf{R}^n)^2$ solution of the* coupled system (2.27)*, there exists a unique pair $(\hat{u}, \hat{v})$ that verifies* (2.28)*, and the convexity-concavity condition* (2.29) *is verified.*

2.7. Open loop saddle point and Riccati differential equation. Under the assumption of the finiteness of the open loop value of the game in $[0, T]$ for each initial state, we can *unexpectedly* use *invariant embedding* and introduce a *decoupling symmetrical matrix* solution of the *matrix Riccati differential equation* (2.12).

THEOREM 2.9. *Assume that the open loop value $v(x_0)$ is finite for all $x_0 \in \mathbf{R}^n$.*

(i) *There exists a unique symmetrical solution with elements in $H^1(0, T)$ of the matrix Riccati differential equation*

$$P' + PA + A^* P - PRP + Q = 0, \quad P(T) = F, \tag{2.43}$$

where $R = B_1B_1^* - B_2B_2^*$. *Moreover,*

$$\hat{p}(t) = P(t)\,\hat{x}(t),\ 0 \le t \le T, \text{ and } C_{x_0}(\hat{u},\hat{v}) = P(0)x_0 \cdot x_0, \tag{2.44}$$

where $(\hat{x},\hat{p}) \in H^1(0,T;\mathbf{R}^n)^2$ *is the unique solution of the coupled system* (2.27).

(ii) *The optimal strategies of the two players are closed loop*

$$\hat{u} = -B_1^*P\hat{x} \quad \text{and} \quad \hat{v} = B_2^*P\hat{x}, \tag{2.45}$$

and they achieve a closed loop-closed loop saddle point in the sense of [4].

(iii) *For all* $x_0 \in \mathbf{R}^n$ *the function* $C_{x_0}(u,v)$ *is convex-concave.*

Proof. For the proof, see section 5.3. □

The existence of a symmetrical solution to the matrix Riccati differential equation (2.43) implies that, for all $x_0 \in \mathbf{R}^n$, there exists a solution $(\hat{x},\hat{p}) \in H^1(0,T;\mathbf{R}^n) \times H^1(0,T;\mathbf{R}^n)$ of the coupled system (2.27). However, as we have seen in Example 2.2, this is not sufficient to get an open loop saddle point of the utility function $C_{x_0}(u,v)$.

THEOREM 2.10. *A set of necessary and sufficient conditions for the existence of an open loop saddle point of the utility function* $C_{x_0}(u,v)$ *for all* $x_0 \in \mathbf{R}^n$ *is*

(a) *the utility function* $C_{x_0}(u,v)$ *is convex in* u *and concave in* v *for some* x_0, *and*

(b) *there exists a (unique) symmetrical solution in* $H^1(0,T)$ *to the matrix Riccati differential equation* (2.43).

Proof. For the proof, see section 5.4. □

Remark 2.5. The method of *completion of the squares* (cf., for instance, Başar and Bernhard [1, Chap. 9, Thm. 9.4]) can also be used here to obtain

$$\sup_{v\in L^2(0,T;\mathbf{R}^k)} \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,v) \le P(0)x_0 \cdot x_0 \le \inf_{u\in L^2(0,T;\mathbf{R}^m)} \sup_{v\in L^2(0,T;\mathbf{R}^k)} C_{x_0}(u,v).$$

So it would be tempting to conclude that there is a saddle point *without* condition (a). But, as illustrated in Example 2.2 where we show that $U(x_0) = \varnothing$ for all x_0, condition (a) is really necessary. In order to get a saddle point, both $v^-(x_0)$ and $v^+(x_0)$ must be finite. Therefore the open loop lower value of the game will be finite if (b) is verified and $V(x_0) \neq \varnothing$; the open loop upper value of the game will be finite if (b) is verified and $U(x_0) \neq \varnothing$.

3. Open loop lower value of the game. We review the three steps: existence and characterization of a minimizer for $v \in V(x_0)$, formulation of the resulting maximization problem with respect to v, and, finally, existence and characterization of the pair that achieves the finite open loop lower value of the game.

3.1. Existence and characterization of the minimizers.

THEOREM 3.1. *Given* $x_0 \in \mathbf{R}^n$ *and* $v \in L^2(0,T;\mathbf{R}^k)$, *the following statements are equivalent.*

(i) *There exists* $\hat{u} \in L^2(0,T;\mathbf{R}^m)$ *such that*

$$C_{x_0}(\hat{u},v) = J_{x_0}^-(v) = \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,v). \tag{3.1}$$

(ii) $\inf_{u\in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,v) > -\infty$ *(that is,* $v \in V(x_0)$*).*

(iii) *There exists a pair* $(x,p) \in H^1(0,T;\mathbf{R}^n)\times H^1(0,T;\mathbf{R}^n)$ *solution of the system*

$$\begin{cases} x' = Ax - B_1B_1^*p + B_2v, & x(0) = x_0, \\ p' + A^*p + Qx = 0, & p(T) = Fx(T), \end{cases} \tag{3.2}$$

$$\hat{u}(t) = -B_1^*(t)p(t), \quad J_{x_0}^-(v) = p(0)\cdot x_0 + \int_0^T B_2^*p\cdot v - |v|^2\,dt, \tag{3.3}$$

and

$$\inf_{u\in L^2(0,T;\mathbf{R}^m)} C_0(u,0) \geq 0. \tag{3.4}$$

(iv) *The convexity inequality* (3.4) *is verified and*

$$\forall q \in N_p, \quad x_0\cdot q(0) + \int_0^T v\cdot B_2^*q\,dt = 0, \tag{3.5}$$

where

$$N_p \overset{\text{def}}{=} \left\{q \in H^1(0,T;\mathbf{R}^n) : \forall (y,q) \in N_{x,p}\right\} \tag{3.6}$$

and $N_{x,p}$ *denotes the set of all solutions* (y,q) *of the homogeneous system*

$$\begin{cases} y' = Ay - B_1\,B_1^*q, & y(0) = 0, \\ q' + A^*q + Qy = 0, & q(T) = Fy(T). \end{cases} \tag{3.7}$$

Proof. The proof follows from the following lemma, the computation of first and second order derivatives (2.19) and (2.22) in section 2.3, and the equivalent condition of Lemma 2.1(ii) for the u-convexity of $C_0(u,v)$.

LEMMA 3.1. *Let* U *be a Hilbert space,* $M : U \to U$ *a continuous linear self-adjoint compact operator,* $f \in U$, c *a constant, and* $j(u) = c + 2(f,u) + ([I+M]u,u)$.

(i) *Then the following conditions are equivalent.*

(a)

$$\exists \hat{u} \in U, \quad j(\hat{u}) = \inf_{u\in U} j(u), \tag{3.8}$$

(b)

$$\inf_{u\in U} j(u) > -\infty, \tag{3.9}$$

(c)

$$\exists \hat{u} \in U \text{ such that } [I+M]\hat{u} + f = 0, \text{ and} \tag{3.10}$$

$$\forall u \in U, \quad ([I+M]u,u) \geq 0. \tag{3.11}$$

(ii) *Condition* (3.10) *is equivalent to*

$$\forall w \in \ker[I+M], \quad (f,w) = 0. \tag{3.12}$$

(iii) *Condition* (3.11) *is equivalent to the convexity of* j.

We omit the proof of the lemma. □

NOTATION 3.1. *Given* $x_0 \in \mathbf{R}^n$ *such that* $V(x_0) \neq \varnothing$ *and* $v \in V(x_0)$, *denote by* $\mathcal{P}(v,x_0)$ *the set of all solutions* (x,p) *of system* (3.2). *It is readily checked that for all* $p \in \mathcal{P}(v,x_0)$, $\mathcal{P}(v,x_0) = p + N_p$.

3.2. Some intermediary results.

THEOREM 3.2.

(i) *The sets $N_{x,p}$, N_p, and $B_2^* N_p$ are finite dimensional linear subspaces of $H^1(0,T;\mathbf{R}^n)^2$, $H^1(0,T;\mathbf{R}^n)$, and $L^2(0,T;\mathbf{R}^k)$, respectively. $\mathcal{P}(v,x_0)$ is a finite dimensional affine subspace of $H^1(0,T;\mathbf{R}^n)$.*

(ii) *If $V(x_0) \neq \varnothing$ for some $x_0 \in \mathbf{R}^n$, then $V(x_0)$ is a closed affine subspace of $L^2(0,T;\mathbf{R}^k)$, $V(0)$ is a nonempty closed linear subspace of $L^2(0,T;\mathbf{R}^k)$,*

$$V(0) = (B_2^* N_p)^\perp, \tag{3.13}$$

$$\forall v \in V(x_0), \quad V(x_0) = v + V(0). \tag{3.14}$$

(iii) *Given $v \in V(x_0)$ and $p \in \mathcal{P}(v,x_0)$, define*

$$v^* \stackrel{\text{def}}{=} v + P_{V(0)}(B_2^* p - v), \tag{3.15}$$

where $P_{V(0)}$ is the orthogonal projection onto $V(0)$ in $L^2(0,T;\mathbf{R}^k)$. Then v^* is independent of the choice of p, v^* is unique in $V(x_0) \cap B_2^* \mathcal{P}(v,x_0)$, and there exists $p^* \in \mathcal{P}(v,x_0)$ such that $v^* = B_2^* p^*$. If, in addition, $B_2^* p - v \in V(0)^\perp$, then $v = v^* = B_2^* p^*$.

Analogues of Theorems 3.1 and 3.2 hold for the open loop upper value.

Remark 3.1. This theorem due to Zhang [10] is a key result in the proof of the existence of a maximizer of the inf problem. We have added part (i) to show that the subspace $B_2^* N_p$ is finite dimensional and hence closed. This is critical in the proof of part (ii). The proof essentially uses the arguments of [10].

Proof of Theorem 3.2. (i) From system (3.7), $N_{x,p}$ is a closed linear subspace of $H^1(0,T;\mathbf{R}^n)^2$ as the kernel of the continuous linear map

$$\begin{aligned}&(x,p) \mapsto \mathcal{A}(x,p) \stackrel{\text{def}}{=} (-x' + Ax - B_1 B_1^* p, -x(0), p' + A^* p + Qx, Fx(T) - p(T))\\ &: H^1(0,T;\mathbf{R}^n)^2 \to \left(L^2(0,T;\mathbf{R}^n) \times \mathbf{R}^n\right)^2.\end{aligned}$$

We now use the fact that a topological vector space is finite dimensional if and only if every closed bounded set is compact. Indeed, let K be a closed bounded subset of points (y,q) in $N_{x,p}$ for the $L^2(0,T;\mathbf{R}^n)^2$-topology. Since all the matrices in system (3.7) are bounded, the right-hand sides are bounded and the derivatives (y',q') are also bounded in $L^2(0,T;\mathbf{R}^n)^2$ and, a fortiori, in $H^1(0,T;\mathbf{R}^n)^2$. Since the injection of $H^1(0,T;\mathbf{R}^n)^2$ into $L^2(0,T;\mathbf{R}^n)^2$ is compact, then the closure of K in $L^2(0,T;\mathbf{R}^n)^2$ is compact. But, by assumption, we already know that K is closed. Thence K is compact in $L^2(0,T;\mathbf{R}^n)^2$ and $N_{x,p}$ is finite dimensional.

(ii) Since $V(x_0) \neq \varnothing$, then, by definition, for all v_1, v_2 in $V(x_0)$, condition (ii) of Theorem 3.1 is verified and condition (iii) is also verified for some pairs (x_1,p_1) and (x_2,p_2) verifying the system (3.2). Therefore, for any $\alpha \in \mathbf{R}$, the pair $(x_\alpha,p_\alpha) = (\alpha x_1 + (1-\alpha)x_2, \alpha p_1 + (1-\alpha)p_2)$ is also a solution of system (3.2) for x_0 and $v_\alpha = \alpha v_1 + (1-\alpha)v_2 \in V(x_0)$. Identity (3.14) follows from the fact that $V(x_0)$ is an affine subspace. Moreover, from (3.14), $V(x_0) \neq \varnothing$ necessarily implies that $V(0) \neq \varnothing$. Finally, from condition (3.5) with $x_0 = 0$

$$v \in V(0) \quad \Leftrightarrow \quad \forall q \in N_p, \quad \int_0^T v \cdot B_2^* q \, dt = 0 \quad \Leftrightarrow \; v \in (B_2^* N_p)^\perp,$$

and $V(0) = (B_2^* N_p)^\perp$, a nonempty closed linear subspace.

(iii) Given p_1, p_2 in $\mathcal{P}(v, x_0)$, $p_2 - p_1 \in N_p$ and

$$v + P_{V(0)}(B_2^* p_2 - v) - (v + P_{V(0)}(B_2^* p_1 - v)) = P_{V(0)}(B_2^*(p_2 - p_1)) = 0,$$

since $B_2^* N_p = V(0)^\perp$. So v^* is independent of the choice of $p \in \mathcal{P}(v, x_0)$. Since $V(x_0)$ is affine, then for all $v \in V(x_0)$,

$$\begin{aligned} v^* &= v + P_{V(0)}(B_2^* p - v) \in v + V(0) = V(x_0), \\ v^* - B_2^* p &= v - B_2^* p - P_{V(0)}(v - B_2^* p) \in V(0)^\perp = B_2^* N_p \\ \Rightarrow\ \exists q \in N_p &\text{ such that } v^* - B_2^* p = B_2^* q \quad \Rightarrow\ v^* = B_2^*(p+q) \in B_2^* \mathcal{P}(v, x_0), \end{aligned} \tag{3.16}$$

and $v^* \in V(x_0) \cap B_2^* \mathcal{P}(v, x_0)$. This element is unique since for v_1^* and v_2^* in $V(x_0) \cap B_2^* \mathcal{P}(v, x_0)$, $v_2^* - v_1^* \in V(0) \cap B_2^* N_p = V(0) \cap V(0)^\perp = \{0\}$. Finally, if $B_2^* p - v \in V(0)^\perp$, then from (3.16) we get $v = v^*$. □

3.3. Existence and characterization of maximizers. Assume that $v^-(x_0)$ is finite. By definition of $V(x_0)$, it is not empty and

$$v^-(x_0) = \sup_{v \in L^2(0,T;\mathbf{R}^k)} \inf_{u \in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u, v) = \sup_{v \in V(x_0)} J_{x_0}^-(v), \tag{3.17}$$

where $V(x_0)$ is a closed affine subspace of $L^2(0, T; \mathbf{R}^k)$ and by (3.3) and condition (3.5)

$$J_{x_0}^-(v) = \inf_{u \in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u, v) = p(0) \cdot x_0 + \int_0^T B_2^* p \cdot v - |v|^2 \, dt, \tag{3.18}$$

or, equivalently,

$$J_{x_0}^-(v) = Fx(T) \cdot x(T) + \int_0^T Q(t)x(t) \cdot x(t) + |B_1^*(t)p(t)|^2 - |v(t)|^2 \, dt \tag{3.19}$$

for all solutions (x, p) of system (3.2). Define the equivalence class $[(x, p)] = (x, p) + N_{x,p}$. Then for each pair $v \in V(x_0)$, $[(x, p)]$ is the unique solution in $H^1(0, T; \mathbf{R}^n) \times H^1(0, T; \mathbf{R}^n)/N_{x,p}$ of system (3.2). Thus the map

$$v \mapsto [(x, p)] : V(x_0) \to \frac{H^1(0, T; \mathbf{R}^n) \times H^1(0, T; \mathbf{R}^n)}{N_{x,p}} \tag{3.20}$$

is affine and continuous, and the map

$$\begin{aligned} &(x, p) \mapsto (x(T), x, p) \\ &: H^1(0, T; \mathbf{R}^n) \times H^1(0, T; \mathbf{R}^n) \to \mathbf{R}^n \times L^2(0, T; \mathbf{R}^n) \times L^2(0, T; \mathbf{R}^n) \end{aligned} \tag{3.21}$$

is continuous and compact.

So we are back to a continuous linear quadratic function $J_{x_0}^-(v)$ that is to be maximized over the closed affine subspace $V(x_0)$. The state is now the pair (x, p) solution of (3.2), but the structure is the same. Lemma 3.1 readily extends to the case of a sup over a closed affine subspace and the following conditions are equivalent:

(a)

$$\exists \hat{v} \in V(x_0), \quad J_{x_0}^-(\hat{v}) = \sup_{v \in V(x_0)} J_{x_0}^-(v), \tag{3.22}$$

(b)

$$\sup_{v\in V(x_0)} J^-_{x_0}(v) < +\infty, \tag{3.23}$$

(c)

$$\exists \hat v \in V(x_0) \text{ such that } [I+M]\hat v + f \in V(0)^\perp, \text{ and} \tag{3.24}$$

$$\forall w \in V(0), \quad ([I+M]w, w) \le 0 \tag{3.25}$$

for the new compact operator M corresponding to the new state (x,p).

It remains to compute the directional derivative of $J^-_{x_0}(v)$ at $v \in V(x_0)$ in the direction $w \in V(0)$. By direct computation from formula (2.19)

$$\frac{1}{2} dC_{x_0}(-B_1^* p, v; 0, w) = \int_0^T (B_2^* p - v)\cdot w\, dt, \tag{3.26}$$

which is independent of $p \in \mathcal{P}(v, x_0)$ for all $w \in V(0)$ by Theorem 3.1(iv). Hence

$$dJ^-_{x_0}(v; w) = dC_{x_0}(-B_1^* p, v; 0, w) = 2\int_0^T (B_2^* p - v)\cdot w\, dt, \quad \forall p \in \mathcal{P}(v, x_0). \tag{3.27}$$

As for the second order derivative,

$$\begin{aligned} &\frac{1}{2} d^2 C_{x_0}(-B_1^* p, v; 0, w; 0, w') \\ &= Fy_{w'}(T)\cdot y_w(T) + \int_0^T Qy_{w'}\cdot y_w + B_1^* q_{w'}\cdot B_1^* q_w - w'\cdot w\, dt \end{aligned} \tag{3.28}$$

$$\begin{aligned} \Rightarrow \frac{1}{2} d^2 J^-_{x_0}(v; w; w) &= \frac{1}{2} d^2 C_{x_0}(-B_1^* p, v; 0, w; 0, w) \\ &= \frac{1}{2} J_0^-(w) = \frac{1}{2} \inf_{u\in L^2(0.T;\mathbf{R}^m)} C_0(u, w), \end{aligned} \tag{3.29}$$

where the last term must be negative or zero for all $w \in V(0)$. But, from Theorem 3.1(iii), $C_0(u, 0)$ is convex in u. By using the equivalent condition of Lemma 2.1(ii) for the u-convexity of $C_0(u, 0)$, we finally get the two-part condition

$$\sup_{v\in V(0)} \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_0(u, v) \le 0 \le \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_0(u, 0).$$

This condition is equivalent to condition (2.35) since $C_0(0, 0) = 0$.

Proof of Theorem 2.2. (i) $\Rightarrow$ (ii) is obvious. (ii) $\Rightarrow$ (iii). From the previous discussion, the finiteness of $v^-(x_0)$ is equivalent to

$$\exists \hat v \in V(x_0) \text{ such that } dJ^-_{x_0}(\hat v; w) = 2\int_0^T (B_2^* \hat p - \hat v)\cdot w\, dt = 0, \quad \forall w \in V(0),$$

$$d^2 J^-_{x_0}(\hat v; w; w) = 2 \inf_{u\in L^2(0.T;\mathbf{R}^m)} C_0(u, w) \le 0, \quad \forall w \in V(0).$$

The second order condition combined with the fact that $V(x_0) \neq \varnothing$ (Theorem 3.1(iii)) yields condition (2.35). The first order condition says that $B_2^*\hat p - \hat v \in V(0)^\perp$. By Theorem 3.2(iii) there exists $\hat p^* \in \mathcal{P}(\hat v, x_0)$ such that $\hat v = B_2^* \hat p^*$, where $(\hat x^*, \hat p^*)$ is a solution of (3.2). Since $\hat v = B_2^* \hat p^*$, system (2.27) has a solution unique up to an

element of $\mathcal{N}_{x,p}$. After substitution of $\hat{v} = B_2^* \hat{p}^*$ in (3.2), $(\hat{x}^*, \hat{p}^*)$ becomes a solution of the coupled system (2.27). This also yields the identities (2.34). (iii) $\Rightarrow$ (i). By assumption $\hat{v} = B_2^* p \in V(x_0)$. The existence of a solution (x, p) to system (2.27) yields the existence of a solution to system (3.2) of Theorem 3.1(iii) with $\hat{u} = -B_1^* p$ as a minimizer. For all $v \in V(x_0)$,

$$J_{x_0}^-(v) = J_{x_0}^-(B_2^* p) + dJ_{x_0}^-(B_2^* p; v - B_2^* p) + \frac{1}{2} d^2 J_{x_0}^-(B_2^* p; v - B_2^* p; v - B_2^* p).$$

The second order term is negative by condition (2.35) since, by assumption, $B_2^* p \in V(x_0)$ and hence $v - B_2^* p \in V(0)$ for all $v \in V(x_0)$. As for the first order term, recall that, in view of (2.34), for all $w \in V(0)$

$$dJ_{x_0}^-(B_2^* p; w) = \int_0^T (B_2^* p - v) \cdot w \, dt = 0.$$

Thus $dJ_{x_0}^-(B_2^* p; v - B_2^* p) = 0$ since $v - B_2^* p \in V(0)$: $B_2^* p$ is a maximizer of $J_{x_0}^-$. ∎

4. Invariant embedding and convexity-concavity. Consider the LQ game on the time interval $[s, T]$, $0 \le s < T$, with initial state $h \in \mathbf{R}^n$ at time s:

$$C_h^s(u, v) \stackrel{\text{def}}{=} Fx(T) \cdot x(T) + \int_s^T Qx \cdot x + |u|^2 - |v|^2 \, dt, \tag{4.1}$$

$$x' = Ax + B_1 u + B_2 v \quad \text{a.e. in } [s, T], \quad x(s) = h. \tag{4.2}$$

DEFINITION 4.1. *Let $h \in \mathbf{R}^n$ be an initial state at time s, $0 \le s < T$.*

(i) *The game is said to achieve its* open loop lower value *(resp.,* upper value*) if*

$$v_s^-(h) \stackrel{\text{def}}{=} \sup_{v \in L^2(s.T; \mathbf{R}^k)} \inf_{u \in L^2(s.T; \mathbf{R}^m)} C_h^s(u, v) \tag{4.3}$$

$$(\textit{resp.}, \; v_s^+(h) \stackrel{\text{def}}{=} \inf_{u \in L^2(s.T; \mathbf{R}^m)} \sup_{v \in L^2(s, T; \mathbf{R}^k)} C_h^s(u, v)) \tag{4.4}$$

is finite.

(ii) *The game is said to achieve its* open loop value *if its open loop lower value $v_s^-(h)$ and upper value $v_s^+(h)$ are achieved and $v_s^-(h) = v_s^+(h)$. The open loop value of the game will be denoted by $v_s(h)$.*

(iii) *A pair $(\bar{u}, \bar{v})$ in $L^2(s, T; \mathbf{R}^m) \times L^2(s, T; \mathbf{R}^k)$ is an* open loop saddle point *of $C_h^s(u, v)$ if for all u in $L^2(s, T; \mathbf{R}^m)$ and all v in $L^2(s, T; \mathbf{R}^k)$*

$$C_h^s(\bar{u}, v) \le C_h^s(\bar{u}, \bar{v}) \le C_h^s(u, \bar{v}). \tag{4.5}$$

The first result is that, if the $C_{x_0}(u, v)$ is convex, concave, or convex-concave for some x_0, so is $C_h^s(u, v)$ for all $h \in \mathbf{R}^n$ and all s, $0 \le s < T$.

THEOREM 4.1.

(i) *If, for all $(x_0, v) \in \mathbf{R}^n \times L^2(0, T; \mathbf{R}^k)$, the map $u \mapsto C_{x_0}(u, v)$ is convex, then for all s, $0 \le s < T$, and all $(h, v) \in \mathbf{R}^n \times L^2(s, T; \mathbf{R}^k)$ the map $u \mapsto C_h^s(u, v)$ is convex.*

(ii) *If, for all $(x_0, u) \in \mathbf{R}^n \times L^2(0, T; \mathbf{R}^m)$, the map $v \mapsto C_{x_0}(u, v)$ is concave, then for all s, $0 \le s < T$, and all $(h, u) \in \mathbf{R}^n \times L^2(s, T; \mathbf{R}^m)$ the map $v \mapsto C_h^s(u, v)$ is concave.*

Proof. We prove only (i). From (2.20)–(2.22) for all $(u,v) \in L^2(0,T;\mathbf{R}^m) \times L^2(0,T;\mathbf{R}^k)$,

$$\begin{aligned} &\forall \bar{u} \in L^2(0,T;\mathbf{R}^m), \quad d^2C_{x_0}(u,v;\bar{u},0;\bar{u},0) \\ &\qquad = F\bar{y}(T)\cdot\bar{y}(T) + (Q\bar{y},\bar{y}) + (\bar{u},\bar{u}) \geq 0, \end{aligned} \tag{4.6}$$

where $\bar{y}$ is the solution of

$$\bar{y}' = A\bar{y} + B_1\bar{u}, \quad \bar{y}(0) = 0. \tag{4.7}$$

To prove the same result on $[s,T]$, associate with each $\bar{u} \in L^2(s,T;\mathbf{R}^m)$ its extension by zero $\tilde{\bar{u}}$ from $[s,T]$ to $[0,T]$. Therefore

$$\forall \bar{u} \in L^2(s,T;\mathbf{R}^m), \quad F\bar{y}(T)\cdot\bar{y}(T) + \int_0^T Q\bar{y}\cdot\bar{y} + \tilde{\bar{u}}\cdot\tilde{\bar{u}}\,dt \geq 0, \tag{4.8}$$

where $\bar{y}$ is the solution of

$$\bar{y}' = A\bar{y} + B_1\tilde{\bar{u}}, \quad \bar{y}(0) = 0. \tag{4.9}$$

Notice that, since $\tilde{\bar{u}}$ is zero in $[0,s]$, $\bar{y} = 0$ in $[0,s]$ and $\bar{y}$ is also the solution of

$$\bar{y}' = A\bar{y} + B_1\bar{u}, \quad \bar{y}(s) = 0 \tag{4.10}$$

$$\Rightarrow \forall \bar{u} \in L^2(s,T;\mathbf{R}^m), \quad F\bar{y}(T)\cdot\bar{y}(T) + \int_s^T Q\bar{y}\cdot\bar{y} + \bar{u}\cdot\bar{u}\,dt \geq 0. \tag{4.11}$$

Hence for all $h \in \mathbf{R}^n$, all $(u,v) \in L^2(s,T;\mathbf{R}^m)\times L^2(s,T;\mathbf{R}^k)$, and all $\bar{u} \in L^2(s,T;\mathbf{R}^m)$,

$$\begin{aligned} d^2C_h^s(u,v;\bar{u},0;\bar{u},0) &= F\bar{y}(T)\cdot\bar{y}(T) + \int_s^T Q\bar{y}\cdot\bar{y} + \bar{u}\cdot\bar{u}\,dt \\ &= d^2C_{x_0}(0,0;\tilde{\bar{u}},0;\tilde{\bar{u}},0) \geq 0. \end{aligned}$$

Thus for all s and all (h,v), the map $u \mapsto C_h^s(u,v)$ is convex. ∎

5. Decoupling and Riccati differential equation in the saddle point case.

5.1. Open loop saddle point optimality principle. At this juncture, it is important to notice that the necessary conditions (2.35) and (2.38) associated with the respective finiteness of the lower and upper values of the game on $[0,T]$ do not generally survive on $[s,T]$. However the convexity-concavity condition (2.29) does.

THEOREM 5.1. *Assume that $v(x_0)$ is finite for some $x_0 \in \mathbf{R}^n$, let $x(\cdot;x_0), p(\cdot;x_0)$ be a solution of the coupled system (2.27) in $[0,T]$, and let s, $0 \leq s < T$.*

(i) *The value $v_s(x(s;x_0))$ of the game is finite.*

(ii) *The restriction of (x,p) to $[s,T]$ is a solution of the coupled system*

$$\begin{cases} x_s' = Ax_s - B_1B_1^*p_s + B_2B_2^*p_s \text{ a.e. in } [s,T], & x_s(s) = x(s;x_0), \\ p_s' + A^*p_s + Qx_s = 0, & p_s(T) = Fx_s(T), \end{cases} \tag{5.1}$$

the restrictions $(u_s,v_s) = (u|_{[s,T]}, v_{[s,T]})$ of the controls (u,v) on $[0,T]$ to $[s,T]$ verify

$$u_s = -B_1^*p_s \text{ and } v_s = B_2^*p_s, \quad v_s(x(s;x_0)) = p_s(s)\cdot x(s;x_0), \tag{5.2}$$

$$v(x_0) = v_s(x(s;x_0)) + \int_0^s Qx\cdot x + |u|^2 - |v|^2\,dt, \tag{5.3}$$

and

$$\sup_{v\in L^2(s,T;\mathbf{R}^k)} C_0^s(0,v) = C_0^s(0,0) = \inf_{u\in L^2(s,T;\mathbf{R}^m)} C_0^s(u,0). \tag{5.4}$$

Proof. From Theorem 2.4 on $[s,T]$, part (i) is equivalent to part (ii), and thus it is sufficient to prove part (ii). From Theorem 4.1, the convexity-concavity conditions on $[0,T]$ survive on $[s,T]$, and we get (5.4). Moreover, if $(x(\cdot;x_0),p(\cdot;x_0))$ is a solution of the coupled system (2.27) in $[0,T]$ with initial state x_0 at time 0 and the controls (u,v) verify identities (2.28), then the restrictions $(x_s,p_s)=(x|_{[s,T]},p|_{[s,T]})$ are a solution to the coupled system (5.1), and the restrictions $(u_s,v_s)=(u|_{[s,T]},v_{[s,T]})$ of the controls on $[0,T]$ verify (5.2). Thus, by the analogue of Theorem 2.4, we get the finiteness of the value of the game on $[s,T]$. □

THEOREM 5.2. *Assume that $v(x_0)$ is finite for all $x_0\in\mathbf{R}^n$.*

(i) *The solution (x_s,p_s) of the coupled system* (5.1) *and the controls (u_s,v_s) on $[s,T]$ in* (5.2) *are unique.*

(ii) *The map*

$$x_0 \mapsto X(s)x_0 \stackrel{\text{def}}{=} x(s;x_0) : \mathbf{R}^n \to \mathbf{R}^n \tag{5.5}$$

is a linear bijection, where $x(\cdot;x_0),p(\cdot;x_0)$ is the unique solution of the coupled system (2.27) *in $[0,T]$.*

(iii) *For all $h\in\mathbf{R}^n$, the utility function $C_h^s(u,v)$ has a unique open loop saddle point $(\hat u_s,\hat v_s)\in L^2(s,T;\mathbf{R}^m)\times L^2(s,T;\mathbf{R}^k)$, and there exists a unique solution $(\hat x_s,\hat p_s)$ of the coupled system*

$$\begin{cases} \hat x_s' = A\hat x_s - B_1B_1^*\hat p_s + B_2B_2^*\hat p_s \ \text{a.e. in } [s,T], & \hat x_s(s)=h,\\ \hat p_s' + A^*\hat p_s + Q\hat x_s = 0 \ \text{a.e. in } [s,T], & \hat p_s(T) = F\hat x_s(T), \end{cases} \tag{5.6}$$

$$\text{such that } \hat u_s = -B_1^*\hat p_s \text{ and } \hat v_s = B_2^*\hat p_s. \tag{5.7}$$

Proof. (i) Assume that the pair $(\hat u_s,\hat v_s)$ is a saddle point of $C^s_{\hat x(s)}$ on the time interval $[s,T]$. Denote by $(\hat x_s,\hat p_s)$ the corresponding solution to the coupled system (5.1). Consider the new pair on the interval $[0,T]$,

$$\tilde u \stackrel{\text{def}}{=} \begin{cases} \hat u & \text{in } [0,s],\\ \hat u_s & \text{in } [s,T], \end{cases} \quad \tilde v \stackrel{\text{def}}{=} \begin{cases} \hat v & \text{in } [0,s],\\ \hat v_s & \text{in } [s,T], \end{cases} \tag{5.8}$$

and the corresponding solution $(\tilde x,\tilde p)$ to the state-adjoint state system (2.4)–(2.18). If it can be shown that the pair $(\tilde u,\tilde v)$ is a saddle point of $C_{x_0}(u,v)$ on $[0,T]$, then by uniqueness of the saddle point on $[0,T]$ we can conclude that $(\tilde u,\tilde v)=(\hat u,\hat v)$ and hence $(\hat u_s,\hat v_s)=(\hat u|_{[s,T]},\hat v|_{[s,T]})$. From this we get the uniqueness of the saddle point of $C^s_{\hat x(s)}$ on $[s,T]$ and the uniqueness of solution to the coupled system (5.1). The first remark is that $\tilde x(s)=\hat x(s)$ and from (5.3)

$$\begin{aligned} C_{x_0}(\hat u,\hat v) = v(x_0) &= \int_0^s Q\hat x\cdot\hat x + |\hat u|^2 - |\hat v|^2\,dt + v_s(\hat x(s))\\ &= \int_0^s Q\hat x\cdot\hat x + |\hat u|^2 - |\hat v|^2\,dt + F\hat x_s(T)\cdot\hat x_s(T) + \int_s^T Q\hat x_s\cdot\hat x_s + |\hat u_s|^2 - |\hat v_s|^2\,dt\\ &\Rightarrow C_{x_0}(\hat u,\hat v) = C_{x_0}(\tilde u,\tilde v). \end{aligned}$$

Yet, this is not sufficient to conclude that $(\tilde{u},\tilde{v})$ is a saddle point of $C_{x_0}(u,v)$. We must show that

$$\sup_{v\in L^2(0,T;\mathbf{R}^k)} C_{x_0}(\tilde{u},v) = C_{x_0}(\tilde{u},\tilde{v}) = \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_{x_0}(u,\tilde{v}). \tag{5.9}$$

The second remark is that, since $(\tilde{u}-\hat{u},\tilde{v}-\hat{v})$ is equal to $(0,0)$ on $[0,s]$, $(\hat{u}_s-\hat{u},\hat{v}_s-\hat{v})$ is a saddle point of $C_0^s(u_s,v_s)$. Combining this with the fact that, by (5.4), $(0,0)$ is also a saddle point of $C_0^s(u_s,v_s)$, the pairs $(\hat{u}_s-\hat{u},0)$ and $(0,\hat{v}_s-\hat{v})$ are also saddle points of $C_0^s(u_s,v_s)$ and $C_0^s(\hat{u}_s-\hat{u},0) = C_0^s(0,\hat{v}_s-\hat{v}) = 0$. The third remark is that

$$C_{x_0}(\hat{u},\tilde{v}) = C_{x_0}(\hat{u},\hat{v}) + dC_{x_0}(\hat{u},\hat{v};0,\tilde{v}-\hat{v}) + C_0(0,\tilde{v}-\hat{v}) = C_{x_0}(\hat{u},\hat{v}) + C_0(0,\tilde{v}-\hat{v}).$$

But, since $\tilde{v}-\hat{v}$ is equal to 0 on $[0,s]$,

$$C_0(0,\tilde{v}-\hat{v}) = C_0^s(0,\hat{v}_s-\hat{v}) = 0 \quad \Rightarrow C_{x_0}(\hat{u},\tilde{v}) = C_{x_0}(\hat{u},\hat{v}) = C_{x_0}(\tilde{u},\tilde{v}).$$

We now prove the second part of identity (5.9):

$$C_{x_0}(u,\tilde{v}) = C_{x_0}(\hat{u},\tilde{v}) + dC_{x_0}(\hat{u},\tilde{v};u-\hat{u},0) + C_0(u-\hat{u},0). \tag{5.10}$$

Since $(0,0)$ is a saddle point of $C_0(u,v)$,

$$\inf_{u\in L^2(0,T;\mathbf{R}^m)} C_0(u-\hat{u},0) = \inf_{u\in L^2(0,T;\mathbf{R}^m)} C_0(u,0) = 0.$$

It remains to prove that for all $u\in L^2(0,T;\mathbf{R}^m)$, $dC_{x_0}(\hat{u},\tilde{v};u-\hat{u},0) = 0$. First observe that

$$\begin{aligned} dC_{x_0}(\hat{u},\tilde{v};u-\hat{u},0) &= dC_{x_0}(\hat{u},\hat{v};u-\hat{u},0) + dC_0(0,\tilde{v}-\hat{v};u-\hat{u},0) \\ &= dC_0(0,\tilde{v}-\hat{v};u-\hat{u},0). \end{aligned}$$

Since $(0,\hat{v}_s-\hat{v})$ is a saddle point of C_0^s on $[s,T]$, there exists a pair (ξ,π) solution of the coupled system

$$\begin{cases} \xi' = A\xi - B_1B_1^*\pi + B_2B_2^*\pi \text{ a.e. in } [s,T], & \xi(s) = 0, \\ \pi' + A^*\pi + Q\xi = 0, & \pi(T) = F\xi(T), \end{cases} \tag{5.11}$$

$$0 = -B_1^*\pi, \quad \hat{v}_s - \hat{v} = B_2^*\pi. \tag{5.12}$$

The first equation can also be written

$$\xi' = A\xi + B_2(\hat{v}_s-\hat{v}) \text{ a.e. in } [s,T], \quad \xi(s) = 0.$$

Denote by $\tilde{\xi}$ the solution of the state equation (2.4) on $[0,T]$ corresponding to the initial state 0 and the control pair $(0,\tilde{v}-\hat{v})$:

$$\tilde{\xi}' = A\tilde{\xi} + B_2(\tilde{v}-\hat{v}) \text{ a.e. in } [0,T], \quad \tilde{\xi}(0) = 0.$$

Then observe that, since the restriction of $\tilde{v}-\hat{v}$ to $[0,s]$ is 0, $\tilde{\xi} = 0$ on $[0,s]$ and $\tilde{\xi} = \xi$ on $[s,T]$. Denoting by y the solution of

$$y' = Ay + B_1(u-\hat{u}) \text{ a.e. in } [0,T], \quad y(0) = 0,$$

we get the expression (cf. (2.16) and (2.19) for the directional derivative)

$$\begin{aligned} dC_0(0,\tilde{v}-\hat{v};u-\hat{u},0) &= F\tilde{\xi}(T)\cdot y(T)+\int_0^T Q\tilde{\xi}\cdot y+0\cdot(u-\hat{u})+(\tilde{v}-\hat{v})\cdot 0\,dt \\ &= F\tilde{\xi}(T)\cdot y(T)+\int_0^T Q\tilde{\xi}\cdot y\,dt = F\tilde{\xi}(T)\cdot y(T)+\int_s^T Q\tilde{\xi}\cdot y\,dt \\ &= F\xi(T)\cdot y(T)+\int_s^T Q\xi\cdot y\,dt = \int_s^T B_1^*\pi\cdot(u-\hat{u})\,dt = 0, \end{aligned}$$

since $B_1^*\pi = 0$ on $[s,T]$ from (5.12). This establishes the second part of expression (5.9). The proof of the first part is dual to the proof of the second part. This yields the uniqueness and completes the proof of part (i).

(ii) The map (5.5) is clearly linear (and continuous). Assume that it is not bijective; then there exists some $x_0 \in \mathbf{R}^n$, $x_0 \neq 0$, such that $\hat{x}(s) = 0$. The restriction of $(\hat{x},\hat{p})$ to the interval $[s,T]$ is a solution of the system

$$\text{(5.13)}\qquad \begin{cases} \xi' = A\xi - B_1B_1^*\pi + B_2B_2^*\pi \text{ a.e. in } [s,T], & \xi(s) = 0 = \hat{x}(s), \\ \pi' + A^*\pi + Q\xi = 0 \text{ a.e. in } [s,T], & \pi(T) = F\xi(T). \end{cases}$$

But from part (i) the unique solution of system (5.13) is $(0,0)$. Hence

$$\begin{aligned} &(\hat{x},\hat{p}) = (0,0) \text{ in } [s,T] \quad \Rightarrow (\hat{x}(s),\hat{p}(s)) = (0,0) \\ \Rightarrow &\begin{cases} \hat{x}' = A\hat{x} - B_1B_1^*\hat{p} + B_2B_2^*\hat{p} \text{ a.e. in } [0,s], & \hat{x}(s) = 0, \\ \hat{p}' + A^*\hat{p} + Q\hat{x} = 0 \text{ a.e. in } [0,s], & \hat{p}(s) = 0, \end{cases} \\ &\Rightarrow (\hat{x},\hat{p}) = (0,0) \text{ in } [0,s] \quad \Rightarrow x_0 = \hat{x}(0) = 0. \end{aligned}$$

This contradicts our initial conjecture that $x_0 \neq 0$, and we conclude that the linear map (5.5) is injective and, a fortiori, bijective.

(iii) From part (i) for each $h \in \mathbf{R}^n$ and each s, $0 \leq s < T$, there exists a unique $h_0 \in \mathbf{R}^n$ such that $h = X(s)h_0$. But $C_{h_0}(u,v)$ has a unique open loop saddle point in $[0,T]$. From part (i), $C^s_{X(s)h_0}(u,v)$ has a unique open loop saddle point in $[s,T]$. The result now follows from the fact that $h = X(s)h_0$. The equations and the identities follow from Theorem 5.1(ii). □

Remark 5.1. The proof of part (i) is not trivial. It is one of the key elements needed to get the result of part (iii) that says that $C^s_h(u,v)$ has a saddle point for all initial state h and all initial times s.

5.2. Decoupling of the coupled system. We need the following lemma.

LEMMA 5.1. *Assume that the open loop saddle point value $v(x_0)$ is finite for all $x_0 \in \mathbf{R}^n$. Let s, $0 \leq s < T$, and $(\hat{x}_s,\hat{p}_s)$ be the unique solution of the coupled system* (5.6) *with initial state h at time s. Then the map $P(s)$*

$$\text{(5.14)}\qquad h \mapsto P(s)h \stackrel{\text{def}}{=} \hat{p}_s(s) : \mathbf{R}^n \to \mathbf{R}^n$$

is linear, continuous, and symmetrical.

Proof. By definition, $P(s)$ is linear and continuous. For the symmetry, let (x,p) and $(\bar{x},\bar{p})$ be the solutions of the coupled system (5.6) for the respective initial states

h and $\bar{h}$ at time s. By symmetry of F, $Q(t)$, and $B_1(t)B_1^*(t) - B_2(t)B_2^*(t)$,

$$\begin{aligned} P(s)h\cdot\bar{h} = p(s)\cdot\bar{x}(s) &= p(T)\cdot\bar{x}(T) - \int_s^T p'\cdot\bar{x} + p\cdot\bar{x}'\,dt \\ &= Fx(T)\cdot\bar{x}(T) - \int_s^T -(A^*p+Qx)\cdot\bar{x} + p\cdot(A\bar{x} - B_1B_1^*\bar{p} + B_2B_2^*\bar{p})\,dt \\ &= Fx(T)\cdot\bar{x}(T) + \int_s^T Qx\cdot\bar{x} + p\cdot(B_1B_1^* - B_2B_2^*)\bar{p}\,dt = P(s)\bar{h}\cdot h, \end{aligned}$$

and $P(s)^* = P(s)$. □

Remark 5.2. At this juncture the matrix Riccati differential equation can be readily obtained from Lemma 3.1 in [4] since, from Theorem 5.2(ii), the matrix function $X(s)$ is invertible for all s. However, in view of Lemma 5.1, we use invariant embedding to get more a priori information on the decoupling matrix function $P(s)$.

THEOREM 5.3. *Assume that $v(x_0)$ is finite for all $x_0 \in \mathbf{R}^n$.*

(i) *Given the solution of the coupled system* (2.27) *in* $[0,T]$ *for* $x_0 \in \mathbf{R}^n$,

$$\hat{p}(s) = P(s)\hat{x}(s), \quad 0 \le s \le T. \tag{5.15}$$

(ii) *The elements of the matrix function P are $H^1(0,T)$-functions, the elements of the matrix functions*

$$A_P \overset{\text{def}}{=} A - RP, \quad R \overset{\text{def}}{=} B_1B_1^* - B_2B_2^* \tag{5.16}$$

belong to $L^\infty(0,T)$, and the closed loop system

$$\hat{x}' = [A - (B_1B_1^* - B_2B_2^*)\,P]\,\hat{x} \text{ a.e. in } [0,T], \quad \hat{x}(0) = x_0, \tag{5.17}$$

has a unique solution in $H^1(0,T;\mathbf{R}^n)$. For all (t,s), $0 \le s \le t \le T$, the fundamental matrix solution $\Phi_P(t,s)$ associated with the closed loop system (5.17) *and its inverse $\Phi_P(t,s)^{-1}$ are continuous in $\{(t,s) : 0 \le s \le t \le T\}$. For all pairs $0 \le s \le t \le T$*

$$\frac{\partial}{\partial s}\Phi_P(t,s) + \Phi_P(t,s)A_P(s) = 0 \text{ a.e. in } [0,t], \quad \Phi_P(t,t) = I. \tag{5.18}$$

(iii) *For all h and $\overline{h}$ in $\mathbf{R}^n$*

$$\begin{aligned} h\cdot P(s)\,\overline{h} &= \Phi_P(T,s)h\cdot F\Phi_P(T,s)\bar{h} \\ &\quad + \int_s^T \Phi_P(t,s)h\cdot[Q(t) + P(t)R(t)P(t)]\,\Phi_P(t,s)\bar{h}\,dt. \end{aligned} \tag{5.19}$$

Proof. (i) From Theorems 5.1 and 5.2(i)

$$\hat{x}_s = \hat{x}|_{[s,T]}, \quad \hat{p}_s = \hat{p}|_{[s,T]} \quad \Rightarrow \hat{p}(s) = \hat{p}_s(s) = P(s)\hat{x}_s(s) = P(s)\hat{x}(s),$$

and we get (5.15). The closed loop system is obtained by direct substitution of the identity (5.15) for $\hat{p}$ into the first equation of the coupled system (2.27) in $[0,T]$.

(ii) Associate with the solution of the coupled system (2.27) in $[0,T]$ the matrix function

$$\Lambda(s)x_0 \overset{\text{def}}{=} \hat{p}(s;x_0), \quad \forall x_0 \in \mathbf{R}^n,\ 0 \le s \le T. \tag{5.20}$$

From (5.15) in part (i) and the invertibility of $X(s)$

$$\Lambda(s)x_0 \overset{\text{def}}{=} P(s)X(s)x_0, \quad \forall x_0 \in \mathbf{R}^n,\ 0 \le s \le T$$
$$\Rightarrow\ \Lambda(s) = P(s)X(s),\ \Rightarrow\ P(s) = \Lambda(s)X(s)^{-1}, \quad 0 \le s \le T.$$

Since $X(s)$ is invertible and the elements of the matrices X and Λ are $H^1(0,T)$-functions,

$$P'(s) = \Lambda(s)'X(s)^{-1} - \Lambda(s)X(s)^{-1}X(s)'X(s)^{-1}. \tag{5.21}$$

In particular the elements of the matrix function P are $H^1(0,T)$-functions. Then the matrix function $A_P(t)$ in (5.16) belongs to $L^\infty(0,T)$, and the closed loop system (5.17) has a unique solution in $H^1(0,T;\mathbf{R}^n)$. From this Φ_P has the usual properties of a fundamental matrix solution Φ_P in $\{(t,s) : 0 \le s \le t \le T\}$, $\Phi_P(t,0) = \Phi_P(t,s)\Phi_P(s,0)$, and

$$\frac{\partial \Phi_P}{\partial s}(t,s) + \Phi_P(t,s)A_P(s) \text{ a.e. in } [0,T], \quad \Phi_P(t,t) = I. \tag{5.22}$$

(iii) Let (ψ,φ) (resp., $(\overline{\psi},\overline{\varphi})$) be the solution of the coupled system (5.1) for the initial state h (resp., $\overline{h}$). Then by direct computation

$$h \cdot \overline{\psi}(s) = \varphi(T) \cdot F\overline{\varphi}(T) + \int_s^T \varphi(t) \cdot Q(t)\,\overline{\varphi}(t) + \psi(t) \cdot R(t)\,\overline{\psi}(t)\,dt,$$
$$h \cdot P(s)\,\overline{h} = \Phi_P(T,s)h \cdot F\Phi_P(T,s)\bar{h} + \int_s^T \Phi_P(t,s)h \cdot [Q(t) + P(t)R(t)P(t)]\,\Phi_P(t,s)\bar{h}\,dt. \tag{5.23}$$

□

5.3. Proof of Theorem 2.9. (i) From identity (5.21) in the proof of part (ii) of Theorem 5.3 a straightforward computation yields that the matrix function P is a solution of the matrix Riccati differential equation (2.43). This solution is unique. Indeed if $\bar{P}$ is another solution of the Riccati equation, the closed loop system with $\bar{P}$ has a unique solution $\bar{x}$ and it is easy to check that $\bar{p} = \bar{P}\bar{x}$ is a solution of the associated adjoint equation. But there is a unique solution to the coupled system. By definition of P via invariant embedding we get that $\bar{P} = P$. (ii) and (iii) The proof follows from identities (2.34) and (2.35) in Theorem 2.2.

5.4. Proof of Theorem 2.10. From Theorem 2.9 we get (a) and (b). Conversely, from (a) if P is a solution of the Riccati differential equation, the closed loop system has a unique solution x_P and $p_P = Px_P$ is the solution of the adjoint system. It is then easy to check that the pair (x_P, p_P) is indeed a solution of the coupled system (2.27) in $[0,T]$. Finally from the convexity-concavity property (b) we get the existence of the open loop saddle point.

REFERENCES

[1] T. Başar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Birkhäuser Boston, Boston, 1991 (2nd ed., 1995).

[2] A. Bensoussan and P. Bernhard, *Remarks on the theory of robust control*, in Optimization, Optimal Control and Partial Differential Equations (Iaşi, 1992), Internat. Ser. Numer. Math. 107, Birkhäuser, Basel, Switzerland, 1992, pp. 149–166.

[3] P. Bernhard, *Contribution à l'étude des jeux différentiels à deux personnes, somme nulle et information parfaite*, Thesis, Université de Paris VI, Paris, France, 1978.

[4] P. Bernhard, *Linear-quadratic, two-person, zero-sum differential games: Necessary and sufficient conditions*, J. Optim. Theory Appl. 27 (1979), pp. 51–69.
[5] M. C. Delfour and S. K. Mitter, *Reachability of perturbed systems and min sup problems*, SIAM J. Control, 7 (1969), pp. 521–533.
[6] I. Ekeland and R. Temam, *Analyse convexe et problèmes variationnels*, Dunod, Gauthier–Villars, Paris, 1974.
[7] P. Faurre, *Jeux différentiels à stratégie complètement optimale et principe de séparation*, Fourth IFAC World Congress, Warsaw, Poland 1969.
[8] Y. Ho, A. E. Bryson, and S. Baron, *Differential games and optimal pursuit-evasion strategies*, IEEE Trans. Automat. Control, AC-10 (1965), pp. 385–389.
[9] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, Heidelberg, Berlin, 1971.
[10] P. Zhang, *Some results on two-person zero-sum linear quadratic differential games*, SIAM J. Control Optim., 43 (2005), pp. 2157–2165.
[11] P. Zhang, G. Zheng, Y. Xu, and J. Xi, *The Riccati equation of game type*, in Proceedings of the 24th Chinese Control Conference, Guangzhou, China, 2005, pp. 573–577.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 775–791

OPTIMAL CONTROL OF A NUTRIENT-PHYTOPLANKTON-ZOOPLANKTON-FISH SYSTEM*

MARCUS R. GARVIE† AND CATALIN TRENCHEA‡

Abstract. We consider the mathematical formulation, analysis, and numerical solution of an optimal control problem for a nonlinear "nutrient-phytoplankton-zooplankton-fish" reaction-diffusion system. We study the existence of optimal solutions, derive an optimality system, and determine optimal solutions. In the original spatially homogeneous formulation [M. Scheffer, *Oikos*, 62 (1991), pp. 271–282] the dynamics of plankton were investigated as a function of parameters for nutrient levels and fish predation rate on zooplankton. In our paper the model is spatially extended and the parameter for fish predation treated as a multiplicative control variable. The model has implications for the biomanipulation of food-webs in eutrophic lakes to help improve water quality. In order to illustrate the control of irregular spatiotemporal dynamics of plankton in the model we implement a semi-implicit (in time) finite element method with "mass lumping" and present the results of numerical experiments in two space dimensions.

Key words. chaos, optimal control, biomanipulation, predator-prey interaction, finite element method

AMS subject classifications. 49K20, 92D25, 35K57

DOI. 10.1137/050645415

1. Introduction.

1.1. Model equations.

In this paper we study the following nutrient-phytoplankton-zooplankton-fish reaction-diffusion system:

$$\begin{cases} \dfrac{\partial A}{\partial \tau} = D_A \Delta A + \widehat{r}\dfrac{n}{n+h_n}A - cA^2 - pZ\dfrac{A}{A+h_a}, \\ \dfrac{\partial Z}{\partial \tau} = D_Z \Delta Z + peZ\dfrac{A}{A+h_a} - \widehat{m}Z - F\dfrac{Z^2}{Z^2+h_Z^2}, \end{cases} \tag{1.1}$$

where A is phytoplankton biomass, Z is zooplankton biomass, D_a and D_z are the diffusion coefficients of phytoplankton and zooplankton, respectively, and n is the nutrient level of the system. F is the rate of zooplankton biomass consumed by fish per unit volume of water per day (average predation rate times the density of fish). It is important to note that in our formulation $F \equiv F(x,t)$, i.e., F is defined at every point in the lake and at every point in time. For definitions of the positive parameters c, e, $\widehat{m}$, $\widehat{r}$, p, h_a, h_z, and h_n, see [39]. The symbol τ denotes time in days, and $\Delta = \sum_{i=1}^{d} \partial^2/\partial X_i^2$ is the usual Laplacian operator in $d = 2$ or 3 space dimensions. The grazing rate of zooplankton on phytoplankton is of a type II functional response, while the predation rate of fish on zooplankton is of type III [17]. In the absence of zooplankton the phytoplankton are assumed to grow logistically.

*Received by the editors November 16, 2005; accepted for publication (in revised form) November 6, 2006; published electronically May 29, 2007.
http://www.siam.org/journals/sicon/46-3/64541.html

†Department of Mathematics and Statistics, University of Guelph, Guelph, ON N1G 2W1, Canada (mgarvie@uoguelph.ca).

‡Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (trenchea@pitt.edu).

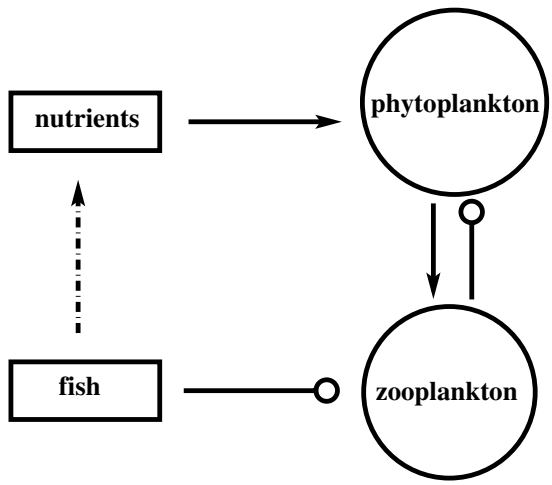


FIG. 1.1. *Interactions incorporated into a nutrient-plankton-fish model. Arrows indicate positive effects, circles indicate negative effects (redrawn from* [39]*).*

The nutrient-phytoplankton-zooplankton-fish model was originally formulated as a system of ordinary differential equations (ODEs) by Scheffer [39] and has since been spatially extended (see, for example, [32, 29, 30, 33, 34, 45]).

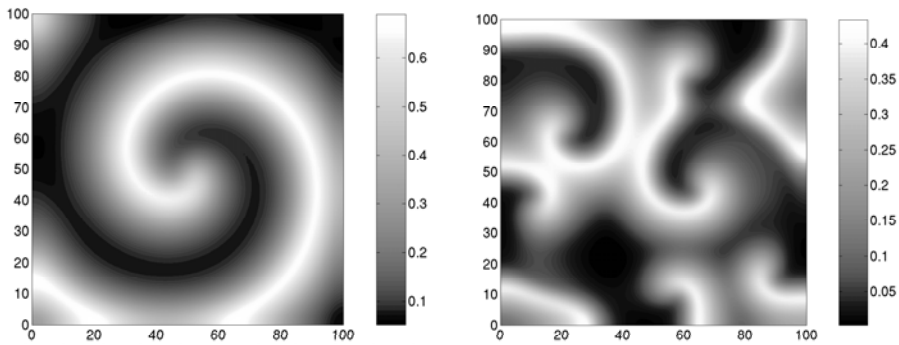
The reaction-diffusion system models a simple food-web in lakes where planktivorous fish feed on zooplankton, and the zooplankton feed on phytoplankton (algae). The basic interactions in the model are illustrated in Figure 1.1. The model is "minimal" in the sense that only a few important interactions are taken into account. For example, the positive effect that fish have on the nutrients of the system is omitted in this model (indicated by the dotted arrow in Figure 1.1). Nevertheless, such minimal predator-prey systems display a wide range of ecologically relevant behavior, for example, spiral waves [34], target waves [41], diffusion-induced instability [30], and chaos [36, 34]. See [31] for a historical overview of modeling plankton dynamics and pattern formation mechanisms.

It is simpler to work with equations that have been scaled to nondimensional form; thus after letting $N := n/(n + h_n)$ in (1.1) we define dimensionless phytoplankton densities, zooplankton densities, spatial coordinates, and time via

$$u = \frac{cA}{\widehat{r}N}, \quad v = \frac{cZ}{\widehat{r}eN}, \quad x_i = \frac{kX_i}{L}, \quad t = R_0\tau$$

(cf. [33] and [34]), where R_0 is the characteristic (or typical) growth rate of phytoplankton, k is a factor related to the scale of expected patchy patterns, and L is the maximum diameter of the lake in the coordinate direction x_2, or x_3. Note that in the case of a square domain, L is the side length. We also rescale the parameters via

$$r = \frac{R}{R_0}, \qquad a = \frac{C_1K}{C_2R_0}, \qquad b = \frac{K}{C_2}, \qquad m = \frac{M}{R_0},$$
$$f = \frac{F}{C_3R_0}, \qquad g = \frac{K}{C_3A}, \qquad d_1 = \frac{k^2D_A}{L^2R_0}, \qquad d_2 = \frac{k^2D_Z}{L^2R_0},$$



(a) (b)

FIG. 1.2. *Numerical solution v of (1.2) at time $t = 1000$ with $d_1 = d_2 = 0.05$, $f = 0$, $g = 10$, solved on a space-time grid of $101 \times 101 \times 12000$. A semi-implicit (in time) Galerkin finite element method with piecewise linear continuous basis functions was employed with homogeneous Neumann boundary conditions. In both cases initial data of the form $u_0 = A(x/100) + B$, $v_0 = C(100 - y)/100 + D$ was employed with the following parameter values:* (a) $A = 0.02$, $B = 0.19$, $C = 0.02$, $D = 0.31$, $r = 1$, $a = b = 5$, $m = 0.5$; (b) $A = 0.2$, $B = 0.1$, $C = 0.2$, $D = 0.1$, $r = 1$, $a = b = 20$, $m = 0.8$.

which leads to the dimensionless system

$$\begin{cases} \dfrac{\partial u}{\partial t} = d_1 \Delta u + ru(1-u) - \dfrac{auv}{1+bu}, \\ \dfrac{\partial v}{\partial t} = d_2 \Delta v + \dfrac{auv}{1+bu} - mv - f\dfrac{gv^2}{1+g^2v^2}. \end{cases} \tag{1.2}$$

Any 2-component reaction-diffusion system with reaction kinetics close to a super critical Hopf bifurcation, with equal diffusion coefficients, can be transformed into a generic reaction-diffusion system of "λ-ω" form [40]. Thus as spiral wave solutions of λ-ω reaction-diffusion systems have been proved to exist [9], for appropriate parameter values and initial data we also expect spiral wave solutions to exist for system (1.1). In Figure 1.2 we present snapshots at $t = 1000$ for the uncontrolled system (1.2) representing spiral wave solutions that persist indefinitely (Figure 1.2(a)) or, after initialization, rapidly break up into irregular patterns (Figure 1.2(b)). In both cases we checked that this behavior persists up to $t = 10{,}000$. This behavior is important to our study as we apply controls that drive the system from the unstable regime to form regular patterns.

In this paper we consider the above model in the context of eutrophication of lakes. Eutrophication is the process where excessive input of nutrients in lakes leads to high levels of phytoplankton (algae) and hence degraded water quality (see [6, 18] and the references therein). The most common approach to improving water quality in this situation is to either reduce the external nutrient loading or enhance zooplankton by reducing planktivorous fish, thereby reducing algal biomass. We focus on the latter approach ("top-down" control), where the nutrient level in the system is determined by a parameter. In practice, planktivorous fish can be reduced by fish removal, or by piscivore[1] stocking. This manipulation of the food-web is called biomanipulation and

[1]Species that feed on fish.

is an important approach for improving water quality in eutrophic lakes (e.g., [23, 25, 28, 38, 44, 50]). Evidence from the limnological literature supports the hypothesis that the most effective biomanipulation strategy for improving water quality is the partial removal of fish, and that there may be an optimum harvesting rate of planktivorous fish [11, 21, 48, 49]. This observation helped motivate the work in this paper, where we consider the optimal control of phytoplankton u and zooplankton v densities, where f is the distributed control. We aim to minimize the quadratic cost functional

$$J(u,v,f) = \frac{1}{2}\int_Q \left(|u-\bar{u}|^2 + |v-\bar{v}|^2\right) dxdt + \frac{\alpha}{2}\int_Q \left|\frac{\partial f}{\partial t}\right|^2 dxdt, \tag{1.3}$$

where $(\bar{u},\bar{v})$ are the desired phytoplankton-zooplankton densities, and $Q = \Omega \times (0,T)$ is the space-time domain of interest. The first two terms in (1.3) measure, with respect to the L^2-norm in space and time, the difference between the given target densities $(\bar{u},\bar{v})$ and the state densities (u,v). The last term in (1.3) reflects the fact that we want to avoid changing the control (adding or removing fish) too often. The constant α can be chosen to adjust the relative importance of this cost. We assume the state equations (1.2) are augmented with appropriate initial and boundary conditions. Note that as there is no forcing in the first equation of (1.2), the control of phytoplankton must result indirectly through the coupling with the second equation. Furthermore, the forcing in (1.2) enters not merely as an additive inhomogeneous source term, but rather in a multiplicative manner. Additionally, as the control f is a *rate*, it is strictly nonnegative.

We emphasize that although this work has implications for the control of eutrophication in lakes, the main focus of this work is on the rigorous mathematical analysis of the optimal control problem. We do not provide practical implementational details for the improvement of water quality in lakes, but we do provide a theoretical basis on which such a task would be based. We also remark that the mathematical problem and results from numerical simulations may provide insights into the field situation.

It is important to distinguish between the mathematical (optimal control) problem and the practical problem in the field. Mathematically, we assume that f can be manipulated at every point in space and time. However, from a practical point of view, we only have direct control of the net density of fish in the lake at any instant. For example, fish released into a lake may distribute themselves uniformly throughout the lake, or move in schools. Nevertheless, there is significant overlap between the mathematical problem and the field situation, for example, in the case where we wish to reduce the (net) algal growth. In addition to improving water quality we are also interested in the more fundamental aim of maintaining a stable equilibrium between the plankton, thus avoiding the extinction of one or more species.

The aim of this paper is to undertake the mathematical analysis of the optimal control problem introduced above, namely, to minimize (1.3) subject to (1.2). We also provide some numerical results that illustrate the theoretical results. We remark that there are few optimal control studies in the literature for interacting species involving space and time (e.g., [1, 2, 3, 8, 13, 24, 26]) or for reaction-diffusion equations applied in other contexts (e.g., [5, 7, 12, 19, 43]).

2. Mathematical preliminaries.

2.1. Local analysis. We present some details of the local dynamics of the reaction-diffusion system for the state equations (1.2). This is important for deriving necessary conditions on the system parameters for the kinetics to possess biologically

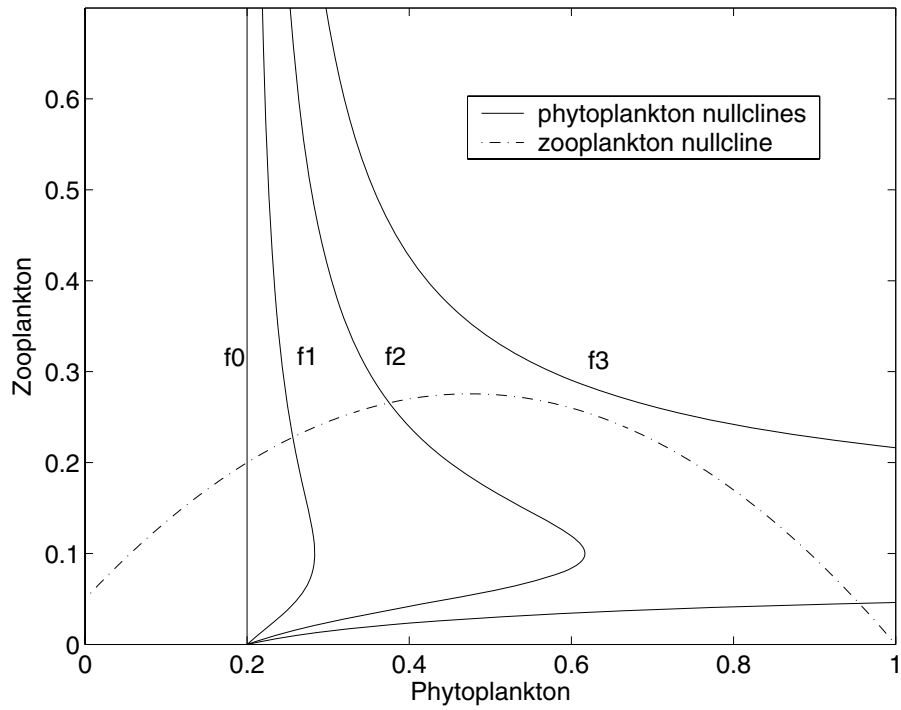


FIG. 2.1. *Nullclines for the local kinetics of* (1.2) *with* $r = 1$, $a = b = 20$, $m = 0.8$, $g = 10$, *for fish predation rates* $f_0 = 0$ *(limit cycle, no fish),* $f_1 = 0.1$ *(limit cycle),* $f_2 = 0.25$ *(limit cycle), and* $f_3 = 0.4$ *(phytoplankton dominance).*

meaningful equilibria. Furthermore, this acts as a guide in the appropriate choice of parameters for the numerical simulation of optimal solutions.

The local dynamics can be analyzed by considering the nullclines ("zero-isoclines") of this system, which are the solution curves for

$$v = \psi(u) := \frac{r}{a}(1-u)(1+bu),$$
$$u = \phi(v) := \frac{g^2 m v^2 + fgv + m}{(a-bm)g^2v^2 - bfgv + a - bm}, \tag{2.1}$$

corresponding to the first and second equations of (1.2). It is easy to show that there are saddle points at $(0,0)$ and $(0,1)$, and a stationary point (u_s, v_s) (stable or unstable) corresponding to the coexistence of phytoplankton and zooplankton. Note that for positive u_s and v_s we must have

$$m < \frac{a}{1+b}, \tag{2.2}$$

which follows from the restriction $0 < u_s < 1$. The nullclines are illustrated in Figure 2.1 for a specific parameter set and increasing predation rate. With no fish present, or a low fish predation rate, there is a limit cycle in the reaction kinetics surrounding the unstable stationary point, while at higher predation rates the system is dominated by a phytoplankton-only state [33, 39]. In general there are no closed

form expressions for u_s and v_s and consequently the analysis must be done numerically in each case. However, in the special case with no fish present, the analysis is straightforward and is briefly outlined below.

With $f = 0$ the intersection of the nullclines $v = \psi(u)$ and $u = \phi(v)$ in (2.1) intersect at (u_s, v_s), where

$$u_s = \frac{m}{a - bm}, \quad v_s = \frac{r}{a}(1 - u_s)(1 + bu_s). \tag{2.3}$$

To find conditions that guarantee limit cycle kinetics in the positive quadrant of phase space we first prove the existence of a positively invariant region in the sense of [47, Definition 1.1.4] that contains (u_s, v_s), and then apply standard theory of dynamical systems. We claim that the trapezoidal region $\Sigma \in [0, \infty)^2$ defined by

$$\Sigma := \left\{(u, v) \,:\, u \leq l, \quad v - \frac{l}{m}(r + m) + u \leq 0, \quad u \geq 0, \quad v \geq 0\right\}, \quad l > 1, \tag{2.4}$$

is positively invariant. To see this, first observe that the reaction kinetics do not point out of Σ along $u = 0$, $v = 0$, and $u = l$. To show that this is also true along the line $v = \frac{l}{m}(r + m) - u$, set $G(u, v) := v - \frac{l}{m}(r + m) + u$ and denote the outward normal to Σ along this line by $\nabla G := (\partial G/\partial u, \partial G/\partial v)^T = (1, 1)^T$. Then denoting the vector of reaction kinetics by $\widehat{\mathbf{f}} = (\widehat{f}, \widehat{g})^T$, observe that

$$\nabla G \cdot \widehat{\mathbf{f}}|_{v = \frac{l}{m}(r+m) - u} \leq (r + m)(u - l) \leq 0,$$

which proves the assertion. We derive a simple condition that ensures the critical point given by (2.3) is either an unstable node or an unstable focus, and thus by the Poincaré–Bendixson theorem [47, Theorem 1.1.19] there exists a limit cycle solution surrounding this point. Define

$$A := \begin{pmatrix} \widehat{f}_u & \widehat{f}_v \\ \widehat{g}_u & \widehat{g}_v \end{pmatrix}\Big|_{(u_s, v_s)};$$

direct calculation then leads to

$$\operatorname{tr} A := mr\left[\frac{(b + 1)}{a} - \frac{2}{(a - bm)}\right], \quad |A| = \frac{mr}{a}\left[a - (b + 1)m\right].$$

Now from (2.2) we have $|A| > 0$, and with the condition

$$m < \frac{a(b - 1)}{b(b + 1)}, \tag{2.5}$$

it also follows that $\operatorname{tr} A > 0$. Thus if (2.5) is satisfied, then the critical point (u_s, v_s) is an unstable node or focus (e.g., [46, p. 107]), as required. Thus, to summarize, in the special case with $f = 0$, if the system parameters satisfy (2.5), then there exists a limit cycle solution in the positive quadrant of phase space surrounding the unstable stationary state given by (2.3) (note that condition (2.5) implies condition (2.2)). It is easy to check that (2.5) is satisfied for the parameter sets in Figure 1.2, and also for the numerical results of the uncontrolled problem in section 4.

2.2. Well-posedness of the state equations. We use results from semigroup theory and the abstract theoretical setup of Morgan [35], which is based on the kinetics satisfying a Lyapunov-type condition, to infer the global existence and uniqueness of classical solutions of the state equations (1.2).

Before proving well-posedness of the equations we need to establish the formal setting and restate the fish-plankton system with appropriate initial and boundary data. Let Ω be a bounded and open subset of $\mathbb{R}^d$, $d \leq 3$, with a boundary $\partial\Omega$ of class C^{2+s}, $s > 0$, i.e., $\partial\Omega$ is a $(d-1)$-dimensional $C^{2+\nu}$ manifold on which Ω lies locally on one side. The model problem is formulated as follows:

Find the phytoplankton $u(\mathbf{x}, t)$ and zooplankton $v(\mathbf{x}, t)$ densities such that

$$\frac{\partial u}{\partial t} = d_1 \Delta u + ru(1-u) - \frac{auv}{1+bu} \qquad \text{in} \quad Q := \Omega \times (0,T), \tag{2.6a}$$

$$\frac{\partial v}{\partial t} = d_2 \Delta v + \frac{auv}{1+bu} - m\, v - f\frac{gv^2}{1+g^2v^2} \qquad \text{in} \quad Q, \tag{2.6b}$$

$$u(\mathbf{x},0) = u_0(\mathbf{x}), \qquad v(\mathbf{x},0) = v_0(\mathbf{x}), \quad \mathbf{x} \in \Omega \tag{2.6c}$$

$$\frac{\partial u}{\partial \boldsymbol{\nu}} = \frac{\partial v}{\partial \boldsymbol{\nu}} = 0 \quad \text{on} \quad \partial\Omega \times (0,T), \tag{2.6d}$$

where the parameters a, b, m, and g are real and strictly positive, and $\boldsymbol{\nu}$ denotes the outward normal to $\partial\Omega$. We assume that the control $f \equiv f(\mathbf{x}, t)$ is a Lipschitz continuous function on Q, which we denote by $f \in \mathrm{Lip}(Q)$, and that the initial data is bounded, i.e., $u_0(\mathbf{x}), v_0(\mathbf{x}) \in L^\infty(\Omega)$. It will be convenient to denote the vector of reaction kinetics by $\widehat{\mathbf{f}}(\mathbf{u}) := (\widehat{f}(\mathbf{u}), \widehat{g}(\mathbf{u}))^T$, where $\mathbf{u} := (u, v)^T$.

THEOREM 2.1. *Let $f \in \mathrm{Lip}(Q)$ and $u_0(\mathbf{x}), v_0(\mathbf{x}) \in L^\infty(\Omega)$. Then there exists a unique nonnegative classical solution of the fish-plankton system* (2.6a)–(2.6d) *for all $(\mathbf{x}, t) \in \Omega \times [0, \infty)$. Furthermore, if $d_1 = d_2$ and the initial data is chosen in the positively invariant region $\Sigma \in [0, \infty)^2$ given by* (2.4)*, then $(u, v) \in \Sigma$ for all $(\mathbf{x}, t) \in \Omega \times [0, \infty)$.*

Proof. Local existence of solutions is based on well-known semigroup theory (see, for example, Pazy [37] or Henry [20]). From Proposition 1 in [22] it follows immediately that (2.6a)–(2.6d) has a unique noncontinuable classical solution (u, v) for $(\mathbf{x}, t) \in \Omega \times [0, T_{max})$. Moreover, if $T_{max} < \infty$, then

$$\lim_{t \uparrow T_{max}} \sup_{\mathbf{x} \in \Omega} \{|u(\mathbf{x}, t)| + |v(\mathbf{x}, t)|\} = \infty. \tag{2.7}$$

To prove the nonnegativity of solutions observe that the reaction kinetics satisfy

$$\widehat{f}(0, v),\ \widehat{g}(u, 0) \geq 0 \quad \forall u, v \geq 0,$$

and the initial data $(u_0(\mathbf{x}), v_0(\mathbf{x}))$ is in $[0, \infty)^2$ for all (or almost every) $\mathbf{x} \in \Omega$. Thus by a maximum principle (see, e.g., [42, Lemma 14.20]) the solution $(u(\mathbf{x}, t), v(\mathbf{x}, t))$ lies in $[0, \infty)^2$ for all $\mathbf{x} \in \Omega$ and for all $t > 0$ for which the solution of (2.6a)–(2.6d) exists. In other words $[0, \infty)^2$ is positively invariant for the system. Proving global existence of solutions from local existence in the equal diffusion case is straightforward. The invariant region yields an L^∞ a priori bound [42] which contradicts nonglobal existence as solutions either exist for all time or blow up in the sup-norm in finite time (see (2.7)) [4]. The proof of global existence results in the distinct diffusion coefficient case requires additional theory as the only admissible invariant regions are products of intervals [42]. We apply the theoretical framework of Morgan [35] to prove global

existence and uniqueness, which involves verifying "intermediate sum" conditions and polynomial growth conditions on the kinetics.

We first define a so-called Lyapunov-type function given by

$$H(\mathbf{u}) := h_1(u) + h_2(v), \quad \text{where} \quad h_1(u) = u, \quad h_2(v) = v.$$

Then with $a_{11} = a_{22} = a_{21} = \mu = q = 1$, $K_2 = K_4 = K_6 = 0$, $K_1 = K_5 = r$, $K_3 = \max\{r, a/b\}$ the following conditions are easily verified for all $\mathbf{u} \in [0,\infty)^2$, corresponding to conditions (H1), (H3), (H4)(i), (H5), and (H6) in [35], respectively:

$$\begin{gathered} a_{11}h_1^{'}(u)\widehat{f}(\mathbf{u}) \leq K_1(H(\mathbf{u}))^{\mu} + K_2, \\ a_{21}h_1^{'}(u)\widehat{f}(\mathbf{u}) + a_{22}h_2^{'}(v)\widehat{g}(\mathbf{u}) \leq K_1(H(\mathbf{u}))^{\mu} + K_2, \\ h_1^{'}(u)\widehat{f}(\mathbf{u}),\ h_2^{'}(v)\widehat{g}(\mathbf{u}) \leq K_3(H(\mathbf{u}))^{q} + K_4, \\ \nabla H(\mathbf{u}) \cdot \widehat{\mathbf{f}}(\mathbf{u}) \leq K_5 H(\mathbf{u}) + K_6. \end{gathered}$$

Thus Theorems 3.2 and 2.2 in [35] hold, which implies $T_{max} = \infty$, i.e., we have global existence of nonnegative, classical solutions. ∎

3. The optimal control problem. Let Ω be a bounded, open subset of $\mathbb{R}^2$ with smooth boundary $\partial\Omega$ and let the set of all possible target densities $L^2_{\text{loc}}(0,T;L^2(\Omega))$ be denoted by $\mathcal{T}_{ad}$. There are no particular requirements on the target densities $(\bar{u},\bar{v})$ other than the fact that the cost functional (1.3) must be bounded. The target densities need not to be solutions of (2.6a)–(2.6d).

Let $\mathcal{U}_{ad}$ be the set of admissible controls

$$\mathcal{U}_{ad} = \{f \in \text{Lip}(Q);\ 0 \leq f(x,t)\ \forall (x,t) \in Q\}.$$

Given $T > 0$, $u_0, v_0 \in H^1(\Omega) \cap L^\infty(\Omega)$, and $(\bar{u},\bar{v}) \in \mathcal{T}_{ad}$, then (u,v,f) is said to be an admissible element if $u, v \in L^2(0,T;H^1(\Omega))$, $f \in \mathcal{U}_{ad}$, the functional $J(u,v,f)$ is bounded, and (u,v,f) satisfies (2.6a)–(2.6d). Let $\mathcal{A}_{ad}$ be the set of admissible states and controls. With this notation, the formulation of the optimal control problem is given by the following:

$$\text{(P)} \qquad \begin{array}{l} \text{Given } T > 0,\ u_0, v_0 \in H^1(\Omega) \cap L^\infty(\Omega), \text{ and } (\bar{u},\bar{v}) \in \mathcal{T}_{ad}, \\ \text{find } (u^*,v^*,f^*) \in \mathcal{A}_{ad} \text{ such that } J(u^*,v^*,f^*) \leq J(u,v,f) \\ \forall (u,v,f) \in \mathcal{A}_{ad}. \end{array}$$

THEOREM 3.1. *Given $u_0, v_0 \in H^1(\Omega) \cap L^\infty(\Omega)$ and $(\bar{u},\bar{v}) \in \mathcal{T}_{ad}$, there exists a solution (u^*,v^*,f^*) of the optimal control problem* (P).

Proof. The admissible set is bounded and nonempty, e.g., the system (2.6a)–(2.6d) has a solution for $f = 0$ (see [14]). Let $\{(u_n,v_n,f_n)\}_n$ be a minimizing sequence in $\mathcal{A}_{ad}$. The sequence $\{\frac{\partial f_n}{\partial t}\}$ is bounded in $L^2(0,T;L^2(\Omega))$, and therefore with a subsequence, again indexed by n, we have

$$f_n \to f^* \qquad \text{weakly in } H^1([0,T];L^2(\Omega)).$$

On the other hand, from (2.6a)–(2.6b) we have

$$\|u_n(t)\|^2_{L^2(\Omega)} + \|v_n(t)\|^2_{L^2(\Omega)} + \int_0^t \left(d_1\|\nabla u_n(t)\|^2_{L^2(\Omega)} + d_2\|\nabla v_n(t)\|^2_{L^2(\Omega)}\right) dt \leq C$$

for all $t \in [0,T]$. Here and in what follows we denote by C a positive constant independent of u, v, f, and n. Next we multiply system (2.6a)–(2.6b) by $(-\Delta u_n, -\Delta v_n)$ and integrate over $(0,t)$ to get, after some calculation involving Gronwall's lemma,

$$\|\nabla u_n(t)\|^2_{L^2(\Omega)} + \|\nabla v_n(t)\|^2_{L^2(\Omega)} + \int_0^t \left(d_1 \|\Delta u_n(s)\|^2_{L^2(\Omega)} + d_2 \|\Delta v_n(s)\|^2_{L^2(\Omega)} \right) dt \le C,$$

$$\left\| \frac{d}{dt} u_n(t) \right\|^2_{L^2(\Omega)} + \left\| \frac{d}{dt} v_n(t) \right\|^2_{L^2(\Omega)} \le C \quad \forall t \in [0,T].$$

Using the classical Lions–Aubin compactness lemma [27] and Ascoli's theorem [10], and noting that the injection of $H^1(\Omega)$ into $L^2(\Omega)$ is compact, we infer that $\{u_n\}, \{v_n\}$ are compact in $L^2(0,T;H^1(\Omega)) \cap C([0,T];L^2(\Omega))$. Thus selecting further subsequences, if necessary, we have

$$\begin{aligned}
&u_n \to u^*,\ v_n \to v^* && \text{strongly in } L^2(0,T;H^1(\Omega)) \cap C([0,T];L^2(\Omega)),\\
&u_n \to u^*,\ v_n \to v^* && \text{weakly in } L^2(0,T;H^2(\Omega)),\\
&\frac{d}{dt}u_n \to \frac{d}{dt}u^*,\ \frac{d}{dt}v_n \to \frac{d}{dt}v^* && \text{weakly in } L^2(0,T;L^2(\Omega)).
\end{aligned}$$

Moreover we have that

$$\begin{aligned}
&u_n(1-u_n) \to u^*(1-u^*) && \text{strongly in } L^2(0,T;L^2(\Omega)),\\
&\frac{u_n v_n}{1+bu_n} \to \frac{u^* v^*}{1+bu^*} && \text{strongly in } L^2(0,T;L^2(\Omega)),\\
&f_n \frac{v_n^2}{1+g^2 v_n^2} \to f^* \frac{{v^*}^2}{1+g^2 {v^*}^2} && \text{weakly in } L^2(0,T;L^2(\Omega)).
\end{aligned}$$

Letting n go to ∞ we see that (u^*, v^*, f^*) satisfies the system (2.6a)–(2.6d) and $J(u^*, v^*, f^*) = \inf_{\mathcal{A}_{ad}} J$. ∎

3.1. First-order necessary conditions. We show that the optimal solution must satisfy the first-order necessary condition associated with the optimal control problem (P).

We introduce the tangential (contingent) cone to $\mathcal{A}_{ad}$ at $(u,v,f) \in \mathcal{A}_{ad}$:

$$\begin{aligned}
\operatorname{Tan} \mathcal{A}_{ad}(u,v,f) = \Bigg\{ & (y,z,h) \,|\, y,z \in L^2(Q), h \in \operatorname{Tan} \mathcal{U}_{ad}(f) \quad \text{and}\\
& \frac{\partial y}{\partial t} = d_1 \Delta y + ry(1-2u) - \frac{av}{(1+bu)^2} y - \frac{au}{1+bu} z \text{ in } Q,\\
& \frac{\partial z}{\partial t} = d_2 \Delta z + \frac{av}{(1+bu)^2} y + \left(\frac{au}{1+bu} - m - \frac{2gvf}{(1+g^2v^2)^2} \right) z - \frac{gv^2 h}{1+g^2v^2} \text{ in } Q,\\
& \frac{\partial y}{\partial \nu} = \frac{\partial z}{\partial \nu} = 0 \qquad \text{on } \partial\Omega \times (0,T),\\
& y(x,0) = z(x,0) = 0 \qquad \text{in } \Omega \Bigg\}.
\end{aligned} \tag{3.1}$$

Recall that if

$$J(u^*, v^*, f^*) = \inf_{(u,v,f) \in \mathcal{A}_{ad}} J(u,v,f)$$

and the functional $J(u,v,f)$ is Gâteaux differentiable, then necessarily

$$\partial J(u^*,v^*,f^*)(y,z,h) \geq 0 \quad \forall (y,z,h) \in \operatorname{Tan} \mathcal{A}_{ad}(u^*,v^*,f^*), \tag{3.2}$$

where $\partial J(u^*,v^*,f^*)$ denotes the Gâteaux derivative of J at $(u^*,v^*,f^*) \in \mathcal{A}_{ad}$. Applying the optimum principle given by (3.2), it follows that

$$\int_Q \left[(u^*-\bar{u})y + (v^*-\bar{v})z + \alpha \frac{\partial f^*}{\partial t}\frac{\partial h}{\partial t} \right] dxdt \geq 0 \quad \forall (y,z,h) \in \operatorname{Tan} \mathcal{A}_{ad}(u^*,v^*,f^*). \tag{3.3}$$

THEOREM 3.2. *Let $u_0, v_0 \in H^1(\Omega) \cap L^\infty(\Omega)$. The mapping $(u,v) = (u(f), v(f))$ from $\mathcal{U}_{ad}$ to $L^2(0,T;H^1(\Omega))$, defined as the solution of the system* (2.6a)–(2.6d)*, has a Gâteaux derivative $(D(u,v)/Df)\cdot h$ in every direction $h \in \operatorname{Tan}\mathcal{U}_{ad}(f)$. Furthermore, $(y(h),z(h)) = (D(u,v)/Df)\cdot h$ is the classical solution of problem* (3.1).

Proof. Let $h \in \operatorname{Lip}(Q)$ be such that $f + \lambda h \in \mathcal{U}_{ad}$ for all $\lambda \in \mathbb{R}$ sufficiently small and let (u_f, v_f) and $(u_{f+\lambda h}, v_{f+\lambda h})$ denote the solutions of (2.6a)–(2.6d) with fish predation rates f and $f+\lambda h$, respectively. To simplify the notation in this proof we use $\|\cdot\|$ to denote $\|\cdot\|_{L^2(\Omega)}$.

We need to prove that

$$\lim_{\lambda \to 0} \frac{1}{\lambda} \left\{ \|u_{f+\lambda h} - u_f - \lambda y(h)\|_{L^2(0,T;H^1(\Omega))} + \|v_{f+\lambda h} - v_f - \lambda z(h)\|_{L^2(0,T;H^1(\Omega))} \right\} = 0.$$

We set $\widetilde{u} = u_{f+\lambda h} - u_f - \lambda y(h)$, $\widetilde{v} = v_{f+\lambda h} - v_f - \lambda z(h)$ so that $(\widetilde{u}, \widetilde{v})$ satisfies the system

$$\frac{\partial \widetilde{u}}{\partial t} = d_1 \Delta \widetilde{u} + r\widetilde{u} - r\Lambda_1 - a\Lambda_2 \qquad \text{in } Q, \tag{3.4a}$$

$$\frac{\partial \widetilde{v}}{\partial t} = d_2 \Delta \widetilde{v} + a\Lambda_2 - m\widetilde{v} - g\Lambda_3 \qquad \text{in } Q, \tag{3.4b}$$

$$\frac{\partial \widetilde{u}}{\partial \nu} = \frac{\partial \widetilde{v}}{\partial \nu} = 0 \qquad \text{on } \partial\Omega \times (0,T), \tag{3.4c}$$

$$\widetilde{u}(x,0) = \widetilde{v}(x,0) = 0 \qquad \text{in } \Omega, \tag{3.4d}$$

where

$$\Lambda_1 = u_{f+\lambda h}^2 - u_f^2 - 2\lambda y u_f, \quad \Lambda_2 = \frac{u_{f+\lambda h}v_{f+\lambda h}}{1+bu_{f+\lambda h}} - \frac{u_f v_f}{1+bu_f} - \lambda \frac{v_f y}{(1+bu_f)^2} - \lambda \frac{u_f z}{1+bu_f},$$

$$\Lambda_3 = (f+\lambda h)\frac{v_{f+\lambda h}^2}{1+g^2 v_{f+\lambda h}^2} - f\frac{v_f^2}{1+g^2v_f^2} - \lambda \frac{2v_f f}{(1+g^2v_f^2)^2} z - \lambda \frac{v_f^2 h}{1+g^2v_f^2}.$$

Now we multiply (3.4a)–(3.4b) by $\widetilde{u}$ and $\widetilde{v}$, respectively, and get

$$\frac{d}{dt}\left(\|\widetilde{u}\|^2 + \|\widetilde{v}\|^2\right) + \left(\|\nabla\widetilde{u}\|^2 + \|\nabla\widetilde{v}\|^2\right) \leq C\left(\|\widetilde{u}\|^2 + \|\widetilde{v}\|^2 + \|\Lambda_1\|^2 + \|\Lambda_2\|^2 + \|\Lambda_3\|^2\right).$$

We denote $\widehat{u} = u_{f+\lambda h} - u_f$, $\widehat{v} = v_{f+\lambda h} - v_f$ and use the fact that (u_f, v_f) is a classical solution to (2.6a)–(2.6d), yielding

$$\int_\Omega \Lambda_1^2\, dx = \int_\Omega \left(\widehat{u}^2 + 2u_f\widetilde{u}\right)^2 dx \leq 2\|\widehat{u}\|_{L^4(\Omega)}^4 + C\|\widetilde{u}\|^2,$$

$$\int_\Omega \Lambda_2^2\, dx \leq C\int_\Omega \left(\widehat{u}^4 + \widehat{v}^4 + \widetilde{u}^2 + \widetilde{v}^2\right) dx,$$

$$\int_\Omega \Lambda_3^2\, dx \leq C\int_\Omega \left(\widehat{v}^4 + \widehat{v}^6 + \widetilde{v}^2 + \lambda^2 h^2 \widehat{v}^2\right) dx \leq C\int_\Omega \left(\widehat{v}^4 + \widetilde{v}^2 + \lambda^4\right) dx.$$

Hence from the above estimates and the Gronwall inequality, this leads to

$$\|\widetilde{u}(t)\|^2 + \|\widetilde{v}(t)\|^2 + \int_0^T \left(\|\nabla\widetilde{u}(s)\|^2 + \|\nabla\widetilde{v}(s)\|^2\right) ds \\ \leq C\left(\lambda^4 + \int_0^T (\|\widehat{u}\|^2\|\widehat{u}\|^2_{H^1(\Omega)} + \|\widehat{v}\|^2\|\widehat{v}\|^2_{H^1(\Omega)})\, dt\right) \quad \forall t \in [0,T].$$

To estimate the norm of $\widehat{u}, \widehat{v}$ in $L^2(0,T;H^1(\Omega))$ we note that

$$\frac{\partial\widehat{u}}{\partial t} = d_1\Delta\widehat{u} + r\widehat{u} - r\widehat{u}(\widehat{u}+2u_f) - a\left(\frac{\widehat{v}(\widehat{u}+u_f)}{1+b(\widehat{u}+u_f)} - \frac{\widehat{u}v_f}{(1+bu_f)[1+b(\widehat{u}+u_f)]}\right),$$

$$\frac{\partial\widehat{v}}{\partial t} = d_2\Delta\widehat{v} + a\left(\frac{\widehat{v}(\widehat{u}+u_f)}{1+b(\widehat{u}+u_f)} - \frac{\widehat{u}v_f}{(1+bu_f)[1+b(\widehat{u}+u_f)]}\right) - m\widehat{v} \\ + g\left(f\frac{\widehat{v}(\widehat{v}+2v_f)}{(1+g^2v_f^2)(1+g^2(\widehat{v}+v_f)^2)} + \lambda h\frac{(\widehat{v}+v_f)^2}{1+g^2(\widehat{v}+v_f)^2}\right),$$

$$\frac{\partial\widehat{u}}{\partial\nu} = \frac{\partial\widehat{v}}{\partial\nu} = 0 \qquad \text{on } \partial\Omega\times(0,T),$$

$$\widehat{u}(x,0) = \widehat{v}(x,0) = 0 \qquad \text{in } \Omega.$$

We then easily obtain

$$\|\widehat{u}(t)\|^2 + \|\widehat{v}(t)\|^2 + \int_0^t \left(\|\nabla\widehat{u}(s)\|^2 + \|\nabla\widehat{v}(s)\|^2\right) ds \leq C\lambda^2\int_0^t \|h(s)\|^2 ds \quad \forall t\in[0,T].$$

From the last two results we obtain the estimate

$$\int_0^T \left(\|\widetilde{u}\|^2_{H^1(\Omega)} + \|\widetilde{v}\|^2_{H^1(\Omega)}\right) dt \leq C\lambda^4,$$

from which our claim follows. The proof of the regularity of y and z is similar to the proof of Theorem 2.1. □

The Gâteaux derivative gives useful information about the sensitivity of the system (2.6a)–(2.6d) at a particular point (u,v) in a particular direction h, but complete information requires the solution of (3.1) for every direction h. Fortunately, in order to minimize the functional we need only an integral over all these directions, which can be obtained by solving a single adjoint equation.

THEOREM 3.3. *Let $u_0, v_0 \in H^1(\Omega)\cap L^\infty(\Omega)$ and $(\bar{u},\bar{v}) \in \mathcal{T}_{ad}$ be given. If (u^*,v^*,f^*) is an optimal solution for* (P)*, then we have*

$$\int_Q \left(\frac{gv^{*2}k}{1+g^2v^{*2}}(f-f^*) + \alpha\frac{\partial f^*}{\partial t}\left(\frac{\partial f}{\partial t} - \frac{\partial f^*}{\partial t}\right)\right) dxdt \geq 0 \quad \forall f\in\mathcal{U}_{ad}, \tag{3.5}$$

where (p,k) is the unique classical solution of the adjoint equation

(3.6a)

$$\frac{\partial p}{\partial t} + d_1\Delta p + rp(1-2u^*) - a\frac{v^*}{(1+bu^*)^2}p + a\frac{v^*}{(1+bu^*)^2}k = u^* - \bar{u} \text{ in } Q,$$

$$\frac{\partial k}{\partial t} + d_2\Delta k + a\frac{u^*}{1+bu^*}k - mk - 2\frac{gf^*v^*}{(1+g^2v^{*2})^2}k - a\frac{u^*}{1+bu^*}p = v^* - \bar{v} \text{ in } Q,$$

$$\frac{\partial p}{\partial\nu} = \frac{\partial k}{\partial\nu} = 0 \qquad \text{on } \partial\Omega\times(0,T),$$

(3.6b)
$$p(x,T) = k(x,T) = 0 \qquad in\ \Omega.$$

Proof. Let (u^*, v^*, f^*) be an optimal solution. From the optimality condition (3.3) and (3.1) and from (3.6a)–(3.6b), we have after integration by parts

$$\begin{aligned}
0 &\leq \int_Q \left[(u^* - \bar{u})y + (v^* - \bar{v})z + \alpha \frac{\partial f^*}{\partial t}\frac{\partial h}{\partial t}\right] dxdt \\
&= \int_Q \left\{ y\left[\frac{\partial p}{\partial t} + d_1\Delta p + rp(1-2u^*) - a\frac{v^*}{(1+bu^*)^2}p + a\frac{v^*}{(1+bu^*)^2}k\right]\right. \\
&\left. + z\left(\frac{\partial k}{\partial t} + d_2\Delta k + a\frac{u^*}{1+bu^*}k - mk - \frac{2gf^*v^*}{(1+g^2{v^*}^2)^2}k - a\frac{pu^*}{1+bu^*}\right) + \alpha\frac{\partial f^*}{\partial t}\frac{\partial h}{\partial t}\right\} dxdt \\
&= \int_Q p\left[-\frac{\partial y}{\partial t} + d_1\Delta y + ry(1-2u^*) - a\frac{v^*}{(1+bu^*)^2}y - a\frac{u^*}{1+bu^*}z\right] dxdt + \int_\Omega yp\Big|_0^T dx \\
&+ \int_Q k\left[-\frac{\partial z}{\partial t} + d_2\Delta z - \frac{2gf^*v^*}{(1+g^2{v^*}^2)^2}z + a\frac{u^*}{1+bu^*}z - mz + a\frac{v^*}{(1+bu^*)^2}y\right] dxdt \\
&+ \int_\Omega zk\Big|_0^T dx + \alpha\int_\Omega \frac{\partial f^*}{\partial t}\frac{\partial h}{\partial t}\, dxdt \\
&= \int_Q k\frac{g{v^*}^2}{1+g^2{v^*}^2}h\, dxdt + \alpha\int_\Omega \frac{\partial f^*}{\partial t}\frac{\partial h}{\partial t}\, dxdt,
\end{aligned}$$

and therefore (3.5). ∎

4. Numerical results. To illustrate the theoretical results of the optimal control problem in the previous section we present results of numerical experiments in two space dimensions. The state equations and adjoint equations were solved using a "lumped mass," semi-implicit (in time) Galerkin finite element method with piecewise linear continuous basis functions. We showed previously that this approach was highly effective in solving the forward-in-time equations of a similar predator-prey system [15].

The control was updated using a variable-step gradient algorithm based on the fully discrete optimality condition (see (4.4) below). At each iteration of the gradient algorithm the method requires the sequential solution of the discrete state and adjoint equations (see (4.1) and (4.3) below). In practice one cannot solve these systems simultaneously. The discrete state equations are solved by marching forward in time starting from an initial condition, while the discrete adjoint equations are solved by marching backward in time (from T) starting from a terminal condition. For further details, see [16].

We employed a (uniform) right-angled triangulation Ω^h of the square $\Omega = [0, L] \times [0, L]$, with space steps h, and numerically solved the optimal control problem up to time T with uniform time steps Δt. We introduce S^h, the standard finite element space

$$S^h := \{v \in C(\overline{\Omega}) \ :\ v|_\tau \quad \text{is linear } \forall \tau \in \Omega^h\} \subset H^1(\Omega).$$

Let $\{x_i\}_{i=0}^J$ be the set of nodes of the triangulation. We introduce $\pi^h : C(\overline{\Omega}) \mapsto S^h$, the Lagrange interpolation operator, such that $\pi^h v(x_j) = v(x_j)$ for all $j = 0, \ldots, J$. In order to formulate our finite element approximation of the reaction-diffusion system we define a discrete L^2 inner product on $C(\overline{\Omega})$ given by $(u, v)^h := \int_\Omega \pi^h(u(x)v(x))\, dx$,

which approximates the usual L^2 inner product (u, v). Given $\mathbf{f}_h \in \mathrm{Lip}(\Omega)$ and $u_0, v_0 \in H^1(\Omega) \cap L^\infty(\Omega)$, $(\mathbf{u}_h, \mathbf{v}_h)$ is a solution of the fully discrete, semi-implicit (in time) fish-plankton system if $u_h^{(n)}, v_h^{(n)} \in S^h$ satisfies the system

$$
\begin{aligned}
&\frac{1}{\Delta t}\left(u_h^{(n)} - u_h^{(n-1)}, \chi_h\right)^h + d_1\left(\nabla u_h^{(n)}, \nabla \chi_h\right) \\
&= \left(r u_h^{(n)}(1 - |u_h^{(n-1)}|) - \frac{a u_h^{(n-1)} v_h^{(n)}}{1 + b|u_h^{(n-1)}|}, \chi_h\right)^h, \\
&\frac{1}{\Delta t}\left(v_h^{(n)} - v_h^{(n-1)}, \chi_h\right)^h + d_2\left(\nabla v_h^{(n)}, \nabla \chi_h\right) \\
&= \left(\frac{a u_h^{(n-1)} v_h^{(n)}}{1 + b|u_h^{(n-1)}|} - m v_h^{(n)} - \frac{g f_h^{(n)} v_h^{(n-1)2}}{1 + g^2 v_h^{(n-1)2}}, \chi_h\right)^h
\end{aligned}
\tag{4.1}
$$

for all $\chi_h \in S^h$, $n = 1, 2, \ldots, N$, with initial densities $u_h^{(0)} = \pi^h u_0(x)$, $v_h^{(0)} = \pi^h v_0(x)$. The discrete cost functional used in the optimal control problem is given by

$$
\mathcal{J}_h^N(\mathbf{u}_h, \mathbf{v}_h, \mathbf{f}_h) = \frac{\Delta t}{2}\sum_{n=1}^{N}\left(\|u_h^{(n)} - \bar{u}^{(n)}\|^2 + \|v_h^{(n)} - \bar{v}^{(n)}\|^2\right) + \frac{\alpha}{2\Delta t}\sum_{n=1}^{N}\|f_h^{(n)} - f_h^{(n-1)}\|^2. \tag{4.2}
$$

Thus we can now formulate the fully discrete optimal control problem as follows:

($\mathrm{P}^{h,\Delta t}$) Given $\Delta t = T/N$, $h = L/J$, $u_0, v_0 \in H^1(\Omega) \cap L^\infty(\Omega)$ and $(\bar{u}, \bar{v}) \in \mathcal{T}_{ad}$, find $(\mathbf{u}_h, \mathbf{v}_h, \mathbf{f}_h) \in S^h \times S^h \times S^h$ such that (4.1) is satisfied for $n = 1, 2, \ldots, N$ and the cost functional (4.2) is minimized.

To complete the fully discrete optimality system we also need the following fully discrete adjoint system: The adjoint functions $p_h^{(n)}, k_h^{(n)} \in S^h$ satisfy

$$
\begin{aligned}
&-\frac{1}{\Delta t}\left(p_h^{(n)} - p_h^{(n-1)}, \chi_h\right)^h + d_1\left(\nabla p_h^{(n-1)}, \nabla \chi_h\right) \\
&= \left(r(1 - 2|u_h^{(n)}|)p_h^{(n-1)} - a\frac{v_h^{(n)}}{(1+b|u_h^{(n)}|^2)^2}p_h^{(n-1)} + a\frac{v_h^{(n)}}{1+b|u_h^{(n)}|^2}k_h^{(n-1)} - u_h^{(n)} - \bar{u}^{(n)}, \chi_h\right)^h \\
&-\frac{1}{\Delta t}\left(k_h^{(n)} - k_h^{(n-1)}, \chi_h\right)^h + d_2\left(\nabla k_h^{(n-1)}, \nabla \chi_h\right) \\
&= \left(a\frac{u_h^{(n)}}{1+b|u_h^{(n)}|^2}p_h^{(n-1)} + a\frac{u_h^{(n)}}{1+b|u_h^{(n)}|^2}k_h^{(n-1)} - m k_h^{(n-1)} - g f_h^{(n)}\frac{2v_h^{(n)}}{1+g^2 v_h^{(n)2}} - v_h^{(n)} - \bar{v}^{(n)}, \chi_h\right)^h
\end{aligned}
\tag{4.3}
$$

for all $\chi_h \in S^h$, $n = 1, 2, \ldots, N$, with the terminal conditions $p_h^{(N)} = k_h^{(N)} = 0$. The fully discrete optimality condition is

$$
\begin{aligned}
0 \le\ &\Delta t \sum_{n=1}^{N-1}\left(\tilde{f}_h^{(n)}, g\frac{v_h^{(n)2}}{1 + g^2 v_h^{(n)2}}k_h^{(n-1)} - \alpha\frac{f_h^{(n+1)} - 2f_h^{(n)} + f_h^{(n-1)}}{\Delta t}\right) \\
&- \frac{\alpha}{\Delta t}\left(\tilde{f}_h^{(0)}, f_h^{(1)} - f_h^{(0)}\right) + \left(\tilde{f}_h^{(n)}, \Delta t g\frac{v_h^{(N)2}}{1 + g^2 v_h^{(N)2}}k_h^{(n-1)} + \alpha\frac{f_h^{(N)} - f_h^{(N-1)}}{\Delta t}\right)
\end{aligned}
\tag{4.4}
$$

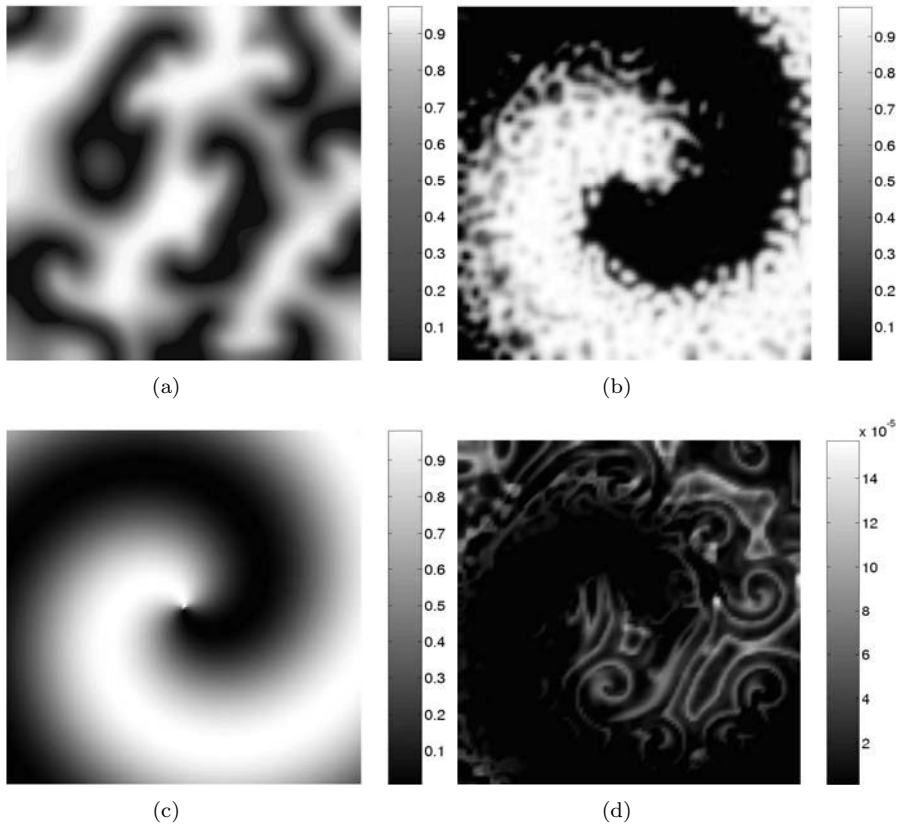


FIG. 4.1. *Uncontrolled* (a)*, optimally controlled* (b)*, target* (c)*, and the control* f (d) *for phytoplankton densities* u *at time* $T = 100$ *for* (1.2)*. Parameter values:* $d_1 = d_2 = 0.05$*,* $r = 1$*,* $a = b = 20$*,* $m = 0.8$*,* $g = 10$*,* $\alpha = 10^{-5}$*. For details of initial data, see text.*

for all $\tilde{\mathbf{f}}_h \in \mathrm{Tan}\mathcal{U}_{ad}(\mathbf{f}_h)$.

In the numerical experiments we chose the domain $\Omega = [0, 100] \times [0, 100]$ and initial conditions of the optimal control problem to be the freely evolving ($f = 0$) system (1.2) at time $t = 1000$, with parameter values and initial data corresponding to Figure 1.2(b) (see caption). The target functions are one-armed Archimedean spirals with period of rotation equal to 20 and are given by

$$\bar{u}(R, \theta, t) = 0.495 \cos(\theta + R/12 + \pi t/10) + 0.495,$$
$$\bar{v}(R, \theta, t) = 0.21 \sin(\theta + R/12 + \pi t/10) + 0.22,$$

where $R := \sqrt{(x-50)^2 + (y-50)^2}$ and $\theta := \arctan[(y-50)/(x-50)]$. The numerical results for the optimal control of phytoplankton densities u are shown in Figure 4.1 at time $T = 100$. A plot of the reduction in cost functional with increasing iteration count suggests the near optimality of the system (see [16]). The results show that at time $T = 100$ the controlled system is close to the desired state, and that we were successful in driving the system from a disordered state to an ordered one.

5. Conclusions. The mathematical formulation, analysis, and numerical solution of an optimal control problem for a nonlinear plankton-fish reaction-diffusion

system was presented. The model was discussed in the context of biomanipulation of eutrophic lakes. After considering the local dynamics of the system and proving the global existence and uniqueness of the classical solutions of the state equations, we presented the mathematical analysis of the plankton-fish optimal control problem. Numerical solutions were obtained with the aid of a semi-implicit Galerkin finite element method with piecewise linear continuous basis functions. The numerical results illustrate the ability of a variable step-size gradient algorithm to drive the plankton dynamics from a chaotic regime to an (arbitrary) ordered state. The time taken to achieve the ordered distribution of phytoplankton in nondimensional units was $t = 100$, which with an assumed maximum growth rate for phytoplankton in eutrophic conditions of $R_0 = 0.5$ per day [39] gives the time taken to achieve this state to be $\tau = 200$ days. The theoretical results in this paper provide the basis for a numerical analysis of the optimal control problem [16]. Furthermore, our results can be generalized in numerous ways to include, for example, convection driven flows, forcing via nutrient inputs, and stochastic influences. We leave these tasks for future work.

REFERENCES

[1] S. Aniţa, *Optimal control of a nonlinear population dynamics with diffusion*, J. Math. Anal. Appl., 152 (1990), pp. 176–208.

[2] S. Aniţa, *A fractional step scheme for the optimal control of a nonlinear population dynamics with diffusion*, An. Ştiinţ. Univ. Al. I. Cuza Iaşi. Mat., 46 (2000), pp. 157–168.

[3] N. Apreutesei, *An optimal control problem of Lotka-Volterra system with diffusion*, Bul. Inst. Politeh. Iaşi. Secţ. I. Mat. Mec. Teor. Fiz., 44 (1998), pp. 31–41.

[4] J. Ball, *Remarks on blow-up and nonexistence theorems for nonlinear evolution equations*, Quart. J. Math. Oxford, 2 (1977), pp. 473–486.

[5] A. Borzì and R. Griesse, *Experiences with a space-time multigrid method for the optimal control of a chemical turbulence model*, Internat. J. Numer. Methods Fluids, 47 (2005), pp. 879–885.

[6] S. Carpenter, D. Christensen, J. Cole, K. Cottingham, X. He, J. Hodgson, J. Kitchell, S. Knight, M. Pace, D. Post, D. Schindler, and N. Voichick, *Biological control of eutrophication in lakes*, Environ. Sci. Technol., 29 (1995), pp. 784–786.

[7] S. Cerrai, *Optimal control problems for stochastic reaction-diffusion systems with non-Lipschitz coefficients*, SIAM J. Control Optim., 39 (2001), pp. 1779–1816.

[8] R. Z. Chen, D. S. Zhang, and J. Q. Li, *Existence and uniqueness of solution and boundary control for population systems with spatial diffusion*, J. Systems Sci. Math. Sci., 22 (2002), pp. 1–13.

[9] D. S. Cohen, J. C. Neu, and R. R. Rosales, *Rotating spiral wave solutions of reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 536–547.

[10] J. Dieudonné, *Foundations of Modern Analysis*, Pure Appl. Math., Academic Press, New York, 1960.

[11] R. Drenner and K. Hambright, *Biomanipulation of fish assemblages as a lake restoration technique*, Arch. Hydrobiol., 146 (1999), pp. 129–165.

[12] K. Fister and C. M. McCarthy, *Optimal control of a chemotaxis system*, Quart. Appl. Math., 61 (2003), pp. 193–211.

[13] R. Fister, *Optimal control of harvesting in a predator-prey parabolic system*, Houston J. Math., 23 (1997), pp. 341–355.

[14] M. Garvie and C. Trenchea, *Analysis of two generic spatially extended predator-prey models*, Nonlinear Anal. Real World Appl., submitted.

[15] M. Garvie and C. Trenchea, *Finite element approximation of spatially extended predator-prey interactions with the Holling type* II *functional response*, Numer. Math., to appear.

[16] M. Garvie and C. Trenchea, *Numerical Analysis of a Nutrient–Plankton Optimal Control Problem*, in preparation, 2006.

[17] W. Gentleman, A. Leising, B. Frost, S. Strom, and J. Murray, *Functional responses for zooplankton feeding on multiple resources: A review of assumptions and biological dynamics*, Deep-Sea Res. Pt. II, 50 (2003), pp. 2847–2875.

[18] L.-A. Hansson, M. Gyllström, A. Stahl-Delbanco, and M. Svensson, *Responses to fish predation and nutrients by plankton at different levels of taxonomic resolution*, Freshwater Biol., 49 (2004), pp. 1538–1550.

[19] J. He and R. Glowinski, *Neumann control of unstable parabolic systems: Numerical approach*, J. Optim. Theory Appl., 96 (1998), pp. 1–55.

[20] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, New York, 1981.

[21] J. Hietala, K. Vakkilainen, and T. Kairesalo, *Community resistance and change to nutrient enrichment and fish manipulation in a vegetated lake littoral*, Freshwater Biol., 49 (2004), pp. 1525–1537.

[22] S. L. Hollis, R. H. Martin, Jr., and M. Pierre, *Global existence and boundedness in reaction-diffusion systems*, SIAM J. Math. Anal., 18 (1987), pp. 744–761.

[23] J. Horppila, H. Peltonen, T. Malinen, E. Luokkanen, and T. Kairesalo, *Top-down or bottom-up effects by fish: Issues of concern in biomanipulation of lakes*, Restor. Ecol., 6 (1998), pp. 20–28.

[24] I. Hrinca, *An optimal control problem for the Lotka-Volterra system with diffusion*, Panamer. Math. J., 12 (2002), pp. 23–46.

[25] A. Kozak and R. Goldyn, *Zooplankton versus phyto- and bacterioplankton in the Maltański Resevoir (Poland) during an extensive biomanipulation experiment*, J. Plankton Res., 26 (2004), pp. 37–48.

[26] S. Lenhart, M. Liang, and V. Protopopescu, *Optimal control of boundary habitat hostility for interacting species*, Math. Method. Appl. Sci., 22 (1999), pp. 1061–1077.

[27] J. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires*, Dunod Gauthier-Villars, Paris, 1969.

[28] M. Lund and J. Davis, *Seasonal dynamics of plankton communities and water chemistry in a eutrophic wetland (Lake Monger, Western Australia): Implications for biomanipulation*, Mar. Freshwater Res., 51 (2000), pp. 321–332.

[29] H. Malchow, *Motional instablities in prey-predator systems*, J. Theoret. Biol., 204 (2000), pp. 639–647.

[30] H. Malchow, F. Hilker, and S. Petrovskii, *Noise and productivity dependence of spatio-temporal pattern formation in a prey-predator system*, Discrete Contin. Dyn. Syst. Ser. B, 4 (2004), pp. 705–711.

[31] H. Malchow, S. Petrovskii, and A. Medvinsky, *Pattern formation in models of plankton dynamics. A synthesis*, Oceanol. Acta, 24 (2001), pp. 479–487.

[32] H. Malchow, S. Petrovskii, and A. Medvinsky, *Numerical study of plankton-fish dynamics in a spatially structured and noisy environment*, Ecol. Model., 149 (2002), pp. 247–255.

[33] H. Malchow, B. Radtke, M. Kallache, A. Medvinsky, D. Tikhonov, and S. Petrovskii, *Spatio-temporal pattern formation in coupled models of plankton dynamics and fish school motion*, Nonlinear Anal.-Real., 1 (2000), pp. 53–67.

[34] A. B. Medvinsky, S. V. Petrovskii, I. A. Tikhonova, H. Malchow, and B.-L. Li, *Spatio-temporal complexity of plankton and fish dynamics*, SIAM Rev., 44 (2002), pp. 311–370.

[35] J. Morgan, *Global existence for semilinear parabolic systems*, SIAM J. Math. Anal., 20 (1989), pp. 1128–1144.

[36] M. Pascual, *Diffusion-induced chaos in a spatial predator-prey system*, Proc. Roy. Soc. London Ser. B, 251 (1993), pp. 1–7.

[37] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.

[38] D. Rejas, S. Declerck, J. Auwerkerken, P. Tak, and L. Meester, *Plankton dynamics in a tropical floodplain lake: Fish, nutrients, and the relative importance of bottom-up and top-down control*, Freshwater Biol., 50 (2005), pp. 52–69.

[39] M. Scheffer, *Fish and nutrients interplay determines algal biomass: A minimal model*, Oikos, 62 (1991), pp. 271–282.

[40] J. Sherratt, *Invading wave fronts and their oscillatory wakes are linked by a modulated travelling phase resetting wave*, Phys. D, 117 (1998), pp. 145–166.

[41] J. Sherratt, B. Eagan, and M. Lewis, *Oscillations and chaos behind predator-prey invasion: Mathematical artifact or ecological reality?*, Philos. Trans. Roy. Soc. London Ser. B, 352 (1997), pp. 21–38.

[42] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Grundlehren Math. Wiss. 258, Springer-Verlag, New York, 1983.

[43] T. Staib, *Control of a chemical reactor (reaction-diffusion equation)*, Z. Angew. Math. Mech., 76 (1996), pp. 681–682.

[44] I. Tátrai, K. Mátyás, J. Korponai, G. Paulovits, P. Pomogyi, and F. Pekár, *Manage-*

ment of fish communities and its impacts on the lower trophic levels in shallow ecosystems in Hungary, Hydrobiologia, 506–509 (2003), pp. 489–496.

[45] I. Tikhonova, B. Li, H. Malchow, and A. Medvinsky, *The impact of the phytoplankton growth rate on spatial and temporal dynamics of plankton communities in a heterogeneous environment*, Biofizika, 48 (2003), pp. 891–899.

[46] P. Tu, *Dynamical Systems: An Introduction with Applications in Economics and Biology*, Springer-Verlag, Berlin, 1994.

[47] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Texts Appl. Math. 2, Springer-Verlag, New York, 1990.

[48] A. Wissel and J. Benndorf, *Contrasting effects of the invertebrate predator* Chaoborus obscuripes *and planktivorous fish on plankton communities of a long term biomanipulation experiment*, Arch. Hydrobiol., 143 (1998), pp. 129–146.

[49] B. Wissel, K. Freier, B. Müller, J. Koop, and J. Benndorf, *Moderate planktivorous fish biomass stabilizes biomanipulation by suppressing large invertebrate predators of* Daphnia, Arch. Hydrobiol., 149 (2000), pp. 177–192.

[50] K. Wysujack and T. Mehner, *Comparison of losses of planktivorous fish by predation and seine-fishing in a lake undergoing long-term biomanipulation*, Freshwater Biol., 47 (2002), pp. 2425–2434.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 792–815

INEXACT CENTRAL PATH FOLLOWING ALGORITHMS FOR OPTIMAL CONTROL PROBLEMS*

MARTIN WEISER† AND PETER DEUFLHARD‡

Abstract. A new approach to the numerical solution of optimal control problems including control and state constraints is presented. Like hybrid methods, the approach aims at combining the advantages of direct and indirect methods. Unlike hybrid methods, however, our method is directly based on interior point concepts in function space—realized via an adaptive multilevel scheme applied to the complementarity formulation and numerical continuation along the central path. An adaptive stepsize control with respect to the duality gap parameter is worked out in the framework of affine invariant inexact Newton methods. Finally, the performance of our new type of algorithm is documented by a simple example within the range of our present theory, and by the successful treatment of the well-known intricate windshear problem outside this range.

Key words. numerical optimal control, interior point methods in function space, affine invariant Newton methods

AMS subject classifications. 65K10, 49M15, 90C48, 90C51

DOI. 10.1137/S0363012903396851

1. Introduction.

In the last decade, the numerical solution of optimal control problems has reached a high level of sophistication. Present methods are able to treat important classes of large-scale real-life problems in science and engineering. Two types of method are in common use:

- *direct methods*, mostly based on some *robust collocation* including an ad hoc parametrization of the controls (see Bock and Plitt [8] or, for more recent publications, the nice survey paper by Hager [23] or his more specialized work on a Runge–Kutta approach [22]);
- *indirect methods*, typically based on either *multiple shooting* techniques (see Bulirsch [9], Stoer and Bulirsch [34], Deuflhard [16, 17], and Bock [7]) or *adaptive collocation methods* (see Ascher, Christiansen, and Russell [2], Ascher and Bader [4], and Ascher, Mattheij, and Russell [3]).

Whenever the necessary Euler–Lagrange conditions give a sufficient description of the problem, then indirect methods lead to a provably optimal solution [9]. However, they require rather detailed a priori knowledge about the sequence of optimal subtrajectories and a significant amount of problem-specific analytical preparation. In contrast to that, direct methods may dispense of this severe constraint, but they have a tendency to lead to nonoptimal solutions now and then. For this reason, *hybrid methods* seem to be the state of the art (see von Stryk and Bulirsch [39] and Bulirsch et al. [12]): in a first step, some direct method, wherein the control variables are parameterized

*Received by the editors November 5, 2003; accepted for publication (in revised form) November 17, 2006; published electronically May 29, 2007. This research was supported by the DFG Research Center MATHEON "Mathematics for Key Technologies" in Berlin. A preprint of this paper appeared as ZIB-Report 01-12, 2001.

http://www.siam.org/journals/sicon/46-3/39685.html

†Corresponding author. Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), 14195 Berlin, Germany (weiser@zib.de). The work of this author was supported by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 273.

‡Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB) and Freie Universität Berlin (FU), 14195 Berlin, Germany (deuflhard@zib.de).

ad hoc, supplies a rough idea about the optimal subarc sequence; then, in the second step, an indirect method is employed to finally solve the problem to high accuracy.

The present paper advocates a *unified function space approach* realizing ideas of both direct and indirect methods in infinite dimension rather than in finite dimension—as opposed to the above hybrid methods. There are several possibilities of such an extension. In a recent monograph Pytlak [32] proposed an approach that he claims is a function space approach. However, in our wording above, that approach is of the "robust collocation" type, leaving a considerable gap between the presented theory and the rather heuristic algorithmic realization. A genuine function space SQP method has been analyzed by Alt and Malanowski [1]. The solution of the infinite-dimensional QP subproblems, however, is not addressed.

The present paper advocates a function space interior point (IP) approach realized as a *nested* reduction of mesh size and duality gap parameter. The simultaneous progress in mesh refinement and duality gap reduction eliminates the need for warm start capabilities, which IP methods are lacking. The extension of IP-type methods from finite to infinite dimension is not at all straightforward: after all, the concept of logarithmic barrier functions is no longer useful in the infinite-dimensional setting (cf. Jarre [26]). A first attempt in this direction has been made by Ito, Kelley, and Sachs [25] in 1995. An approach closer to our present suggestion, and, in fact, in the same spirit, was recently published by M. Ulbrich and S. Ulbrich in [36] as well as with coauthor Heinkenschloss in [37].

The present paper is based on the *complementarity* version of IP methods, including the *central path* concept, which carries over naturally to infinite dimension. In the opinion of the authors, forged in a PDE environment, a function space–oriented approach is a convenient means to exploit similarity of the solution on different discretization levels. In particular, it allows for a natural use of adaptive multilevel methods—a technique well established in PDE applications. Compared with the indirect methods, the benefit of such an approach is the avoidance of human preparatory work, as opposed to the PDE aspect, where typically a fast solution is the focus. Compared with the direct methods, adaptive multilevel techniques establish appropriate grids using a posteriori error estimates. Of course, mixtures of tools are possible and have been worked out, e.g., by Schulz [33]. It seems worth mentioning that the method we suggest herein differs clearly from his finite-dimensional multigrid techniques, where the adaptive refinement of the *control variables* is done in the *outer loop*, as opposed to our infinite-dimensional technique, where the refinement is performed in the *innermost loop*. Recent suggestions by Ulbrich [35] concerning a semismooth Newton method represent a true alternative to our method as suggested below. In certain situations, this method is equivalent to a primal-dual active set strategy. A comparison of these methods with IP methods in a finite-dimensional setting has been published by Bergounioux et al. [5].

The paper is organized as follows. In section 2, we motivate the central path in function space as the mathematical concept. In section 3, details of an algorithm based on this concept are worked out. The basic scheme is an adaptive Newton-continuation method along the central path, realized as an inexact continuation method (predictor) with an inexact Newton method (corrector). Affine invariant norms are used to control the iteration process towards the numerical solution (see Volkwein and Weiser [38], Potra [31], or the research monograph by Deuflhard [18]). Our computational approach actually *exploits function space* via an *adaptive multilevel refinement* of all variables, *including the controls*. Its present realization is done within the setting of collocation methods. An adaptive stepsize control along the central path is

worked out as the infinite-dimensional extension of finite-dimensional suggestions due to Deuflhard [17]. An intriguing feature of our suggested algorithm is that it requires only the solution of *linear* operator equations in each step. In section 4, we give two numerical examples, a first one, which is covered by our theoretical frame, and a second one, which exceeds this frame. The second one is a well-known intricate optimal control problem, the abort landing in the presence of windshear (cf. Miele et al. [30] and Bulirsch, Montrone, and Pesch [10, 11]). Even though our approach does not need any cumbersome analytic preparation, our numerical results are in full agreement with those obtained by multiple shooting [11].

2. Central path continuation in function space. On the time interval $\Omega = [0,1]$ we consider the optimal control problem

$$\min J(x) \quad \text{subject to} \quad \begin{aligned} c(x) &= 0 \quad \text{a.e.},\\ r(x) &= 0,\\ g(x) &\geq 0 \quad \text{a.e.} \end{aligned} \tag{2.1}$$

with a partitioning of the variable $x = (u, y) \in X = L_\infty^{n_u}(\Omega) \times (W_\infty^1)^{n_y}(\Omega)$ into controls and states, a Lagrange-type cost functional

$$J(x) = \int_0^1 \tilde{f}(x(t))\, dt\,,$$

ODEs with boundary conditions

$$c(x) = \begin{bmatrix} \bar{c}(x) \\ y(0) - y_0 \end{bmatrix}, \quad \bar{c}(x)(t) = \tilde{c}(x(t)) - \dot{y}(t)\,,$$
$$r(x) = \tilde{r}(y(1))$$

as equality constraints, and pointwise state and control constraints

$$g(x)(t) = \begin{bmatrix} \tilde{g}_u(u(t)) \\ \tilde{g}_y(y(t)) \end{bmatrix}.$$

For the whole paper we will restrict the discussion to the fixed time interval Ω and hence simplify the notation by omitting it from the function spaces. We assume all the functions $\tilde{f} : \mathbb{R}^{n_u} \times \mathbb{R}^{n_y} \to \mathbb{R}$, $\tilde{c} : \mathbb{R}^{n_u} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$, $\tilde{r} : \mathbb{R}^{n_y} \to \mathbb{R}^{n_r}$, $\tilde{g}_u : \mathbb{R}^{n_u} \to \mathbb{R}^{n_\eta^u}$, and $\tilde{g}_y : \mathbb{R}^{n_y} \to \mathbb{R}^{n_\eta^y}$ to be twice Lipschitz continuously differentiable on arbitrary bounded sets. This implies that also $J : X \to \mathbb{R}$, $c : X \to L_\infty^{n_y} \times \mathbb{R}^{n_y}$, $r : X \to \mathbb{R}^{n_r}$, and $g : X \to L_\infty^{n_\eta^u} \times (W_\infty^1)^{n_\eta^y}$ are twice Lipschitz continuously differentiable (see [40]).

The inequality constraints can also be written as $g(x) \in K$, where $K = \{z \in L_\infty^{n_\eta^u} \times (W_\infty^1)^{n_\eta^y} : z(t) \geq 0 \text{ a.e.}\}$ is the closed convex cone of nonnegative functions. Its dual cone is given by $K^+ = \{\zeta \in (L_\infty^{n_\eta^u} \times (W_\infty^1)^{n_\eta^y})^* : \langle \zeta, z \rangle \geq 0 \text{ for all } z \in K\}$.

Note that state constrained problems (i.e., $n_\eta^y > 0$) in general are notoriously difficult to solve, and that both analysis and numerics are usually much more demanding than for control constrained problems ($n_\eta^y = 0$). This extends to this article, where most of the theoretical results are valid only for control problems. State constrained problems are nevertheless included in the setting as far as possible because numerical experience suggests that the algorithm presented here performs reasonably well for such problems.

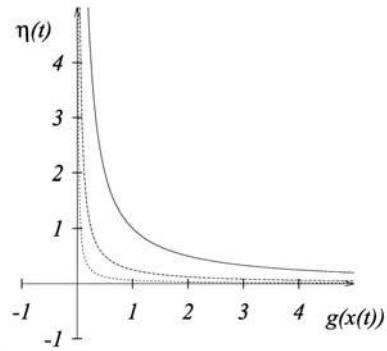


FIG. 2.1. *Zero level sets of the smoothed complementarity function* (2.8) *for* $\mu \in \{1, 1/4, 1/16\}$.

The aim of the IP method presented here is to approximate Kuhn–Tucker points x_*. These are feasible points characterized by the existence of Lagrange multipliers $\lambda_c \in (L_\infty^{n_y})^* \times \mathbb{R}^{n_y}$, $\lambda_r \in \mathbb{R}^{n_r}$, and $\eta \in K^+$, such that the following conditions are satisfied:

$$J'(x_*) - c'(x_*)^*\lambda_c - r'(x_*)^*\lambda_r - g'(x_*)^*\eta = 0 \in X^*, \tag{2.2}$$
$$c(x_*) = 0\,, \quad r(x_*) = 0,$$
$$g(x_*) \geq 0\,, \quad \eta \geq 0\,, \quad \langle \eta, g(x_*)\rangle = 0. \tag{2.3}$$

Under certain assumptions (see, e.g., [28, 29]) these conditions are necessary for x_* to be a local solution of (2.1), such that Kuhn–Tucker points are promising candidates for solutions of (2.1).

If $g(x_*)$ and η are sufficiently smooth, condition (2.3) is equivalent to *pointwise complementarity* in the sense that

$$\eta(t)g(x)(t) = 0, \quad \eta(t) \geq 0, \quad g(x)(t) \geq 0 \quad \text{for almost all } t \in \Omega. \tag{2.4}$$

Informally, the idea of *interior point methods* is to replace the unwieldy complementarity condition (2.4) with a relaxed substitute condition of the type

$$\eta(t)g(x)(t) = \mu \quad \text{for } \mu > 0\,, \tag{2.5}$$

where μ is the *duality gap* parameter. μ may also be interpreted as a regularization parameter (see Figure 2.1) or a barrier parameter in the primal IP formulation. The connection between (2.4) and (2.5) is via a homotopy with respect to μ, the *central path*. Since (2.5) allows solutions with g and η both positive as well as g and η both negative, the additional *feasibility condition*

$$\eta(t) \geq 0, \quad g(x)(t) \geq 0 \tag{2.6}$$

has to be satisfied, such that the central path is well-defined. For $\mu \to 0$ we arrive at the condition (2.4)—see Figure 2.1.

Replacing condition (2.4) with a condition of the type

$$\Psi(g(x), \eta; \mu) = 0 \quad \text{for } \mu > 0, \tag{2.7}$$

where the feasibility of the central path is guaranteed by the construction of Ψ, leads to so-called *complementarity methods*. Thus, infeasible intermediate iterates can be accepted—a feature which increases the overall robustness of the method. Throughout the paper we specify Ψ to be defined by pointwise application $\Psi(g(x),\eta;\mu)(t) = \psi(g(x)(t),\eta(t);\mu)$ of the Fischer–Burmeister function [21]

$$\psi(a,b;\mu) = a + b - \sqrt{a^2+b^2+2\mu}\,, \tag{2.8}$$

the zero level set of which is characterized by the IP conditions (2.5) and (2.6). Different complementarity functions are studied, e.g., in [13, 14, 15, 27].

For ease of writing, we introduce the extended variables $v = (x, \lambda_c, \lambda_r, \eta)$. Defining the Lagrangian as

$$L(v) = J(x) - \langle \lambda_c, c(x)\rangle - \lambda_r^T r(x) - \langle \eta, g(x)\rangle,$$

the adjoint equation (2.2) can be written as $\partial_x L(x,\lambda_c,\lambda_r,\eta) = 0$. Relaxing the necessary conditions (2.2)–(2.3) by (2.7), we arrive at the formulation

$$F(v;\mu) = \begin{bmatrix} \partial_x L(x,\lambda_c,\lambda_r,\eta) \\ -c(x) \\ -r(x) \\ \Psi(g(x),\eta;\mu) \end{bmatrix} = 0 \tag{2.9}$$

defining the central path $v(\mu)$. This formulation will actually be treated numerically by a continuation method. In order to justify this approach, we have to establish the existence of the central path in appropriate function spaces and to study under which conditions convergence to a KKT point can be shown. This is the topic of the following two subsections.

Notation. Some variables and operators are constructed such that they have a natural block partitioning corresponding to the components u and y of x. The individual blocks are denoted by the corresponding component as a superscript, e.g.,

$$g(x) = \begin{bmatrix} g^u(u) \\ g^y(y) \end{bmatrix} \quad \text{and} \quad \Psi(g(x),\eta) = \begin{bmatrix} \Psi^u(g^u(u),\eta^u) \\ \Psi^y(g^y(y),\eta^y) \end{bmatrix}.$$

2.1. Existence of the central path. Existence of the central path can be shown by using the implicit function theorem. For this to be applicable to (2.9), however, F has to be continuously Fréchet-differentiable. For both the IP condition (2.5) and the complementarity condition (2.7) to be well-defined and continuously differentiable, the multiplier η has to be sufficiently regular, i.e., $\eta \in L_\infty$. This is not covered by the necessary conditions (2.3) stating only $\eta \in K^+$, and, in general, does not hold for the solution η_*. In the presence of *state* constraints, their Lagrange multipliers usually contain Dirac distributions at touch points and boundary points of state constrained subarcs.

During the homotopy, however, i.e., for $\mu > 0$, the multipliers are indeed contained in L_∞. In [42], existence of the central path and boundedness of the Lagrange multipliers $\eta(\mu) \in L_\infty \times L_\infty \subset (L_\infty \times W^1_\infty)^*$ associated with central path solutions $x(\mu)$ is shown for $\mu > 0$; see Theorem 2.1.

As it turns out, the appropriate setting to show existence of the central path is to consider F mapping

$$V \times \mathbb{R}_+ = (L_\infty^{n_u} \times (W_\infty^1)^{n_y}) \times (\mathbb{R}^{n_y} \times L_\infty^{n_y}) \times \mathbb{R}^{n_r} \times (L_\infty^{n_\eta^u} \times L_\infty^{n_\eta^y}) \times \mathbb{R}_+$$

into

$$Z = (L_1^{n_u} \times (W_1^1)^{n_y})^* \times (\mathbb{R}^{n_y} \times L_\infty^{n_y}) \times \mathbb{R}^{n_r} \times (L_\infty^{n_\eta^u} \times L_\infty^{n_\eta^y}).$$

The following theorem has been compiled from Theorems 3.3 and 3.4 given in [42]. It proves that the homotopy can be performed for $\mu > 0$ in this setting.

THEOREM 2.1. *Suppose there are an open bounded set $D \subset V$ and $v_0 \in D$ and $\mu_0 > 0$ with $F(v_0; \mu_0) = 0$. Assume there are constants $\beta > 0$ and $\alpha > 0$, such that the following conditions hold uniformly for all $v = (x, \lambda_c, \lambda_r, \eta) \in D$ and $0 < \mu \le \mu_0$.*

1. *The state equation satisfies the following* inf-sup *condition:*

$$\inf_{\xi \in \mathbb{R}^{n_r}} \sup_{\delta u \in L_1^{n_u}} \frac{\xi^T \partial_y r(x) \partial_y c(x)^{-1} \partial_u c(x) \delta u}{|\xi| \, \|\delta u\|_{L_1^{n_u}}} \ge \beta.$$

2. *A strengthened Legendre–Clebsch-type condition holds:*

$$\xi^T M_u(t) \xi \ge \alpha |\xi|^2$$

for all $\xi \in \mathbb{R}^{n_u}$ and almost all $t \in \Omega$. Here,

$$M_u(t) := \partial_u^2 \tilde{f}(x(t)) - \partial_u^2 \tilde{c}(x(t))^T \lambda_c(t) - (\tilde{g}^u)''(u(t))^T \eta^u(t)$$
$$+ (\tilde{g}^u)'(u(t))^T \partial_\eta \psi(\tilde{g}^u(u(t)), \eta^u(t); \mu)^{-1} \partial_g \psi(\tilde{g}^u(u(t)), \eta^u(t); \mu)(\tilde{g}^u)'(u(t)).$$

3. *The augmented second derivative of the Lagrangian is positive definite on the nullspace of the state equation:*

$$\langle \xi, (\partial_x^2 L(v) + g'(x)^* \partial_\eta \Psi(g(x), \eta)^{-1} \partial_g \Psi(g(x), \eta) g'(x)) \xi \rangle \ge \alpha \|\xi\|^2_{L_2^{n_u} \times (W_2^1)^{n_y}}$$

for all $\xi \in \ker c'(x)$.

Then there exist a maximal open interval $I \subset R_+$ around μ_0 and a continuously differentiable central path $v : I_\mu \to D$ with the following properties:

1. *$v(\mu_0) = v_0$.*
2. *$F(v(\mu); \mu) = 0$ for all $\mu \in I_\mu$.*
3. *$\operatorname{dist}(v(I_\mu), \partial D) = 0$ or $\inf I_\mu = 0$ holds.*

In the presence of state constraints, the associated multipliers η^y may rapidly increase in order to approximate Dirac distributions occurring in the solution ${\eta^y}_*$. Therefore, convergence of $v(\mu)$ to a KKT point cannot be expected in the setting of Theorem 2.1. This is reflected by massive grid refinements in the vicinity of critical points and numerical difficulties for very small μ when $\eta(\mu)$ approximates Dirac measures. Nevertheless, numerical experience suggests that convergence may actually occur in a weaker norm.

In contrast, proving convergence to a KKT point requires stronger assumptions and in particular the more restricted setting of control constrained problems. This is addressed in section 2.2 below.

2.2. Convergence for control constrained problems. In order to prove convergence of $v(\mu)$ to a KKT point, we need at least to bound $\eta^y(\mu)$ in L_∞ and therefore restrict our attention to the control constrained case $n_\eta^y = 0$. Moreover, proving convergence needs a more subtle setup than Theorem 2.1. We therefore introduce the notion of nearly active sets.

DEFINITION 2.2. *For some $\rho > 0$ and functions $u \in L_\infty^{n_u}$ and $\eta \in L_\infty^{n_\eta^u}(\Omega)$, define the characteristic function $\chi^A = \chi^A(t; u, \eta, \mu)$ of the nearly active set vector Ω^A componentwise as*

$$\chi_i^A(t) = \begin{cases} 1, & \tilde{g}_i^u(u_i(t)) \le \rho \eta_i^u(t), \\ 0, & \text{otherwise.} \end{cases}$$

The corresponding characteristic function χ^I of the nearly inactive set vector Ω^I is defined as $\mathbf{1} - \chi^A$, where $\mathbf{1} \in L_\infty^{n_\eta^u}$ is the constant function with value 1.

Compared to Theorem 2.1 above, the strengthened Legendre–Clebsch-type condition is only required to hold on a smaller subspace, but the inf-sup condition needs to hold on a larger one. The difference is just the nearly active set. The following theorem is compiled from [42, Thms. 3.6, 3.8, 3.10].

THEOREM 2.3. *Suppose $n_\eta^y = 0$, i.e., there are no state constraints. Assume that the following conditions are satisfied.*

1. *The feasible region $D_u := \{u \in L_\infty^{n_u} : g(u) \ge 0\}$ is bounded.*
2. *The state contribution function in the state equation is linearly bounded:*

$$|\tilde{c}(u, y)| \le \text{const}(1 + |u|^p + |y|) \quad \text{for all } (u, y) \in \mathbb{R}^{n_u} \times \mathbb{R}^{n_y}.$$

Then there is a bounded set $D_y \subset (W_\infty^1)^{n_y}$ such that for all $\mu > 0$ every solution v of $F(v; \mu) = 0$ satisfies $u \in D_u$ and $y \in D_y$.

If, in addition, there is a constant $\beta > 0$ and some $\mu_0 > 0$ such that the equality constraints and nearly active control constraints satisfy the inf-sup *condition*

$$\inf_{h \in \mathbb{R}^{n_r}, \xi \in L_p^{n_\eta^u}} \sup_{\delta u \in L_q^{n_u}} \frac{h^T \partial_y r(x) \partial_y c(x)^{-1} \partial_u c(x) \delta u + \langle \chi^A \xi, g'(u) \delta u \rangle}{(|h| + \|\chi^A \xi\|_{L_p^{n_\eta^u}}) \, \|\delta u\|_{L_q^{n_u}}} \ge \beta \tag{2.10}$$

with $(p, q) = (\infty, 1)$ and $(p, q) = (2, 2)$ uniformly for central path solutions v with $x \in D_u \times D_y$ and $\mu \le \mu_0$, then there is a bounded set $D_0 \subset V$ such that $v \in D_0$. Define $D = \bigcup_{v \in D_0} S(v, \epsilon)$ for some $\epsilon > 0$, where $S(v, \epsilon)$ is the open ball around v with radius ϵ.

Suppose there is a constant $\alpha > 0$, such that the following conditions hold uniformly for all central path solutions $v = v(\mu) \in D$ with $F(v(\mu); \mu) = 0$ and $0 < \mu \le \mu_0$.

3. *The augmented second derivative of the Lagrangian,*

$$M = \partial_x^2 L(v) + g'(x)^* \partial_\eta \Psi(g(x), \eta)^{-1} \partial_g \Psi(g(x), \eta) g'(x),$$

satisfies the following positivity conditions:

$$\begin{aligned} &\langle \xi, M\xi \rangle \ge \alpha \|\xi\|^2_{L_2^{n_u} \times (W_2^1)^{n_y}} && \text{for } \xi \in \ker c'(x) \cap \ker \chi^A g'(u), \\ &\langle \xi, M\xi \rangle \ge 0 && \text{for } \xi \in \ker c'(x). \end{aligned}$$

Then the central path $v(\mu)$ converges to a Kuhn–Tucker point $v(0) \in D$:

$$\|v(\mu) - v(0)\|_V \le \text{const} \sqrt{\mu}.$$

The numerical scheme, however, carries further than the currently available theory. Even our key example in section 4.2 falls out of the present analytic setting, since it involves state constraints as well. However, in this case, the *adaptive multilevel algorithm* to be worked out in section 3 produces successively sharper local peaks on successively finer meshes—thus realizing a multilevel approximation of the Dirac distribution. For an illustration of this effect see Figure 4.7 below.

3. Numerical algorithm. For the numerical computation of the solution point $v(0)$ we employ a Newton-type continuation method following the central path $v(\mu)$ defined by (2.9). When applied to $F(v) = 0$ and $AF(Bv) = 0$, where A and B are invertible linear transformations, Newton's method generates equivalent sequences of iterates. This *invariance* property should be inherited by numerical algorithms and accompanying convergence theory. Unfortunately, full invariance is impossible due to the necessity of measuring convergence in some appropriate norm. Fixing $B = I$ one obtains affine covariant (error-oriented) methods [19], whereas setting $A = I$ yields affine contravariant (residual-oriented) methods [24]. Coupling $A = B^*$ results in affine conjugate (energy-oriented) methods for convex unconstrained optimization problems [20]. For an in-depth treatment of affine invariance we refer to the research monograph [18].

3.1. Affine invariant norms. Neither of the above-mentioned invariance classes reflects the structure of the optimization problem (2.1). A new class of affine invariance and a corresponding invariant norm for equality constrained problems ($n_\eta = 0$) that has been worked out by Volkwein and Weiser [38] needs to be adapted to the L_∞ setting and to include inequality constraints as well. Due to the pointwise nature of the constraints we have to restrict ourselves to *pointwise* transformations when studying affine invariance.

DEFINITION 3.1. *For $(z_a, z_c, z_r, z_\eta) \in Z$ and $(v, \mu) \in V \times \mathbb{R}_+$ we define*

(3.1)
$$\|(z_a, z_c, z_r, z_\eta)\|^2_{(v,\mu)} := \|\phi\|_\infty + \|\xi_{\lambda_c}\|^2_{L_\infty} + |\xi_{\lambda_r}|^2 + \|\xi_\eta\|^2_{L_\infty} + \|z_c\|^2_{L_\infty} + |z_r|^2 + \|z_\eta\|^2_{L_\infty}$$

with ξ and ϕ given by

$$F'(v;\mu)(\xi_x, \xi_{\lambda_c}, \xi_{\lambda_r}, \xi_\eta)^T = (z_a, 0, 0, 0)^T \tag{3.2}$$

and

$$\phi(t) = \xi_x(t)^T \partial_x^2 L(v;\mu)(t)\xi_x(t). \tag{3.3}$$

Equation (3.2) *defines the subspace $\mathcal{N}(v;\mu) \subset V$.*

For the easily computable expression (3.1) to be a norm we have to modify the assumptions on the problem slightly compared to Theorem 2.3. Instead of augmenting the Hessian of the Lagrangian with the almost inactive part of the IP regularization when requesting pointwise positivity, we directly assume positivity of the Hessian itself. This will in general be only a slight modification because the augmentation is rather small and bounded for $\mu \to 0$. The positivity is not required to hold on the nullspace $\mathcal{N}(v;\mu)$ of the almost active constraints, but rather on the nullspace of state equation and regularized complementarity condition. Again, these subspaces are rather close to each other, in particular for $\mu \to 0$.

In this section we will need the following properties of $\partial_v F$ shown in [42, Thms. 3.2 and 3.8].

LEMMA 3.2. *Assume that Theorem* 2.3 *holds. Then there is a constant C such that the following estimates hold uniformly for all $v \in D$ and $\mu > 0$:*

$$\|\partial_v F(v;\mu)^{-1}\| \le C, \quad \|\partial_v F(v_1;\mu) - \partial_v F(v_2;\mu)\| \le \frac{C}{\sqrt{\mu}}\|v_1 - v_2\|.$$

THEOREM 3.3. *Assume Theorem* 2.3 *holds and there is some $\alpha > 0$ such that*

$$\langle \xi_x, \partial_x^2 L(v;\mu)\xi_x\rangle \ge \alpha \|\xi_x\|^2_{L_2 \times W_2^1} \tag{3.4}$$

for all $\xi \in \mathcal{N}(v;\mu)$. Then (3.1) *defines a norm on Z that is invariant under pointwise invertible transformations of X. Moreover, this norm is equivalent to the usual norm on Z.*

Proof. For brevity we will omit any function arguments. First we will show that $\partial_u^2 L \succeq \alpha$ for almost all $t \in \Omega$. From this we will conclude that there is a constant $c > 0$, such that $\|\phi\|_{L_\infty} \geq c\|\xi_x\|_{L_\infty}^2$ for all $\xi \in \mathcal{N}(v;\mu)$. Due to the bounded invertibility of $F'(v;\mu)$ as asserted by Lemma 3.2, it is then obvious that (3.1) defines a norm that is indeed equivalent to the usual norm of Z.

Let us approximate the effect of an impulse control at $t \in \Omega$ by a constant control with shrinking support, i.e., $\chi_{[t-\epsilon,t+\epsilon]}$. Since the homogeneous terminal condition $\partial_y r(x)y = 0$ has to be satisfied by the associated state, we need to complement the approximate impulse control with some compensation $\epsilon\tilde{u}_{t,\epsilon}$. Note that due to (2.10), $\tilde{u}_{t,\epsilon}$ can be chosen to be uniformly bounded in L_∞ for all t and ϵ. We define $\hat{u}_\epsilon = \chi_{[t-\epsilon,t+\epsilon]} + \epsilon\tilde{u}_{t,\epsilon}$ with associated state $\hat{y}_\epsilon$ and multipliers such that $\hat{\xi}_\epsilon \in \mathcal{N}(v;\mu)$. From standard ODE theory it follows that $\|\hat{y}_\epsilon\|_{L_\infty} \leq c\|\hat{u}_\epsilon\|_{L_1} = c\epsilon$. From (3.4) we know that

$$\begin{aligned}\langle \hat{u}_\epsilon, \partial_u^2 L \hat{u}_\epsilon\rangle &= \langle \hat{\xi}_{x\epsilon}, \partial_x^2 L \hat{\xi}_{x\epsilon}\rangle - \langle \hat{y}_\epsilon, \partial_y^2 L \hat{y}_\epsilon\rangle - 2\langle \hat{u}_\epsilon, \partial_{yu} L \hat{y}_\epsilon\rangle \\ &\geq \alpha\|\hat{\xi}_{x\epsilon}\|_{L_2\times W_2^1}^2 - c\|\hat{y}_\epsilon\|_{L_\infty}^2 - c\|\hat{y}_\epsilon\|_{L_\infty}\|\hat{u}_\epsilon\|_{L_1} \\ &\geq 2\alpha\epsilon - c\epsilon^2\end{aligned}$$

for some generic constant c independent of ϵ, and hence $\partial_u^2 L \succeq \alpha$ for almost all $t \in \Omega$.

Now assume that for any $\epsilon > 0$ there is some $\xi_\epsilon \in \mathcal{N}(v;\mu)$ with control and state component u_ϵ and y_ϵ, respectively, such that

$$\|\phi_\epsilon\|_{L_\infty} \leq \epsilon\|\xi_\epsilon\|_{L_\infty}^2 \tag{3.5}$$

and $\|u_\epsilon\|_{L_\infty} = 1$. From (3.4) we know that

$$\begin{aligned}\|y_\epsilon\|_{L_\infty}^2 &\leq c\|u_\epsilon\|_{L_1}^2 \leq c\|\xi_\epsilon\|_{L_2}^2 \leq \frac{c}{\alpha}\|\phi_\epsilon\|_{L_2} \\ &\leq \frac{c}{\alpha}\|\phi_\epsilon\|_{L_\infty} \leq \epsilon\frac{c}{\alpha}\|\xi_\epsilon\|_{L_\infty}^2 \leq \epsilon\frac{c}{\alpha}\|y_\epsilon\|_{L_\infty}^2 + \epsilon\frac{c}{\alpha}\|u_\epsilon\|_{L_\infty}^2\end{aligned}$$

for some generic constant c independent of ϵ. For sufficiently small ϵ we have $\epsilon c/\alpha < 1$ and may conclude that $(1 - \epsilon c/\alpha)\|y_\epsilon\|_{L_\infty}^2 \leq c\epsilon/\alpha\|u_\epsilon\|_{L_\infty}^2$ and hence $\|y_\epsilon\|_{L_\infty}^2 \leq c\epsilon$. Using the boundedness of $\partial_x^2 L$, we obtain

$$\phi_\epsilon = u_\epsilon^T \partial_u^2 L u_\epsilon + y_\epsilon^T \partial_y^2 L y_\epsilon + 2u_\epsilon^T \partial_{yu} L y_\epsilon \geq \alpha|u_\epsilon|^2 - c\epsilon - c\sqrt{\epsilon}|u_\epsilon|$$

for almost all $t \in \Omega$. For all sufficiently small ϵ this implies $\|\phi_\epsilon\|_{L_\infty} \geq c\|\xi_x\|_{L_\infty}^2$ for some constant c independent of ϵ, which contradicts (3.5).

Now we turn to affine invariance. Let us consider a transformation $\hat{x} = B^{-1}x$ of X. With the pointwise transformation $D = \mathrm{diag}(B, I, I, I)$ the transformed problem is $D^*\hat{F}(\hat{v};\mu) = F(B\hat{x}, \lambda_c, \lambda_r, \eta)$ with derivative $\partial_{\hat{v}}\hat{F}(\hat{v};\mu) = D^*\partial_v F(v;\mu)D$. Computing the norm of $(\hat{z}_a, \hat{z}_c, \hat{z}_r, \hat{z}_\eta) = (B^* z_a, z_c, z_r, z_\eta)$, we obtain

$$\begin{aligned}(\hat{\xi}_x, \hat{\xi}_{\lambda_c}, \hat{\xi}_{\lambda_r}, \hat{\xi}_\eta) &= \partial_{\hat{v}}\hat{F}(\hat{v};\mu)^{-1}(B^* z_a, z_c, z_r, z_\eta) \\ &= D^{-1}\partial_v F(v;\mu)^{-1}(z_a, z_c, z_r, z_\eta) \\ &= (B^{-1}\xi_x, \xi_{\lambda_c}, \xi_{\lambda_r}, \xi_\eta)\end{aligned}$$

and

$$\begin{aligned}\hat{\phi}(t) &= \hat{\xi}_a(t)^T \partial_{\hat{x}}^2 \hat{L}(\hat{v};\mu)(t)\hat{\xi}_a(t)\\ &= \xi_a(t)^T B(t)^{-T} B(t)^T \partial_x^2 L(v;\mu)(t)B(t)B(t)^{-1}\xi_a(t)\\ &= \phi(t).\end{aligned}$$

Since all terms in the sum (3.1) coincide, the norm is invariant. If B was not a pointwise transformation, $\hat{\phi}$ could not be defined in a pointwise manner. This is the reason why the invariance class has to be restricted to pointwise transformations. ∎

Formulating convergence results in terms of local norms, the norms need to be Lipschitz continuous with respect to the evaluation point to which they are attached.

THEOREM 3.4. *There are constants $\gamma_v(\mu)$ and $\gamma_\mu(\mu)$ independent of v such that*

$$\big|\|z\|_{(v_1,\mu)} - \|z\|_{(v_2,\mu)}\big| \le \gamma_v(\mu)\|\partial_v F(v_1;\mu)(v_1 - v_2)\|_{(v_1,\mu)}\|z\|_{(v_1,\mu)} \tag{3.6}$$

and

$$\big|\|z\|_{(v,\mu_1)} - \|z\|_{(v,\mu_2)}\big| \le \gamma_\mu(\mu)|\mu_1 - \mu_2|\|z\|_{(v,\mu_1)} \tag{3.7}$$

for all $z \in Z$.

Proof. With $z = (z_a, z_c, z_r, z_\eta)$ and $\hat{z} = (z_a, 0, 0, 0)$ we define $\xi_i = \partial_v F(v_i;\mu)^{-1}\hat{z} \in \mathcal{N}(v_i;\mu)$ for $i = 1, 2$. Using the Lipschitz continuity and bounded invertibility of $\partial_v F$ as given by Lemma 3.2, we first obtain

$$\begin{aligned}\|\xi_1 - \xi_2\| &= \|(\partial_v F(v_1;\mu)^{-1} - \partial_v F(v_2;\mu)^{-1})\hat{z}\|\\ &\le \|F(v_2;\mu)^{-1}(\partial_v F(v_1;\mu) - \partial_v F(v_2;\mu))\partial_v F(v_1;\mu)^{-1}\hat{z}\|\\ &\le c\|\partial_v F(v_1;\mu) - \partial_v F(v_2;\mu)\|\|\hat{z}\|\\ &\le \frac{c}{\sqrt{\mu}}\|v_1 - v_2\|\|\hat{z}\|_{(v_1,\mu)} \qquad (3.8)\end{aligned}$$

for some generic constant c independent of μ and v_i. Second we have

$$\begin{aligned}|\xi_{x1}^T\partial_x^2 L(v_2;\mu)\xi_{x1}| &= |\xi_{x1}^T\partial_x^2 L(v_1;\mu)\xi_{x1}| + |\xi_{x1}^T(\partial_x^2 L(v_2;\mu) - \partial_x^2 L(v_1;\mu))\xi_{x1}|\\ &\le |\xi_{x1}^T\partial_x^2 L(v_1;\mu)\xi_{x1}| + |\xi_{x1}|^2|\partial_x^2 L(v_2;\mu) - \partial_x^2 L(v_1;\mu)|\end{aligned}$$

for almost all $t \in \Omega$ and hence

$$\|\xi_{x1}^T\partial_x^2 L(v_2;\mu)\xi_{x1}\|_{L_\infty} \le \left(1 + \frac{c}{\sqrt{\mu}}\|v_2 - v_1\|\right)\|\phi_1\|_{L_\infty}. \tag{3.9}$$

Writing the norm induced by (3.1) on $\mathcal{N}(v;\mu)$ as $\|\cdot\|_v$ (note that the expression does not depend on μ), we utilize (3.9) and (3.8) in order to obtain

$$\begin{aligned}\|\hat{z}\|_{(v_2,\mu)} &= \|\xi_2\|_{v_2}\\ &\le \|\xi_1\|_{v_2} + c\|\xi_2 - \xi_1\|\\ &\le \left(\|\xi_{x1}^T\partial_x^2 L(v_2;\mu)\xi_{x1}\|_{L_\infty} + \|(\xi_{\lambda_c 1}, \xi_{\lambda_r 1}, \xi_{\eta 1})\|_{L_\infty}^2\right)^{1/2} + c\|\xi_2 - \xi_1\|\\ &\le \left(\sqrt{1 + \frac{c}{\sqrt{\mu}}\|v_2 - v_1\|}\,\|\phi_1\|_{L_\infty} + \|(\xi_{\lambda_c 1}, \xi_{\lambda_r 1}, \xi_{\eta 1})\|_{L_\infty}^2\right)^{1/2} + c\|\xi_2 - \xi_1\|\\ &\le \left(1 + \frac{c}{\sqrt{\mu}}\|v_2 - v_1\|\right)^{1/4}\|\xi_1\|_{v_1} + \frac{c}{\sqrt{\mu}}\|v_1 - v_2\|\|\hat{z}\|_{(v_1,\mu)}\\ &\le \left(1 + \frac{c}{\sqrt{\mu}}\|v_1 - v_2\|\right)\|\hat{z}\|_{(v_1,\mu)}\end{aligned}$$

for some generic constant c independent of μ. This and the bounded invertibility of $\partial_v F(v;\mu)$ imply

$$\begin{aligned}\big|\|z\|_{(v_2,\mu)} - \|z\|_{(v_1,\mu)}\big| \le \big|\|\hat z\|_{(v_2,\mu)} - \|\hat z\|_{(v_1,\mu)}\big| \\ \le \frac{c}{\sqrt{\mu}}\|\partial_v F(v_1;\mu)(v_1-v_2)\|_{(v_1,\mu)}\|\hat z\|_{(v_1,\mu)}\end{aligned}$$

and verify the claim (3.6) for $\gamma_v = c\mu^{-1/2}$. The corresponding estimate (3.7) can be shown analogously. ∎

Remark. We would like to point out that for local norms with structures simpler than the one considered here, useful analytical bounds for γ_v can be found in terms of a corresponding affine invariant Lipschitz constant. In the affine contravariant setting $\|\cdot\|_v = \|\cdot\|$, $\gamma_v = 0$ is trivially obtained, $\gamma_v \le \omega$ holds in the affine covariant case $\|\cdot\|_v = \|F'(v)^{-1}\cdot\|$, and $\gamma_v \le \frac{\omega}{2}$ in the affine conjugate case $\|\cdot\|_v = \|F'(v)^{-1/2}\cdot\|$. Based on an assumption similar to (3.12), a rather crude estimate of $\gamma_v \le \frac{3}{2}\omega$ was derived in [38] for equality constrained problems. Numerical experience with computational estimates suggests that even if available, the analytical bounds are often too large to describe the actual norm variation correctly. Therefore, using computational estimates may be advantageous. For the norm (3.1), however, no analytical quantitative estimate is known up to now, such that we have to resort to computational estimates (see (3.29) below).

Remark. Recently, Potra [31] employed an affine invariant norm for proving $\mathcal{O}(\sqrt{n}L)$-iteration complexity of an IP algorithm applied to finite-dimensional horizontal linear complementarity problems, which include linear and linear-quadratic optimization problems. In the notation of the current paper, that author uses a diagonal scaling like $D := \sqrt{\partial_w \psi^{-1} \partial_\eta \psi}$. A similar norm has been used to analyze the convergence rate of an IP method applied to infinite-dimensional optimal control problems in a restricted setting with bang-bang control [41].

3.2. Adaptive central path following. Once the central path homotopy is theoretically established, a numerical continuation scheme for following the path towards the solution $v(0)$ must be developed. For *numerical* pathfollowing, an adaptive tangential predictor/Newton-type corrector algorithm is worked out. The method is applied directly to the infinite-dimensional function space formulation, involving discretization only in the innermost loop when solving linear subproblems. Since a reduction of the discretization error is expensive, we substitute both the tangential predictor and the Newton corrector by their inexact counterparts and aim for linear convergence only—in the spirit of complexity estimates of [20].

As illustrated in Figure 3.1, the local convergence domain of the Newton corrector can be expected to collapse for $\mu \to 0$ as a consequence of an ever increasing Lipschitz constant $\omega(\mu)$. Nevertheless, Theorem 2.3 provides a qualitative upper bound on the error incurred by a premature termination at $\mu_{\text{final}} > 0$ of the numerical continuation along the central path in the order of $\mathcal{O}(\mu_{\text{final}}^{1/2})$. Experience shows that feasible and acceptably suboptimal solutions can indeed be obtained by following the central path up to some $\mu_{\text{final}} > 0$—see section 4.

Inexact Newton corrector. The corrector operates with constant duality gap parameter; thus we drop μ in order to simplify notation and write $F'(v)$ instead of $\partial_v F(v;\mu)$. Due to the prohibitive cost of reducing the discretization error, we cannot strive for highly accurate Newton corrections. Instead, we will employ an *inexact*

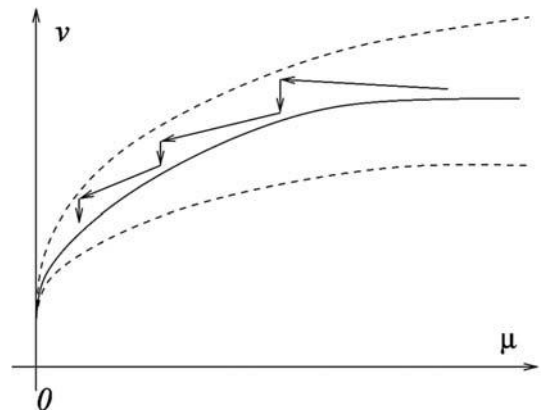


FIG. 3.1. *Inexact tangential continuation along the central path including local convergence domain of Newton's method.*

Newton method, where an *inner residual* remains:

$$F'(v^k)\delta v^k = -F(v^k) + r^k, \qquad v^{k+1} = v^k + \delta v^k\,. \tag{3.10}$$

The relative accuracy δ_k of the *inexact Newton correction* δv^k, given by

$$\delta_k := \frac{\|r^k\|_{v^k}}{\|F(v^k)\|_{v^k}}\,, \tag{3.11}$$

will play a crucial role. In actual computation, the inexact simplified Newton correction $\overline{\delta v}^{k+1}$ approximating the exact simplified Newton correction

$$F'(v^k)\overline{\Delta v}^{k+1} = -F(v^{k+1})$$

will also be used. Again, an inner residual $\overline{r}^{k+1}$ remains:

$$F'(v^k)\overline{\delta v}^{k+1} = -F(v^{k+1}) + \overline{r}^{k+1}.$$

In view of convergence of Newton's method, we need to have a Lipschitz constant for $\partial_v F$ formulated in the affine invariant norm. This affine invariant Lipschitz condition comprises two usual assumptions for Newton convergence theorems: that F' is Lipschitz continuous, and that $F'(v^k)$ has a bounded inverse (cf. [18]).

LEMMA 3.5. *If* (3.4) *holds for* $\mu > 0$*, then there exists an affine invariant Lipschitz constant* $\omega(\mu) \le c(1 + \mu^{-1/2})$*, such that*

$$\|F'(v_1) - F'(v_2)\|_\zeta \le \omega(\mu)\|F'(v_1)(v_1 - v_2)\|_{v_1} \tag{3.12}$$

for all v_1, v_2 *such that* $\zeta \in \mathrm{co}\{v_1, v_2\} \subset D$.

Proof. The bound on ω is a direct consequence of the corresponding Lipschitz estimate for the complementarity function ψ (see [42, Theorem 3.2]) and the equivalence of the norms $\|\cdot\|_{v_1}$ and $\|\cdot\|$ on Z. ∎

THEOREM 3.6. *Let* V *be a Banach space and* $D \subset V$ *an open set. Let* Z *be a Banach space equipped with a family of equivalent local norms* $\|\cdot\|_v$ *parameterized*

over D. Assume $F : D \to Z$ is a continuously differentiable function. Let γ_v and ω be constants such that the local norms $\|\cdot\|_v$ on Z satisfy

$$|\|r\|_{v_1} - \|r\|_{v_2}| \le \gamma \|F'(v_1)(v_1 - v_2)\|_{v_1} \|r\|_{v_1} \quad \text{for all } r \in Z, \tag{3.13}$$

and the affine invariant Lipschitz condition

$$\|(F'(\xi) - F'(v))(\xi - v)\|_\zeta \le \omega \|F'(v)(\xi - v)\|_v^2 \tag{3.14}$$

holds for all collinear $v, \xi, \zeta \in V$ such that the convex hull $\mathrm{co}\{v, \xi, \zeta\}$ is contained in D. Let $\Theta < 1$ and

$$\mathcal{L}(v) := \Big\{ \xi \in D : \|F(\xi)\|_\xi \le \Big(1 + \frac{\gamma_v \Theta}{2\omega}\Big) \|F(v)\|_v \Big\}.$$

Assume that $v^0 \in D$ and that the level set $\mathcal{L}(v^0)$ is closed. If $\omega \|F(v^0)\|_{v_0} < 2\Theta$ and the relative error δ_k in computing the Newton correction is controlled such that

$$\frac{1 + \delta_k}{2} \omega \|F'(v^k)\delta v^k\|_{v^k} + (1 + \gamma_v \|F'(v^k)\delta v^k\|_{v^k})\delta_k \le \Theta \tag{3.15}$$

(which is possible for all k), then the iterates are well-defined for all $k \in \mathbb{N}$ and stay in $\mathcal{L}(v^0)$, and the residuals converge to zero at a rate of

$$\|F(v^{k+1})\|_{v^{k+1}} \le \Theta \|F(v^k)\|_{v^k}. \tag{3.16}$$

Furthermore, if the inexact simplified Newton correction is computed with relative accuracy $\overline{\delta_{k+1}}$,

$$\|F'(v^k)\overline{\delta v^{k+1}}\|_{v^k} \le (1 + \overline{\delta_{k+1}}) \Big(\frac{\delta_k}{1 - \delta_k} + \frac{\omega}{2} \|F'(v^k)\delta v^k\|_{v^k} \Big) \|F'(v^k)\delta v^k\|_{v^k} \tag{3.17}$$

holds.

Proof. By induction, let $\mathcal{L}(v^k)$ be closed and $\omega \|F(v^k)\|_{v^k} < 2$. Then

$$F(v^k + s\delta v^k) = F(v^k) + \int_0^s F'(v^k + t\delta v^k)\delta v^k \, dt \tag{3.18}$$

$$= (1 - s)F(v^k) + sr^k + \int_0^s (F'(v^k + t\delta v^k) - F'(v^k))\delta v^k \, dt \tag{3.19}$$

for all $s \in [0, 1]$ with $\mathrm{co}\{v^k, v^k + s\delta v^k\} \subset D$.

Using the Lipschitz continuity (3.14) of F' and the norm continuity (3.13), for $\sigma \in [0, s]$ we have

$$\begin{aligned}
&\|F(v^k + s\delta v^k)\|_{v^k + \sigma\delta v^k} \\
&\quad \le (1 - s)\|F(v^k)\|_{v^k + \sigma\delta v^k} + s\|r^k\|_{v^k + \sigma\delta v^k} \\
&\qquad + \int_0^s \|(F'(v^k + t\delta v^k) - F'(v^k))\delta v^k\|_{v^k + \sigma\delta v^k} \, dt \\
&\quad \le (1 - s)(1 + \sigma\gamma_v \|F'(v^k)\delta v^k\|_{v^k})\|F(v^k)\|_{v^k} \\
&\qquad + s(1 + \sigma\gamma_v \|F'(v^k)\delta v^k\|_{v^k})\|r^k\|_{v^k} + \int_0^s t\omega \|F'(v^k)\delta v^k\|_{v^k}^2 \, dt \\
&\quad = (1 + \sigma\gamma_v \|F'(v^k)\delta v^k\|_{v^k})((1 - s)\|F(v^k)\|_{v^k} + s\delta_k \|F(v^k)\|_{v^k}) \\
&\qquad + \frac{s^2}{2} \omega \|F'(v^k)\delta v^k\|_{v^k}^2.
\end{aligned} \tag{3.20}$$

From (3.11) we have

$$(1-\delta_k)\|F(v^k)\|_{v^k} \le \|F'(v^k)\delta v^k\|_{v^k} = \|F(v^k) - r^k\|_{v^k} \le (1+\delta_k)\|F(v^k)\|_{v^k}\,. \tag{3.21}$$

Thus we arrive at

$$\begin{aligned}\frac{\|F(v^k + s\delta v^k)\|_{v^k+s\delta v^k}}{\|F(v^k)\|_{v^k}} \\ \le (1+s\gamma_v\|F'(v^k)\delta v^k\|_{v^k})(1-s+s\delta_k) + \frac{1+\delta_k}{2}s^2\omega\|F'(v^k)\delta v^k\|_{v^k}\,.\end{aligned}$$

The ultimate goal is to establish a contraction property for the undamped Newton method ($s=1$). Thus we have to require

$$(1+\gamma_v\|F'(v^k)\delta v^k\|_{v^k})\delta_k + \frac{1+\delta_k}{2}\omega\|F'(v^k)\delta v^k\|_{v^k} \le \Theta < 1\,,$$

which is the accuracy condition (3.15). Defining $\chi := \gamma_v\|F'(v^k)\delta v^k\|_{v^k}$ and using $s \le 1$ and (3.15), we have

$$\begin{aligned}\frac{\|F(v^k + s\delta v^k)\|_{v^k+s\delta v^k}}{\|F(v^k)\|_{v^k}} &\le (1+s\chi)(1-s+s\delta_k) + \frac{1+\delta_k}{2}s^2\omega\|F'(v^k)\delta v^k\|_{v^k} \\ &= (1+s\chi)(1-s) + (1+s\chi)s\delta_k + \frac{1+\delta_k}{2}s^2\omega\|F'(v^k)\delta v^k\|_{v^k} \\ &\le (1+s\chi)(1-s) + s\Theta \\ &= 1 - s + s\Theta + s(1-s)\chi \\ &\le 1 + \frac{\chi}{4}\,.\end{aligned} \tag{3.22}$$

From (3.15) we infer $\|F'(v^k)\delta v^k\|_{v^k} \le 2\Theta/\omega$, and hence

$$\|F(v^k + s\delta v^k)\|_{v^k+s\delta v^k} \le \left(1 + \frac{\gamma_v\Theta}{2\omega}\right)\|F(v^k)\|_{v^k}$$

holds. Since D is open and $\mathcal{L}(v^k) \subset D$ is closed, $\mathrm{co}\{v^k, v^k + \delta v^k\} \not\subset D$ implies the existence of some $s^* \in [0,1)$ with $\mathrm{co}\{v^k, v^k + s^*\delta v^k\} \subset D$ but $v^k + s^*\delta v^k \notin \mathcal{L}(v^k)$, i.e.,

$$\|F(v^k + s^*\delta v^k)\|_{v^k+s^*\delta v^k} > \left(1 + \frac{\gamma_v\Theta}{2\omega}\right)\|F(v^k)\|_{v^k}\,,$$

which is a contradiction. Thus, $v^{k+1} \in D$. Furthermore, setting $s=1$ in (3.22) yields

$$\|F(v^{k+1})\|_{v^{k+1}} \le \Theta\|F(v^k)\|_{v^k} \tag{3.23}$$

and therefore $\mathcal{L}(v^{k+1}) \subset \mathcal{L}(v^k)$. Since $\mathcal{L}(v^k)$ is closed, every Cauchy sequence in $\mathcal{L}(v^{k+1})$ converges to a limit point in $\mathcal{L}(v^k)$, which is, by continuity of the norm, also contained in $\mathcal{L}(v^{k+1})$. Hence, $\mathcal{L}(v^{k+1})$ is closed. Moreover, $\omega\|F(v^{k+1})\|_{v^{k+1}} \le \omega\Theta^{k+1}\|F(v^0)\|_{v^0} < 2\Theta^{k+2} < 2\Theta$, such that (3.15) can be satisfied in the next iteration by choosing a sufficiently small δ_{k+1}.

Inserting $\sigma = 0$, $s = 1$ into (3.20) yields

$$\begin{aligned}\|F'(v^k)\overline{\delta v^{k+1}}\|_{v^k} &\le (1+\overline{\delta_{k+1}})\|F(v^{k+1})\|_{v^k} \\ &\le (1+\overline{\delta_{k+1}})\left(\delta_k\|F(v^k)\|_{v^k} + \frac{\omega}{2}\|F'(v^k)\delta v^k\|_{v^k}^2\right) \\ &\le (1+\overline{\delta_{k+1}})\left(\frac{\delta_k}{1-\delta_k} + \frac{\omega}{2}\|F'(v^k)\delta v^k\|_{v^k}\right)\|F'(v^k)\delta v^k\|_{v^k}\,,\end{aligned}$$

which completes the proof. ∎

The boundedness of $F'(v^k)$ as provided by Theorem 2.3 now provides linear convergence of the iterates towards the solution $v(\mu)$.

COROLLARY 3.7. *Suppose that Theorems* 2.3 *and* 3.3 *hold and* (3.4) *is satisfied. If the Newton iteration is controlled according to* (3.15)*, the iterates v^k converge linearly to $v(\mu)$.*

Proof. Under the given conditions, Theorem 3.3 and Lemma 3.5 provide the assumptions of Theorem 3.6 with D given by Theorem 2.3, such that the inexact Newton iteration is well-defined and the residuals converge to 0 according to (3.16). Due to (3.10), (3.11), and the equivalence of $\|\cdot\|_v$ and $\|\cdot\|$ we have $\|\delta v^k\| \le \|F'(v^k)^{-1}\|\|F(v^k) - r^k\| \le c(1+\delta_k)\|F(v^k)\|_{v^k}$ with a constant c independent of μ. From (3.15) we infer $\delta_k < 2$ and with (3.16) holding obtain $\|\delta v^k\| \le 2c\Theta^k\|F(v^0)\|_{v^0}$. Finally,

$$\|v^k - v(\mu)\| \le 2c\|F(v^0)\|_{v^0}\frac{\Theta^k}{1-\Theta}$$

proves r-linear convergence. ∎

Inexact prediction step. From numerical experience, we expect more or less constant reduction factors for μ, translating into constantly decreasing continuation stepsizes. In order to avoid this biased stepsize behavior, the predictor is formulated in terms of $\tau = -\log\mu$. The inexact tangential predictor $\hat{v}(\tau)$ is defined by

$$F'(v_0;\tau_0)\phi = -\partial_\tau F(v_0;\tau_0) + r\,, \quad \hat{v}(\tau) = v_0 + (\tau - \tau_0)\phi\,, \tag{3.24}$$

where again a residual r remains.

LEMMA 3.8. *Assumptions of Theorem* 3.6. *Let γ_τ and β be nonnegative constants such that the local norm $\|\cdot\|_{(v,\tau)}$ satisfies*

$$\big|\|\rho\|_{(v_1,\tau_1)} - \|\rho\|_{(v_0,\tau_0)}\big| \le \gamma_\tau(\tau_1 - \tau_0)\|\rho\|_{(v_0,\tau_0)} \tag{3.25}$$

and

$$\|F(v_1;\tau_1)\|_{(v_0,\tau_0)} \le \|F(v_0;\tau_0)\|_{(v_0,\tau_0)} + \|r\|_{(v_0,\tau)}(\tau_1 - \tau_0) + \beta(\tau_1 - \tau_0)^2 \tag{3.26}$$

for all $\rho \in Z_\infty$, v_0, v_1 such that $F'(v_0;\tau_0)(v_1 - v_0) = -(\tau_1 - \tau_0)(\partial_\mu F(v_0;\tau_0) + r)$, and $\mathrm{co}\{v_0, v_1\} \subset D$. Then the inexact Newton corrector with starting point v_1 converges to the central path $v(\tau)$ for all stepsizes $\Delta\tau = \tau_1 - \tau_0$ satisfying

$$(1 + \gamma_\tau\Delta\tau)(\|F(v_0;\tau_0)\|_{(v_0,\tau_0)} + \|r\|_{(v_0,\tau_0)}\Delta\tau + \beta\Delta\tau^2) < \frac{2}{\omega}\,. \tag{3.27}$$

Proof. Combining the convergence condition $\omega\|F(v;\tau)\|_{(v,\tau)} < 2$ from Theorem 3.6 with assumptions (3.25) and (3.26) yields the result. ∎

Note that since (3.27) represents a monotone convex function of $\Delta\tau$, the maximum permitted stepsize can be easily computed by an ordinary Newton method starting from $\sqrt{2/(\omega\beta)}$.

Remark. Again, the constant γ_τ is needed because of the inexactness of the tangential predictor. In exact Newton continuation algorithms (see Deuflhard [17]), the change of local norms can be subsumed under the second order term β.

Computable Lipschitz estimates. For actual computation we need easily computable estimates of the theoretical quantities ω, γ_v, γ_τ, and β to be inserted into

conditions (3.15) and (3.27). From (3.17) and (3.6), (3.21), respectively, we derive the computable estimates

$$[\omega]_k = \frac{2}{\|F'(v^k)\delta v^k\|_{v^k}} \left(\frac{\|F'(v^k)\overline{\delta v^{k+1}}\|_{v^k}}{(1+\overline{\delta_{k+1}})\|F'(v^k)\delta v^k\|_{v^k}} - \frac{\delta_k}{1-\delta_k} \right) \leq \omega \tag{3.28}$$

and

$$[\gamma_v]_k = \frac{d\Big(\Phi\big(\delta_{k+1}, \|F'(v^{k+1})\delta v^{k+1}\|_{v^{k+1}}\big), \Phi\big(\overline{\delta_{k+1}}, \|F'(v^k)\overline{\delta v^{k+1}}\|_{v^k}\big)\Big)}{(1+\delta_{k+1})\|F'(v^{k+1})\delta v^{k+1}\|_{v^{k+1}}\|F'(v^k)\delta v^k\|_{v^k}} \leq \gamma_v, \tag{3.29}$$

where $\Phi(a,b) = [\frac{b}{1+a}, \frac{b}{1-a}]$ denotes the inaccuracy interval and

$$d(A,B) = \inf_{a\in A, b\in B} |a-b|$$

is the usual set distance. Furthermore, computable estimates for γ_τ and β can be derived from (3.25) and (3.26) as

$$[\gamma_\tau] = \frac{d\Big(\Phi\big(\delta_2, \|F'(v_2;\tau_2)\delta v_2\|_{(v_2;\tau_2)}\big), \Phi\big(\overline{\delta_2}, \|F'(v_1;\tau_1)\overline{\delta v_2}\|_{(v_1;\tau_1)}\big)\Big)}{(1+\delta_2)\|F'(v_2;\tau_2)\delta v_2\|_{(v_2;\tau_2)}(\tau_2-\tau_1)} \leq \gamma_\tau \tag{3.30}$$

and

$$[\beta] = \max\{0, \tilde{\beta}\} \leq \beta, \tag{3.31}$$

where

$$\tilde{\beta} = \frac{d\Big(\Phi\big(\overline{\delta_2}, \|F'(v_1;\tau_1)\overline{\delta v_2}\|_{(v_1,\tau_1)}\big), \Phi\big(\delta_1, \|F'(v_1;\tau_1)\delta v_1\|_{(v_1,\tau_1)}\big)\Big)}{(\tau_2-\tau_1)^2} - \frac{\|r\|_{(v_1,\tau_1)}}{\tau_2-\tau_1},$$

respectively. Of course, *reliable* estimates are obtained only if δ_k is sufficiently small, which imposes additional accuracy requirements on the computation of the predictor and corrector.

Since the computable estimates are based on *local sampling* only, they are necessarily too small. Therefore, the computed continuation stepsize $\Delta\tau$ is larger than intended and may even be too large for the corrector to converge. In this case, a *stepsize reduction* has to be performed on the basis of updated estimates. In view of computational efficiency, an early detection of violation of the theoretical assumptions is preferable. Putting it all together, we arrive at the following inexact continuation algorithm.

ALGORITHM 3.9.

initialize $v, \tau, [\omega], [\beta], [\gamma_\tau], [\gamma_v]$

choose δ_{cor}, $\rho < 1$, $\Theta < 1$

1: *while* $\tau < \tau_{\mathrm{final}}$*:*

compute predictor $\partial_v F(v;\tau)\hat{v} = -\partial_\tau F(v;\tau) + r$ *without mesh refinement*

choose $\delta_{\mathrm{TOL}} < 2/[\omega]$

compute a stepsize $\Delta\tau > 0$ *such that*

$(1 + [\gamma_\tau]\Delta\tau\|\partial_v F(v;\tau)\hat{v}\|_{(v,\tau)})(\|F(v;\tau)\|_{(v,\tau)} + \Delta\tau\|r\|_{(v,\tau)} + [\beta]\Delta\tau^2) \leq \rho^3\delta_{\mathrm{TOL}}$

update $[\beta]$ *according to* (3.31)

if $\|F(v+\Delta\tau\hat{v};\tau+\Delta\tau)\|_{(v,\tau)} > \rho^2\delta_{\mathrm{TOL}}$ *go to* 1:

update $[\gamma_\tau]$ *according to* (3.30)
if $\|F(v+\Delta\tau\hat{v},\tau+\Delta\tau)\|_{(v+\Delta\tau\hat{v},\tau+\Delta\tau)}$
$> \rho^{-1}(1+[\gamma_\tau]\Delta\tau\|\partial_v F(v;\tau)\hat{v}\|_{(v,\tau)})(\|F(v;\tau)\|_{(v,\tau)}+\Delta\tau\|r\|_{(v,\tau)})$ *go to* 1:
initialize corrector $v_0 \leftarrow v+\Delta\tau\hat{v}$
while $\|F(v_k;\tau+\Delta\tau)\|_{(v_k,\tau+\Delta\tau)} > \delta_{\mathrm{cor}}\delta_{\mathrm{TOL}}$
if not in first Newton iteration
compute $\partial_v F(v_{k-1};\tau+\Delta\tau)\overline{\delta v} = -F(v_k;\tau+\Delta\tau)+\bar{r}$ *without refinement*
update $[\omega]$ *according to* (3.28)
if (3.17) *is violated go to* 1:
compute corrector $\partial_v F(v_k;\tau+\Delta\tau)\delta v = -F(v_k;\tau+\Delta\tau)+r$ *satisfying* (3.15)
update $[\gamma_v]$ *according to* (3.29)
$v_{k+1} = v_k+\delta v$
advance $v \leftarrow v_k$, $\tau \leftarrow \tau+\Delta\tau$

Termination of this stepsize reduction scheme has been studied in [40]. In actual computation, the estimated values $[\gamma_v]$ and $[\gamma_\mu]$ are very small, which suggests that using analytical bounds as mentioned in section 3.1 would be inefficient. Instead, the continuation algorithm mainly depends on $[\omega]$ and $[\beta]$.

Solution of linear subproblems. Applying Newton continuation methods in function space requires solving a *sequence of linear perturbed saddle point problems* of the same structure as the nonlinear complementarity problem (2.9). In principle, any standard linear BVP solver technique can be employed. If there is a stable direction for the initial value problem, a multiple shooting discretization of the state equation is certainly appropriate for the linear problem. For the numerical example in section 4 and further ones in [40], a Gauss collocation method has been employed. A fixed but arbitrary polynomial order p has been used for the states and $p-1$ for the controls, with p in the range $4, 5, 6$. The successive grid adaptation is based on a not very sophisticated ad hoc error estimator comparing approximations of different order. Standard band and sparse solvers have been used to solve the finite-dimensional linear systems. These and related algorithmic issues will be worked out to a higher level of sophistication in the near future.

4. Numerical examples. This section is devoted to numerical examples for the function space–oriented IP method developed above. In all the examples shown below, the continuation process has been run until either the memory requirement induced by successive refinement or insufficient accuracy of the linear system solver induced by increasing condition numbers limited the progress on the central path.

4.1. A simple class of optimal control problems. First we turn to a class of simple problems which are covered by Theorems 2.1 and 2.3:

$$\min \int_0^1 \left(\tilde{f}^y(y(t)) + \frac{\alpha}{2}|u(t)|^2\right) dt$$

$$\text{subject to} \quad \dot{y}(t) = Ay(t)+Bu(t),$$
$$y(0) = y_0,$$
$$a \le u(t) \le b.$$

Since the case of vector-valued controls is notationally more complex but provides no additional insight, we restrict the presentation to scalar controls.

THEOREM 4.1. *Suppose that on arbitrary bounded sets in* $\mathbb{R}^{n_y}$, $\tilde{f}^y$ *is convex and twice Lipschitz continuously differentiable,* $\alpha > 0$, $a < b$, $A \in \mathbb{R}^{n_y \times n_y}$, *and* $B \in \mathbb{R}^{n_y}$.

Assume there are v_0 and $\mu_0 > 0$ such that $F(v_0; \mu_0) = 0$. Then the central path $v(\mu)$ converges to a Kuhn–Tucker point $v(0) \in D$ at a rate of

$$\|v(\mu) - v(0)\| \leq \text{const}\, \sqrt{\mu}.$$

Proof. We start with Theorem 2.3, choosing

$$\rho < \frac{1}{\mu_0} \left(\frac{b-a}{2} \right)^2 \tag{4.1}$$

for separating nearly active and nearly inactive constraints. Obviously, conditions 1 and 2 are satisfied. Since no terminal boundary conditions are given, the inf-sup condition (2.10) simplifies to

$$\inf_{\xi \in L_p^2} \sup_{\delta u \in L_q^1} \frac{\langle \chi^A \xi, g'(u) \delta u \rangle}{\|\chi^A \xi\|_{L_p^2} \|\delta u\|_{L_q^1}} \geq \beta.$$

Assume that for a central path solution (v, μ) with $\mu \leq \mu_0$, the lower constraint $u \geq a$ is nearly active at t, i.e., $\rho \eta^a(t) \geq u(t) - a$. For simplicity, we will omit the argument t in the following. Together with (4.1) and the IP condition $\eta^a(u-a) = \mu = \eta^b(b-u)$ holding for all central path solutions, this implies

$$\begin{aligned} b - u = b - a - (u - a) &\geq b - a - \sqrt{\rho \eta^a (u-a)} = b - a - \sqrt{\rho \mu} \\ &\geq b - a - \frac{b-a}{2} = \frac{b-a}{2} > \sqrt{\rho \mu} = \sqrt{\rho \eta^b (b-u)}. \end{aligned}$$

Squaring and dividing by $b-u$ finally yields $b - u > \rho \eta^b$, which implies that the upper constraint $u \leq b$ is nearly inactive whenever the lower constraint is nearly active. Analogously, the converse can be shown, such that at most one of the two constraints is active. Always choosing the nearly active constraint whenever possible, we see that

$$\inf_{\xi \in L_p^2} \sup_{\delta u \in L_q^1} \frac{\langle \chi^A \xi, g'(u) \delta u \rangle}{\|\chi^A \xi\|_{L_p^2} \|\delta u\|_{L_q^1}} \geq \inf_{\xi \in L_p^1} \sup_{\delta u \in L_q^1} \frac{\langle \xi, \delta u \rangle}{\|\xi\|_{L_p^1} \|\delta u\|_{L_q^1}} \geq 1$$

for both $(p, q) = (\infty, 1)$ and $(p, q) = (2, 2)$, which confirms the inf-sup condition.

Now we verify the assumptions of Theorem 2.1 on the whole space $D = V$. Assumption 1 is trivially inapplicable due to $n_r = 0$. The Legendre–Clebsch condition 2 is obviously satisfied due to $\alpha > 0$ and the linearity of the constraints, as is the positive definiteness of $\partial_x^2 L(v)$.

Finally, the remaining assumptions of Theorem 2.3, conditions 3 and 4, are satisfied uniformly for $v \in D$ as their counterparts from Theorem 2.1. Thus, Theorem 2.3 can be applied and yields the claim. ∎

For a brief discussion on how to extend the argumentation to more complex nonlinear problems we refer to [42].

As an illustrative example of the problem class studied in Theorem 4.1 we choose

$$\begin{aligned} \min \int_0^1 \left(t^2 y(t)^2 + \frac{\alpha}{2} u(t)^2 \right) dt \\ \text{subject to} \qquad \ddot{y}(t) = u(t), \\ y(0) = 1,\ \dot{y}(0) = 0, \\ -1 \leq u(t) \leq 1 \end{aligned} \tag{4.2}$$

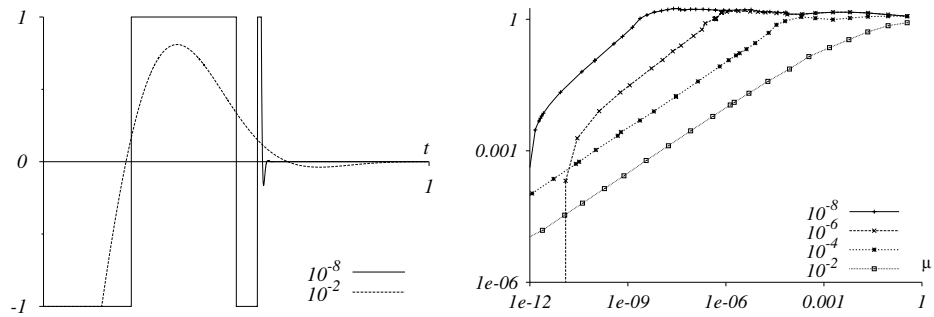


FIG. 4.1. *The control u of problem* (4.2) *for different values of α. Left: numerical solutions. Right: the approximate error $\|u(\mu) - u(\mu_{\text{final}})\|_{L_\infty}$.*

with control regularization values $\alpha = 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}$. Rewriting (4.2) as a system of two first order ODEs, the example is covered by Theorem 4.1. Clearly visible in Figure 4.1 is the convergence of order $\mathcal{O}(\sqrt{\mu})$ in the control, which is the most critical variable. The quantitative value of μ where this asymptotics sets in depends, of course, on the control regularization parameter α. For $\alpha \to 0$ we approach the bang-bang case, for which no convergence in L_∞ can be expected at all.

4.2. Abort landing in the presence of windshear. Here we will consider a well-known intricate optimal control problem, the abort landing in the presence of windshear. Our numerical results are based on the precise model given in [6]. The problem is of Chebyshev type, maximizing the minimal altitude. The optimal solution consists of control and state constrained subarcs as well as touch points and singular subarcs, which makes the problem difficult to tackle by means of the maximum principle. It contains a third order state constraint and a nondifferentiable wind model based on spline representation—and is therefore not covered by the theoretical presentation in [42]. Nevertheless, as will be reported now, already the first version of our algorithm developed herein worked satisfactorily.

Originally, the problem has been modeled by Miele et al. [30]; as for the numerical solution, these authors seem to have applied a *robust collocation* method based on a finite-dimensional parametrization of the control and combined with a gradient restoration technique to find the corresponding optimal finite-dimensional solution. Their paper does not present any numerical results for the control, which is the most sensitive and numerically critical variable.

As preparation for the application of *multiple shooting*, Bulirsch, Montrone, and Pesch [10] required 11 pages to present a brief outline of the analytic derivation of the necessary conditions. In contrast to that, the method we propose here does not require any analytical preprocessing—thus saving considerable human effort. In a second paper [11], the application of the multiple shooting method has been described along with the homotopy necessary to obtain the correct switching structure. In 1995, Berkmann and Pesch [6] solved the same problem even more accurately and claimed that "a competing direct method is unlikely to be able to produce solutions with such high resolution." In fact, our direct function space method did require a substantial computational effort to reach a comparably high accuracy. A comparison of computing times, however, would be too early, since our first focus was on developing a working algorithm within the rather new conceptual frame. There is enough space left for

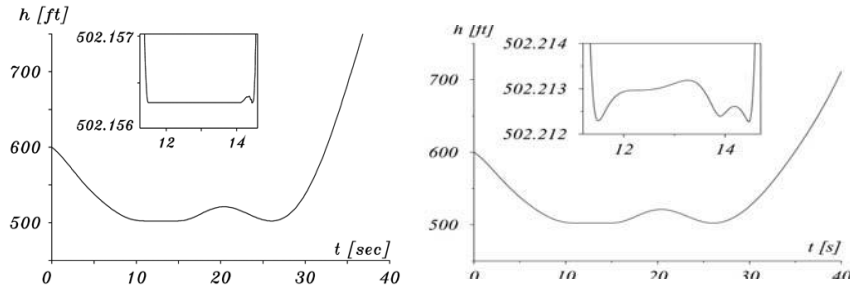


FIG. 4.2. *Altitude h for windshear problem. Left: multiple shooting result from* [6]. *Right: central path result at $\mu = 2.1 \cdot 10^{-4}$ (this paper).*

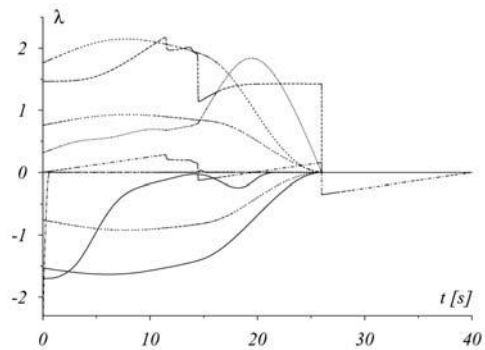


FIG. 4.3. *Lagrange multipliers λ_i corresponding to equality constraints (scaled).*

perfecting our algorithm, which will be filled in the near future.

Figure 4.2 shows a comparison of our altitude results with those obtained in [6]. The agreement is perfect within the interval up to the last touch point $t^* = 25.997$s. Beyond that point, there exists a continuum of optimal solutions. This can be understood from the fact that for $t > t^*$ all relevant Lagrange multipliers vanish (Figure 4.3) and none of the inequality constraints is active. Only the multiplier corresponding to the Chebyshev reformulation of the minimization problem does not vanish.

As already mentioned above, the most critical variable is the control u, the angle of attack rate. That is why we present its rather complex behavior in Figure 4.4. As can be seen, our results once again are in perfect agreement with the multiple shooting results from [6] at least in the relevant interval $[0, t^*]$. Only one slight deviation occurs before $t = t^*$, the reason for which is not yet clear. The second deviation, the downward spike starting at t^*, can be attributed to the nonuniqueness of the solution. In the interval $[t^*, 1]$, both methods happen to choose different, but equivalent, solutions. The pronounced peak of u at t^* reflects the tendency of the IP method to drive a previously actively constrained state variable towards the interior of the feasible set. Of course, this peak could have been suppressed by adding some penalty term ϵu^2 to the cost functional on the interval $[t^*, 1]$, without changing the uniquely determined part of the solution in the interval $[0, t^*]$.

In passing we note that Pytlak [32] has also attacked this problem using his method and documented his results, but he obtained a control behavior quite different from the one given in Figure 4.4.

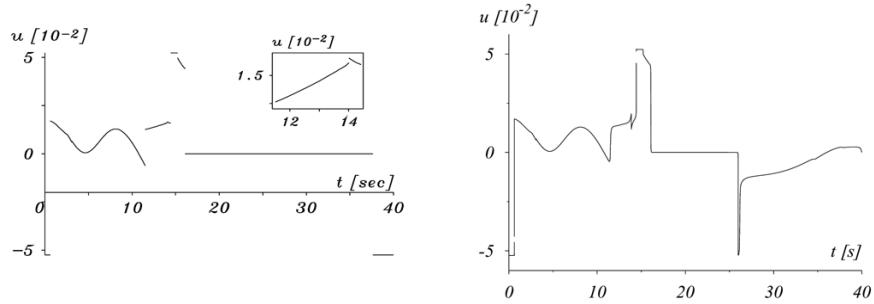


FIG. 4.4. *Control u for windshear problem. Left: multiple shooting result from* [6]. *Right: central path result at $\mu = 2.1 \cdot 10^{-4}$ (this paper).*

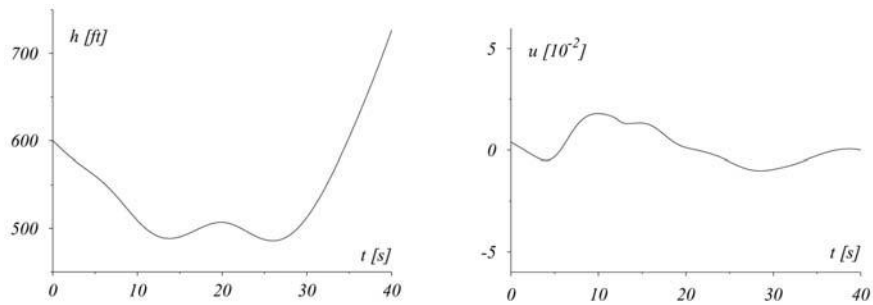


FIG. 4.5. *Initial trajectory for windshear problem. Left: altitude h. Right: control u.*

As for the obtained functional value (minimum altitude $h_{\min}$), Bulirsch, Montrone, and Pesch [11] report an improvement of 10ft over Miele et al. [30]. On top of that, Berkmann and Pesch [6] achieved a further improvement of $2.7 \cdot 10^{-6}$ ft to a value of $h_{\min} = 502.1562810$ ft. Our method led to an even better minimal altitude of $h_{\min} = 502.210661$ ft. In order to assess this value, we solved the initial value problem in both forward and backward directions using the computed control from our algorithm. As a numerical integrator we selected the MATLAB implementation of Dormand–Prince RK45. In the forward direction we obtained $h_{\min} = 502.210433$ ft, in the backward direction $h_{\min} = 502.210438$ ft. This seems to confirm that our results are an improvement even over [6]—within the discretization error level, of course.

In order to shed some light on the performance of our new algorithm, we next give some details about the continuation process with respect to the duality gap parameter μ and the adaptive multilevel scheme.

The computations were started on a uniform initial grid with mesh size $h_0 = 1/25$. On this grid, the nonlinear KKT equations $F(v; 1) = 0$ with dimension 2748 have been solved using a Newton method with damping. The corresponding initial trajectory is depicted in Figure 4.5.

An illustration of the adaptive continuation process along the central path is given in Figure 4.6. Assuming a convergence of the form

$$J(v(\mu)) - J(v(0)) \sim \mu^{\alpha}\,, \tag{4.3}$$

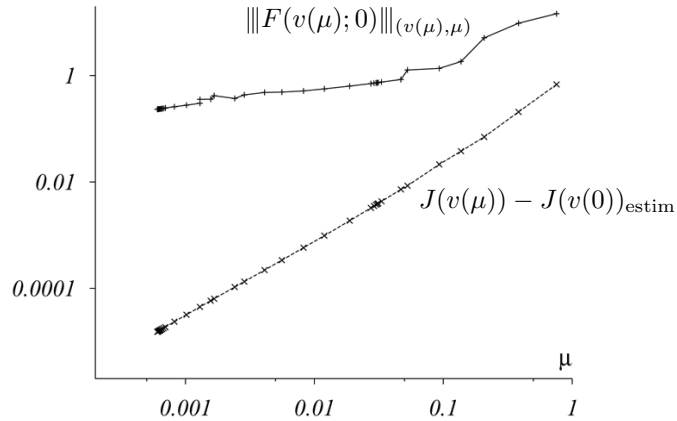


FIG. 4.6. *Central path continuation for windshear problem.*

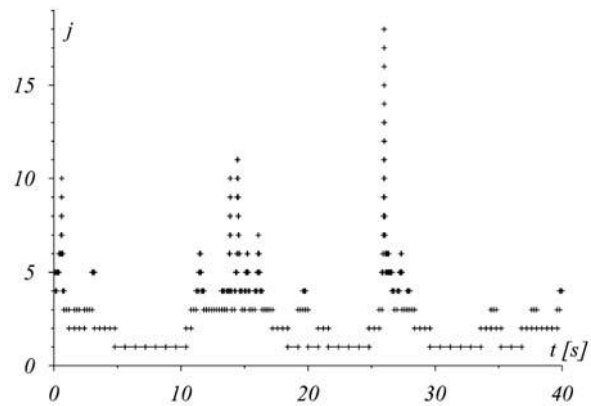


FIG. 4.7. *Adaptive mesh refinement in windshear problem.*

a simple parameter fitting yields $J(v(0))_{\text{estim}} \approx 5.022118$ and $\alpha \approx 1.44$. The log-log scale indicates that (4.3) is indeed quite reasonable.

Finally, the adaptive mesh refinement structure for this problem is presented in Figure 4.7. Successive refinement led to mesh sizes

$$h_j = 2^{-j} h_0 \,.$$

Obviously, the highly dynamic structure of the solution is captured reasonably well by the adaptive refinement procedure.

Conclusion. In this paper we present a direct function space method for optimal control problems based on the complementarity formulation of IP methods. The new method essentially dispenses with any analytical preprocessing—thus saving considerable human effort. In its algorithmic realization, function space is exploited via an adaptive multilevel method in combination with an adaptive central path following algorithm. A theoretical justification of the algorithm has been achieved only for control constrained problems. However, numerical results for a well-known intricate

optimal control problem with both control and state constraints seem to indicate that a much wider class of problems should be tractable by our algorithm. Even though a lot remains to be done both in theoretical justification and in algorithmic realization, we are confident to have opened a promising alternative path towards the numerical solution of complex optimal control problems from science and engineering.

Acknowledgments. The authors gratefully acknowledge invaluable helpful discussions with F. Tröltzsch on a former version of the paper. Moreover, they would like to thank H. Maurer and C. Helmberg for helpful comments.

REFERENCES

[1] W. Alt and K. Malanowski, *The Lagrange-Newton method for state constrained optimal control problems*, Comput. Optim. Appl., 4 (1995), pp. 217–239.

[2] U. Ascher, J. Christiansen, and R. Russell, *Collocation software for boundary-value ODEs*, ACM Trans. Math. Softw., 7 (1981), pp. 209–222.

[3] U. Ascher, R. Mattheij, and R. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1988.

[4] G. Bader and U. Ascher, *A new basis implementation for a mixed order boundary value ODE solver*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 483–500.

[5] M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch, *A comparison of a Moreau–Yosida-based active set strategy and interior point methods for constrained optimal control problems*, SIAM J. Optim., 11 (2000), pp. 495–521.

[6] P. Berkmann and H. Pesch, *Abort landing in windshear: An optimal control problem with third-order state constraint and varied switching structure*, J. Optim. Theory Appl., 85 (1995), pp. 21–57.

[7] H. Bock, *Numerische Behandlung von zustandsbeschränkten und Chebychef-Steuerungs-Problemen*, tech. rep., Carl-Cranz-Gesellschaft, Oberpfaffenhofen, Germany, 1981.

[8] H. Bock and K.-J. Plitt, *A multiple shooting algorithm for direct solution of optimal control problems*, in Proceedings of the 9th IFAC World Congress, Budapest, Pergamon Press, Oxford, 1984, pp. 242–247.

[9] R. Bulirsch, *Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen und Aufgaben der optimalen Steuerung*, tech. rep., Carl-Cranz-Gesellschaft, Oberpfaffenhofen, Germany, 1971.

[10] R. Bulirsch, F. Montrone, and H. Pesch, *Abort landing in the presence of windshear as a minimax optimal control problem.* I: *Necessary conditions*, J. Optim. Theory Appl., 70 (1991), pp. 1–23.

[11] R. Bulirsch, F. Montrone, and H. Pesch, *Abort landing in the presence of windshear as a minimax optimal control problem.* II: *Multiple shooting and homotopy*, J. Optim. Theory Appl., 70 (1991), pp. 223–254.

[12] R. Bulirsch, E. Nerz, H. Pesch, and O. von Stryk, *Combining direct and indirect methods in optimal control: Range maximization of a hang glider*, in Optimal Control, Internat. Ser. Numer. Math. 111, R. Bulirsch, A. Miele, J. Stoer, and K. H. Well, eds., Birkhäuser, Basel, 1993, pp. 273–288.

[13] J. Burke and S. Xu, *The global linear convergence of a noninterior path-following algorithm for linear complementarity problems*, Math. Oper. Res., 23 (1998), pp. 719–734.

[14] B. Chen and N. Xiu, *A global linear and local quadratic noninterior continuation method for nonlinear complementarity problems based on Chen–Mangasarian smoothing functions*, SIAM J. Optim., 9 (1999), pp. 605–623.

[15] C. Chen and O. L. Mangasarian, *A class of smoothing functions for nonlinear and mixed complementarity problems*, Comput. Optim. Appl., 5 (1996), pp. 97–138.

[16] P. Deuflhard, *A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with application to multiple shooting*, Numer. Math., 22 (1974), pp. 289–315.

[17] P. Deuflhard, *A stepsize control for continuation methods and its special application to multiple shooting techniques*, Numer. Math., 33 (1979), pp. 115–146.

[18] P. Deuflhard, *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, Comput. Math. 35, Springer, Berlin, 2004.

[19] P. Deuflhard and G. Heindl, *Affine invariant convergence theorems for Newton's method and extensions to related methods*, SIAM J. Numer. Anal., 16 (1979), pp. 1–10.

[20] P. Deuflhard and M. Weiser, *Local inexact Newton multilevel FEM for nonlinear elliptic problems*, in Computational Science for the 21st Century, M.-O. Bristeau, G. Etgen, W. Fitzgibbon, J.-L. Lions, J. Periaux, and M. Wheeler, eds., Wiley, Chichester, UK, 1997, pp. 129–138.
[21] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.
[22] W. Hager, *Runge-Kutta methods in optimal control and the transformed adjoint system*, Numer. Math., 87 (2000), pp. 247–282.
[23] W. Hager, *Numerical analysis in optimal control*, in Proceedings of the Conference on Optimal Control of Complex Structures, Internat. Ser. Numer. Math. 139, K.-H. Hoffmann, I. Lasiecka, G. Leugering, J. Sprekels, and F. Tröltzsch, eds., Birkhäuser, Basel, 2001, pp. 83–93.
[24] A. Hohmann, *Inexact Gauss Newton Methods for Parameter Dependent Nonlinear Problems*, Ph.D. thesis, Free University of Berlin, Berlin, Germany, 1994.
[25] S. Ito, C. Kelley, and E. Sachs, *Inexact primal-dual interior point iteration for linear programs in function spaces*, Comput. Optim. Appl., 4 (1995), pp. 189–201.
[26] F. Jarre, *Comparing Two Interior-Point Approaches for Semi-infinite Programs*, tech. rep., Universität Trier, Trier, Germany, 1999.
[27] C. Kanzow, *Some noninterior continuation methods for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 851–868.
[28] S. Kurcyusz, *On the existence and nonexistence of Lagrange multipliers in Banach spaces*, J. Optim. Theory Appl., 20 (1976), pp. 81–110.
[29] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Stud., 14 (1981), pp. 163–177.
[30] A. Miele, T. Wang, C. Tzeng, and W. Melvin, *Optimal abort landing trajectories in the presence of a windshear*, J. Optim. Theory Appl., 55 (1987), pp. 165–202.
[31] F. Potra, *A Path-Following Method for Linear Complementarity Problems Based on the Affine Invariant Kantorovich Theorem*, ZIB Report 00-30, Zuse Institute Berlin, Berlin, Germany, 2000.
[32] R. Pytlak, *Numerical Methods for Optimal Control Problems with State Constraints*, Lecture Notes in Math. 1707, Springer, New York, 1999.
[33] V. Schulz, *Solving discretized optimization problems by partially reduced SQP methods*, Comput. Vis. Sci., 1 (1998), pp. 83–96.
[34] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer, New York, 1993.
[35] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–842.
[36] M. Ulbrich and S. Ulbrich, *Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds*, SIAM J. Control Optim., 38 (2000), pp. 1938–1984.
[37] M. Ulbrich, S. Ulbrich, and M. Heinkenschloss, *Global convergence of trust-region interior-point algorithms for infinite-dimensional nonconvex minimization subject to pointwise bounds*, SIAM J. Control Optim., 37 (1999), pp. 731–764.
[38] S. Volkwein and M. Weiser, *Affine invariant convergence analysis for inexact augmented Lagrangian-SQP methods*, SIAM J. Control Optim., 41 (2002), pp. 875–899.
[39] O. von Stryk and R. Bulirsch, *Direct and indirect methods for trajectory optimization*, Ann. Oper. Res., 37 (1992), pp. 357–373.
[40] M. Weiser, *Function Space Complementarity Methods for Optimal Control Problems*, doctoral thesis, Free University of Berlin, Berlin, Germany, 2001.
[41] M. Weiser, *Linear Convergence of an Interior Point Method for Linear Control Constrained Optimal Control Problems*, ZIB Report 02-13, Zuse Institute Berlin, Berlin, Germany, 2002.
[42] M. Weiser, *Interior point methods in function space*, SIAM J. Control Optim., 44 (2005), pp. 1766–1786.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 816–838

DIFFERENTIAL GAMES WITH ASYMMETRIC INFORMATION*

P. CARDALIAGUET†

Abstract. We investigate a two-player zero-sum differential game in which the players have asymmetric information on the random terminal payoff. We prove that the game has a value and characterize this value in terms of *dual* solutions of some Hamilton–Jacobi equation. We also explain how to adapt the results to differential games where the initial position is random.

Key words. differential game, asymmetric information, viscosity solution

AMS subject classifications. 49N70, 49L25, 91A23

DOI. 10.1137/060654396

1. Introduction.

In this paper we investigate a two-player zero-sum differential game in which the players have asymmetric information on the random terminal payoff. The dynamics of the game is given by

$$\left\{\begin{array}{ll} x'(t) = f(x(t), u(t), v(t)) , & u(t) \in U,\ v(t) \in V, \\ x(t_0) = x_0, \end{array}\right. \tag{1}$$

where U and V are compact subsets of some finite dimensional spaces, and $f : \mathbb{R}^N \times U \times V \to \mathbb{R}^N$ is Lipschitz continuous. We consider a finite horizon problem with a terminal time denoted by T. The game starts at time $t_0 \in [0, T]$ from the initial position x_0.

The description of the game involves $I \times J$ terminal payoffs (where $I, J \geq 1$): $g_{ij} : \mathbb{R}^N \to \mathbb{R}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, a probability $p = (p_i)_{i=1,\ldots,I}$ belonging to the set $\Delta(I)$ of probabilities on $\{1, \ldots, I\}$ and a probability $q = (q_j)_{j=1,\ldots,J}$ of the set $\Delta(J)$ of probabilities on $\{1, \ldots, J\}$.

The game is played in two steps: At time t_0, a pair (i, j) is chosen at random among $\{1, \ldots, I\} \times \{1, \ldots, J\}$ according to the probability $p \otimes q$; the choice of i is communicated to player 1 only, while the choice of j is communicated to player 2 only.

Then the players control system (1) in order for player 1 to minimize the terminal payoff $g_{ij}(x(T))$ and for player 2 to maximize it. We assume that both players observe their opponent's control. Note, however, that the players do not know which g_{ij} they are actually optimizing, because they have only a part of the information on the 91A05 pair (i, j). They can nevertheless try to guess their missing information by observing what their opponent is doing. Indeed, in order to use his information a player necessarily reveals at least a part of it, and any piece of information he reveals can be later exploited by his opponent.

As usual we introduce two value functions associated to this game. We have to take special care of the way we define the strategies of the players, since this definition has to represent the lack of symmetry in the knowledge of the players.

*Received by the editors March 16, 2006; accepted for publication (in revised form) February 1, 2007; published electronically May 29, 2007.
http://www.siam.org/journals/sicon/46-3/65439.html

†Université de Bretagne Occidentale, Laboratoire de Mathématiques, UMR 6205, 6 Av. Le Gorgeu, BP 809, 29285 Brest, France (pierre.cardaliaguet@univ.brest.fr).

The upper value is given by

$$V^+(t_0,x_0,p,q) = \inf_{(\alpha_i)\in(\mathcal{A}_r(t_0))^I} \sup_{(\beta_j)\in(\mathcal{B}_r(t_0))^J} \sum_{i=1}^{I}\sum_{j=1}^{J} p_i q_j \mathbf{E}_{\alpha_i\beta_j}\left(g_{ij}\left(X_T^{t_0,x_0,\alpha_i,\beta_j}\right)\right) ,$$

where the $\alpha_i \in \mathcal{A}_r(t_0)$ (for $i = 1,\dots,I$) are I random strategies for player 1, the $\beta_j \in \mathcal{B}_r(t_0)$ (for $j = 1,\dots,J$) are J random strategies for player 2, and $\mathbf{E}_{\alpha_i\beta_j}(g_{ij}(X_T^{t_0,x_0,\alpha_i,\beta_j}))$ is the payoff associated with the pair of strategies (α_i,β_j) for the terminal payoff g_{ij}: These notions are explained in the next section. The key point in the definition is that player 1 chooses his strategy α_i $(i = 1,\dots,I)$ according to the value of the index i only, while player 2 has a strategy (β_j) which depends only upon the index j. This reflects the asymmetry of information of the players. The sum $\sum_i \sum_j p_i q_j \dots$ is the expectation of the payoff when the pair (i,j) is chosen according to the probability $p \otimes q$, where $p = (p_1,\dots,p_I)$ and $q = (q_1,\dots,q_J)$.

The lower value is defined by the symmetric formula:

$$V^-(t_0,x_0,p,q) = \sup_{(\beta_j)\in(\mathcal{B}_r(t_0))^J} \inf_{(\alpha_i)\in(\mathcal{A}_r(t_0))^I} \sum_{i=1}^{I}\sum_{j=1}^{J} p_i q_j \mathbf{E}_{\alpha_i\beta_j}\left(g_{ij}\left(X_T^{t_0,x_0,\alpha_i,\beta_j}\right)\right) .$$

Obviously we have

$$V^-(t_0,x_0,p,q) \le V^+(t_0,x_0,p,q)$$

for any $(t_0,x_0) \in [0,T]\times\mathbb{R}^N$, any probability $p \in \Delta(I)$ on $\{1,\dots,I\}$, and any probability $q \in \Delta(J)$ on $\{1,\dots,J\}$. Our aim is to show that the equality holds, i.e., that the game has a value, and to provide a PDE characterization of the value.

The game studied in this paper is strongly inspired by *repeated games* with a lack of information introduced by Aumann and Maschler; see [2, 22] for a general presentation. Repeated games with a lack of information on one side (i.e., $I = 1$ or $J = 1$) or on both sides (i.e., $I, J \ge 2$) have a value [2, 18], in the sense that the averaged n-stage games converge to a limit as $n \to +\infty$. This value can be characterized in terms of the value of the "nonrevealing game." In this paper, we prove the existence of a value for differential games with a lack of information on both sides. However, we show in the companion paper [10] that the characterization in terms of a game without information does not hold. In that respect, our game is close to stochastic games with incomplete information, as studied in [21], for instance. Although it is known that stochastic games with a lack of information on one side have a value when the game is controlled by the informed player only [21], the general case is still open.

There are several proofs of Aumann and Maschler's result. In order to show that our game has a value, we use a strategy of proof initiated by De Meyer in [12] and later developed in [13, 14, 17]. We first note that the maps $V^+ = V^+(t,x,p,q)$ and $V^- = V^-(t,x,p,q)$ are convex in p and concave in q (Lemma 3.2). This leads us to introduce, for a generic map $w : [0,T]\times\mathbb{R}^N\times\Delta(I)\times\Delta(J) \mapsto \mathbb{R}$, the convex Fenchel conjugate w^* of w with respect to the variable p and its concave conjugate $w^\sharp$ with respect to q: for all $(t,x,\hat{p},q) \in [0,T]\times\mathbb{R}^N\times\mathbb{R}^I\times\Delta(J)$

$$w^*(t,x,\hat{p},q) = \sup_{p\in\Delta(I)} p.\hat{p} - w(t,x,p,q)$$

and for all $(t,x,p,\hat{q}) \in [0,T]\times\mathbb{R}^N\times\Delta(I)\times\mathbb{R}^J$

$$w^\sharp(t,x,p,\hat{q}) = \inf_{q\in\Delta(J)} q.\hat{q} - w(t,x,p,q) .$$

Then the proof of the equality $V^+ = V^-$ runs as follows: We first check (Lemma 4.2) that V^{-*} satisfies a subdynamic programming principle and thus (Corollary 4.3) that $(t,x) \mapsto V^{-*}(t,x,\hat{p},q)$ is a viscosity subsolution of the (dual) Hamilton–Jacobi (HJ) equation

$$w_t + H^*(x, Dw) = 0 \text{ in } [0,T] \times \mathbb{R}^N \tag{2}$$

for any $(\hat{p},q) \in \mathbb{R}^I \times \Delta(J)$. The map H^* is defined through the standard Hamiltonian H of the game

$$H(x,\xi) = \inf_{u\in U} \sup_{v\in V} f(x,u,v).\xi = \sup_{v\in V} \inf_{u\in U} f(x,u,v).\xi$$

via the relation by $H^*(x,\xi) = -H(x,-\xi)$. Note that we assume that Isaacs' condition holds. We recall that the notion of viscosity solutions was introduced by Crandall and Lions in [11] and first used in the framework of differential games in [15] (see also [3, 4] for a general presentation). We also establish a symmetric result for $V^{+\sharp}$ (Corollary 4.4): For any $(p,\hat{q}) \in \Delta(I) \times \mathbb{R}^J$, the map $(t,x) \mapsto V^{+\sharp}(t,x,p,\hat{q})$ is a viscosity supersolution of the same equation (2). A new comparison principle (Theorem 5.1) then implies that $V^+ \le V^-$. Since inequality $V^+ \ge V^-$ is obvious, the game has a value: $V^+ = V^-$. We also have the following characterization of this value: $\mathbf{V} := V^+ = V^-$ is the unique Lipschitz continuous function which is convex in p, concave in q, such that $(t,x) \mapsto \mathbf{V}^*(t,x,\hat{p},q)$ is a subsolution of the HJ equation (2), while $(t,x) \mapsto \mathbf{V}^\sharp(t,x,p,\hat{q})$ is a supersolution of (2). We call such a function the *dual solution* to the Hamilton–Jacobi equation

$$\begin{cases} w_t + H(x,Dw) = 0 & \text{in } [0,T) \times \mathbb{R}^N, \\ w(T,x) = \sum_{ij} p_i q_j g_{ij}(x) & \text{in } \mathbb{R}^N. \end{cases}$$

We discuss this terminology below.

We explain in section 6 how to adapt our approach to differential games with a lack of information on the initial positions. As previously, the game is played in two steps. At time t_0, the initial position of the game is chosen at random among $I \times J$ possible initial positions x^0_{ij} according to a probability $p \otimes q$, where $p \in \Delta(I)$ and $q \in \Delta(J)$; the index i is communicated to player 1, while the index j is communicated to player 2. Then the players control system (1) in order, for player 1, to minimize a terminal payoff $g(x(T))$ and, for player 2, to maximize it. The key assumption is that the players observe their opponent's behavior but not the state of the system $x(\cdot)$. We prove that this game has a value, which can be characterized as the unique dual solution of some HJ equation in $[0,T] \times \mathbb{R}^{NIJ}$.

Although there have been several attempts to formalize differential games with lack of information [5, 6, 7, 8], there are only very few papers in which a game is proved to have a value; see in particular [19], [20], which discuss interesting examples. In [9] we consider a game with a lack of information on the current position, but with symmetric information. To the best of our knowledge, our result is the first one showing the existence of a value for differential games with asymmetry in the information in a general setting.

The kind of characterization proposed in this paper for the value function (as a dual solution of some Hamilton–Jacobi equations) is also new. It relies upon a new comparison principle (Theorem 5.1) stating the following: Assume that w_1 and w_2 defined on $[0,T] \times \mathbb{R}^N \times \Delta(I) \times \Delta(J)$ are convex in p, concave in q, that $(t,x) \mapsto w_1^\sharp(t,x,p,\hat{q})$ is a supersolution of the dual HJ equation (2) for any $(p,\hat{q})$, and that

$(t,x) \mapsto w_2^*(t,x,\hat{p},q)$ is a subsolution of this HJ equation for any $(\hat{p},q)$. If futhermore $w_1(T,x,p,q) \leq w_2(T,x,p,q)$ for any (x,p,q), then $w_1 \leq w_2$.

Note that the function w_2, for instance, is a kind of supersolution for our problem. For this reason we call it a dual supersolution of the orginal HJ equation

$$w_t + H(x, Dw) = 0 \text{ in } [0,T] \times \mathbb{R}^N, \tag{3}$$

and we see the HJ equation (2) as a dual one. Let us recall that, although the Fenchel conjugate of a supersolution of (3) is a subsolution of the dual equation (2) (see [1]), the converse does not hold in general. In fact we show through several examples in [10] that the value function $\mathbf{V} := V^+ = V^-$ of our game is *not* a solution of the original HJ equation (3), nor are its Fenchel conjugates $\mathbf{V}^*$ and $\mathbf{V}^\sharp$ solutions of the dual one (2). The particular structure of our problem leads us to replace the classical notion of sub- and supersolutions by a weaker one, involving families of sub- and supersolutions in some dual spaces (see also Lemma 5.4, where an equivalent definition for a dual subsolution is discussed).

We complete this introduction by describing the organization of the paper. In section 2, we introduce the main notations: In particular we explain the notions of random strategies and define the value functions of our game. Section 3 is mainly devoted to the proof of the convexity properties of the value functions. In section 4 we show that V^{-*} satisfies a subdynamic programming principle and is a subsolution of the dual HJ equation and give the corresponding results for $V^{+\sharp}$. Section 5 is devoted to the comparison principle and to the existence of a value. In the last section, we extend our results to differential games with a lack of information on the initial position.

2. Definition of the value functions.

Notations. Throughout the paper, $x.y$ denotes the scalar product in the space $\mathbb{R}^N$, $\mathbb{R}^I$, or $\mathbb{R}^J$ (depending on the context) and $|\cdot|$ the Euclidean norm. The ball of center x and radius r will be denoted by $B_r(x)$. If E is a set, then $\mathbf{1}_E$ is the indicatrix function of E (equal to 1 if E and to 0 outside of E). The set $\Delta(I)$ is the set of probability measures on $\{1,\dots,I\}$, always identified with the simplex of $\mathbb{R}^I$:

$$p = (p_1,\dots,p_I) \in \Delta(I) \quad \Leftrightarrow \quad \sum_{i=1}^{I} p_i = 1 \text{ and } p_i \geq 0 \text{ for } i = 1,\dots,I\,.$$

The set $\Delta(J)$ of probability measures on $\{1,\dots,J\}$ is defined symmetrically.

The dynamics of the game is given by

$$\begin{cases} x'(t) = f(x(t),u(t),v(t))\,, & u(t) \in U,\ v(t) \in V, \\ x(t_0) = x_0. \end{cases} \tag{4}$$

Throughtout the paper we assume that

$$\begin{cases} \text{(i)} & U \text{ and } V \text{ are compact subsets of some finite dimensional spaces;} \\ \text{(ii)} & f : \mathbb{R}^N \times U \times V \to \mathbb{R}^N \text{ is bounded, continuous, and uniformly} \\ & \text{Lipschitz continuous with respect to the } x \text{ variable;} \\ \text{(iii)} & \text{for } i = 1,\dots,I \text{ and } j = 1,\dots,J,\ g_{ij} : \mathbb{R}^N \to \mathbb{R} \text{ is Lipschitz} \\ & \text{continuous and bounded.} \end{cases} \tag{5}$$

We also assume that Isaacs' condition holds:

$$H(x,\xi) := \inf_{u\in U} \sup_{v\in V} f(x,u,v).\xi = \sup_{v\in V} \inf_{u\in U} f(x,u,v).\xi \tag{6}$$

for any $(x,\xi) \in \mathbb{R}^N \times \mathbb{R}^N$. We note that the Hamilton–Jacobi equation naturally associated with the dynamics is the so-called primal Hamilton–Jacobi equation

$$w_t + H(x, Dw) = 0 \qquad \text{in } [0,T) \times \mathbb{R}^N. \tag{7}$$

For any $t_0 < t_1 \le T$, the set of open-loop controls for player 1 on $[t_0,t_1]$ is defined by

$$\mathcal{U}(t_0,t_1) = \{u : [t_0,t_1] \mapsto U \text{ Lebesgue measurable}\} .$$

If $t_1 = T$, we simply set $\mathcal{U}(t_0) := \mathcal{U}(t_0,T)$. Open-loop controls on the interval $[t_0,t_1]$ for player 2 are defined symmetrically and denoted by $\mathcal{V}(t_0,t_1)$ (and by $\mathcal{V}(t_0)$ if $t_1 = T$).

If $u \in \mathcal{U}(t_0)$ and $t_0 \le t_1 < t_2 \le T$, we denote by $u_{|_{[t_1,t_2]}}$ the restriction of u to the interval $[t_1,t_2]$. We note that $u_{|_{[t_1,T]}}$ belongs to $\mathcal{U}(t_1)$.

For any $(u,v) \in \mathcal{U}(t_0) \times \mathcal{V}(t_0)$ and any initial position $x_0 \in \mathbb{R}^N$, we denote by $t \mapsto X_t^{t_0,x_0,u,v}$ the solution to (4).

Next we introduce the notions of pure and mixed strategies. The definition of mixed strategies involves a set $\mathcal{S}$ of (nontrivial) probability spaces, which has to be stable by a finite product. To fix the ideas we choose from now on

$$\mathcal{S} = \{([0,1]^n, B([0,1]^n), \mathcal{L}^n) \text{ for some } \ n \in \mathbb{N}^*\} \ ,$$

where $B([0,1]^n)$ is the class of Borel sets and $\mathcal{L}^n$ is the Lebesgue measure on $\mathbb{R}^n$. As the reader can easily check, the results presented in this paper do not depend on this particular choice of $\mathcal{S}$.

DEFINITION 2.1 (pure and random strategies). *A pure strategy for player* 1 *at time* t_0 *is a map* $\alpha : \mathcal{V}(t_0) \mapsto \mathcal{U}(t_0)$ *which satisfies the following conditions:*

(i) α *is a measurable map from* $\mathcal{V}(t_0)$ *to* $\mathcal{U}(t_0)$*, where* $\mathcal{U}(t_0)$ *and* $\mathcal{V}(t_0)$ *are endowed with the Borel* σ*-field associated with the* L^1 *distance;*

(ii) α *is nonanticipative with delay; i.e., there is some* $\tau > 0$ *such that, for any* $v_1, v_2 \in \mathcal{V}(t_0)$*, if* $v_1 \equiv v_2$ *a.e. on* $[t_0,t]$ *for some* $t \in (t_0, T-\tau)$*, then* $\alpha(v_1) \equiv \alpha(v_2)$ *a.e. on* $[t_0, t+\tau]$.

A random strategy for player 1 *is a pair* $((\Omega_\alpha, \mathcal{F}_\alpha, \mathbf{P}_\alpha), \alpha)$*, where* $(\Omega_\alpha, \mathcal{F}_\alpha, \mathbf{P}_\alpha)$ *belongs to the set of probability spaces* $\mathcal{S}$ *and* $\alpha : \Omega_\alpha \times \mathcal{V}(t_0) \mapsto \mathcal{U}(t_0)$ *satisfies the following conditions:*

(i) α *is a measurable map from* $\Omega_\alpha \times \mathcal{V}(t_0)$ *to* $\mathcal{U}(t_0)$*, with* Ω_α *endowed with the* σ*-field* $\mathcal{F}_\alpha$ *and* $\mathcal{U}(t_0)$ *and* $\mathcal{V}(t_0)$ *with the Borel* σ*-field associated with the* L^1 *distance;*

(ii) *there is some delay* $\tau > 0$ *such that, for any* $v_1, v_2 \in \mathcal{V}(t_0)$*, any* $t \in (t_0, T-\tau)$*, and any* $\omega \in \Omega_\alpha$*,*

$$v_1 \equiv v_2 \text{ on } [t_0,t) \ \Rightarrow \ \alpha(\omega, v_1) \equiv \alpha(\omega, v_2) \text{ on } [t_0, t+\tau) \ .$$

We denote by $\mathcal{A}(t_0)$ the set of pure strategies and by $\mathcal{A}_r(t_0)$ the set of random strategies for player 1. By abuse of notations, an element of $\mathcal{A}_r(t_0)$ is simply noted α—instead of $((\Omega_\alpha, \mathcal{F}_\alpha, \mathbf{P}_\alpha), \alpha)$—the underlying probability space being always denoted by $(\Omega_\alpha, \mathcal{F}_\alpha, \mathbf{P}_\alpha)$. Let us point out the inclusion $\mathcal{A}(t_0) \subset \mathcal{A}_r(t_0)$.

In order to take into account the fact that player 1 knows the index i of the terminal payoff, a strategy for player 1 is actually an I-uplet $\hat{\alpha} = (\alpha_1, \dots, \alpha_I) \in (\mathcal{A}_r(t_0))^I$.

Pure and random strategies for player 2 are defined symmetrically: At time t_0, a pure strategy β is a measurable map which is nonanticipative with delay from $\mathcal{U}(t_0)$ to $\mathcal{V}(t_0)$, while a random strategy is a map $\beta : \Omega_\beta \times \mathcal{U}(t_0) \mapsto \mathcal{V}(t_0)$, where $(\Omega_\beta, \mathcal{F}_\beta, \mathbf{P}_\beta)$ belongs to $\mathcal{S}$, which satisfies the conditions:

(i) β is measurable from $\Omega_\beta \times \mathcal{U}(t_0)$ to $\mathcal{V}(t_0)$;

(ii) there is some delay $\tau > 0$ such that, for any $u_1, u_2 \in \mathcal{U}(t_0)$, any $t \in (t_0, T-\tau)$, and any $\omega \in \Omega_\beta$,

$$u_1 \equiv u_2 \text{ on } [t_0, t) \;\Rightarrow\; \beta(\omega, u_1) \equiv \beta(\omega, u_2) \text{ on } [t_0, t+\tau) \; .$$

The set of pure and random strategies for player 2 are denoted by $\mathcal{B}(t_0)$ and $\mathcal{B}_r(t_0)$, respectively. Elements of $\mathcal{B}_r(t_0)$ are denoted simply by β and the underlying probability space by $(\Omega_\beta, \mathcal{F}_\beta, \mathbf{P}_\beta)$. Since player 2 knows the index j of the terminal payoff, a strategy for player 2 is a J-uplet $\hat{\beta} = (\beta_1, \dots, \beta_J) \in (\mathcal{B}_r(t_0))^J$.

Let us now comment upon these definitions of strategies. In the recent literature on differential games, one generally uses "nonanticipative strategies" or "Varayia–Roxin–Elliott–Kalton" strategies, i.e., pure strategies without delay (see [3, 15], for instance). In this setting, no measurability condition is required upon the strategy. We have chosen here to work with "nonanticipative strategies with delay" because using these strategies allows us to put the game under normal form: See Lemma 2.2 below; standard nonanticipative strategies do not enjoy this crucial property. The introduction of random strategies is new in the context of differential games, in which one usually has the existence of a value in pure strategies provided Isaacs' condition (6) holds. In the game considered in this paper, the players have to hide their private information. To do this they have to use some randomness in their strategies. From a technical point of view, this randomness is the key argument for the convexity properties of the value functions to hold; see the proof of Lemmas 3.2 and 4.1 in particular. The measurability requirement of a random strategy with respect to ω is necessary to be able to define an expectation. The joint measurability in (i) of the definition is a technical requirement needed in the proof of the dynamic programming. One also has to face such a difficulty for stochastic differential games; see [16], for instance.

LEMMA 2.2. *For any pair $(\alpha, \beta) \in \mathcal{A}_r(t_0) \times \mathcal{B}_r(t_0)$ and any $\omega := (\omega_1, \omega_2) \in \Omega_\alpha \times \Omega_\beta$, there is a unique pair $(u_\omega, v_\omega) \in \mathcal{U}(t_0) \times \mathcal{V}(t_0)$ such that*

$$\alpha(\omega_1, v_\omega) = u_\omega \text{ and } \beta(\omega_2, u_\omega) = v_\omega \; . \tag{8}$$

Furthermore the map $\omega \to (u_\omega, v_\omega)$ is measurable from $\Omega_\alpha \times \Omega_\beta$ endowed with $\mathcal{F}_\alpha \otimes \mathcal{F}_\beta$ into $\mathcal{U}(t_0) \times \mathcal{V}(t_0)$ endowed with the Borel σ-field associated with the L^1 distance.

Notations. Given any pair $(\alpha, \beta) \in \mathcal{A}_r(t_0) \times \mathcal{B}_r(t_0)$, we denote by $(X_t^{t_0,x_0,\alpha,\beta})$ the map $(t, \omega) \mapsto X_t^{t_0,x_0,u_\omega,v_\omega}$ defined on $[t_0, T] \times \Omega_\alpha \times \Omega_\beta$, where (u_ω, v_ω) satisfies (8). We also define the expectation $\mathbf{E}_{\alpha\beta}$ as the integral over $\Omega_\alpha \times \Omega_\beta$ against the probability measure $\mathbf{P}_\alpha \otimes \mathbf{P}_\beta$. In particular, if $\phi : \mathbb{R}^N \to \mathbb{R}$ is some bounded continuous map and $t \in (t_0, T]$, we have

$$\mathbf{E}_{\alpha\beta}\left(\phi\left(X_t^{t_0,x_0,\alpha,\beta}\right)\right) := \int_{\Omega_\alpha \times \Omega_\beta} \phi\left(X_t^{t_0,x_0,u_\omega,v_\omega}\right) d\mathbf{P}_\alpha \otimes \mathbf{P}_\beta(\omega) \; , \tag{9}$$

where (u_ω, v_ω) is defined by (8). Note that (9) makes sense because, the map $(u, v) \mapsto X_t^{t_0,x_0,u,v}$ being continuous in L^1, the map $\omega \mapsto \phi\left(X_t^{t_0,x_0,u_\omega,v_\omega}\right)$ is measurable in $\Omega_\alpha \times \Omega_\beta$ and bounded. If either α or β is a pure strategy, then we simply drop α or β in the expectation $\mathbf{E}_{\alpha\beta}$, which then becomes $\mathbf{E}_\beta$ or $\mathbf{E}_\alpha$.

Proof of Lemma 2.2. The existence of (u_ω, v_ω) is simply due to the delay and proved in detail in [9]. We show here only the measurability of $\omega \to (u_\omega, v_\omega)$. For this we argue by induction by proving that the map $\omega \to (u_\omega, v_\omega)_{|[t_0,t_0+n\tau]}$ from $\Omega_\alpha \times \Omega_\beta$ into $L^1([t_0, t_0 + n\tau])$ is measurable, where τ denotes the minimum of the delays for α and β (see condition (ii) in Definition 2.1).

Let us start with $n = 1$. It is enough to show that, for any Borel subsets B_1 and B_2 of $\mathcal{U}(t_0, t_0 + \tau)$ and $\mathcal{V}(t_0, t_0 + \tau)$, the set

$$\Omega := \{\omega \in \Omega_\alpha \times \Omega_\beta \mid (u_\omega, v_\omega)_{|[t_0,t_0+\tau)} \in B_1 \times B_2\}$$

is measurable. Let us fix $\hat{u}$ and $\hat{v}$ in $\mathcal{U}(t_0)$ and $\mathcal{V}(t_0)$. Since $\alpha(\omega_1, \cdot)$ and $\beta(\omega_2, \cdot)$ are nonanticipative with delay τ, the restrictions of $\alpha(\omega_1, \hat{v})$ and $\beta(\omega_2, \hat{u})$ to $[t_0, t_0 + \tau]$ do not depend on $\hat{u}$ and $\hat{v}$. Hence $(u_\omega, v_\omega) \equiv (\alpha(\omega, \hat{v}), \beta(\omega, \hat{u}))$ a.e. in $[t_0, t_0 + \tau)$. Therefore

$$\Omega = \{\omega_1 \in \Omega_\alpha \mid \alpha(\omega_1, \hat{v})_{|[t_0,t_0+\tau)} \in B_1\} \times \{\omega_2 \in \Omega_\beta \mid \beta(\omega_2, \hat{u})_{|[t_0,t_0+\tau)} \in B_2\} ,$$

which is measurable since α and β are measurable. So the result holds true for $n = 1$.

Let us now assume that $\omega \to (u_\omega, v_\omega)_{|[t_0,t_0+n\tau]}$ from $\Omega_\alpha \times \Omega_\beta$ into $L^1([t_0, t_0 + n\tau])$ is measurable, and let us show that this still holds true for $n+1$. It is again enough to show that, for any Borel subsets B_1 and B_2 of $\mathcal{U}(t_0, t_0+(n+1)\tau)$ and $\mathcal{V}(t_0, t_0+(n+1)\tau)$, the set

$$\Omega := \{\omega \in \Omega_\alpha \times \Omega_\beta \mid (u_\omega, v_\omega)_{|[t_0,t_0+(n+1)\tau)} \in B_1 \times B_2\}$$

is measurable. Let us fix again $\hat{u}$ and $\hat{v}$ in $\mathcal{U}(t_0)$ and $\mathcal{V}(t_0)$. For any $(u, v) \in \mathcal{U}(t_0, t_0 + n\tau) \times \mathcal{V}(t_0, t_0+n\tau)$, we denote by $\tilde{u}$ and $\tilde{v}$ the maps equal to u and v on $[t_0, t_0+n\tau]$ and to $\hat{u}$ and $\hat{v}$ on $[t_0+n\tau, T]$. Note that $(u, v) \mapsto (\tilde{u}, \tilde{v})$ is continuous from L^1 to L^1. Since α and β are nonanticipative with delay τ, $u_\omega \equiv \alpha(\omega_1, \widetilde{v_\omega})$ on $[t_0, t_0 + (n + 1)\tau)$ and $v_\omega \equiv \beta(\omega_1, \widetilde{u_\omega})$ on $[t_0, t_0 + (n + 1)\tau)$. Therefore Ω is the preimage of the set $B_1 \times B_2$ by the map $\omega \to (\alpha(\omega_1, \widetilde{v_\omega}), \beta(\omega_2, \widetilde{u_\omega}))$, which is measurable as the composition of the measurable maps $\omega \mapsto (u_\omega, v_\omega)_{|[t_0,t_0+n\tau]}$, the map $(u, v) \mapsto (\tilde{u}, \tilde{v})$, and the maps α and β. Hence Ω is measurable, and the result is proved. □

We now define the payoff associated with a strategy $\hat{\alpha}$ of player 1 and a strategy $\hat{\beta}$ of player 2.

Definition of the payoff: Let $(p, q) \in \Delta(I) \times \Delta(J)$, $(t_0, x_0) \in [0, T) \times \mathbb{R}^N$, $\hat{\alpha} = (\alpha_i)_{i=1,\dots,I} \in (\mathcal{A}_r(t_0))^I$, and $\hat{\beta} = (\beta_j) \in (\mathcal{B}_r(t_0))^J$. We set

$$\mathcal{J}(t_0, x_0, \hat{\alpha}, \hat{\beta}, p, q) = \sum_{i=1}^{I} \sum_{j=1}^{J} p_i q_j \mathbf{E}_{\alpha_i \beta_j} \left(g_{ij} \left(X_T^{t_0, x_0, \alpha_i, \beta_j} \right) \right) , \tag{10}$$

where $\mathbf{E}_{\alpha_i \beta_j}$ is defined by (9). Note that $\hat{\alpha}$ does not depend on j, while $\hat{\beta}$ does not depend on i, which formalizes the asymmetry of information.

Definition of the value functions: The upper value function is given by

$$V^+(t_0, x_0, p, q) = \inf_{\hat{\alpha} \in (\mathcal{A}_r(t_0))^I} \sup_{\hat{\beta} \in (\mathcal{B}_r(t_0))^J} \mathcal{J}(t_0, x_0, \hat{\alpha}, \hat{\beta}, p, q),$$

while the lower value function is defined by

$$V^-(t_0, x_0, p, q) = \sup_{\hat{\beta} \in (\mathcal{B}_r(t_0))^J} \inf_{\hat{\alpha} \in (\mathcal{A}_r(t_0))^I} \mathcal{J}(t_0, x_0, \hat{\alpha}, \hat{\beta}, p, q) .$$

Let us emphasize that, because of the special form of the payoff, the value functions defined above *cannot* be recasted in terms of usual value functions of a zero-sum differential game with perfect information. For instance, they do not satisfy the standard dynamic programming principle, as we show in the companion paper [10].

3. Convexity properties of the value functions. The main result of this section is Lemma 3.2, which states that the value functions V^+ and V^- are convex in p and concave in q. We also investigate some regularity properties of the value functions.

LEMMA 3.1 (regularity of V^+ and V^-). *Under assumption* (5), V^+ *and* V^- *are Lipschitz continuous.*

Proof. We first note that the Lipschitz continuity of V^- and V^+ with respect to p and q just comes from the boundness of the g_{ij}. Using standard arguments, one easily shows that, for any $t_0 \in [0,T]$, $(u,v) \in \mathcal{U}(t_0) \times \mathcal{V}(t_0)$, the map

$$x \to g_{ij}\left(X_T^{t_0,x,u,v}\right)$$

is Lipschitz continuous with a Lipschitz constant independant of $t_0 \in [0,T]$. Hence for any pair of strategies $(\hat{\alpha},\hat{\beta}) \in (\mathcal{A}_r(t_0))^I \times (\mathcal{B}_r(t_0))^J$ the map

$$x \to \mathcal{J}(t,x,\hat{\alpha},\hat{\beta},p,q) = \sum_{i=1}^{I}\sum_{j=1}^{J} p_i q_j \mathbf{E}_{\alpha_i\beta_j}\left(g_{ij}(X_T^{t_0,x,\alpha_i,\beta_j})\right)$$

is C-Lipschitz continuous for some constant C independent of $t \in [0,T]$, of $p \in \Delta(I)$, and of $q \in \Delta(J)$. From this one easily deduces that V^+ and V^- are C-Lipschitz continuous with respect to x (see, for instance, [15]).

We now consider the time regularity of V^- and V^+. We do only the proof for V^-, since the case of V^+ can be treated similarly. Let $x_0 \in \mathbb{R}^N$, $(p,q) \in \Delta(I)\times\Delta(J)$, and $t_0 < t_1 < T$ be fixed. Let $\hat{\beta} = (\beta_j) \in (\mathcal{B}_r(t_0))^J$ be ϵ-optimal for $V^-(t_0,x_0,p,q)$ and $\alpha \in \mathcal{A}_r(t_1)$. Let us define, for any $j = 1,\dots,J$, $\tilde{\beta}_j \in \mathcal{B}_r(t_1)$ and $\alpha' \in \mathcal{A}_r(t_0)$ by setting (for some $\bar{u} \in U$ fixed)

$$\tilde{\beta}_j(\omega,u) = \beta_j(\omega,\tilde{u}), \text{ where } \tilde{u}(t) = \begin{cases} \bar{u} & \text{if } t \in [t_0,t_1), \\ u & \text{otherwise} \end{cases}$$

for any $\omega \in \Omega_{\tilde{\beta}_j} := \Omega_{\beta_j}$ and $u \in \mathcal{U}(t_1)$, and

$$\alpha'(\omega,v) = \begin{cases} \bar{u} & \text{if } t \in [t_0,t_1), \\ \alpha(\omega, v_{|_{[t_1,T]}}) & \text{otherwise} \end{cases} \qquad \forall \omega \in \Omega_{\alpha'} := \Omega_\alpha,\ \forall v \in \mathcal{V}(t_0)\ .$$

We note that, for any $\alpha \in \mathcal{A}_r(t_1)$ and $j = 1,\dots,J$, we have

$$\left|X_t^{t_0,x_0,\alpha',\beta_j} - X_t^{t_1,x_0,\alpha,\tilde{\beta}_j}\right| \le M|t_0 - t_1|e^{L(t-t_1)} \qquad \forall t \ge t_1$$

(where $M = \|f\|_\infty$ and f is L-Lipschitz continuous), because the pair (u_ω, v_ω) satisfying

$$\alpha'(\omega_1, v_\omega) = u_\omega \text{ and } \beta_j(\omega_2, u_\omega) = v_\omega$$

is given by $u_\omega = \bar{u}$ and $v_\omega = \beta_j(\omega_2,\bar{u})$ on $[t_0,t_1]$ and coincides on $[t_1,T]$ with the pair (u'_ω, v'_ω) satisfying

$$\alpha(\omega_1, v'_\omega) = u'_\omega \text{ and } \tilde{\beta}_j(\omega_2, u'_\omega) = v'_\omega \text{ on } [t_1,T]\ .$$

Therefore, for any $\hat\alpha = (\alpha_i) \in (\mathcal{A}_r(t_1))^I$, we have

$$\begin{aligned}
&\mathcal{J}(t_1, x_0, \hat\alpha, (\tilde\beta_j), p, q)\\
&\quad \geq\ \mathcal{J}(t_0, x_0, \hat\alpha', \hat\beta, p, q) - LM|t_0 - t_1|e^{L(T-t_1)}\\
&\quad \geq\ \inf_{\hat\alpha" \in (\mathcal{A}_r(t_0))^I} \mathcal{J}(t_0, x_0, \hat\alpha", \hat\beta, p, q) - LM|t_0 - t_1|e^{L(T-t_1)}\\
&\quad \geq\ V^-(t_0, x_0, p, q) - \epsilon - LM|t_0 - t_1|e^{L(T-t_1)}
\end{aligned}$$

(where L is also a Lipchitz constant for the g_i), because $\hat\beta$ is ϵ-optimal for $V^-(t_0, x_0, p, q)$. Since this holds for any $\hat\alpha = (\alpha_i) \in (\mathcal{A}_r(t_1))^I$ and any $\epsilon > 0$, we get

$$V^-(t_1, x_0, p, q) \geq V^-(t_0, x_0, p, q) - LM|t_0 - t_1|e^{L(T-t_1)}\ .$$

The reverse inequality can be proved in a similar way: We choose some ϵ-optimal strategy $\hat\beta = (\beta_j) \in (\mathcal{B}_r(t_1))^J$ for $V^-(t_1, x_0, p, q)$, and we extend it to a strategy $(\tilde\beta_j) \in (\mathcal{B}_r(t_0))^J$ by setting (for some $\bar v \in V$ fixed)

$$\tilde\beta_j(\omega, u) = \begin{cases} \bar v & \text{if } t \in [t_0, t_1),\\ \beta_j(\omega, u_{|_{[t_1,T]}}) & \text{otherwise} \end{cases} \qquad \forall \omega \in \Omega_{\tilde\beta_j} := \Omega_{\beta_j},\ \forall u \in \mathcal{U}(t_0)\ .$$

Then similar estimates as above show that for any $\hat\alpha \in (\mathcal{A}_r(t_0))^I$ we have

$$\mathcal{J}(t_0, x_0, \hat\alpha, (\tilde\beta_j), p, q) \geq V^-(t_1, x_0, p, q) - \epsilon - LM|t_0 - t_1|e^{L(T-t_1)}$$

from the ϵ-optimality of $\hat\beta$ for $V^-(t_1, x_0, p, q)$. Then we get

$$V^-(t_0, x_0, p, q) \geq V^-(t_1, x_0, p, q) - LM|t_0 - t_1|e^{L(T-t_1)}\ . \qquad \square$$

LEMMA 3.2 (convexity properties of V^- and V^+). *For any $(t, x) \in [0, T) \times \mathbb{R}^N$, the maps $V^+ = V^+(t, x, p, q)$ and $V^- = V^-(t, x, p, q)$ are convex in p and concave in q on $\Delta(I)$ and $\Delta(J)$, respectively.*

Remark. This result is well known for repeated games with a lack of information. The procedure we use in the proof is usually called "splitting"; see [22], for instance.

Proof of Lemma 3.2. We do only the proof for V^+; the proof for V^- can be achieved by reversing the roles of the players. One first easily checks that

$$V^+(t_0, x_0, p, q) = \inf_{(\alpha_i) \in (\mathcal{A}_r(t_0))^I} \sum_{j=1}^J q_j \sup_{\beta \in \mathcal{B}_r(t_0)} \left[\sum_{i=1}^I p_i \mathbf{E}_{\alpha_i \beta} \left(g\left(X_T^{t_0, x_0, \alpha_i, \beta} \right) \right) \right]\ .$$

Hence $q \to V^+(t, x, p, q)$ is concave for any (t, x, p).

We now prove the convexity of V^+ with respect to p. Let $(t, x, q) \in [0, T) \times \mathbb{R}^N \times \Delta(J)$, $p^0, p^1 \in \Delta(I)$, $\lambda \in (0, 1)$, and let $\hat\alpha^0 = (\alpha_i^0) \in (\mathcal{A}_r(t))^I$ and $\hat\alpha^1 = (\alpha_i^1) \in (\mathcal{A}_r(t))^I$ be ϵ-optimal for $V^+(t, x, p^0, q)$ and $V^+(t, x, p^1, q)$, respectively ($\epsilon > 0$). Let us set $p^\lambda = (1 - \lambda)p^0 + \lambda p^1$. We can assume without loss of generality that $p_i^\lambda \neq 0$ for any i (because if $p_i^\lambda = 0$, then $p_i^0 = p_i^1 = 0$, so that this index i plays no role in our computation). We now define the strategy $\hat\alpha^\lambda = (\alpha_i^\lambda) \in (\mathcal{A}_r(t))^I$ by setting

$$\Omega_{\alpha_i^\lambda} = [0, 1] \times \Omega_{\alpha_i^0} \times \Omega_{\alpha_i^1},\ \mathcal{F}_{\alpha_i^\lambda} = B([0, 1]) \otimes \mathcal{F}_{\alpha_i^0} \otimes \mathcal{F}_{\alpha_i^1},\ \mathbf{P}_{\alpha_i^\lambda} = \mathcal{L}^1 \otimes \mathbf{P}_{\alpha_i^0} \otimes \mathbf{P}_{\alpha_i^1}$$

(where $B([0, 1])$ is the Borel σ-field and $\mathcal{L}^1$ the Lebesgue measure on $[0, 1]$) and

$$\alpha_i^\lambda(\omega_1, \omega_2, \omega_3, v) = \begin{cases} \alpha_i^0(\omega_2, v) & \text{if } \omega_1 \in [0, \frac{(1-\lambda)p_i^0}{p_i^\lambda}),\\ \alpha_i^1(\omega_3, v) & \text{if } \omega_1 \in [\frac{(1-\lambda)p_i^0}{p_i^\lambda}, 1] \end{cases}$$

for any $(\omega_1, \omega_2, \omega_3) \in \Omega_{\alpha_i^\lambda}$ and $v \in \mathcal{V}(t)$. We note that $(\Omega_{\alpha_i^\lambda}, \mathcal{F}_{\alpha_i^\lambda}, \mathbf{P}_{\alpha_i^\lambda})$ belongs to the set of probability spaces $\mathcal{S}$ and that α_i^λ belongs to $\mathcal{A}_r(t_0)$ for any $i = 1, \dots, I$.

The interpretation of the strategy $\hat{\alpha}^\lambda$ is the following: If the index i is chosen according to the probability p^λ, then player 1 chooses α_i^0 with probability $\frac{(1-\lambda)p_i^0}{p_i^\lambda}$ and α_i^1 with probability $1 - \frac{(1-\lambda)p_i^0}{p_i^\lambda} = \frac{\lambda p_i^1}{p_i^\lambda}$. Hence the probability for the strategy α_i^0 to be chosen is $p_i^\lambda \frac{(1-\lambda)p_i^0}{p_i^\lambda} = (1-\lambda)p_i^0$, while the strategy α_i^1 appears with probability $p_i^\lambda \frac{\lambda p_i^1}{p_i^\lambda} = \lambda p_i^1$. Therefore

$$\begin{aligned}
\sup_{\hat{\beta}} \mathcal{J}(t,x,\hat{\alpha}^\lambda,\hat{\beta},p^\lambda,q) &= \sum_j q_j \sup_\beta \sum_i p_i^\lambda \mathbf{E}_{\alpha_i^\lambda,\beta}\left(g_{ij}(X_T^{t,x,\alpha_i^\lambda,\beta})\right) \\
&= \sum_j q_j \sup_\beta \sum_i p_i^\lambda \left[\frac{(1-\lambda)p_i^0}{p_i^\lambda}\mathbf{E}_{\alpha_i^0,\beta}\left(g_{ij}(X_T^{t,x,\alpha_i^0,\beta})\right) + \frac{\lambda p_i^1}{p_i^\lambda}\mathbf{E}_{\alpha_i^1,\beta}\left(g_{ij}(X_T^{t,x,\alpha_i^1,\beta})\right)\right] \\
&\le (1-\lambda)\sum_j q_j \sup_\beta \sum_i p_i^0 \mathbf{E}_{\alpha_i^0,\beta}\left(g_{ij}(X_T^{t,x,\alpha_i^0,\beta})\right) \\
&\qquad + \lambda \sum_j q_j \sup_\beta \sum_i p_i^1 \mathbf{E}_{\alpha_i^1,\beta}\left(g_{ij}(X_T^{t,x,\alpha_i^1,\beta})\right) \\
&\le (1-\lambda)V^+(t,x,p^0,q) + \lambda V^+(t,x,p^1,q) + \epsilon,
\end{aligned}$$

because $\hat{\alpha}^0$ and $\hat{\alpha}^1$ are ϵ-optimal for $V^+(t,x,p^0,q)$ and $V^+(t,x,p^1,q)$, respectively. Therefore

$$\begin{aligned}
V^+(t,x,p^\lambda,q) &\le \sup_{\hat{\beta}} \mathcal{J}(t,x,\hat{\alpha}^\lambda,\hat{\beta},p^\lambda,q) \\
&\le (1-\lambda)V^+(t,x,p^0) + \lambda V^+(t,x,p^1) + \epsilon\,,
\end{aligned}$$

which proves the desired claim because ϵ is arbitrary. □

The convexity properties of the value functions lead one naturally to consider their Fenchel conjugates. Let $w : [0,T] \times \mathbb{R}^N \times \Delta(I) \times \Delta(J) \mapsto \mathbb{R}$ be some function. We denote by w^* its convex conjugate with respect to variable p:

$$w^*(t,x,\hat{p},q) = \sup_{p \in \Delta(I)} \hat{p}.p - w(t,x,p,q) \qquad \forall (t,x,\hat{p},q) \in [0,T] \times \mathbb{R}^N \times \mathbb{R}^I \times \Delta(J)\,.$$

In particular V^{-*} and V^{+*} denote the convex conjugate with respect to the p-variable of the functions V^- and V^+.

For a function $w = w(t,x,\hat{p},q)$ defined on the dual space $[0,T] \times \mathbb{R}^N \times \mathbb{R}^I \times \Delta(J)$, we also denote by w^* its convex conjugate with respect to $\hat{p}$ defined on $[0,T] \times \mathbb{R}^N \times \Delta(I) \times \Delta(J)$:

$$w^*(t,x,p,q) = \sup_{\hat{p} \in \mathbb{R}^I} p.\hat{p} - w(t,x,p,q) \qquad \forall (t,x,p,q) \in [0,T] \times \mathbb{R}^N \times \Delta(I) \times \Delta(J)\,.$$

In a symmetric way, we denote by $w^\sharp = w^\sharp(t,x,p,\hat{q})$ the concave conjugate with respect to q of w:

$$w^\sharp(t,x,p,\hat{q}) = \inf_{q \in \Delta(J)} \hat{q}.q - w(t,x,p,q) \qquad \forall (t,x,p,\hat{q}) \in [0,T] \times \mathbb{R}^N \times \Delta(I) \times \mathbb{R}^J\,.$$

4. The subdynamic programming. The main result of this section is that $V^{+\sharp}$ and V^{-*} are subsolutions of the dual HJ equation. To fix the ideas, we study here the case of V^{-*} and deduce at the very end of the section the symmetric results for $V^{+\sharp}$.

LEMMA 4.1 (reformulation of V^{-*}). *We have*

$$V^{-*}(t,x,\hat{p},q) = \inf_{(\beta_j)\in(\mathcal{B}_r(t_0))^J} \sup_{\alpha\in\mathcal{A}_r(t_0)} \max_{i\in\{1,\dots,I\}} \left\{ \hat{p}_i - \sum_{j=1}^J q_j \mathbf{E}_{\alpha\beta_j}\left(g_{ij}(X_T^{t,x,\alpha,\beta_j})\right)\right\}. \tag{11}$$

Proof of Lemma 4.1. Let us note for later use that

$$\begin{aligned} V^-(t,x,p,q) &:= \sup_{\beta_j}\inf_{\alpha_i}\sum_{i,j} p_i q_j \mathbf{E}_{\alpha_i\beta_j}\left(g_{ij}(X_T^{t,x,\alpha_i,\beta_j})\right) \\ &= \sup_{\beta_j}\sum_i p_i \inf_\alpha \sum_j q_j \mathbf{E}_{\alpha\beta_j}\left(g_{ij}(X_T^{t,x,\alpha,\beta_j})\right), \end{aligned} \tag{12}$$

because player 1 can choose—and indeed chooses—his strategy $\hat{\alpha}=(\alpha_i)$ such that α_i minimizes $\sum_j q_j \mathbf{E}_{\alpha\beta_j}\left(g_{ij}(X_T^{t,x,\alpha_i,\beta_j})\right)$.

Let us denote by $z = z(t,x,\hat{p},q)$ the right-hand side of equality (11). We first claim that

$$z \text{ is convex with respect to } \hat{p}. \qquad \square \tag{13}$$

Proof of (13). The proof mimics the proof of the convexity of V^+. Let $(t,x,q)\in[0,T)\times\mathbb{R}^N\times\Delta(J)$, $\hat{p}^0,\hat{p}^1\in\mathbb{R}^I$, $\lambda\in(0,1)$, and let $(\beta_j^0)\in(\mathcal{B}_r(t))^J$ and $(\beta_j^1)\in(\mathcal{B}_r(t))^J$ be ϵ-optimal for $z(t,x,\hat{p}^0,q)$ and $z(t,x,\hat{p}^1,q)$, respectively ($\epsilon>0$). Let us set $\hat{p}^\lambda = (1-\lambda)\hat{p}^0+\lambda\hat{p}^1$. We define the strategies $\beta_j^\lambda\in\mathcal{B}_r(t)$ by setting

$$\Omega_{\beta_j^\lambda} = [0,1]\times\Omega_{\beta_j^0}\times\Omega_{\beta_j^1},\ \mathcal{F}_{\beta_j^\lambda} = B([0,1])\otimes\mathcal{F}_{\beta_j^0}\otimes\mathcal{F}_{\beta_j^1},\ \mathbf{P}_{\beta_j^\lambda} = \mathcal{L}^1\otimes\mathbf{P}_{\beta_j^0}\otimes\mathbf{P}_{\beta_j^1},$$

and

$$\beta_j^\lambda(\omega_1,\omega_2,\omega_3,u) = \begin{cases} \beta_j^0(\omega_2,u) & \text{if } \omega_1\in[0,(1-\lambda)),\\ \beta_j^1(\omega_3,u) & \text{if } \omega_1\in[(1-\lambda),1]\end{cases}$$

for any $(\omega_1,\omega_2,\omega_3)\in\Omega_{\beta_j^\lambda}$ and $u\in\mathcal{U}(t)$. Then $(\Omega_{\beta_j^\lambda},\mathcal{F}_{\beta_j^\lambda},\mathbf{P}_{\beta_j^\lambda})$ belongs to $\mathcal{S}$ and $(\beta_j^\lambda)\in(\mathcal{B}_r(t_0))^J$. For any $\alpha\in\mathcal{A}_r(t)$, we have by using the convexity of the map $(s_i)\mapsto\max_i\{s_i\}$:

$$\begin{aligned} &\max_i\left\{\hat{p}_i^\lambda - \sum_j q_j\mathbf{E}_{\alpha,\beta_j^\lambda}\left(g_{ij}(X_T^{t,x,\alpha,\beta_j^\lambda})\right)\right\} \\ &= \max_i\left\{(1-\lambda)\left(\hat{p}_i^0 - \sum_j q_j\mathbf{E}_{\alpha\beta_j^0}\left(g_{ij}(X_T^{t,x,\alpha,\beta_j^0})\right)\right)\right. \\ &\qquad\left. +\lambda\left(\hat{p}_i^1 - \sum_j q_j\mathbf{E}_{\alpha\beta_j^1}\left(g_{ij}(X_T^{t,x,\alpha,\beta_j^1})\right)\right)\right\} \\ &\le (1-\lambda)\sup_\alpha\max_i\left\{\hat{p}_i^0 - \sum_j q_j\mathbf{E}_{\alpha\beta_j^0}\left(g_{ij}(X_T^{t,x,\alpha,\beta_j^0})\right)\right\} \\ &\qquad +\lambda\sup_\alpha\max_i\left\{\hat{p}_i^1 - \sum_j q_j\mathbf{E}_{\alpha\beta_j^1}\left(g_{ij}(X_T^{t,x,\alpha,\beta_j^1})\right)\right\} \\ &\le (1-\lambda)z(t,x,\hat{p}^0,q) + \lambda z(t,x,\hat{p}^1,q) + \epsilon, \end{aligned}$$

because β^0 and β^1 are ϵ-optimal for $z(t,x,\hat{p}^0,q)$ and $z(t,x,\hat{p}^1,q)$, respectively. Hence

$$\begin{aligned} z(t,x,\hat{p}^\lambda,q) &\le \sup_\alpha \max_i \left\{ \hat{p}_i^\lambda - \sum_j q_j \mathbf{E}_{\alpha,\beta_j^\lambda} \left(g_{ij}(X_T^{t,x,\alpha,\beta_j^\lambda}) \right) \right\} \\ &\le (1-\lambda) z(t,x,q^0) + \lambda z(t,x,q^1) + \epsilon \,, \end{aligned}$$

which proves the desired claim because ϵ is arbitrary.

Next we show that $V^{-*} = z$. Indeed we have by definition of z:

$$\begin{aligned} z^*(t,x,p,q) &= \sup_{\hat{p}} p.\hat{p} - \inf_{(\beta_j)} \max_i \left\{ \hat{p}_i - \inf_\alpha \sum_j q_j \mathbf{E}_{\alpha\beta_j} \left(g_{ij}(X_T^{t,x,\alpha,\beta_j}) \right) \right\} \\ &= \sup_{(\beta_j)} \sup_{\hat{p}} \min_i \left\{ p.\hat{p} - \hat{p}_i + \inf_\alpha \sum_j q_j \mathbf{E}_{\alpha\beta_j} \left(g_{ij}(X_T^{t,x,\alpha,\beta_j}) \right) \right\} . \end{aligned}$$

In this last expression, the $\sup_{\hat{p}}$ is attained by

$$\hat{p}_i = \inf_\alpha \sum_j q_j \mathbf{E}_{\alpha\beta_j} \left(g_{ij}(X_T^{t,x,\alpha,\beta_j} \right) \,,$$

for which all of the arguments of the $\min_i$ are equal. Hence

$$\begin{aligned} z^*(t,x,p,q) &= \sup_{\beta_j} \sum_i p_i \inf_\alpha \sum_j q_j \mathbf{E}_{\alpha\beta_j} \left(g_{ij}(X_T^{t,x,\alpha,\beta_j}) \right) \\ &= V^-(t,x,p,q) \end{aligned}$$

because of (12). Since we have proved that z is convex with respect to $\hat{p}$, we get by duality $V^{-*} = z^{**} = z$. □

LEMMA 4.2 (subdynamic principle for V^{-*}). *We have for any* $(t_0,x_0,\hat{p},q) \in [0,T) \times \mathbb{R}^N \times \mathbb{R}^I \times \Delta(J)$ *and any* $t_1 \in (t_0,T]$

$$V^{-*}(t_0,x_0,\hat{p},q) \le \inf_{\beta \in \mathcal{B}(t_0)} \sup_{\alpha \in \mathcal{A}(t_0)} V^{-*}(t_1, X_{t_1}^{t_0,x_0,\alpha,\beta}, \hat{p}, q) \,.$$

Proof. Let us denote by $V_1^{-*}(t_0,t_1,x_0,\hat{p},q)$ the right-hand side of the above inequality. Arguing as in Lemma 3.1 one can prove that V_1^{-*} is Lipschitz continuous with respect to x. We also note that player 1 can play in pure strategies in V^{-*}, namely,

$$V^{-*}(t,x,\hat{p},q) = \inf_{(\beta_j) \in (\mathcal{B}_r(t))^J} \sup_{\alpha \in \mathcal{A}(t)} \max_{i \in \{1,\dots,I\}} \left\{ \hat{p}_i - \sum_j q_j \mathbf{E}_{\beta_j} \left[g_{ij}(X_T^{t,x,\alpha,\beta_j}) \right] \right\} \tag{14}$$

for any $(t,x,\hat{p},q) \in [0,T) \times \mathbb{R}^N \times \mathbb{R}^I \times \Delta(J)$. Indeed, we have from Lemma 4.1 that

$$\begin{aligned} &V^{-*}(t,x,\hat{p},q) = \\ &\inf_{(\beta_j) \in (\mathcal{B}_r(t_0))^J} \sup_{\alpha \in \mathcal{A}_r(t_0)} \max_{i \in \{1,\dots,I\}} \left\{ \hat{p}_i - \sum_{j=1}^J q_j \mathbf{E}_{\alpha\beta_j} \left(g_{ij}(X_T^{t,x,\alpha,\beta_j}) \right) \right\} . \end{aligned}$$

Hence the inequality "$\geq$" in (14) is obvious because $\mathcal{A}(t) \subset \mathcal{A}_r(t)$. To prove the reverse inequality we first note that, for any $\alpha \in \mathcal{B}_r(t_0)$ and for any $\omega_1 \in \Omega_\alpha$, $\alpha(\omega_1, \cdot)$ belongs to $\mathcal{A}(t_0)$. Let us fix $(\beta_j) \in (\mathcal{B}(t))^J$. We have, from the convexity of $(s_i) \mapsto \max_i\{s_i\}$,

$$\begin{aligned}
&\sup_{\alpha\in\mathcal{A}_r(t)} \max_i \left\{ \hat{p}_i - \sum_j q_j \mathbf{E}_{\alpha\beta_j}(g_{ij}(X_T^{t,x,\alpha,\beta_j})) \right\} \\
&\quad \leq \sup_{\alpha\in\mathcal{A}_r(t)} \int_{\Omega_\alpha} \max_i \left\{ \hat{p}_i - \sum_j q_j \mathbf{E}_{\beta_j}(g_{ij}(X_T^{t,x,\alpha(\omega_1,\cdot),\beta_j})) \right\} dP_\alpha(\omega_1) \\
&\quad \leq \sup_{\alpha\in\mathcal{A}_r(t)} \sup_{\omega_1\in\Omega_\alpha} \max_i \left\{ \hat{p}_i - \sum_j q_j \mathbf{E}_{\beta_j}(g_{ij}(X_T^{t,x,\alpha(\omega_1,\cdot),\beta_j})) \right\} \\
&\quad \leq \sup_{\alpha\in\mathcal{A}(t)} \max_i \left\{ \hat{p}_i - \sum_j q_j \mathbf{E}_{\beta_j}(g_{ij}(X_T^{t,x,\alpha,\beta_j})) \right\}.
\end{aligned}$$

Taking the infimum over $(\beta_j) \in (\mathcal{B}(t))^J$ gives (14).

Let $\epsilon > 0$ and $\beta^0 \in \mathcal{B}(t_0)$ be some pure ϵ-optimal strategy for $V_1^{-*}(t_0, t_1, x_0, \hat{p}, q)$. For any $x \in \mathbb{R}^N$, we can find some ϵ-optimal strategy $\hat{\beta}^x = (\beta_j^x) \in \mathcal{B}_r(t_1)$ for player 2 in the game $V^{-*}(t_1, x, \hat{p}, q)$. From the Lipschitz continuity of the map

$$y \to \sup_{\alpha\in\mathcal{A}(t)} \max_{i\in\{1,\ldots,I\}} \left\{ \hat{p}_i - \sum_j q_j \mathbf{E}_{\beta_j^x} \left[g_{ij}(X_T^{t,y,\alpha,\beta_j^x}) \right] \right\},$$

and of $y \to V^{-*}(t_1, y, \hat{p}, q)$, β^x is also (2ϵ)-optimal for $V^{-*}(t_1, y, \hat{p}, q)$ if $y \in B_r(x)$ for some radius $r > 0$. Using the fact that f is bounded, one can show that the reachable states from (t_0, x_0) by using the differential equation (1) is bounded and contained in some ball $B_R(0)$. Let us set $M = \|f\|_\infty$, and let us fix $\sigma > 0$ small such that $M\sigma \leq r/2$. Then we choose $(x_l)_{l=1,\ldots,l_0}$ such that $\bigcup_{l=1}^{l_0} B_{r/2}(x_l)$ contains the ball $B_R(0)$. Let $(E_l)_{l=1,\ldots,l_0}$ be a Borel partition of $B_R(0)$ such that, for any l, $E_l \subset B_{r/2}(x_l)$. We set

$$\beta_j^l = \beta_j^{x_l},\ \Omega_j^l = \Omega_{\beta_j^l},\ \mathcal{F}_j^l = \mathcal{F}_{\beta_j^l},\ \text{and } \mathbf{P}_j^l = \mathbf{P}_{\beta_j^l}$$

for $j = 1, \ldots, J$ and $l = 1, \ldots, l_0$. We choose some delay $\tau \in (0, \sigma]$ common to all of the strategies β_j^l.

We note for later use that, if for some controls $(u, v) \in \mathcal{U}(t_0) \times \mathcal{V}(t_0)$ and for some l, we have $X_{t_1-\tau}^{t_0,x_0,u,v} \in E_l$, then

$$|X_{t_1-\tau}^{t_0,x_0,u,v} - X_{t_1}^{t_0,x_0,u,v}| \leq \|f\|_\infty \tau \leq M\sigma \leq r/2\,,$$

so that $X_{t_1}^{t_0,x_0,u,v}$ belongs to $B_r(x_l)$. In particular $(\beta_j^l)_j$ is (2ϵ)-optimal for V^+ at $(t_1, X_{t_1}^{t_0,x_0,u,v}, \hat{p}, q)$. To summarize

$$X_{t_1-\tau}^{t_0,x_0,u,v} \in E_l \ \Rightarrow\ (\beta_j^l)_j \text{ is } (2\epsilon)\text{-optimal for } V^{-*} \text{ at } (t_1, X_{t_1}^{t_0,x_0,u,v}, \hat{p}, q). \tag{15}$$

Let us now define a new strategy $\hat{\beta} = (\beta_j) \in (\mathcal{B}_r(t_0))^J$ in the following way: Set

$$\Omega_{\beta_j} = \Pi_{l=1}^{l_0} \Omega_j^l,\ \mathcal{F}_{\beta_j} = \mathcal{F}_j^1 \otimes \ldots \otimes \mathcal{F}_j^{l_0}, \text{and } \mathbf{P}_{\beta_j} = \mathbf{P}_j^1 \otimes \ldots \otimes \mathbf{P}_j^{j_0},$$

and, for any $\omega = (\omega^1, \dots, \omega^{l_0}) \in \Omega_{\beta_j}$ and $u \in \mathcal{U}(t_0)$, set

$$\beta_j(\omega, u)(t) = \begin{cases} \beta^0(u)(t) & \text{if } t \in [t_0, t_1), \\ \beta_j^l(\omega^l, u_{|_{[t_1,T]}})(t) & \text{if } t \in [t_1, T] \text{ and } X_{t_1-\tau}^{t_0,x_0,u,\beta^0(u)} \in E_l. \end{cases}$$

Then $(\Omega_{\beta_j}, \mathcal{F}_{\beta_j}, \mathbf{P}_{\beta_j})$ belongs to $\mathcal{S}$ and $\hat{\beta} = (\beta_j) \in (\mathcal{B}_r(t_0))^J$.

For any pure strategy $\alpha \in \mathcal{A}(t_0)$, we have

$$g_{ij}(X_T^{t_0,x_0,\alpha,\beta_j}) = \sum_{l=1}^{l_0} g_{ij}\left(X_T^{t_1, X_{t_1}^{t_0,x_0,\alpha,\beta^0}, \tilde{\alpha}, \beta_j^l}\right) \mathbf{1}_{\{X_{t_1-\tau}^{t_0,x_0,\alpha,\beta^0} \in E_l\}},$$

where $\tilde{\alpha} \in \mathcal{A}(t_1)$ is a restriction of α to the time interval $[t_1, T]$ defined by

$$\tilde{\alpha}(v) = \alpha(v') \qquad \forall v \in \mathcal{V}(t_1), \text{ where } v'(t) = \begin{cases} \bar{v}(t) & \text{if } t \in [t_0, t_1], \\ v(t) & \text{otherwise,} \end{cases}$$

the controls $(\bar{u}, \bar{v})$ being the pair associated with (α, β^0) as in (8). Hence

$$\max_{i \in \{1,\dots,I\}} \left\{ \hat{p}_i - \sum_j q_j \mathbf{E}_{\beta_j} \left(g_{ij}(X_T^{t_0,x_0,\alpha,\beta_j}) \right) \right\} =$$

$$\max_{i \in \{1,\dots,I\}} \left\{ \hat{p}_i - \sum_j q_j \sum_{l=1}^{l_0} \left(\int_{\Omega_j^l} g_{ij}\left(X_T^{t_1, X_{t_1}^{t_0,x_0,\alpha,\beta^0}, \tilde{\alpha}, \beta_j^l}\right) d\mathbf{P}_j^l(\omega^l) \right) \mathbf{1}_{O^l} \right\}$$

(where we have set $O^l = \{X_{t_1-\tau}^{t_0,x_0,\alpha,\beta^0} \in E_l\}$)

$$\leq \sum_{l=1}^{l_0} \sup_{\alpha' \in \mathcal{B}(t_1)} \max_{i \in \{1,\dots,I\}} \left\{ \hat{p}_i - \sum_j q_j \left(\int_{\Omega_j^l} g_{ij}\left(X_T^{t_1, X_{t_1}^{t_0,x_0,\alpha,\beta^0}, \alpha', \beta_j^l}\right) d\mathbf{P}_j^l(\omega^l) \right) \right\} \mathbf{1}_{O^l}$$

(because of the convexity of the map $s = (s_i) \mapsto \max\{s_i\}$)

$$\leq \sum_{l=1}^{l_0} \left(V^{-*}\left(t_1, X_{t_1}^{t_0,x_0,\alpha,\beta^0}, \hat{p}, q\right) + 2\epsilon \right) \mathbf{1}_{O^l}$$

(because of (15))

$$= V^{-*}\left(t_1, X_{t_1}^{t_0,x_0,\alpha,\beta^0}, \hat{p}, q\right) + 2\epsilon$$

$$\leq V_1^{-*}(t_0, t_1, x_0, \hat{p}, q) + 3\epsilon \,,$$

because β^0 is ϵ-optimal for $V_1^{-*}(t_0, t_1, x_0, \hat{p}, q)$.

From this we conclude easily that

$$V^{-*}(t_0, x_0, \hat{p}, q) \leq V_1^{-*}(t_0, t_1, x_0, \hat{p}, q) \,. \qquad \square$$

COROLLARY 4.3 (V^{-*} is a subsolution of HJ). *For any $(\hat{p}, q) \in \mathbb{R}^I \times \Delta(J)$, the map $(t, x) \mapsto V^{-*}(t, x, \hat{p}, q)$ is a viscosity subsolution of the dual Hamilton–Jacobi equation:*

$$w_t + H^*(x, Dw) = 0 \text{ in } [0, T] \times \mathbb{R}^N \,, \tag{16}$$

where H is defined by (6) *and $H^*(x, \xi) = -H(x, -\xi)$.*

Remark. From the definition of H, we have

$$H^*(x,\xi) := \sup_{u\in U}\inf_{v\in V} f(x,u,v).\xi = \inf_{v\in V}\sup_{u\in U} f(x,u,v).\xi. \tag{17}$$

Proof of Corollary 4.3. It is well known that a function satisfying a subdynamic programming principle is a subsolution of the associated HJ equation when the game is played with classical nonanticipative strategies (see [15]). We give a short proof of this fact in the framework of nonanticipative strategies with delay. Let $(\hat{p},q)\in\mathbb{R}^I\times\Delta(J)$ be fixed, and let ϕ be a smooth test function such that

$$\phi(t,x)\geq V^{-*}(t,x,\hat{p},q) \qquad \forall (t,x)\in[0,T]\times\mathbb{R}^N\ , \tag{18}$$

with an equality at (t_0,x_0), where $t_0\in[0,T)$. For any $v\in V$, let us define the pure strategy $\beta\in\mathcal{B}(t_0)$ by setting

$$\beta(u)(t) = v \qquad \forall u\in\mathcal{U}(t_0),\ t\in[t_0,T]\ .$$

Let us fix $\epsilon>0$ and $h>0$ small.

Since V^{-*} satisfies the subdynamic programming principle of Lemma 4.2, there is some strategy $\alpha_h\in\mathcal{A}(t_0)$ such that

$$V^{-*}(t_0,x_0,\hat{p},q)\leq V^{-*}(t_0+h,X_{t_0+h}^{t_0,x_0,\alpha_h,\beta},\hat{p},q)+\epsilon h\ . \tag{19}$$

Let us set $u_h(s)=\alpha_h(v)(s)$ and $x_h(s)=X_s^{t_0,x_0,\alpha_h,\beta}=X_s^{t_0,x_0,u_h,v}$. Then

$$x_h(t_0+h)=x_0+\int_{t_0}^{t_0+h} f(x_h(s),u_h(s),v)ds = x_0+\int_{t_0}^{t_0+h} f(x_0,u_h(s),v)ds+h\epsilon(h),$$

where $\epsilon(h)\mapsto 0$ as $h\to 0^+$. From (18) and (19) we have

$$\begin{aligned}
0 \ &\leq\ V^{-*}(t_0+h,X_{t_0+h}^{t_0,x_0,\alpha_h,\beta},\hat{p},q)-V^{-*}(t_0,x_0,\hat{p},q)+\epsilon h\\
&\leq\ \phi\left(t_0+h,x_0+\int_{t_0}^{t_0+h} f(x_0,u_h(s),v)ds+h\epsilon(h)\right)ds-\phi(t_0,x_0)+\epsilon h\\
&\leq\ h\phi_t(t_0,x_0)+\int_{t_0}^{t_0+h} D\phi(t_0,x_0).f(x_0,u_h(s),v)ds+h\epsilon_1(h)+\epsilon h\\
&\leq\ h\phi_t(t_0,x_0)+h\sup_{u\in U} D\phi(t_0,x_0).f(x_0,u,v)+h\epsilon_1(h)+\epsilon h
\end{aligned}$$

where $\epsilon_1(h)\mapsto 0$ as $h\to 0^+$. Dividing the last inequality by $h>0$ and letting $h\to 0^+$ gives

$$\phi_t(t_0,x_0)+\sup_{u\in U} D\phi(t_0,x_0).f(x_0,u,v)\geq -\epsilon\ .$$

Then we let $\epsilon\to 0^+$, take the minimum over $v\in V$, and use (17) to get the desired inequality:

$$\phi_t(t_0,x_0,p)+H^*(x_0,D\phi(t_0,x_0))\geq 0\ . \qquad \square$$

To state the symmetric results for $V^{+\sharp}$, we need only to note that

$$-V^+(t,x,p,q)=\sup_{\hat{\alpha}\in(\mathcal{A}_r(t_0))^I}\inf_{\hat{\beta}\in(\mathcal{B}_r(t_0))^J}\sum_{i=1}^I\sum_{j=1}^I p_iq_j\mathbf{E}_{\alpha_i\beta_j}\left((-g_{ij})\left(X_T^{t_0,x_0,\alpha_i,\beta_j}\right)\right)\ ,$$

which is of the same form as V^- when one changes the roles of the players. In particular the convex Fenchel conjugate of $(-V^+)$ with respect to q, i.e., $-V^{+\sharp}(-\hat{q})$, satisfies a subdynamic programming principle and is therefore a subsolution of some associated Hamilton–Jacobi equation. From this we easily deduce the following.

COROLLARY 4.4 ($V^{+\sharp}$ is a supersolution of HJ). *For any* $(t_0, t_1, x_0, p, \hat{q}) \in [0,T] \times [0,T] \times \mathbb{R}^N \times \Delta(I) \times \mathbb{R}^J$, *we have*

$$V^{+\sharp}(t_0, x_0, p, \hat{q}) \geq \sup_{\alpha \in \mathcal{A}(t_0)} \inf_{\beta \in \mathcal{B}(t_0)} V^{+\sharp}(t_1, X_{t_1}^{t_0,x_0,\alpha,\beta}, p, \hat{q}) \ .$$

Hence $V^{+\sharp}$ *is a supersolution of the dual Hamilton–Jacobi equation* (16).

Remark. We use here Isaacs' assumption (6). Indeed, if V^{-*} is a subsolution of the HJ equation (16) with $H^*(x,\xi) = \inf_u \sup_v f(x,u,v).\xi$, $V^{+\sharp}$ is actually a supersolution of (16) with a Hamiltonian H^* defined by $H^*(x,\xi) = \sup_v \inf_u f(x,u,v).\xi$.

5. Existence of the value and solutions of the primal/dual HJ equations. In this section we prove that our game has a value: $V^+ = V^-$. This value can be characterized in terms of dual solutions of some HJ equations.

The key argument for this is the following comparison principle, which we state for later use for a general Hamiltonian H. We assume that $H : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is continuous, and we suppose that there is a constant C such that, for any $x_1, x_2 \in \mathbb{R}^N$ and $\theta \geq 0$,

$$|H(x_1, \theta(x_1 - x_2)) - H(x_2, \theta(x_1 - x_2))| \leq C|x_1 - x_2|(1 + \theta|x_1 - x_2|) \ . \tag{20}$$

Let us point out that the map H defined by (6) satisfies the above assumptions under conditions (5) on the dynamics.

Recall that, for any map $w = w(t,x,p,q)$ defined on $[0,T] \times \mathbb{R}^N \times \Delta(I) \times \Delta(J)$, w^* denotes the convex Fenchel conjugate of w with respect to p, while $w^\sharp$ denotes its concave Fenchel conjugate with respect to q.

We now consider a Hamilton–Jacobi equation of the form:

$$z_t + H(x, Dz) = 0 \ . \tag{21}$$

We say that a function $w : [0,T] \times \mathbb{R}^N \times \Delta(I) \times \Delta(J) \mapsto \mathbb{R}$ is a *dual subsolution of* (21) if w is Lipschitz continuous, convex with respect to p, and concave with respect to q and if, for any $(p, \hat{q}) \in \Delta(I) \times \mathbb{R}^J$, $(t,x) \mapsto w^\sharp(t,x,p,\hat{q})$ is a supersolution of the dual HJ equation

$$z_t + H^*(x, Dz) = 0 \ , \tag{22}$$

where $H^*(x,\xi) = -H(x,-\xi)$. In a symmetric way, w is a *dual supersolution of the HJ equation* (21) if w is Lipschitz continuous, convex with respect to p, and concave with respect to q and if, for any for any $(\hat{p}, q) \in \mathbb{R}^I \times \Delta(J)$, $(t,x) \mapsto w^*(t,x,\hat{p},q)$ is a subsolution of the dual HJ equation (22). We say that w is a *dual solution of* (21) if w is at the same time a dual subsolution and a dual supersolution of (21).

THEOREM 5.1 (comparison principle). *Let* $w_1, w_2 : [0,T] \times \mathbb{R}^N \times \Delta(I) \times \Delta(J) \mapsto \mathbb{R}$ *be, respectively, a dual subsolution and a dual supersolution of the HJ equation* (21). *We assume that for any* $(x,p,q) \in \mathbb{R}^N \times \Delta(I) \times \Delta(J)$, $w_1(T,x,p,q) \leq w_2(T,x,p,q)$. *Then* $w_1 \leq w_2$ *in* $[0,T] \times \mathbb{R}^N \times \Delta(I) \times \Delta(J)$.

Remarks.

1. We cannot compare $w_1^\sharp$ and w_2^* at time $t = T$. So this result is *not* an application of the classical comparison principle.

2. It is known that, if w_2 is a supersolution of the HJ equation (21), then w_2^* is a subsolution of the dual HJ equation (22) (see, for instance, [1]). The converse does not hold true in general, and so we cannot rephrase the assumptions in term of sub- and supersolutions of (21) for w_1 and w_2. However, it turns out that w_2, for instance, is a supersolution at "some suitable points," related with its convexity property with respect to p. We explain this more precisely in Lemma 5.4 below.
3. The result can be extended to bounded uniformly continuous subsolutions by standard techniques (see [3], for instance).

The comparison principle is proved at the end of the section. Let us now state the main result of this paper.

THEOREM 5.2 (existence of the value). *Assume that conditions* (5) *on f and on the g_i hold and that Isaacs' assumption* (6) *is satisfied. Then we have*

$$V^+(t,x,p,q) = V^-(t,x,p,q) \qquad \forall (t,x,p) \in [0,T] \times \mathbb{R}^N \times \Delta(I) \times \Delta(J)\ .$$

Proof of Theorem 5.2. From Lemma 3.1, V^- and V^+ are Lipschitz continuous. From Lemma 3.2, we know that V^+ and V^- are convex with respect to p and concave with respect to q. Corollary 4.3 states that, for any $(\hat{p},q) \in \mathbb{R}^I \times \Delta(J)$, $V^{-*}(\cdot,\cdot,\hat{p},q)$ is a subsolution of the dual HJ equation (16). Hence V^- is a dual supersolution of (7). Corollary 4.4 states that $V^{+\sharp}(\cdot,\cdot,p,\hat{q})$ is a supersolution of the HJ equation (16) for any $(p,\hat{q}) \in \Delta(I) \times \mathbb{R}^J$ and therefore a dual subsolution of (7). Since $V^+(T,\cdot,p,q) = V^-(T,\cdot,p,q) = \sum_{i,j} p_i q_j g_{ij}$, the comparison principle states that $V^+ \le V^-$. But the reverse inequality always holds. Hence $V^- = V^+$ and the game has a value. □

The above proof also shows the following.

COROLLARY 5.3 (characterization of the value). *Under the assumptions of Theorem* 5.2*, the value function $V := V^+ = V^-$ is the unique dual solution of the HJ equations* (7)*, such that $V(T,x,p,q) = \sum_{ij} p_i q_j g_{ij}(x)$.*

We complete this section by an equivalent formulation of the notion of dual supersolution. Although the result is not needed in the rest of the text, we think that it can help to enlighten the notion.

LEMMA 5.4. *Let $w : [0,T] \times \mathbb{R}^N \times \Delta(I) \times \Delta(J) \mapsto \mathbb{R}$ be Lipschitz continuous, convex with respect to p, and concave with respect to q. Then the following statements are equivalent:*

(i) *w is a dual supersolution of* (21)*;*

(ii) *for any $q \in \Delta(J)$, for any test function $\phi = \phi(t,x,p)$ which is $\mathcal{C}^1$ and convex in p and such that*

$$(t,x,p) \mapsto w(t,x,p,q) - \phi(t,x,p)$$

has a strict global minimum at some point $(t_0,x_0,p_0) \in [0,T) \times \mathbb{R}^N \times \Delta(I)$, we have

$$\phi_t(t_0,x_0,p_0) + H(x_0, D\phi(t_0,x_0,p_0)) \le 0\ . \tag{23}$$

Remarks.

1. This result means that a dual supersolution of (21)—originally defined in terms of subsolution of the dual HJ equation—is indeed a supersolution of the primal HJ equation (21) in weak sense. However, it is not a classical supersolution. For instance, if $I = 1$, $f = f(u,v)$, and $g_j(x) = a_j.x$ for some

$a_j \in \mathbb{R}^N$ $(j = 1, \dots, J)$, then we prove in [10] that

$$V^+(t,x,p) = V^-(t,x,p) = (T-t)Cav(h)(p) + \sum_j p_j x.a_j,$$

where $h(p) = H(\sum_j p_j a_j)$ and $Cav(h)$ is the concave hull of h with respect to $p \in \Delta(I)$. Then

$$V_t^- + H(DV^-) = -Cav(h)(p) + h(p) \geq 0\ ,$$

with a strict inequality in general. In particular, V^- is not a classical supersolution of the primal HJ equation.

2. Note carefully that we require the minimum $w(t,x,p,q) - \phi(t,x,p)$ at (t_0,x_0,p_0) to be *strict*. This point is absolutely crucial for the equivalence. It is related with similar definition in repeated games, where some function has to be tested only at extreme points (see [17]). Let us point out that a general minimum of $w - \phi$ cannot be made artificially strict by substracting $\epsilon|(t,x,p) - (t_0,x_0,p_0)|^2$ to ϕ (as is usually done in viscosity solutions) because one then loses the convexity of ϕ with respect to p.
3. A symmetric result holds for subsolutions: w is a dual subsolution of (21) if and only if, for any $p \in \Delta(I)$, for any test function $\phi = \phi(t,x,q)$ which is $\mathcal{C}^1$ and concave in q and such that $w - \phi$ has a strict global maximum at some point $(t_0,x_0,q_0) \in [0,T) \times \mathbb{R}^N \times \Delta(J)$, we have

$$\phi_t(t_0,x_0,q_0) + H(x_0, D\phi(t_0,x_0,q_0)) \geq 0\ .$$

Proof of Lemma 5.4. Let us first assume that w is a dual supersolution of (21). Let $q \in \Delta(J)$, $\phi = \phi(t,x,p)$ be a test function which is $\mathcal{C}^1$ and convex in p and such that $w - \phi$ has a strict global minimum at some point $(t_0,x_0,p_0) \in [0,T) \times \mathbb{R}^N \times \Delta(I)$. This means that

$$w(t,x,p,q) \leq \phi(t,x,p) + w(t_0,x_0,p_0,q) - \phi(t_0,x_0,p_0) \tag{24}$$

for any $(t,x,p) \in [0,T] \times \mathbb{R}^N \times \Delta(I)$, with an equality only at (t_0,x_0,p_0). By using the fact that the minimum of $w - \phi$ is strict and the standard perturbation argument (consisting in replacing ϕ by $\phi + \epsilon|p|^2$ if necessary), we can assume that ϕ is strictly convex in p. Then, for any $\hat{p} \in \mathbb{R}^I$, p being the unique element of the subdifferential of $\phi^*(t_0,x_0,\cdot)$ at $\hat{p}$, $\phi^*(\cdot,\cdot,\hat{p})$ is differentiable at (t_0,x_0), and one easily checks that

$$\phi_t^*(t_0,x_0,\hat{p}) = -\phi_t(t_0,x_0,p) \text{ and } D\phi^*(t_0,x_0,\hat{p}) = -D\phi(t_0,x_0,p)\ . \tag{25}$$

Let $\hat{p}_0$ belong to the subdifferential with respect to p of w at (t_0,x_0,p_0). Then inequality (24) shows that $\hat{p}_0$ belongs to the subdifferential of ϕ with respect to p at (t_0,x_0,p_0). Since w and ϕ are convex in p, we have

$$w^*(t_0,x_0,\hat{p}_0,q) = p_0.\hat{p}_0 - w(t_0,x_0,p_0,q) \text{ and } \phi^*(t_0,x_0,\hat{p}_0) = p_0.\hat{p}_0 - \phi(t_0,x_0,p_0)\ .$$

Thus

$$w(t_0,x_0,p_0,q) - \phi(t_0,x_0,p_0) = w^*(t_0,x_0,\hat{p}_0,q) - \phi^*(t_0,x_0,\hat{p}_0)\ . \tag{26}$$

We note that (24) can be rewritten as

$$p.\hat{p}_0 - w(t,x,p,q) \geq p.\hat{p}_0 - \phi(t,x,p) - w(t_0,x_0,p_0,q) + \phi(t_0,x_0,p_0)$$

for all $(t,x,p) \in [0,T] \times \mathbb{R}^N \times \Delta(I)$. Taking the sup over $p \in \Delta(I)$ and taking into account (26) gives

$$w^*(t,x,\hat{p}_0,q) \geq \phi^*(t,x,\hat{p}_0) + w^*(t_0,x_0,\hat{p}_0,q) - \phi^*(t_0,x_0,\hat{p}_0) \ .$$

Therefore $(t,x) \mapsto w^*(t,x,\hat{p}_0,q) - \phi^*(t,x,\hat{p}_0)$ has a maximum at (t_0,x_0). Since w^* is a subsolution of the dual HJ equation, we have

$$\phi^*_t(t_0,x_0,\hat{p}_0) + H^*(x_0, D\phi^*(t_0,x_0,\hat{p}_0)) \geq 0 \ ,$$

which implies the desired inequality (23) thanks to (25).

Conversely, let us assume that (ii) holds. Let ϕ be a $\mathcal{C}^1$ test function such that $(t,x) \mapsto w^*(t,x,\hat{p}_0,q) - \phi(t,x)$ has a local minimum at (t_0,x_0) for some $(\hat{p}_0,q) \in \mathbb{R}^I \times \Delta(I)$. Without loss of generality, we can assume that this minimum is a global one and that $\phi(t_0,x_0) = w^*(t_0,x_0,\hat{p}_0,q)$ (see [3]). Let $\tilde{\phi}(t,x,\hat{p}) = \phi(t,x)$ if $\hat{p} = \hat{p}_0$ and $\tilde{\phi}(t,x,\hat{p}) = +\infty$ otherwise. Then $\tilde{\phi} \geq w^*(\cdot,\cdot,\cdot,q)$ on $[0,T] \times \mathbb{R}^N \times \mathbb{R}^I$, with an equality at $(t_0,x_0,\hat{p}_0)$. Thus, by duality,

$$p.\hat{p}_0 - \phi(t,x) = \tilde{\phi}^*(t,x,p) \leq w^{**}(t,x,p,q) = w(t,x,p,q)$$

for any $(t,x,p) \in [0,T] \times \mathbb{R}^N \times \Delta(I)$, with an equality at (t_0,x_0,p_0) for any $p_0 \in \partial w^*(t_0,x_0,\hat{p}_0,q)$ (where $\partial w^*(t_0,x_0,,\hat{p}_0,q)$ denotes the superdifferential of the convex function $\hat{p} \mapsto w^*(t_0,x_0,\hat{p},q)$ at $\hat{p}_0$). Hence $(t,x,p) \mapsto w(t,x,p,q) - (p.\hat{p}_0 - \phi(t,x))$ has a minimum at (t_0,x_0,p_0) for any $p_0 \in \partial w^*(t_0,x_0,\hat{p}_0,q)$. In order to get a *strict* minimum, we have to introduce some perturbation term. Let $\gamma > 0$, $\epsilon > 0$, and $(t_\epsilon,x_\epsilon,p_\epsilon)$ be a point of minimum of $w - \psi_{\epsilon,\gamma}$, where

$$\psi_{\epsilon,\gamma}(t,x,p) = p.\hat{p}_0 + \epsilon|p|^2 - \phi(t,x) - \gamma|(t,x) - (t_0,x_0)|^2 \ .$$

Then $(t_\epsilon,x_\epsilon,p_\epsilon)$ converges (up to some subsequence) to (t_0,x_0,p_0) for some $p_0 \in \partial w^*(t_0,x_0,\hat{p}_0,q)$ as $\epsilon \to 0^+$ (we use here the penalization term in γ). Moreover, we have

$$\begin{array}{rcl} \tilde{\psi}(t,x,p) & := & \psi_{\epsilon,\gamma}(t,x,p) - \epsilon|p-p_\epsilon|^2 - \epsilon|(t,x)-(t_\epsilon,x_\epsilon)|^2 \\ & < & \psi_{\epsilon,\gamma}(t,x,p) \\ & \leq & w(t,x,p) - w(t_\epsilon,x_\epsilon,p_\epsilon) + \tilde{\psi}(t_\epsilon,x_\epsilon,p_\epsilon) \end{array}$$

for any $(t,x,p) \neq (t_\epsilon,x_\epsilon,p_\epsilon)$, with an equality at $(t_\epsilon,x_\epsilon,p_\epsilon)$. This means that $w - \tilde{\psi}$ has a strict minimum at $(t_\epsilon,x_\epsilon,p_\epsilon)$. Since $\tilde{\psi}$ is still convex in p we get from assumption (ii) that

$$\tilde{\psi}_t(t_\epsilon,x_\epsilon,p_\epsilon) + H(x_\epsilon, D\tilde{\psi}(t_\epsilon,x_\epsilon,p_\epsilon)) \leq 0 \ .$$

Using the definition of $\tilde{\psi}$ and letting $\epsilon \to 0^+$, we then obtain

$$\phi_t(t_0,x_0) + H^*(x_0, D\phi(t_0,x_0)) \geq 0 \ ,$$

which proves that w is a dual supersolution of (21). □

Proof of Theorem 5.1. We follow the proof of Theorem 3.7, p. 152, in [3]. Let us argue by contradiction, by assuming that there is some (t_1,x_1,p_1,q_1) such that $w_1(t_1,x_1,p_1,q_1) > w_2(t_1,x_1,p_1,q_1)$. This means that, for some $\sigma > 0$, we have

$$\sup_{t,x,p,q} \ w_1(t,x,p,q) - w_2(t,x,p,q) - \sigma(T-t) \ > \ 0 \ . \tag{27}$$

We now use the standard method of separation of variables. In order to avoid burdensome details, we do the proof under the additional assumption that there is some $R > 0$ such that $w_1(t,x,p,q) \leq w_2(t,x,p,q)$ for any (t,x,p,q) with $|x| \geq R$. This assumption can be omitted by using penalization arguments at infinity (see [3] for the details). Let $\epsilon > 0$ be fixed. From our assumption, the map

$$(28) \qquad (t,x,s,y,p,q) \mapsto w_1(t,x,p,q) - w_2(s,y,p,q) - \frac{1}{\epsilon}|(t,x)-(s,y)|^2 - \sigma(T-t)$$

has a maximum over $[0,T]\times\mathbb{R}^N\times\Delta(I)\times\Delta(J)$, and we denote by $(t_\epsilon, x_\epsilon, s_\epsilon, y_\epsilon, p_\epsilon, q_\epsilon)$ such a point of maximum. From the usual arguments in [3], we have $t_\epsilon < T$ and $s_\epsilon < T$ for small ϵ, because $w_1(T,x,p,q) \leq w_2(T,x,p,q)$ and w_1 and w_2 are Lipschitz continuous. Moreover

$$(29) \qquad \lim_{\epsilon\to 0^+} \frac{1}{\epsilon}|(t_\epsilon,x_\epsilon)-(s_\epsilon,y_\epsilon)|^2 = 0\ .$$

Since, for $(s,y)=(s_\epsilon,y_\epsilon)$, $(t_\epsilon,x_\epsilon,q_\epsilon)$ is a maximum in (28), we have

$$(30) \qquad \begin{aligned} w_1(t,x,p_\epsilon,q) \leq\ & w_1(t_\epsilon,x_\epsilon,p_\epsilon,q_\epsilon) + w_2(s_\epsilon,y_\epsilon,p_\epsilon,q) - w_2(s_\epsilon,y_\epsilon,p_\epsilon,q_\epsilon) \\ & + \frac{1}{\epsilon}\left(|(t,x)-(s_\epsilon,y_\epsilon)|^2 - |(t_\epsilon,x_\epsilon)-(s_\epsilon,y_\epsilon)|^2\right) + \sigma(t_\epsilon - t) \end{aligned}$$

for any (t,x,q), with an equality at $(t_\epsilon,x_\epsilon,q_\epsilon)$. Let $\hat{q}_\epsilon$ belong to the superdifferential $\partial_q^+ w_2(s_\epsilon,y_\epsilon,p_\epsilon,q_\epsilon)$ of w_2 with respect to q at $(s_\epsilon,y_\epsilon,p_\epsilon,q_\epsilon)$. Then the above inequality shows that $\hat{q}_\epsilon \in \partial_q w_1(t_\epsilon,x_\epsilon,p_\epsilon,q_\epsilon)$. From the concavity of w_1 and w_2 with respect to q, we have

$$w_1^\sharp(t_\epsilon,x_\epsilon,p_\epsilon,\hat{q}_\epsilon) = q_\epsilon.\hat{q}_\epsilon - w_1(t_\epsilon,x_\epsilon,p_\epsilon,q_\epsilon)$$

and

$$w_2^\sharp(s_\epsilon,y_\epsilon,p_\epsilon,\hat{q}_\epsilon) = q_\epsilon.\hat{q}_\epsilon - w_2(s_\epsilon,y_\epsilon,p_\epsilon,q_\epsilon)$$

so that

$$(31) \qquad w_1(t_\epsilon,x_\epsilon,p_\epsilon,q_\epsilon) - w_2(s_\epsilon,y_\epsilon,p_\epsilon,q_\epsilon) = w_2^\sharp(s_\epsilon,y_\epsilon,p_\epsilon,\hat{q}_\epsilon) - w_1^\sharp(t_\epsilon,x_\epsilon,p_\epsilon,\hat{q}_\epsilon)\ .$$

Combining (30) with (31) then gives

$$\begin{aligned} q.\hat{q}_\epsilon - w_1(t,x,p_\epsilon,q) \geq\ & \\ & w_1^\sharp(t_\epsilon,x_\epsilon,p_\epsilon,\hat{q}_\epsilon) + q.\hat{q}_\epsilon - w_2(s_\epsilon,y_\epsilon,p_\epsilon,q) - w_2^\sharp(s_\epsilon,y_\epsilon,p_\epsilon,\hat{q}_\epsilon) \\ & - \frac{1}{\epsilon}\left(|(t,x)-(s_\epsilon,y_\epsilon)|^2 - |(t_\epsilon,x_\epsilon)-(s_\epsilon,y_\epsilon)|^2\right) - \sigma(t_\epsilon - t)\ . \end{aligned}$$

Taking the infimum over q in the above expression then gives

$$\begin{aligned} w_1^\sharp(t,x,p_\epsilon,\hat{q}_\epsilon) \geq\ & \\ & w_1^\sharp(t_\epsilon,x_\epsilon,p_\epsilon,\hat{q}_\epsilon) - \frac{1}{\epsilon}\left(|(t,x)-(s_\epsilon,y_\epsilon)|^2 - |(t_\epsilon,x_\epsilon)-(s_\epsilon,y_\epsilon)|^2\right) - \sigma(t_\epsilon - t)\ . \end{aligned}$$

So $(t,x) \mapsto w_1^\sharp(t,x,p_\epsilon,\hat{q}_\epsilon) - \left(-\frac{|(t,x)-(s_\epsilon,y_\epsilon)|^2}{\epsilon} + \sigma t\right)$ has a minimum at (t_ϵ,x_ϵ). Since $w_1^\sharp(\cdot,\cdot,p_\epsilon,\hat{q}_\epsilon)$ is a supersolution of the HJ equation (21), we get

$$(32) \qquad \sigma + \frac{2}{\epsilon}(s_\epsilon - t_\epsilon) + H^*\left(x_\epsilon, \frac{2}{\epsilon}(y_\epsilon - x_\epsilon)\right) \leq 0\ .$$

We now argue in a symmetric way for w_2. Since $(s_\epsilon, y_\epsilon, p_\epsilon)$ is a maximum in (28), we have

$$w_2(s,y,p,q_\epsilon) \geq w_2(s_\epsilon,y_\epsilon,p_\epsilon,q_\epsilon) + w_1(t_\epsilon,x_\epsilon,p,q_\epsilon) - w_1(t_\epsilon,x_\epsilon,p_\epsilon,q_\epsilon) \\ - \frac{1}{\epsilon}\left(|(t_\epsilon,x_\epsilon)-(s,y)|^2 - |(t_\epsilon,x_\epsilon)-(s_\epsilon,y_\epsilon)|^2\right) \tag{33}$$

for any $(s,y,p) \in [0,T]\times\mathbb{R}^N\times\Delta(I)$. Let $\hat{p}_\epsilon$ belong to the subdifferential $\partial_p^- w_1(t_\epsilon,x_\epsilon,p_\epsilon,q_\epsilon)$ of w_1 with respect to p at $(s_\epsilon,y_\epsilon,p_\epsilon,q_\epsilon)$. Then the above inequality shows that $\hat{p}_\epsilon \in \partial_p^- w_2(s_\epsilon,y_\epsilon,p_\epsilon,q_\epsilon)$. Therefore we have as above

$$w_2(s_\epsilon,y_\epsilon,p_\epsilon,q_\epsilon) - w_1(t_\epsilon,x_\epsilon,p_\epsilon,q_\epsilon) = w_1^*(t_\epsilon,x_\epsilon,\hat{p}_\epsilon,q_\epsilon) - w_2^*(s_\epsilon,y_\epsilon,\hat{p}_\epsilon,q_\epsilon)\ .$$

Then we get from (33):

$$w_2^*(s,y,\hat{p}_\epsilon,q_\epsilon) \leq w_2^*(s_\epsilon,y_\epsilon,\hat{p}_\epsilon,q_\epsilon) + \frac{1}{\epsilon}\left(|(t_\epsilon,x_\epsilon)-(s,y)|^2 - |(t_\epsilon,x_\epsilon)-(s_\epsilon,y_\epsilon)|^2\right)$$

for any $(s,y) \in [0,T]\times\mathbb{R}^N$, with an equality at (s_ϵ,y_ϵ). Since $w_2^*(\cdot,\cdot,\hat{p}_\epsilon,q_\epsilon)$ is a subsolution of the HJ equation (21), this gives

$$\frac{2}{\epsilon}(s_\epsilon - t_\epsilon) + H^*\left(y_\epsilon, \frac{2}{\epsilon}(y_\epsilon - x_\epsilon)\right) \geq 0\ . \tag{34}$$

Computing the difference between (32) and (34) and using the assumption (20) on H (recall that $H^*(x,\xi) = -H(x,-\xi)$) gives

$$-\sigma + C|x_\epsilon - y_\epsilon|\left(1 + \frac{2|x_\epsilon - y_\epsilon|}{\epsilon}\right) \geq 0\ ,$$

which is in contradiction with (29) as $\epsilon \to 0^+$. □

6. The case of a lack of information on the initial position. In this section we investigate a two-player zero-sum differential game in which the players have some private information on the random initial position. The dynamics of the game is still given by

$$x'(t) = f(x,u(t),v(t))\ , \qquad u(t)\in U,\ v(t)\in V, \tag{35}$$

where U, V, and f satisfy (5). The terminal time of the game is denoted by T, and the payoff is a terminal payoff $g(x(T))$, where $g : \mathbb{R}^N \to \mathbb{R}$ is Lipschitz continuous and bounded. The game starts at time $t_0 \in [0,T]$.

The description of the game involves $I \times J$ initial positions x_{ij}^0, $i = 1,\dots,I$, $j = 1,\dots,J$, a probability $p \in \Delta(I)$, and a probability $q \in \Delta(J)$. As before, the game is played in two steps: At time t_0, the pair (i,j) is chosen according to the probability $p \otimes q$; the index i is communicated to player 1 only and the index j to player 2 only.

Then the players control system (35) with initial position x_{ij}^0 in order for player 1 to minimize the terminal payoff $g(x(T))$ and for player 2 to maximize it. The players observe their opponent's behavior and try to deduce from this behavior their missing information. They cannot compute the actual position of the system in general.

As before we define the upper and lower value functions associated to this game. For this we introduce the new state of the system: $\mathbf{x} = (x_{ij})$, which denotes the $I \times J$-uplet of possible positions. The upper value is given for $t_0 \in [0,T)$, $\mathbf{x}^0 = (x_{ij}^0) \in \mathbb{R}^{NIJ}$,

$p \in \Delta(I)$, and $q \in \Delta(J)$ by

$$V^+(t_0, \mathbf{x}^0, p, q) = \inf_{(\alpha_i)\in(\mathcal{A}_r(t_0))^I} \sup_{(\beta_j)\in(\mathcal{B}_r(t_0))^J} \sum_{i=1}^{I}\sum_{j=1}^{J} p_i q_j \mathbf{E}_{\alpha_i\beta_j}\left(g\left(X_T^{t_0,x_{ij}^0,\alpha_i,\beta_j}\right)\right),$$

where $t \to X_t^{t_0,x_{ij}^0,\alpha_i,\beta_j}$ is the random solution to (35) with initial position x_{ij}^0 at time t_0 when the players play the random strategies α_i and β_j (see section 2). The lower value is defined by the symmetric formula:

$$V^-(t_0, \mathbf{x}^0, p, q) = \sup_{(\beta_j)\in(\mathcal{B}_r(t_0))^J} \inf_{(\alpha_i)\in(\mathcal{A}_r(t_0))^I} \sum_{i=1}^{I}\sum_{j=1}^{J} p_i q_j \mathbf{E}_{\alpha_i\beta_j}\left(g\left(X_T^{t_0,x_{ij}^0,\alpha_i,\beta_j}\right)\right).$$

Obviously we have

$$V^-(t_0, \mathbf{x}^0, p, q) \le V^+(t_0, \mathbf{x}^0, p, q)\ \forall (t_0, \mathbf{x}^0, p, q) \in [0,T] \times \mathbb{R}^{NIJ} \times \Delta(I) \times \Delta(J).$$

Our main result is that the equality holds.

THEOREM 6.1. *Assume that f, U, and V satisfy* (5), *that the payoff $g : \mathbb{R}^N \to \mathbb{R}$ is Lipschitz continuous and bounded, and that the following generalized Isaacs condition holds:*

$$\mathbf{H}(\mathbf{x}, \xi) = \inf_{u\in U}\sup_{v\in V} \sum_{i=1}^{I}\sum_{j=1}^{J} f(x_{ij}, u, v).\xi_{ij} = \sup_{v\in V}\inf_{u\in U} \sum_{i=1}^{I}\sum_{j=1}^{J} f(x_{ij}, u, v).\xi_{ij} \tag{36}$$

for any $\mathbf{x} = (x_{ij}) \in \mathbb{R}^{NIJ}$ and $\xi = (\xi_{ij}) \in \mathbb{R}^{NIJ}$.

Then the game has a value:

$$V^-(t_0, \mathbf{x}^0, p, q) = V^+(t_0, \mathbf{x}^0, p, q)\ \forall (t_0, \mathbf{x}^0, p, q) \in [0,T] \times \mathbb{R}^{NIJ} \times \Delta(I) \times \Delta(J).$$

Furthermore this value is the dual solution of the HJ equation

$$\begin{cases} z_t + \mathbf{H}(\mathbf{x}, Dz) = 0 & \text{in } [0,T) \times \mathbb{R}^{NIJ}, \\ z(T, \mathbf{x}, p, q) = \sum_{i=1}^{I}\sum_{j=1}^{J} p_i q_j g(x_{ij}) & \text{for } \mathbf{x} = (x_{ij}) \in \mathbb{R}^{NIJ}. \end{cases} \tag{37}$$

Proof of Theorem 6.1. The proof is mainly the same as the proof of Theorem 5.2 and Corollary 5.3, and we give only an outline of it. We first note that V^+ and V^- are Lipschitz continuous in their arguments, convex in p, and concave in q as in Lemmas 3.1 and 3.2. Then, following Lemma 4.1, one proves that

$$V^{-*}(t, \mathbf{x}^0, \hat{p}, q) = \inf_{(\beta_j)\in(\mathcal{B}_r(t_0))^J} \sup_{\alpha\in\mathcal{A}_r(t_0)} \max_{i=1,\dots,I} \left\{ \hat{p}_i - \sum_j q_j \mathbf{E}_{\alpha\beta_j}\left[g\left(X_T^{t,x_{ij}^0,\alpha,\beta_j}\right)\right]\right\}$$

for any $t \in [0,T]$, $\mathbf{x}^0 = (x_{ij}^0) \in \mathbb{R}^{NIJ}$, $\hat{p} \in \mathbb{R}^I$, and $q \in \Delta(J)$. Using this, one obtains as in Lemma 4.2 that V^{-*} satisfies the subdynamic programming principle

$$V^{-*}(t_0, \mathbf{x}^0, \hat{p}, q) \le \inf_{\beta\in\mathcal{B}(t_0)} \sup_{\alpha\in\mathcal{A}(t_0)} V^{-*}(t_1, \mathbf{X}_{t_1}^{t_0,\mathbf{x}^0,\alpha,\beta}, \hat{p}, q)$$

for any $0 \le t_0 < t_1 \le T$, $\mathbf{x}^0 \in \mathbb{R}^{NIJ}$, $\hat{p} \in \mathbb{R}^I$, and $q \in \Delta(J)$, where

$$\mathbf{X}_{t_1}^{t_0,\mathbf{x}^0,\alpha,\beta} = \left(X_{t_1}^{t_0,x_{ij}^0,\alpha,\beta}\right)_{\substack{i=1,\dots,I\\ j=1,\dots,J}}.$$

Hence $V^{-*}(\cdot,\cdot,\hat{p},q)$ is a subsolution of the dual HJ equation

$$z_t + \mathbf{H}^*(x, Dz) = 0 \qquad \text{in } [0,T]\times\mathbb{R}^{NIJ}$$

for any $(\hat{p},q)$, which means that V^- is a dual supersolution of (37). One proves in the same way that V^+ is a dual subsolution of (37). The comparison Theorem 5.1 then implies that $V^+ \le V^-$. Since the inequality $V^- \le V^+$ is obvious, we get the equality and the characterization of the value function. □

REFERENCES

[1] O. Alvarez, J.-M. Lasry, and P.-L. Lions, *Convex viscosity solutions and state-constraints*, J. Math. Pures Appl., 9 (1997), pp. 265–288.

[2] R. J. Aumann and M. B. Maschler, *Repeated Games with Incomplete Information*, with the collaboration of Richard E. Stearns, MIT Press, Cambridge, MA, 1995.

[3] M. Bardi and I. Capuzzo-Dolcetta, *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*, Systems Control Found. Appli., Birkhäuser, Boston, 1997.

[4] G. Barles, *Solutions de viscosité des équations de Hamilton-Jacobi*, Math. Appl. (Berlin) 17, Springer-Verlag, Paris, 1994.

[5] J. S. Baras and N. S. Patel, *Robust control of set-valued discrete-time dynamical systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 61–75.

[6] M. R. James and J. S. Baras, *Partially observed differential games, infinite-dimensional Hamilton–Jacobi–Isaacs equations, and nonlinear H_∞ control*, SIAM J. Control Optim., 34 (1996), pp. 1342–1364.

[7] T. Basar and P. Bernhard, *H^∞- Optimal Control and Related Minimax Design Problems*, Birkhäuser, Boston, 1995.

[8] P. Bernhard, *A discrete-time min-max certainty equivalence principle*, Systems Control Lett., 24 (1995), pp. 229–234.

[9] P. Cardaliaguet and M. Quincampoix, *Zero-sum Differential Games with Probabilistic Knowledge of the State-variable*, International Game Theory Review, to appear.

[10] P. Cardaliaguet, *Representation Formulas for Differential Games with Asymmetric Information*, J. Optim. Theory and Appl., to appear.

[11] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton-Jacobi Equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[12] B. De Meyer, *Repeated games, duality and the central limit theorem*, Math. Oper. Res., 21 (1996), pp. 237–251.

[13] B. De Meyer and D. Rosenberg, *"Cav u" and the dual game*, Math. Oper. Res., 24 (1999), pp. 619–626.

[14] B. De Meyer and H. M. Saley, *On the strategic origin of Brownian motion in finance*, Special anniversary issue, Internat. J. Game Theory, 31 (2002), pp. 285–319.

[15] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton-Jacobi Equations*, Indiana Univ. Math. J., 282 (1984), pp. 487–502.

[16] W. H. Fleming and P. E. Souganidis, *On the existence of value functions of two-player, zero-sum stochastic differential games*, Indiana Univ. Math. J., 38 (1989), pp. 293–314.

[17] R. Laraki, *Repeated games with lack of information on one side: The dual differential approach*, Math. Oper. Res., 27 (2002), pp. 419–440.

[18] J. F. Mertens and S. Zamir, *The value of two person zero sum repeated games with lack of information on both sides*, Internat. J. Game Theory, 1 (1994), pp. 39–64.

[19] L. A. Petrosjan, *Cooperation in games with incomplete information*, Nonlinear Analysis and Convex Analysis, Yokohama, Yokohama, Japan, 2004, pp 469–479.

[20] A. Rapaport and P. Bernhard, *On a planar pursuit game with imperfect knowledge of a coordinate*, Automat. Prod. Inform. Ind., 29 (1995), pp. 575–601.

[21] D. Rosenberg, E. Solan, and N. Vieille, *Stochastic games with a single controller and incomplete information*, SIAM J. Control Optim., 43 (2004), pp. 86–110

[22] S. Sorin, *A first course on zero-sum repeated games*, Math. Appl. (Berlin) 37, Springer-Verlag, Berlin, 2002.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 839–876

A MODEL FOR REVERSIBLE INVESTMENT CAPACITY EXPANSION*

AMAL MERHI† AND MIHAIL ZERVOS†

Abstract. We consider the problem of determining the optimal investment level that a firm should maintain in the presence of random price and/or demand fluctuations. We model market uncertainty by means of a geometric Brownian motion, and we consider general running payoff functions. Our model allows for capacity expansion as well as for capacity reduction, with each of these actions being associated with proportional costs. The resulting optimization problem takes the form of a singular stochastic control problem that we solve explicitly. We illustrate our results by means of the so-called Cobb–Douglas production function. The problem that we study presents a model in which the associated Hamilton–Jacobi–Bellman equation admits a classical solution that conforms with the underlying economic intuition but does not necessarily identify with the corresponding value function, which may be identically equal to ∞. Thus, our model provides a situation that highlights the need for rigorous mathematical analysis when addressing stochastic optimization applications in finance and economics, as well as in other fields.

Key words. singular stochastic control, investment capacity expansion, variational inequalities

AMS subject classifications. 93E20, 49J40, 91B32, 91B70

DOI. 10.1137/050640758

1. Introduction.

We consider the problem of determining the optimal capacity level of a given investment project operating within a random economic environment in a dynamical way. In particular, we consider an investment project that yields payoff at a rate that is dependent on its installed capacity level and on an underlying economic indicator such as the price of or the demand for the project's unique output commodity, which we model by a geometric Brownian motion. The project's capacity level can be increased or decreased at any time and at given proportional costs. The objective is to determine the project's capacity level that maximizes the associated expected, discounted payoff flow.

Irreversible capacity expansion models have attracted considerable interest in the literature; e.g., see Davis et al. [6] (see also Davis [5]), Kobila [10], Øksendal [11], Wang [12], Chiarolla and Haussmann [4], Bank [2], and the references therein. Recently, Bentolila and Bertola [3] and Abel and Eberly [1] considered models involving both expansion and reduction of a project's capacity level. These authors assume that the rate at which the project yields payoff is modelled by a constant elasticity Cobb–Douglas production function. Our model considers many more general running payoff functions that include the whole family of the Cobb–Douglas production functions as special cases and allow for the situation where a running cost proportional to the project's installed capacity (reflecting, e.g., labor costs) is also included (see Examples 1 and 2). Also, Guo and Pham [7] consider a related partially reversible investment model with entry decisions and a general running payoff function. The model that these authors consider is fundamentally different from the ones considered

*Received by the editors September 20, 2005; accepted for publication (in revised form) February 21, 2007; published electronically June 12, 2007. This research was supported by EPSRC grant GR/S22998/01 and the Isaac Newton Institute, Cambridge.
http://www.siam.org/journals/sicon/46-3/64075.html
†Department of Mathematics, London School of Economics, Houghton Street, London WC2A 2AE, UK (amal.merhi@googlemail.com, m.zervos@lse.ac.uk).

by Bentolila and Bertola [3] and Abel and Eberly [1], or the one that we study here, because, e.g., it is one-dimensional instead of two-dimensional.

Our analysis, which leads to results of an explicit analytic nature, involves the derivation of tight conditions for the project's value function to be finite. The fact that simple choices for the project's running payoff function lead to unique solutions to the associated free-boundary problem that conform with standard economic intuition but are associated with value functions that are identically equal to infinity presents a most interesting feature of our analysis (see Remark 3; also, note that this pathological situation does not arise in the context of the special cases studied by Bentolila and Bertola [3] and Abel and Eberly [1]). Indeed, this possibility stresses the fact that treating optimization models related to investment decision making in a "formal" way, which is often the case in the economics literature, can lead to erroneous conclusions and can suggest the adoption of potentially disastrous policies.

The paper is organized as follows. Section 2 is concerned with a rigorous formulation of the investment decision model that we study. In section 3, we derive tight sufficient conditions, which conform with economic intuition, for the associated optimization problem to possess a finite value function. Assumptions 1 and 2 summarize all of the assumptions that we make about the problem data in the paper. We also establish a number of estimates that we use in our subsequent analysis. Section 4 is concerned with the proof of a verification theorem that provides sufficient conditions for the value function of our control problem to be identified with a solution to the associated dynamic programming or Hamilton–Jacobi–Bellman equation. In section 5, we solve the optimization problem considered. Finally, we illustrate our results by a number of examples in section 6.

2. Problem formulation. We fix a probability space $(\Omega, \mathcal{F}, P)$ equipped with a filtration $(\mathcal{F}_t)$ satisfying the usual conditions of right continuity and augmentation by P-negligible sets and carrying a standard, one-dimensional $(\mathcal{F}_t)$-Brownian motion W. We denote by $\mathcal{A}$ the family of all càglàd, $(\mathcal{F}_t)$-adapted, increasing processes ξ such that $\xi_0 = 0$.

We consider an investment project that produces a given commodity, and we assume that the project's capacity, namely its rate of output, can be controlled at any given time. We denote by Y_t the project's capacity at time t, and we model cumulative capacity increases (resp., decreases) by a process $\xi^+ \in \mathcal{A}$ (resp., $\xi^- \in \mathcal{A}$). In particular, given any times $0 \leq s \leq t$, $\xi^+_{t+} - \xi^+_s$ and $\xi^-_{t+} - \xi^-_s$ are the total capacity increase and decrease, respectively, incurred by the project management's decisions during the time interval $[s, t]$. The project's capacity process Y is therefore given by

$$Y_t = y + \xi^+_t - \xi^-_t, \quad Y_0 = y \geq 0, \tag{1}$$

where $y \geq 0$ is the project's initial capacity. Note that the project's capacity process Y is a finite variation process because it is the difference of two increasing processes. Also, the assumptions that the processes $\xi^{\pm}$ are càglàd and $\xi^{\pm}_0 = 0$ imply that $Y_0 = y$. We make the assumption that the project's management controls only the project's capacity level. Accordingly, we denote by Π_y the set of all admissible decision strategies, which is defined by

$$\Pi_y = \left\{(\xi^+, \xi^-) : \ \xi^+, \xi^- \in \mathcal{A}, \text{ and } Y_t \geq 0, \text{ for all } t \geq 0\right\}.$$

We assume that all randomness associated with the project's operation can be captured by a state process X that satisfies the SDE

$$dX_t = bX_t\,dt + \sqrt{2}\sigma X_t\,dW_t, \quad X_0 = x > 0, \tag{2}$$

for some constants b and σ. In practice, X_t can be the price of one unit of the output commodity or an economic indicator reflecting, e.g., the output commodity's demand, at time t.

To simplify the notation, we define

$$\mathcal{S} = \left\{(x,y) \in \mathbb{R}^2 :\ x > 0,\ y \geq 0\right\},$$

so that $\mathcal{S}$ is the set of all possible initial conditions.

With each decision policy $(\xi^+, \xi^-) \in \Pi_y$ we associate the performance criterion

$$\begin{aligned} J_{x,y}(\xi^+,\xi^-) = E\Bigg[&\int_0^\infty e^{-rt} h(X_t, Y_t)\,dt \\ &- K^+ \int_{[0,\infty[} e^{-rt}\,d\xi_t^+ - K^- \int_{[0,\infty[} e^{-rt}\,d\xi_t^- \Bigg], \end{aligned} \tag{3}$$

where $h : \mathcal{S} \to \mathbb{R}$ is a given function, and $r > 0$ and K^+, K^- are constants. Here, h models the running payoff resulting from the project's operation, and K^+ (resp., K^-) models the costs associated with increasing (resp., decreasing) the project's capacity level.

As it stands in (3), the performance index $J_{x,y}$ is not necessarily well-defined because the random variable inside the expectation may not be integrable or even well-defined. To address this issue, we define

$$U_T = \int_0^T e^{-rt} h(X_t, Y_t)\,dt - K^+ \int_{[0,T]} e^{-rt}\,d\xi_t^+ - K^- \int_{[0,T]} e^{-rt}\,d\xi_t^- \quad \text{for } T \geq 0. \tag{4}$$

In the next section (see Lemma 4, in particular), we are going to impose assumptions on h such that U_T is well-defined, for all $T > 0$, and *either*

$$U_\infty = \lim_{T\to\infty} U_T \text{ exists in } \mathbb{R},\ P\text{-a.s.}, \quad \text{and} \quad U_\infty \in L^1(\Omega, \mathcal{F}, P), \tag{5}$$

in which case we naturally define

$$J_{x,y}(\xi^+,\xi^-) = E\left[U_\infty\right], \tag{6}$$

as in (3), *or* there exists an $(\mathcal{F}_t)$-adapted process Z such that

$$U_T \leq Z_T,\ \text{for all } T \geq 0, \quad \text{and} \quad \limsup_{T\to\infty} E\left[Z_T\right] = -\infty, \tag{7}$$

in which case we define

$$J_{x,y}(\xi^+,\xi^-) = -\infty. \tag{8}$$

The objective is to maximize the performance index $J_{x,y}$ thus defined over all admissible decision strategies $(\xi^+, \xi^-) \in \Pi_y$. The value function of the resulting optimization problem is defined by

$$v(x,y) = \sup_{(\xi^+,\xi^-)\in\Pi_y} J_{x,y}(\xi^+,\xi^-). \tag{9}$$

3. Assumptions and preliminary estimates. The purpose of this section is to establish conditions on the problem data under which our control problem is well-posed and its value function is finite and to prove certain estimates that we will need. Before we address these issues, we first discuss an ODE that will play an instrumental role in the solution of our control problem.

Every solution of the homogeneous ODE

$$\sigma^2 x^2 u''(x) + bxu'(x) - rw(x) = 0$$

is given by

$$u(x) = Ax^n + Bx^m$$

for some $A, B \in \mathbb{R}$. Here, the constants $m < 0 < n$ are the solutions of the quadratic equation

$$\sigma^2\lambda^2 + (b - \sigma^2)\lambda - r = 0, \tag{10}$$

given by

$$m, n = \frac{-(b-\sigma^2) \pm \sqrt{(b-\sigma^2)^2 + 4\sigma^2 r}}{2\sigma^2}. \tag{11}$$

Now, let $k :]0, \infty[\to \mathbb{R}$ be any measurable function such that

$$E\left[\int_0^\infty e^{-rt}\,|k(X_t)|\,dt\right] < \infty \quad \text{for all } x > 0. \tag{12}$$

This integrability condition is equivalent to

$$\int_0^x s^{-m-1}|k(s)|\,ds + \int_x^\infty s^{-n-1}|k(s)|\,ds < \infty, \quad \text{for all } x > 0,$$

and the function $R^{[k]} :]0, \infty[\to \mathbb{R}$ defined by

$$R^{[k]}(x) = \frac{1}{\sigma^2(n-m)}\left[x^m \int_0^x s^{-m-1}k(s)\,ds + x^n \int_x^\infty s^{-n-1}k(s)\,ds\right] \tag{13}$$

is a special solution to the nonhomogeneous ODE

$$\sigma^2 x^2 u''(x) + bxu'(x) - rw(x) + k(x) = 0 \tag{14}$$

and satisfies

$$R^{[k]}(x) = E\left[\int_0^\infty e^{-rt}k(X_t)\,dt\right]. \tag{15}$$

Furthermore,

$$\text{if } k \text{ is increasing, then } R^{[k]} \text{ is increasing,} \tag{16}$$

and

$$\text{if } k \text{ is increasing, then } \lim_{x\downarrow 0}\frac{k(x)}{r} \geq 0 \;\Leftrightarrow\; \lim_{x\downarrow 0} R^{[k]}(x) \geq 0. \tag{17}$$

All of these results are proved in Knudsen, Meister, and Zervos [9]. For future reference, we also note that, given any $\lambda \in \mathbb{R}$,

$$E\left[\int_0^\infty e^{-rt} X_t^\lambda \, dt\right] = x^\lambda \int_0^\infty e^{\left[\sigma^2\lambda^2 + (b-\sigma^2)\lambda - r\right]t} E\left[e^{-\sigma^2\lambda^2 t + \sqrt{2}\sigma\lambda W_t}\right] dt$$

$$= \begin{cases} \infty, & \text{if } \lambda \le m \text{ or } \lambda \ge n, \\ -x^\lambda / \left[\sigma^2\lambda^2 + (b-\sigma^2)\lambda - r\right] & \text{if } \lambda \in]m, n[. \end{cases} \tag{18}$$

We are going to need the following estimate that is related to the definitions above.

LEMMA 1. *Given any $\lambda \in]0, n[$, there exist constants $\varepsilon_1, \varepsilon_2 > 0$ such that*

$$E\left[e^{-rt}\bar{X}_t^\lambda\right] \le \frac{\sigma^2\lambda^2 + \varepsilon_2}{\varepsilon_2} x^\lambda e^{-\varepsilon_1 t} \quad \textit{and} \quad E\left[\sup_{t\ge 0} e^{-rt}\bar{X}_t^\lambda\right] \le \frac{\sigma^2\lambda^2 + \varepsilon_2}{\varepsilon_2} x^\lambda,$$

where $\bar{X}_t = \sup_{s\le t} X_s$.

Proof. Since n is the positive solution of the quadratic equation (10), it follows that there exist $\varepsilon_1, \varepsilon_2 > 0$ such that

$$r - \varepsilon_1 > 0 \quad \text{and} \quad \sigma^2\lambda^2 + (b-\sigma^2)\lambda - (r-\varepsilon_1) = -\varepsilon_2.$$

Given such parameters, we define

$$V = \sup_{t\ge 0}\left[-\frac{\sigma^2\lambda^2 + \varepsilon_2}{\sqrt{2}|\sigma|\lambda} t + W_t\right],$$

we calculate

$$\begin{aligned} e^{-rt}\bar{X}_t^\lambda &= x^\lambda e^{-\varepsilon_1 t} e^{-(r-\varepsilon_1)t} \sup_{s\le t} \exp\big((r-\varepsilon_1)s - (\sigma^2\lambda^2 + \varepsilon_2)s + \sqrt{2}\sigma\lambda W_s\big) \\ &= x^\lambda e^{-\varepsilon_1 t} \sup_{s\le t}\left[\exp\big(-(r-\varepsilon_1)(t-s)\big) \exp\left(-(\sigma^2\lambda^2 + \varepsilon_2)s + \sqrt{2}\sigma\lambda W_s\right)\right] \\ &\le x^\lambda e^{-\varepsilon_1 t} e^{\sqrt{2}|\sigma|\lambda V}, \end{aligned}$$

and we observe that

$$\sup_{t\ge 0} e^{-rt}\bar{X}_t^\lambda \le x^\lambda e^{\sqrt{2}|\sigma|\lambda V}.$$

Since V is exponentially distributed with parameter $2(\sigma^2\lambda^2 + \varepsilon_2)/(\sqrt{2}|\sigma|\lambda)$ (see Karatzas and Shreve [8, Exercise 3.5.9]), the two bounds follow by a simple integration. ∎

The following assumptions on the data of the control problem formulated in section 2 will ensure that the associated free-boundary problem has a unique solution that conforms with economical intuition.

Assumption 1. $r > 0$, and the function h is C^3 and satisfies

$$\int_0^x s^{-m-1}|h(s,y)|\, ds + \int_x^\infty s^{-n-1}|h(s,y)|\, ds < \infty$$

for all $(x, y) \in \mathcal{S}$. If we define

$$H(x,y) = h_y(x,y), \quad \text{for } x, y > 0, \tag{19}$$

then, given any $y > 0$,

$$H_x(x,y) > 0, \text{ for all } x > 0, \quad \text{and} \quad \lim_{x\to\infty} H(x,y) = \infty, \tag{20}$$

and, given any $x > 0$,

$$H_y(x,y) < 0 \quad \text{for all } y > 0. \tag{21}$$

Also, $K^+ + K^- > 0$, and

$$\int_0^x s^{-m-1}\left[|H(s,y)| + |H_y(s,y)|\right] ds + \int_x^\infty s^{-n-1}\left[|H(s,y)| + |H_y(s,y)|\right] ds < \infty$$

for all $x, y > 0$.

It is worth observing that (20) and (21) in this assumption have a natural economic interpretation. Indeed, we can think of $H(x,y)\Delta y$ as the *additional* running payoff that we are faced with if we increase the project's capacity level from y to $y + \Delta y$, for small Δy, and the underlying state process X assumes the value x. In view of this observation, (20) reflects the idea that, given y, a small amount of extra capacity should be associated with increasing values of additional running payoff as the value of x, which, e.g., models the price of or the demand for the project's output commodity, is increasing. Similarly, (21) reflects the fact that, for a given value x of the underlying state process, the extra running payoff resulting from a small amount of additional capacity is decreasing as the level of the already installed capacity y increases. Also, the assumption that $K^+ + K^- > 0$, which is an indispensable one, is a most realistic one. Indeed, the inequality $K^+ + K^- < 0$ gives rise to the unrealistic scenario where the project's management can realize arbitrarily high profits by just sequentially increasing and then decreasing the project's capacity by the same amount sufficiently fast.

At this point, we should also observe that (20) and (21) in Assumption 1 exclude the special case that arises when the running payoff function h does not depend on the capacity level y, i.e., when $h(x,y) = \tilde{h}(x)$, for some function $\tilde{h}$. In this case, it is plainly optimal to never change the project's capacity level. However, the qualitative nature of this strategy is fundamentally different from any of the forms that our analysis allows the optimal strategy to have, which is reflected in our assumptions.

The following additional assumptions will ensure that the value function of the control problem considered is finite and identifies with the solution of the associated Hamilton–Jacobi–Bellman equation. Apart from (26), which can be justified by straightforward economic considerations such as the ones discussed above, the conditions in the assumption are of a technical nature.

Assumption 2. $K^+ > 0$, and there exist constants

$$\alpha > 0, \quad \beta \in]0,1[, \quad \vartheta \in]0, K^+ \wedge (K^+ + K^-) \wedge n[, \quad \text{and } C > 0, \tag{22}$$

where $n > 0$ is as in (11), such that

$$\frac{\alpha}{1-\beta} < n, \tag{23}$$

$$-C(1+y) \le h(x,y) \le C(1 + x^{n-\vartheta} + x^\alpha y^\beta) + r(K^+ - \vartheta)y, \quad \text{for all } (x,y) \in \mathcal{S}, \tag{24}$$

$$-C \le H(x,y) \equiv h_y(x,y) \le \beta C x^\alpha y^{-(1-\beta)} + r(K^+ - \vartheta) \quad \text{for all } x, y > 0. \tag{25}$$

Also,

$$h_x(x,y) \geq 0 \quad \text{for all } (x,y) \in \mathcal{S}. \tag{26}$$

Remark 1. Note that we could have replaced the upper bound in (25) by

$$H(x,y) \leq \begin{cases} C\left(1 + x^{\alpha} y^{-(1-\beta)}\right), & \text{for all } x > 0 \text{ and } y < y_1, \\ \beta C x^{\alpha} y^{-(1-\beta)} + r(K^+ - \vartheta), & \text{for all } x > 0 \text{ and } y \geq y_1, \end{cases}$$

for some constant $y_1 > 0$. Depending on the problem data, such a significant relaxation could result in an optimal policy such as the one depicted by Figure 5 that would qualitatively enrich the class of optimal capacity control strategies (see also Example 3 in section 6). However, we decided against such a relaxation because this would complicate both the presentation and the analysis of our results.

Example 1. A choice for the running payoff function h that has been widely considered in the literature is the so-called Cobb–Douglas production function given by

$$h(x,y) = x^{\alpha} y^{\beta} \quad \text{for some constants } \alpha > 0 \text{ and } \beta \in]0,1[. \tag{27}$$

It is straightforward to verify that this function satisfies all of our assumptions if and only if the parameters α and β satisfy the inequality (23).

Example 2. A choice for the running payoff function h that is a variation of the Cobb–Douglas function and incorporates a running cost proportional to the project's installed capacity is given by

$$h(x,y) = (x+\eta)^{\alpha}(y+\zeta)^{\beta} - Ky \quad \text{for some constants } \alpha, \beta, \eta, \zeta, K > 0. \tag{28}$$

This choice satisfies our assumptions if and only if

$$\alpha, \beta \in]0,1[, \quad \frac{\alpha}{1-\beta} < n, \quad \text{and} \quad \beta \eta^{\alpha} \zeta^{-(1-\beta)} < K + rK^+. \tag{29}$$

To see this claim, fix any $\vartheta > 0$ such that

$$\alpha < n - \vartheta \quad \text{and} \quad \beta \eta^{\alpha} \zeta^{-(1-\beta)} < K + r(K^+ - \vartheta),$$

and observe that there exist constants $\Gamma_1, \Gamma_2, \Gamma_3 > 1$ such that

$$(x+\eta)^{\alpha} \leq \Gamma_1(1 + x^{\alpha}), \quad (y+\zeta)^{\beta} \leq \Gamma_2(1 + y^{\beta}), \quad \text{and} \quad \Gamma_1\Gamma_2 y^{\beta} < \Gamma_3 + r(K^+ - \vartheta)y,$$

because $\alpha, \beta \in]0,1[$. In view of these inequalities, we can see that

$$\begin{aligned} h(x,y) &\leq \Gamma_1\Gamma_2\left(1 + x^{\alpha} + x^{\alpha} y^{\beta}\right) + \Gamma_1\Gamma_2 y^{\beta} \\ &\leq \Gamma_1\Gamma_2\Gamma_3\left(1 + x^{\alpha} + x^{\alpha} y^{\beta}\right) + r(K^+ - \vartheta)y, \end{aligned}$$

and we check that Assumption 1 and (23), (24), and (26) in Assumption 2 all hold true. To verify (25) in Assumption 2, we note that, given a constant $C > 1$,

$$\frac{\partial}{\partial x}\left[H(x,y) - \beta C x^{\alpha} y^{-(1-\beta)}\right] < 0$$

is equivalent to

$$\left(\frac{x}{x+\eta}\right)^{1-\alpha} < C\left(\frac{y+\zeta}{y}\right)^{1-\beta},$$

which is true for all $x, y > 0$. It follows that (25) is satisfied if it is true for $x = 0$, i.e., if

$$\beta\eta^{\alpha}(y+\zeta)^{-(1-\beta)} \leq K + r(K^{+} - \vartheta), \quad \text{for all } y \geq 0,$$

which is true when the associated parameters satisfy (29).

To see that if the last inequality in (29) is not true, then the upper bound in (25) does not hold, we argue by contradiction. Indeed, if there are constants $C, \vartheta > 0$ such that (25) is satisfied, then we can pass to the limit as $x \downarrow 0$ to obtain

$$\beta\eta^{\alpha}(y+\zeta)^{-(1-\beta)} \leq K + r(K^{+} - \vartheta) \quad \text{for all } y > 0.$$

However, this inequality cannot be true for all $y > 0$ if the last inequality in (29) above does not hold.

It is a straightforward exercise to show that the bounds in (24)–(25) imply the following estimates.

LEMMA 2. *With reference to the notation in* (13)*, the bounds provided by* (24) *and* (25) *in Assumption* 2 *imply that there exists a constant* $C_1 > 0$ *such that*

$$-C_1(1+y) \leq R^{[h(\cdot,y)]}(x) \leq C_1\left(1 + y + x^{n-\vartheta} + x^{\alpha}y^{\beta}\right), \quad \textit{for all } (x,y) \in \mathcal{S},$$
$$-C_1 \leq R^{[H(\cdot,y)]}(x) \leq C_1\left(1 + x^{\alpha}y^{-(1-\beta)}\right) \quad \textit{for all } (x,y) \in \mathcal{S}.$$

As we have remarked above, bounds such as the ones appearing in Assumption 2 are essential for the value function to be finite. Indeed, we can prove the following result.

LEMMA 3. *Consider the control problem formulated in section* 2 *that arises if the running payoff function* h *is defined by* (27) *in Example* 1*, and suppose that* $\frac{\alpha}{1-\beta} > n > \alpha$*. Then, under any well-posed definition of the performance index* $J_{x,y}$ *that is consistent with* (3)*,* $v(x,y) = \infty$ *for every initial condition* $(x,y) \in \mathcal{S}$*.*

Proof. Consider the strategy defined by

$$\tilde{\xi}_t^{+} = \bar{X}_t^{(n-\alpha)/\beta} \quad \text{and} \quad \tilde{\xi}_t^{-} = 0, \quad \text{for all } t \geq 0, \tag{30}$$

where $\bar{X}_t = \sup_{s \leq t} X_s$. With regard to (18), we can see that this strategy is associated with

$$E\left[\int_0^{\infty} e^{-rt} X_t^{\alpha} \tilde{Y}_t^{\beta}\, dt\right] \geq E\left[\int_0^{\infty} e^{-rt} X_t^{n}\, dt\right] = \infty. \tag{31}$$

Now, let us assume that $\frac{\alpha}{1-\beta} > n > \alpha$. If we define $\lambda = \frac{n-\alpha}{\beta} > 0$, then such an assumption implies that $\lambda < n$. In view of this observation, we can use the first estimate in Lemma 1, the monotone convergence theorem, and the integration by parts formula to see that the strategy given by (30) satisfies

$$\begin{aligned} E\left[\int_{[0,\infty[} e^{-rt}\, d\tilde{\xi}_t^{+}\right] &= \lim_{T\to\infty} E\left[r\int_0^T e^{-rt}\tilde{\xi}_t^{+}\, dt + e^{-rT}\tilde{\xi}_{T+}^{+}\right] \\ &= \lim_{T\to\infty}\left(r\int_0^T E\left[e^{-rt}\bar{X}_t^{\lambda}\right] dt + E\left[e^{-rT}\bar{X}_T^{\lambda}\right]\right) \\ &\leq r\frac{\sigma^2\lambda^2 + \varepsilon_2}{\varepsilon_1\varepsilon_2}x^{\lambda} \\ &< \infty. \end{aligned}$$

However, this calculation, (30), and (31) imply that

$$E\left[\int_0^\infty e^{-rt}X_t^\alpha \tilde{Y}_t^\beta\,dt - \int_{[0,\infty[} e^{-rt}\,d\tilde{\xi}_t^+ - \int_{[0,\infty[} e^{-rt}\,d\tilde{\xi}_t^-\right]$$

is well-defined and equal to ∞, which proves the result. □

We can now prove that our assumptions are sufficient for the optimization problem considered to be well-posed and for its value function to be finite.

LEMMA 4. *Suppose that the running payoff function h satisfies* (24) *in Assumption* 2 *and that $K^+, K^+ + K^- > 0$. Given any initial condition $(x,y) \in \mathcal{S}$,* (5)–(8) *provide a well-posed definition of the performance criterion $J_{x,y}$, and the following statements hold true:*

(a) *Given any admissible strategy $(\xi^+,\xi^-) \in \Pi_y$, $J_{x,y}(\xi^+,\xi^-) \in \mathbb{R}$ if and only if*

$$E\left[\int_0^\infty e^{-rt}Y_t\,dt + K^+\int_{[0,\infty[} e^{-rt}\,d\xi_t^+ + |K^-|\int_{[0,\infty[} e^{-rt}\,d\xi_t^-\right] < \infty. \tag{32}$$

(b) *Condition* (32) *implies that*

$$\liminf_{T\to\infty} e^{-rT}E\left[Y_{T+}\right] = 0. \tag{33}$$

(c) $v(x,y) \in \mathbb{R}$.

Proof. Fix any initial condition $(x,y) \in \mathcal{S}$ and any admissible strategy $(\xi^+,\xi^-) \in \Pi_y$. Since ξ^+, ξ^- are increasing càglàd processes with $\xi_0^+ = \xi_0^- = 0$, we can use the integration by parts formula to calculate

$$\begin{aligned} &-K^+\int_{[0,T]} e^{-rt}\,d\xi_t^+ - K^-\int_{[0,T]} e^{-rt}\,d\xi_t^- \\ &\qquad = -r\int_0^T e^{-rt}\left[K^+\xi_t^+ + K^-\xi_t^-\right]dt - e^{-rT}\left[K^+\xi_{T+}^+ + K^-\xi_{T+}^-\right]. \end{aligned} \tag{34}$$

With regard to (1) and the inequality $K^+ + K^- > 0$, we can see that

$$-K^+\xi_t^+ - K^-\xi_t^- \le -K^+\left(\xi_t^+ - \xi_t^-\right) = -K^+Y_t + K^+y, \tag{35}$$

which, combined with (34), implies that

$$\begin{aligned} &-K^+\int_{[0,T]} e^{-rt}\,d\xi_t^+ - K^-\int_{[0,T]} e^{-rt}\,d\xi_t^- \\ &\qquad \le -rK^+\int_0^T e^{-rt}Y_t\,dt - e^{-rT}K^+Y_{T+} + K^+y. \end{aligned} \tag{36}$$

However, this inequality and (24) in Assumption 2 imply that the random variables U_T defined by (4) satisfy

$$\begin{aligned} U_T &\le K^+y + \int_0^T e^{-rt}\left[h(X_t,Y_t) - rK^+Y_t\right]dt \\ &\le K^+y + C\int_0^T e^{-rt}\left(1 + X_t^{n-\vartheta}\right) - \hat{Z}_T, \end{aligned} \tag{37}$$

where

$$\hat{Z}_T = \int_0^T e^{-rt}\left[r\vartheta Y_t - CX_t^{\alpha}Y_t^{\beta}\right] dt \quad \text{for } T \geq 0.$$

With reference to (18), we note that

$$\begin{aligned} I_1(x) &:= E\left[C\int_0^{\infty} e^{-rt}\left(1 + X_t^{n-\vartheta}\right) dt\right] \\ &= \frac{C}{r} - \frac{Cx^{n-\vartheta}}{\sigma^2(n-\vartheta)^2 + (b-\sigma^2)(n-\vartheta) - r} \in]0,\infty[. \end{aligned} \tag{38}$$

Now, suppose that the strategy $(\xi^+, \xi^-) \in \Pi_y$ is associated with

$$E\left[\int_0^{\infty} e^{-rt} Y_t\, dt\right] = \infty. \tag{39}$$

With regard to (23) in Assumption 2 and (18), we observe that

$$I_2(x) := E\left[\int_0^{\infty} e^{-rt} X_t^{\alpha/(1-\beta)}\, dt\right] < \infty. \tag{40}$$

Therefore, given any constant $\mu > 0$,

$$E\left[\int_0^{\infty} e^{-rt} Y_t \mathbf{1}_{\left\{Y_t < \mu X_t^{\alpha/(1-\beta)}\right\}}\, dt\right] \leq \mu I_2(x) < \infty. \tag{41}$$

It follows that (39) is true if and only if

$$E\left[\int_0^{\infty} e^{-rt} Y_t \mathbf{1}_{\left\{Y_t \geq \mu X_t^{\alpha/(1-\beta)}\right\}}\, dt\right] = \infty. \tag{42}$$

Now, let any $\mu > 0$ such that $r\vartheta - C\mu^{-(1-\beta)} > 0$, where the constants $\vartheta, C > 0$ and $\beta \in]0,1[$ are as in Assumption 2, and note that

$$\begin{aligned} E\left[\hat{Z}_T\right] \geq\ & -C\mu^{\beta} E\left[\int_0^T e^{-rt} X_t^{\alpha/(1-\beta)} \mathbf{1}_{\left\{Y_t < \mu X_t^{\alpha/(1-\beta)}\right\}}\, dt\right] \\ & + \left(r\vartheta - C\mu^{-(1-\beta)}\right) E\left[\int_0^T e^{-rt} Y_t \mathbf{1}_{\left\{Y_t \geq \mu X_t^{\alpha/(1-\beta)}\right\}}\, dt\right]. \end{aligned}$$

In view of (41)–(42) and the monotone convergence theorem, the right-hand side of this inequality tends to ∞ as $T \to \infty$, which implies that $\lim_{T\to\infty} E[\hat{Z}_T] = \infty$. However, this conclusion, (37), and (38) imply that there exists a process Z such that (7) is satisfied and, therefore, $J_{x,y}(\xi^+, \xi^-) = -\infty$.

To proceed further, let us assume that

$$E\left[\int_0^{\infty} e^{-rt} Y_t\, dt\right] < \infty, \tag{43}$$

which is necessary for condition (32) to be satisfied. Since Y is a finite variation process, its sample paths can have at most countable discontinuities. Using Fubini's theorem, we can see that this observation and (43) imply that

$$\int_0^{\infty} e^{-rt} E\left[Y_{t+}\right] dt = E\left[\int_0^{\infty} e^{-rt} Y_{t+}\, dt\right] = E\left[\int_0^{\infty} e^{-rt} Y_t\, dt\right] < \infty,$$

which proves that (32) implies (33) and establishes part (b) of the lemma.

Now, using Hölder's inequality, we calculate

$$E\left[\int_0^\infty e^{-rt}X_t^\alpha Y_t^\beta\,dt\right] \le I_2^{1-\beta}(x)\left(E\left[\int_0^\infty e^{-rt}Y_t\,dt\right]\right)^\beta < \infty, \tag{44}$$

where $I_2(x)$ is given by (40). This inequality, (38), (43), and the bounds in (24) in Assumption 2 imply that

$$\begin{aligned} E\Big[\int_0^\infty e^{-rt}\,|h(X_t,Y_t)|\,dt\Big] &\le E\left[\int_0^\infty e^{-rt}\left[C\left(1+X_t^{n-\vartheta}+X_t^\alpha Y_t^\beta\right)+\left\{r(K^+-\vartheta)\vee C\right\}Y_t\right]dt\right] \\ &< \infty, \end{aligned}$$

which, combined with the dominated convergence theorem, implies that

$$\lim_{T\to\infty} E\left[\int_0^T e^{-rt}h(X_t,Y_t)\,dt\right] = E\left[\int_0^\infty e^{-rt}h(X_t,Y_t)\,dt\right] \in \mathbb{R}. \tag{45}$$

This observation gives rise to two possibilities. The first one is associated with the inequality

$$E\left[\int_{[0,\infty[} e^{-rt}\,d\xi_t^+ + \int_{[0,\infty[} e^{-rt}\,d\xi_t^-\right] < \infty.$$

In this case, $\lim_{T\to\infty} U_T$ exists, P-a.s., and belongs to $L^1(\Omega,\mathcal{F},P)$, and so $J_{x,y}(\xi^+,\xi^-)$ is finite and is given by (6). The second possibility is associated with

$$E\left[\int_{[0,\infty[} e^{-rt}\,d\xi_t^+ + \int_{[0,\infty[} e^{-rt}\,d\xi_t^-\right] = \infty,$$

which, combined with (43) and (33), implies that

$$E\left[\int_{[0,\infty[} e^{-rt}\,d\xi_t^+\right] = E\left[\int_{[0,\infty[} e^{-rt}\,d\xi_t^-\right] = \infty. \tag{46}$$

If $K^- < 0$, then we can use (1) and the integration by parts formula to calculate

$$\begin{aligned} K^-\int_{[0,T]} e^{-rt}\,d\xi_t^- &= K^-\int_{[0,T]} e^{-rt}\,d\xi_t^+ + |K^-|\int_{[0,T]} e^{-rt}\,dY_t \\ &= K^-\int_{[0,T]} e^{-rt}\,d\xi_t^+ + r|K^-|\int_0^T e^{-rt}Y_t\,dt \\ &\quad + |K^-|e^{-rT}Y_{T+} - |K^-|y \\ &\ge K^-\int_{[0,T]} e^{-rt}\,d\xi_t^+ - |K^-|y, \end{aligned}$$

which implies that

$$\begin{aligned} E\Big[K^+\int_{[0,T]} e^{-rt}\,d\xi_t^+ + K^-\int_{[0,T]} e^{-rt}\,d\xi_t^-\Big] \\ \ge (K^+ + K^-)E\left[\int_{[0,T]} e^{-rt}\,d\xi_t^+\right] - |K^-|y. \end{aligned}$$

This inequality, the assumption that $K^+ + K^- > 0$, (46), and the monotone convergence theorem imply that

$$\lim_{T\to\infty} E\left[K^+ \int_{[0,T]} e^{-rt}\, d\xi_t^+ + K^- \int_{[0,T]} e^{-rt}\, d\xi_t^-\right] = \infty. \tag{47}$$

On the other hand, if $K^- \geq 0$, then (46) plainly implies (47). However, (45) and (47) imply that $\lim_{T\to\infty} E[U_T] = -\infty$, and so (7) is satisfied for $Z = U$ and $J_{x,y}(\xi^+, \xi^-) = -\infty$.

The analysis above establishes the well-posedness of the definition of $J_{x,y}$ given by (5)–(8) as well as parts (a) and (b) of the lemma. To prove part (c) of the lemma, we first note that the first bound in Lemma 2 and (18) imply that

$$R^{[h(\cdot,y)]}(x) = E\left[\int_0^\infty e^{-rt} h(X_t, y)\, dt\right] \in \mathbb{R}.$$

However, this shows that our performance criterion is finite for the strategy that involves no capacity changes at any time, which proves that $v(x,y) > -\infty$. To show that $v(x,y) < \infty$, consider any admissible decision strategy $(\xi^+, \xi^-) \in \Pi_y$ such that $J_{x,y}(\xi^+, \xi^-) > -\infty$. With reference to (43) and (44),

$$\begin{aligned} &E\left[\int_0^\infty e^{-rt}\left[r\vartheta Y_t - CX_t^\alpha Y_t^\beta\right] dt\right] \\ &\qquad \geq r\vartheta E\left[\int_0^\infty e^{-rt} Y_t\, dt\right] - CI_2^{1-\beta}(x)\left(E\left[\int_0^\infty e^{-rt} Y_t\, dt\right]\right)^\beta \\ &\qquad \geq -\frac{(1-\beta)r\vartheta}{\beta}\left(\frac{\beta C}{r\vartheta}\right)^{1/(1-\beta)} I_2(x), \quad \text{for all } T > 0, \end{aligned} \tag{48}$$

the second inequality following because, given any constants $\kappa, \lambda > 0$ and $\beta \in]0,1[$,

$$\kappa Q - \lambda Q^\beta \geq -\frac{(1-\beta)\kappa}{\beta}\left(\frac{\beta\lambda}{\kappa}\right)^{1/(1-\beta)}, \quad \text{for all } Q \geq 0,$$

in particular, for $Q = E\left[\int_0^\infty e^{-rt} Y_t\, dt\right]$. However, (37), (38), and (48) imply that

$$J_{x,y}(\xi^+, \xi^-) \leq I_1(x) + K^+ y + \frac{(1-\beta)r\vartheta}{\beta}\left(\frac{\beta C}{r\vartheta}\right)^{1/(1-\beta)} I_2(x),$$

which proves that $v(x,y) < \infty$ because the right-hand side of this inequality is finite and independent of ξ^+ and ξ^-. □

4. The Hamilton–Jacobi–Bellman (HJB) equation. The problem described in the previous section has the structure of a singular stochastic control problem. With regard to standard theory of singular control, we expect that its value function can be identified with a solution $w : \mathcal{S} \to \mathbb{R}$ to the HJB quasi-variational inequalities

$$\begin{aligned} &\max\{\sigma^2 x^2 w_{xx}(x,y) + bx w_x(x,y) - rw(x,y) + h(x,y), \\ &\qquad w_y(x,y) - K^+,\ -w_y(x,y) - K^-\} = 0, \quad x, y > 0, \end{aligned} \tag{49}$$

$$\max\{\sigma^2 x^2 w_{xx}(x,0) + bx w_x(x,0) - rw(x,0) + h(x,0),\ w_y(x,0) - K^+\} = 0, \quad x > 0, \tag{50}$$

where $w_y(x,0) := \lim_{y\downarrow 0} w_y(x,y)$.

To obtain some qualitative understanding of the origins of this equation, we observe that, at time 0, the project's management has to choose between three options. The first one is to wait for a short time Δt and then continue optimally. With respect to Bellman's principle of optimality, this option is associated with the inequality

$$v(x,y) \geq E\left[\int_0^{\Delta t} e^{-rt}h(X_t,y)\,dt + e^{-r\Delta t}v(X_{\Delta t},y)\right].$$

Applying Itô's formula to the second term in the expectation and dividing by Δt before letting $\Delta t \downarrow 0$, we obtain

$$\sigma^2x^2v_{xx}(x,y) + bxv_x(x,y) - rv(x,y) + h(x,y) \leq 0. \tag{51}$$

The second option is to increase capacity immediately by $\varepsilon > 0$ and then continue optimally. This action is associated with the inequality

$$v(x,y) \geq v(x,y+\varepsilon) - K^+\varepsilon.$$

Rearranging terms and letting $\varepsilon \downarrow 0$, we obtain

$$v_y(x,y) - K^+ \leq 0. \tag{52}$$

Assuming that $y > 0$, the final option is to decrease capacity immediately by $\varepsilon > 0$ and then continue optimally. This option yields the inequality

$$v(x,y) \geq v(x,y-\varepsilon) - K^-\varepsilon,$$

which, in the limit as $\varepsilon \downarrow 0$, implies that

$$-v_y(x,y) - K^- \leq 0. \tag{53}$$

Since these three are the only options available, we expect that one of them should be optimal, so that one of the inequalities (51)–(53) should hold with equality if $y > 0$, while one of the inequalities (51)–(52) should hold with equality if $y = 0$. However, this observation combined with (51)–(53) implies that the value function v should identify with a solution w to (49)–(50).

The following result is concerned with *sufficient* conditions under which the value function v of the control problem considered identifies with a solution to (49)–(50). We impose some of these conditions, (58)–(59) in particular, which are not standard in similar "verification" theorems, with a hindsight relative to our analysis in the next section.

THEOREM 5. *Suppose that the running payoff function h satisfies* (24) *in Assumption* 2 *and that $K^+, K^+ + K^- > 0$. Also, assume that the HJB equation* (49)–(50) *has a C^2 solution $w : \mathcal{S} \to \mathbb{R}$ such that*

$$-C_2\left(1 + y + x^{\alpha/(1-\beta)}\right) \leq w(x,y), \quad \textit{for all } (x,y) \in \mathcal{S}, \tag{54}$$

for some constant $C_2 > 0$. The following statements hold true:

(a) *$v(x,y) \leq w(x,y)$ for all initial conditions $(x,y) \in \mathcal{S}$.*

(b) *Given any initial condition $(x,y) \in \mathcal{S}$, suppose that there exists a decision strategy $(\xi^{o+}, \xi^{o-}) \in \Pi_y$ such that, if Y^o is the associated capacity process, then*

$$(X_t, Y_t^o) \in \left\{(x,y) \in \mathcal{S} :\ \sigma^2x^2w_{xx}(x,y) + bxw_x(x,y) - rw(x,y) + h(x,y) = 0\right\}, \tag{55}$$

Lebesgue-a.e., P-a.s.,

$$\int_{[0,T]} e^{-rs}\left[w_y(X_t,Y_t)-K^+\right]d\xi_s^{o+}=0,\quad \textit{for all } T\geq 0,\ P\textit{-a.s.}, \tag{56}$$

$$\int_{[0,T]} e^{-rs}\left[w_y(X_t,Y_t)+K^-\right]d\xi_s^{o-}=0,\quad \textit{for all } T\geq 0,\ P\textit{-a.s.}, \tag{57}$$

and

$$Y_t^o+X_t^\alpha (Y_t^o)^\beta+\xi_t^{o+}\leq C_3(y)\left(1+\bar{X}_t^{n-\varepsilon_3}\right),\quad \textit{for all } t\geq 0,\ P\textit{-a.s.}, \tag{58}$$

$$w(X_t,Y_t^o)\leq C_3(y)\left(1+\bar{X}_t^{n-\varepsilon_3}\right),\quad \textit{for all } t\geq 0,\ P\textit{-a.s.}, \tag{59}$$

where $\bar{X}_t=\sup_{s\leq t}X_s$, $\varepsilon_3\in\,]0,\vartheta[$ *is a constant, and* $C_3(y)>0$ *is a constant depending on the initial condition* y *only. Then* $v(x,y)=w(x,y)$ *and* (ξ^{o+},ξ^{o-}) *is the optimal strategy.*

Proof. (a) Fix any initial condition (x,y) and any admissible strategy $(\xi^+,\xi^-)\in\Pi_y$ such that $J_{x,y}(\xi^+,\xi^-)>-\infty$, so that $J_{x,y}(\xi^+,\xi^-)=E[U_\infty]$ (see (5)–(6)). Using Itô's formula and the fact that X has continuous sample paths, we obtain

$$\begin{aligned}
&e^{-rT}w(X_T,Y_{T+})\\
&\quad=w(x,y)+\int_0^T e^{-rt}\left[\sigma^2X_t^2w_{xx}(X_t,Y_t)+bX_tw_x(X_t,Y_t)-rw(X_t,Y_t)\right]dt\\
&\qquad+\int_{[0,T]} e^{-rt}\left[w_y(X_t,Y_t)\,d\xi_t^+-w_y(X_t,Y_t)\,d\xi_t^-\right]+M_T\\
&\qquad+\sum_{0\leq t\leq T} e^{-rt}\left[w(X_t,Y_{t+})-w(X_t,Y_t)-w_y(X_t,Y_t)\Delta Y_t\right],
\end{aligned}$$

where

$$M_T=\sqrt{2}\sigma\int_0^T e^{-rt}X_tw_x(X_t,Y_t)\,dW_t,\quad T\geq 0. \tag{60}$$

Recalling the definition of U_T in (4), this implies that

$$\begin{aligned}
&U_T+e^{-rT}w(X_T,Y_{T+})-w(x,y)\\
&=\int_0^T e^{-rt}\left[\sigma^2X_t^2w_{xx}(X_t,Y_t)+bX_tw_x(X_t,Y_t)-rw(X_t,Y_t)+h(X_t,Y_t)\right]dt\\
&\quad+\int_{[0,T]} e^{-rt}\left[w_y(X_t,Y_t)-K^+\right]d\left(\xi^+\right)_t^c\\
&\quad+\int_{[0,T]} e^{-rt}\left[-w_y(X_t,Y_t)-K^-\right]d\left(\xi^-\right)_t^c\\
&\quad+M_T+\sum_{0\leq t\leq T} e^{-rt}\left[w(X_t,Y_{t+})-w(X_t,Y_t)-K^+\Delta Y_t\right]\mathbf{1}_{\{\Delta Y_t>0\}}\\
&\quad+\sum_{0\leq t\leq T} e^{-rt}\left[w(X_t,Y_{t+})-w(X_t,Y_t)+K^-\Delta Y_t\right]\mathbf{1}_{\{\Delta Y_t<0\}}.
\end{aligned}$$

Observing that

$$\begin{aligned}
&\left[w(X_t,Y_{t+})-w(X_t,Y_t)-K^+\Delta Y_t\right]\mathbf{1}_{\{\Delta Y_t>0\}}\\
&\qquad=\mathbf{1}_{\{\Delta Y_t>0\}}\int_0^{\Delta Y_t}\left[w_y(X_t,Y_t+u)-K^+\right]du
\end{aligned}$$

and

$$\begin{aligned}\left[w(X_t,Y_{t+})-w(X_t,Y_t)+K^-\Delta Y_t\right]\mathbf{1}_{\{\Delta Y_t<0\}}&\\ =\mathbf{1}_{\{\Delta Y_t<0\}}\int_0^{|\Delta Y_t|}\left[-w_y\left(X_t,Y_t-|\Delta Y_t|+u\right)-K^-\right]du,&\end{aligned}$$

we can see that, since w satisfies the HJB equation (49)–(50),

$$U_T+e^{-rT}w(X_T,Y_{T+})\le w(x,y)+M_T. \tag{61}$$

Now, in view of (36) and the assumption $K^+>0$,

$$-e^{-rT}Y_{T+}\ge-\int_{[0,T]}e^{-rt}\,d\xi_t^+-\frac{|K^-|}{K^+}\int_{[0,T]}e^{-rt}\,d\xi_t^--y,$$

which, combined with assumption (54), implies that

$$e^{-rT}w(X_T,Y_{T+})\ge-C_{21}\left(1+\int_{[0,T]}e^{-rt}\,d\xi_t^++\int_{[0,T]}e^{-rt}\,d\xi_t^-+e^{-rT}X_T^{\alpha/(1-\beta)}\right)$$

for some constant $C_{21}=C_{21}(y)>0$. Combining this inequality with

$$\int_0^T e^{-rt}h(X_t,Y_t)\,dt\ge-C\int_0^T e^{-rt}Y_t\,dt-\frac{C}{r}\left(1-e^{-rT}\right),$$

which follows from (24) in Assumption 2, we can see that (61) implies

$$\begin{aligned}\inf_{T\ge0}M_T\ge-C_{22}\Bigg(1+\int_0^\infty e^{-rt}Y_t\,dt+\int_{[0,\infty[}e^{-rt}\,d\xi_t^+&\\ +\int_{[0,\infty[}e^{-rt}\,d\xi_t^-+\sup_{T\ge0}e^{-rT}\bar X_T^{\alpha/(1-\beta)}\Bigg),&\end{aligned}$$

where $C_{22}=C_{22}(x,y)>0$ is a constant and $\bar X_t=\sup_{s\le t}X_s$. Recalling the assumption that $\frac{\alpha}{1-\beta}\in]0,n[$, we can see that the second bound in Lemma 1 and (32) in Lemma 4 imply that the random variable on the right-hand side of this inequality has finite expectation. It follows that the stochastic integral M defined by (60) is a supermartingale, and therefore $E[M_T]\le0$, for all $T>0$. Taking expectations in (61), we therefore obtain

$$E[U_T]\le w(x,y)+e^{-rT}E\left[-w(X_T,Y_{T+})\right]. \tag{62}$$

Furthermore, since

$$U_T\ge-C_{22}\left(1+\int_0^\infty e^{-rt}Y_t\,dt+\int_{[0,\infty[}e^{-rt}\,d\xi_t^++\int_{[0,\infty[}e^{-rt}\,d\xi_t^-\right),\quad\text{for all } T\ge0,$$

and the random variable on the right-hand side of this inequality has finite expecta-

tion, Fatou's lemma implies that

$$J_{x,y}(\xi^+,\xi^-) \leq \liminf_{T\to\infty} E\left[U_T\right], \tag{63}$$

while (54) implies that

$$\begin{aligned}\liminf_{T\to\infty} e^{-rT}E\left[-w(X_T,Y_{T+})\right] \leq{}& \lim_{T\to\infty} e^{-rT}C_2 + C_2 \liminf_{T\to\infty} e^{-rT}E\left[Y_{T+}\right]\\ &+ C_2 \lim_{T\to\infty} e^{-rT}E\left[\bar{X}_T^{\alpha/(1-\beta)}\right]\\ ={}& 0,\end{aligned} \tag{64}$$

the equality being true thanks to the first bound in Lemma 1 and (33). However, (62)–(64) imply that $J_{x,y}(\xi^+,\xi^-) \leq w(x,y)$, which establishes part (a) of the theorem.

(b) If $(\xi^{\mathrm{o}+},\xi^{\mathrm{o}-})$ is as in the statement of the theorem, then we can see that the monotone convergence theorem, the integration by parts formula, (58), and the first estimate in Lemma 1 imply that

$$\begin{aligned}&E\left[\int_0^\infty e^{-rt}Y_t^{\mathrm{o}}\,dt + \int_{[0,\infty[} e^{-rt}\,d\xi_t^{\mathrm{o}+}\right]\\ &\qquad= \lim_{T\to\infty} E\left[\int_0^T e^{-rt}Y_t^{\mathrm{o}}\,dt + r\int_0^T e^{-rt}\xi_t^{\mathrm{o}+}\,dt + e^{-rT}\xi_{T+}^{\mathrm{o}+}\right]\\ &\qquad\leq (1+r)C_3(y)\left(\frac{1}{r} + \int_0^\infty e^{-rt}E\left[\bar{X}_t^{n-\varepsilon_3}\right]dt\right) + \lim_{T\to\infty} e^{-rT}E\left[\bar{X}_T^{n-\varepsilon_3}\right]\\ &\qquad< \infty,\end{aligned}$$

which, combined with (1), implies that (32) in Lemma 4 is satisfied, and, therefore,

$$J_{x,y}(\xi^{\mathrm{o}+},\xi^{\mathrm{o}-}) = E\left[\lim_{T\to\infty} U_T^{\mathrm{o}}\right] \in \mathbb{R}, \tag{65}$$

where U^{o} is defined as in (4). Furthermore, we can verify that (61) holds with equality, i.e.,

$$U_T^{\mathrm{o}} + e^{-rT}w(X_T,Y_{T+}^{\mathrm{o}}) = w(x,y) + M_T^{\mathrm{o}}, \tag{66}$$

where the stochastic integral M^{o} is defined as in (60). In view of (24) in Assumption 2 and (58), there exist constants $C_{31} > 0$ and $C_{32} = C_{32}(y) > 0$ such that

$$\begin{aligned}\sup_{T\geq 0}\int_0^T e^{-rt}h(X_t,Y_t^{\mathrm{o}})\,dt &\leq C_{31}\left(1 + \int_0^\infty e^{-rt}\left[X_t^{n-\vartheta} + X_t^{\alpha}(Y_t^{\mathrm{o}})^{\beta} + Y_t^{\mathrm{o}}\right]dt\right)\\ &\leq C_{32}\left(1 + \int_0^\infty e^{-rt}\bar{X}_t^{n-\varepsilon_3}\,dt\right).\end{aligned} \tag{67}$$

With reference to (1), the assumption $K^+ + K^- > 0$, the integration by parts formula,

and (58), we can see that there exists a constant $C_{33} = C_{33}(y) > 0$ such that

$$\begin{aligned}
\sup_{T\geq 0}\left(-K^+\int_{[0,T]} e^{-rt}\,d\xi_t^{o+} - K^-\int_{[0,T]} e^{-rt}\,d\xi_t^{o-}\right)& \\
&\leq \sup_{T\geq 0} K^-\left(\int_{[0,T]} e^{-rt}\,d\xi_t^{o+} - \int_{[0,T]} e^{-rt}\,d\xi_t^{o-}\right)\\
&\leq |K^-|\sup_{T\geq 0}\int_{[0,T]} e^{-rt}\,dY_t^o\\
&\leq |K^-|\sup_{T\geq 0} e^{-rT}Y_{T+}^o + r|K^-|\int_0^\infty e^{-rt}Y_t^o\,dt\\
&\leq |K^-|\sup_{T\geq 0} e^{-rT}Y_{T+}^o + C_{33}\left(1+\int_0^\infty e^{-rt}\bar{X}_t^{n-\varepsilon_3}\,dt\right). \qquad (68)
\end{aligned}$$

Moreover, (58)–(59) imply that

$$\sup_{T\geq 0} e^{-rT}Y_{T+}^o + \sup_{T\geq 0} e^{-rT}w(X_T, Y_{T+}^o) \leq 2C_3(y)\left(1+\sup_{T\geq 0} e^{-rT}\bar{X}_T^{n-\varepsilon_3}\right). \qquad (69)$$

Now, (18) implies that

$$E\left[\int_0^\infty e^{-rt}\bar{X}_t^{n-\varepsilon_3}\,dt\right] < \infty, \qquad (70)$$

while the second estimate in Lemma 1 implies that

$$E\left[\sup_{T\geq 0} e^{-rT}\bar{X}_T^{n-\varepsilon_3}\right] < \infty. \qquad (71)$$

However, (66) and the estimates (67)–(71) imply that $E\left[\sup_{T\geq 0} M_T^o\right] < \infty$, which proves that the stochastic integral M^o is a submartingale. Taking expectations in (66), we therefore obtain

$$E\left[U_T^o\right] \geq w(x,y) + e^{-rT}E\left[-w(X_T, Y_T^o)\right]. \qquad (72)$$

Furthermore, the estimates (67)–(71) imply that the random variables U_T^o, indexed by $T \geq 0$, are all bounded from above by a random variable with finite expectation. This observation, (65), and Fatou's lemma imply that

$$J_{x,y}(\xi^{o+}, \xi^{o-}) \geq \limsup_{T\to\infty} E\left[U_T^o\right]. \qquad (73)$$

Finally, (59) and the first estimate in Lemma 1 imply that

$$\begin{aligned}
\limsup_{T\to\infty} e^{-rT}E\left[-w(X_T, Y_T^o)\right] &\geq -\lim_{T\to\infty} C_3(y)\left(e^{-rT} + E\left[e^{-rT}\bar{X}_T^{n-\varepsilon_3}\right]\right)\\
&= 0,
\end{aligned}$$

which, combined with (72) and (73), implies that $J_{x,y}(\xi^{o+}, \xi^{o-}) \geq w(x,y)$. However, this inequality and part (a) of this theorem complete the proof. □

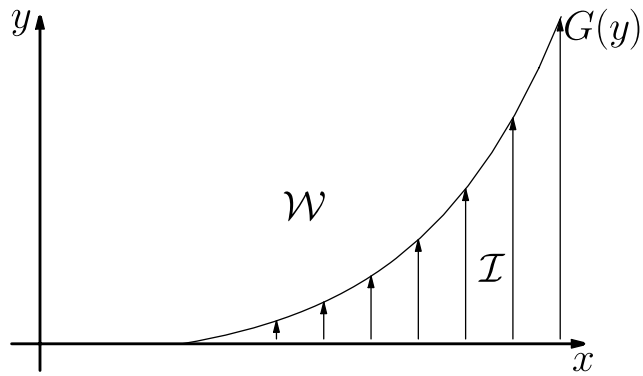


FIG. 1. *A possible optimal capacity control strategy. In this case, it is never optimal to decrease the project's capacity.*

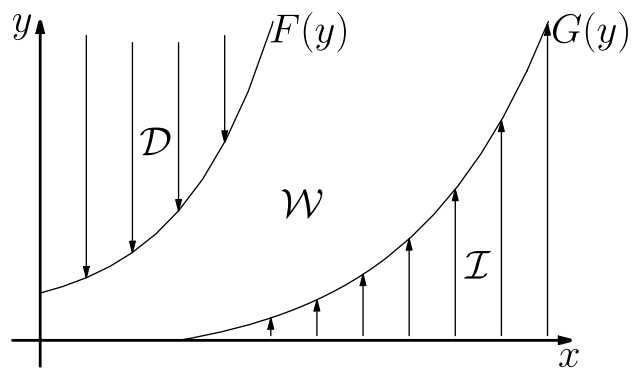


FIG. 2. *A possible optimal capacity control strategy. In this case, increasing the project's capacity, waiting, and decreasing the project's capacity are all parts of the optimal strategy. Also, the point* y^* *defined by* (74) *is strictly positive.*

5. The solution of the control problem. We can now derive an explicit solution to the control problem formulated in section 2 by constructing an appropriate solution w to the HJB equation (49)–(50). With respect to the heuristic arguments in section 4 that led to the derivation of this equation, we start by conjecturing that the optimal strategy is characterized by three disjoint open subsets of $]0,\infty[\times\mathbb{R}_+$: the "wait" region $\mathcal{W}$ where (51) holds with equality, the "investment" region $\mathcal{I}$ where (52) holds with equality, and the "disinvestment" region $\mathcal{D}$ where (53) holds with equality. Also, we conjecture that each of the regions $\mathcal{W}$, $\mathcal{I}$, $\mathcal{D}$ is connected. In particular, we expect that, depending on the problem data, the optimal strategy can take any of the forms depicted by Figures 1–4. Note that one can envisage other possibilities such as the one depicted by Figure 5. However, our assumptions do not allow for the optimality of other such cases under any admissible choice of the problem data (see also Remark 1 in section 3 and Example 3 in section 6).

With regard to Figures 1–4, we denote by $\mathbb{F}$ and $\mathbb{G}$ the boundaries separating the regions $\mathcal{D}$, $\mathcal{W}$ and $\mathcal{W}$, $\mathcal{I}$, respectively, so that

$$\mathbb{F} = \overline{\mathcal{D}} \cap \overline{\mathcal{W}} \quad \text{and} \quad \mathbb{G} = \overline{\mathcal{W}} \cap \overline{\mathcal{I}},$$

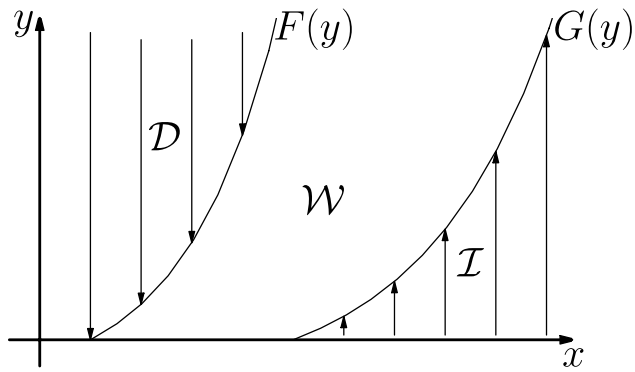


FIG. 3. *A possible optimal capacity control strategy. In this case, increasing the project's capacity, waiting, and decreasing the project's capacity all belong to the set of optimal tactics. Also,* $y^* = 0$*, where* y^* *is defined by* (74)*,* $F(0) > 0$*, and* $\{(x, 0) : \ x \leq F(0)\}$ *is a subset of the "wait" region* $\mathcal{W}$*.*

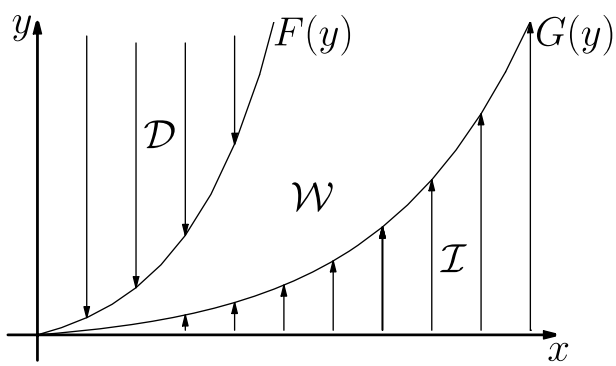


FIG. 4. *A possible optimal capacity control strategy. This case arises when the running payoff function* h *identifies with the Cobb–Douglas production function and* $K^- < 0$*.*

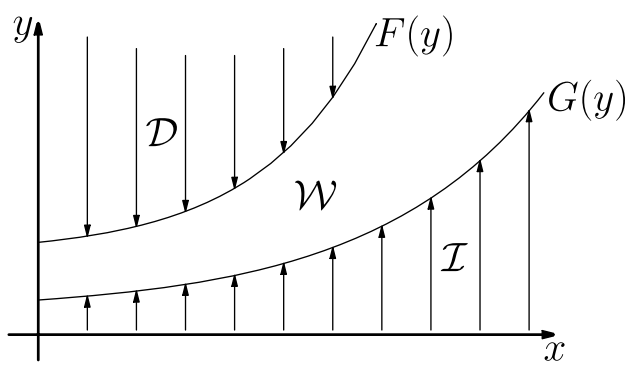


FIG. 5. *A possible optimal capacity control strategy. This case* cannot *arise under our assumptions.*

where $\overline{\mathcal{W}}$, $\overline{\mathcal{I}}$, and $\overline{\mathcal{D}}$ are the closures of $\mathcal{W}$, $\mathcal{I}$, and $\mathcal{D}$ in $\mathbb{R}^2_+$, respectively. Furthermore, we define

$$y^* = \inf\{y \geq 0 : \text{ there exists } x > 0 \text{ such that } (x,y) \in \mathbb{F}\}, \tag{74}$$

with the usual convention that $\inf \emptyset = \infty$. We will prove that

$$\begin{gathered}\text{there exists an increasing function } G : [0,\infty[\to [0,\infty[\text{ such that} \\ \mathbb{G} = \{(G(y), y) : \ y \geq 0\},\end{gathered} \tag{75}$$

and, if $y^* < \infty$, then

$$\begin{gathered}\text{there exists an increasing function } F : [y^*,\infty[\to [0,\infty[\text{ such that} \\ \mathbb{F} \cap (\mathbb{R}_+ \setminus \{0\})^2 = \{(F(y), y) : \ y > y^*\}.\end{gathered} \tag{76}$$

Given such a characterization of $\mathbb{F}$ and $\mathbb{G}$,

$$\begin{aligned}\overline{\mathcal{W}} = & \left\{(x,y) \in \mathbb{R}^2_+ : \ y \leq y^* \text{ and } x \in [0, G(y)]\right\} \\ & \cup \left\{(x,y) \in \mathbb{R}^2_+ : \ y > y^* \text{ and } x \in [F(y), G(y)]\right\}, \\ \overline{\mathcal{I}} = & \left\{(x,y) \in \mathbb{R}^2_+ : \ G(y) \leq x\right\},\end{aligned}$$

while if $y^* < \infty$, then

$$\overline{\mathcal{D}} = \left\{(x,y) \in \mathbb{R}^2_+ : \ y \geq y^* \text{ and } x \in [0, F(y)]\right\}.$$

In view of this structure, it is worth noting that if $y^* = 0$ and $0 < F(0) < G(0)$ (see Figure 3), then $\{(x,0) : \ x < G(0)\} \subset \mathcal{W}$, so that the segment $]0, F(0)]$ is part of the "wait" region $\mathcal{W}$.

Inside the region $\mathcal{W}$, w satisfies the differential equation

$$\sigma^2 x^2 w_{xx}(x,y) + bx w_x(x,y) - rw(x,y) + h(x,y) = 0. \tag{77}$$

In view of the discussion regarding the solvability of (14) in section 3, every solution to this equation is given by

$$w(x,y) = A(y)x^n + B(y)x^m + R(x,y) \tag{78}$$

for some functions A and B. Here, the constants $m < 0 < n$ are given by (11), while the function $R \equiv R^{[h(\cdot,y)]}$ is given by

$$R(x,y) = \frac{1}{\sigma^2(n-m)}\left[x^m \int_0^x s^{-m-1} h(s,y)\, ds + x^n \int_x^\infty s^{-n-1} h(s,y)\, ds\right]. \tag{79}$$

For $y \in [0, y^*] \cap \mathbb{R}$, we must have $B(y) = 0$. This choice is supported by the heuristic observation that, for fixed capacity level $y \geq 0$, the problem's value function should remain bounded as the value x of the underlying state process tends to 0. Also, it eventually turns out that (58)–(59) in the verification theorem, Theorem 5, cannot be satisfied if $B(y) \neq 0$. To determine $A(y)$ and $G(y)$ when $y \in [0, y^*] \cap \mathbb{R}$, we postulate that $w(\cdot, y)$ is C^2 at the free-boundary point $G(y)$. In particular, we postulate that

$$\lim_{x \uparrow G(y)} w_y(x,y) = \lim_{x \downarrow G(y)} w_y(x,y) \quad \text{and} \quad \lim_{x \uparrow G(y)} w_{yx}(x,y) = \lim_{x \downarrow G(y)} w_{yx}(x,y). \tag{80}$$

Since w satisfies

$$w_y(x,y) = K^+, \quad \text{for } (x,y) \in \mathcal{I}, \tag{81}$$

which implies that

$$w_{xy}(x,y) = 0, \quad \text{for } (x,y) \in \mathcal{I}, \tag{82}$$

this requirement yields the system of equations

$$A'(y)G^n(y) = K^+ - R_y(G(y),y), \tag{83}$$

$$A'(y)G^n(y) = -\frac{1}{n}G(y)R_{xy}(G(y),y). \tag{84}$$

Equating the right-hand sides of these equations and using the definition of R in (79), we obtain

$$G^m(y)\int_0^{G(y)} s^{-m-1}H(s,y)\,ds - \sigma^2 nK^+ = 0, \tag{85}$$

where H is the function defined by (19). Using the identity $\sigma^2 mn = -r$, which follows from the definition of the constants m, n in (11), we can see that $G(y)$ should satisfy

$$q(G(y),y) = 0, \tag{86}$$

where

$$q(x,y) = \int_0^x s^{-m-1}\left[H(s,y) - rK^+\right]ds, \quad (x,y) \in \mathcal{S}. \tag{87}$$

Furthermore, adding (83) and (84) side by side and using (79) and (85), we obtain

$$\begin{aligned} A'(y) &= \frac{1}{2}G^{-n}(y)\left[K^+ - R_y(G(y),y) - \frac{1}{n}G(y)R_{xy}(G(y),y)\right] \\ &= -\frac{1}{\sigma^2(n-m)}\int_{G(y)}^{\infty} s^{-n-1}\left[H(s,y) - rK^+\right]ds. \end{aligned} \tag{88}$$

The following result, whose proof is developed in the appendix, is concerned with the solvability of (86).

LEMMA 6. *Suppose that Assumption* 1 *is true. Given any* $y \geq 0$*, the equation* $q(x,y) = 0$ *has a unique solution* $x = x(y) > 0$ *if and only if* $\inf_{x>0} H(x,y) < rK^+$*. If we define*

$$\tilde{y}_* = \inf\left\{y \geq 0 : \inf_{x>0} H(x,y) < rK^+\right\}, \tag{89}$$

then (86) *uniquely defines a function* $\tilde{G} :]\tilde{y}_*, \infty[\to]0, \infty[$ *that is* C^1*, is strictly increasing, and satisfies*

$$H(\tilde{G}(y),y) - rK^+ > 0 \quad \textit{for all } y > \tilde{y}_*. \tag{90}$$

Furthermore, if (25) *in Assumption* 2 *is also true, then* $\tilde{y}_* = 0$ *and*

$$\begin{aligned} &C_4^{-\frac{1-\beta}{\alpha}} y^{\frac{1-\beta}{\alpha}} \leq \tilde{G}(y), \quad \textit{for all } y \geq 0 \\ &\qquad\qquad \Leftrightarrow \quad \tilde{G}^{[-1]}(x) \leq C_4 x^{\frac{\alpha}{1-\beta}}, \quad \textit{for all } x \geq \tilde{G}(0), \end{aligned} \tag{91}$$

where $\tilde{G}(0) := \lim_{y \downarrow 0} \tilde{G}(y)$, $\tilde{G}^{[-1]} : [\tilde{G}(0), \infty[\to \mathbb{R}_+$ *is the inverse function of* $\tilde{G}$, *and* $C_4 > 0$ *is a constant.*

Now, let us consider the case where $\mathcal{D} \neq \emptyset$ and the point y^* defined by (74) is finite (see Figures 2–4). For $y > y^*$, w is given by (77) for x such that $(x, y) \in \mathcal{W}$, by (81) for x such that $(x, y) \in \mathcal{I}$, and by

$$w_y(x, y) = -K^- \tag{92}$$

for x such that $(x, y) \in \mathcal{D}$. Plainly, C^2 continuity of w inside $\mathcal{D}$ implies that

$$w_{xy}(x, y) = 0 \quad \text{for } (x, y) \in \mathcal{D}. \tag{93}$$

To determine $A(y)$, $B(y)$, $F(y)$, and $G(y)$, we postulate that $w(\cdot, y)$ is C^2 at both of the free-boundary points $F(y)$ and $G(y)$. With regard to (78), (81)–(82), (92)–(93), the definition (79) of $R(x, y)$, and the identity $\sigma^2 mn = -r$, this requirement yields

$$A'(y) = -\frac{1}{\sigma^2(n-m)} \int_{F(y)}^{\infty} s^{-n-1} \left[H(s, y) + rK^-\right] ds, \tag{94}$$

$$A'(y) = -\frac{1}{\sigma^2(n-m)} \int_{G(y)}^{\infty} s^{-n-1} \left[H(s, y) - rK^+\right] ds, \tag{95}$$

$$B'(y) = -\frac{1}{\sigma^2(n-m)} \int_{0}^{F(y)} s^{-m-1} \left[H(s, y) + rK^-\right] ds, \tag{96}$$

$$B'(y) = -\frac{1}{\sigma^2(n-m)} \int_{0}^{G(y)} s^{-m-1} \left[H(s, y) - rK^+\right] ds, \tag{97}$$

where H is defined by (19). These calculations imply that the points $F(y)$ and $G(y)$ should satisfy the system of equations

$$f(F(y), G(y), y) = 0, \tag{98}$$

$$g(F(y), G(y), y) = 0, \tag{99}$$

where

$$f(x_1, x_2, y) = \int_0^{x_1} s^{-m-1} \left[H(s, y) + rK^-\right] ds - \int_0^{x_2} s^{-m-1} \left[H(s, y) - rK^+\right] ds, \tag{100}$$

$$g(x_1, x_2, y) = \int_{x_1}^{\infty} s^{-n-1} \left[H(s, y) + rK^-\right] ds - \int_{x_2}^{\infty} s^{-n-1} \left[H(s, y) - rK^+\right] ds. \tag{101}$$

In the appendix, we prove the following result that is concerned with the solvability of the system of equations (98) and (99).

LEMMA 7. *Suppose that Assumption* 1 *holds. Given* $y \geq 0$, *the system of equations* (98) *and* (99) *has a unique solution* $(x_1, x_2) = (x_1(y), x_2(y))$ *such that* $0 < x_1 < x_2$ *if and only if* $\inf_{x>0} H(x, y) < -rK^-$. *Moreover, if we define*

$$\bar{y}^* = \inf \left\{ y \geq 0 : \inf_{x>0} H(x, y) < -rK^- \right\}, \tag{102}$$

with the usual convention that $\inf \emptyset = \infty$, *then, if* $\bar{y}^* < \infty$, *the system of equations* (98) *and* (99) *uniquely defines two functions* $\bar{F}, \bar{G} :]\bar{y}^*, \infty[\to]0, \infty[$ *that are* C^1, *are*

strictly increasing, and satisfy $\bar{F}(y) < \bar{G}(y)$*, for all* $y > \bar{y}^*$*,*

$$\bar{F}(\bar{y}^*) := \lim_{y \downarrow \bar{y}^*} \bar{F}(y) = 0, \quad \text{if } \bar{y}^* > 0, \tag{103}$$

$$\bar{F}(0) := \lim_{y \downarrow 0} \bar{F}(y) \leq \lim_{y \downarrow 0} \bar{G}(y) =: \bar{G}(0), \quad \text{if } \bar{y}^* = 0, \tag{104}$$

$$H(\bar{F}(y), y) + rK^- < 0 \quad \text{and} \quad H(\bar{G}(y), y) - rK^+ > 0 \quad \text{for all } y > \bar{y}^*. \tag{105}$$

Furthermore, if (25) *in Assumption* 2 *also holds, then*

$$\begin{aligned} &C_4^{-\frac{1-\beta}{\alpha}} y^{\frac{1-\beta}{\alpha}} \leq \bar{G}(y), \quad \text{for all } y \geq \bar{y}^* \\ &\qquad\qquad \Leftrightarrow \quad \bar{G}^{[-1]}(x) \leq C_4 x^{\frac{\alpha}{1-\beta}}, \quad \text{for all } x \geq \bar{G}(\bar{y}^*), \end{aligned} \tag{106}$$

where $\bar{G}^{[-1]} : [\bar{G}(0), \infty[\to \mathbb{R}_+$ *is the inverse function of* $\bar{G}$ *and the constant* $C_4 > 0$ *is the same constant as in Lemma* 6*.*

In light of the results above and in the presence of (25) in Assumption 2, $\tilde{y}_* = \infty$, where $\tilde{y}_*$ is defined by (89), and the point $\bar{y}^*$ defined by (102) identifies with the point y^* in (74). Also, the functions $F : [y^*, \infty[\to [0, \infty[$ and $G : [0, \infty[\to [0, \infty[$ separating the three possible regions, as conjectured in (75)–(76), are given by

$$F = \bar{F}, \text{ if } y^* < \infty, \tag{107}$$

$$G = \tilde{G}, \text{ if } y^* = \infty, \quad \text{and} \quad G(y) = \begin{cases} \tilde{G}(y), & \text{for } y \in [0, y^*], \\ \bar{G}(y), & \text{for } y > y^*, \end{cases} \text{ if } y^* < \infty, \tag{108}$$

where $\tilde{G}$ is as in Lemma 6, $\bar{F}$, $\bar{G}$ are as in Lemma 7, and $y^* \equiv \bar{y}^*$, where $\bar{y}^*$ is given by (102).

The results above completely determine the boundaries of the three possible regions. To specify w inside the "wait" region $\mathcal{W}$, we still have to solve (88) and (94)–(97). To this end, it is straightforward to see that if the associated integrals are finite, then the function

$$A(y) = \frac{1}{\sigma^2(n-m)} \int_y^\infty \int_{G(u)}^\infty s^{-n-1} \left[H(s,u) - rK^+ \right] ds\, du > 0, \quad y \geq 0, \tag{109}$$

satisfies (88) as well as (94) and (95). In this expression, the inequality follows thanks to (90) or the second inequality in (105), depending on the case, and the assumption that $H(\cdot, y)$ is increasing. It is worth noting that adding a constant on the right-hand side of (109) would yield a further solution to (88). However, it turns out that (109) gives the *only* solution of (88) that renders w compatible with the requirements of the verification theorem that we proved in section 4.

If $y^* < \infty$, then

$$B(y) = -\frac{1}{\sigma^2(n-m)} \int_{y^*}^y \int_0^{F(u)} s^{-m-1} \left[H(s,u) + rK^- \right] ds\, du > 0, \quad y > y^*, \tag{110}$$

satisfies (96) or (97). Here, the positivity of B follows from the first inequality in (105) and the assumption that $H(\cdot, y)$ is increasing. As above, we have set a possible additive constant to zero because the resulting function w can be identified with the value function of the control problem for *no other* choice.

With reference to (81), w must satisfy

$$w(x,y) = w(x, G^{[-1]}(x)) - K^+ \left(G^{[-1]}(x) - y \right), \quad \text{for } (x,y) \in \mathcal{I},$$

where $G^{[-1]} : [G(0), \infty[\to \mathbb{R}_+$ is the inverse function of G. Also, if $\mathcal{D} \neq \emptyset$, then (92) implies that w should satisfy

$$w(x,y) = w(x, \Phi(x)) - K^-(y - \Phi(x)), \quad \text{for } (x,y) \in \mathcal{D},$$

where the function $\Phi :]0, \infty[\to \mathbb{R}_+$ is defined by

$$\Phi(x) = \begin{cases} F^{[-1]}(x), & \text{if } x \geq F(y^*), \\ 0, & \text{if } y^* = 0 \text{ and } F(0) > x, \end{cases} \tag{111}$$

in which expression $F^{[-1]} : [F(y^*), \infty[\to \mathbb{R}_+$ is the inverse function of F. Summarizing, we have two possibilities. If the point $y^* \equiv \bar{y}^*$ as in (74) or (102) is equal to ∞, then

$$w(x,y) = \begin{cases} A(y)x^n + R(x,y), & \text{for } (x,y) \text{ such that} \\ & 0 < x \leq G(y), \\ w(x, G^{[-1]}(x)) - K^+(G^{[-1]}(x) - y) & \text{for } (x,y) \text{ such that } G(y) < x. \end{cases} \tag{112}$$

On the other hand, if $y^* < \infty$, then

$$w(x,y) = \begin{cases} w(x, \Phi(x)) - K^-(y - \Phi(x)), & \text{for } (x,y) \text{ such that} \\ & y > y^* \text{ and } x < F(y), \\ A(y)x^n + R(x,y), & \text{for } (x,y) \text{ such that} \\ & y \in [0, y^*] \cap \mathbb{R} \text{ and } x \leq G(y), \\ A(y)x^n + B(y)x^m + R(x,y), & \text{for } (x,y) \text{ such that} \\ & y > y^* \text{ and } F(y) \leq x \leq G(y), \\ w(x, G^{[-1]}(x)) - K^+(G^{[-1]}(x) - y) & \text{for } (x,y) \text{ such that } G(y) < x. \end{cases} \tag{113}$$

It is worth noting that if $y^* = 0$ and $F(0) > 0$, then (78) and (110) imply that

$$w(x,0) = A(0)x^n + R(x,0), \quad \text{for } 0 < x \leq G(0),$$

which is consistent with the associated expression resulting from (113).

The next result, which we prove in the appendix, is concerned with proving that the construction above indeed provides a solution to the HJB equation (49)–(50), as well as with certain estimates that we will need.

LEMMA 8. *Suppose that Assumptions* 1 *and* 2 *hold. The function w given by* (112)–(113)*, where F, G and A, B are as in* (107)*,* (108) *and* (109)*,* (110)*, respectively, is C^2 and satisfies the HJB equation* (49)–(50)*. Also, w satisfies*

$$w(x,y) \leq C_5 \left(1 + y + G^{n-\varepsilon_4}(y) + G^{\alpha}(y)y^{\beta} + x^{n-\varepsilon_4}\right), \text{ for all } (x,y) \in \mathcal{S}, \tag{114}$$

for some constants $C_5 > 0$ and $\varepsilon_4 \in]0, n[$, as well as (54) *in the verification theorem, Theorem* 5.

Remark 2. A careful inspection of the proof of this result reveals that, had we perturbed the expressions on the right-hand sides of (109) and (110) by additive constants, we still would have obtained a further solution to the HJB equation (49)–(50).

However, such a solution would not satisfy an estimate such as the one provided by (114) that plays a fundamental role in the proof of the verification theorem, Theorem 5.

We can now prove the main result of the paper.

THEOREM 9. *Consider the capacity control problem formulated in section* 2*, and suppose that Assumptions* 1 *and* 2 *hold. The value function* v *identifies with the function* w *given by* (112)–(113)*, where* F*,* G *and* A*,* B *are as in* (107)*,* (108) *and* (109)*,* (110)*, respectively. The optimal capacity process* Y^{o} *reflects the joint process* (X, Y^{o}) *along the boundaries* G *and* F *in the positive and in the negative* y*-direction, respectively, and can be constructed as follows.*

(a) *If* $y^* = \infty$*, then* Y^{o} *is given by*

$$Y_t^{\mathrm{o}} = y\mathbf{1}_{\{t\le\tau_0\}} + G^{[-1]}\left(\sup_{s\le t} X_s\right)\mathbf{1}_{\{\tau_0<t\}},$$

where $\tau_0 = \inf\{t \ge 0 :\ X_t \ge G(y)\}$ *and* $G^{[-1]} : [G(0), \infty[\to \mathbb{R}_+$ *is the inverse function of* G*.*

(b) *If* $y^* < \infty$*, we first define*

$$\hat{y} = \begin{cases} \Phi(x), & \text{if } y > \Phi(x), \\ y & \text{otherwise,} \end{cases} \qquad \tau_0 = \inf\{t \ge 0 :\ X_t \ge G(\hat{y})\}$$

and

$$Y_t^{(1)} = y\mathbf{1}_{\{t=0\}} + \hat{y}\mathbf{1}_{\{0<t\le\tau_0\}} + G^{[-1]}\left(\sup_{s\le t} X_s\right)\mathbf{1}_{\{\tau_0<t\}},$$

where Φ *is defined by* (111)*. We then recursively define the* $(\mathcal{F}_t)$*-stopping times* τ_n *and the processes* $Y^{(n)}$ *by*

$$\tau_{2k+1} = \inf\left\{t > 0 :\ X_t < \hat{F}\left(Y_t^{(2k+1)}\right)\right\},$$
$$Y_t^{(2k+2)} = Y_t^{(2k+1)}\mathbf{1}_{\{t\le\tau_{2k+1}\}} + \Phi\left(\inf_{\tau_{2k+1}<s\le t} X_s\right)\mathbf{1}_{\{\tau_{2k+1}<t\}}$$

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

$$\hat{F}(y) = \begin{cases} 0, & \text{if } y < y^*, \\ F(y), & \text{if } y \ge y^*, \end{cases}$$

and by

$$\tau_{2k} = \inf\left\{t > 0 :\ X_t > G\left(Y_t^{(2k)}\right)\right\},$$
$$Y_t^{(2k+1)} = Y_t^{(2k)}\mathbf{1}_{\{t\le\tau_{2k}\}} + G^{[-1]}\left(\sup_{\tau_{2k}<s\le t} X_s\right)\mathbf{1}_{\{\tau_{2k}<t\}}$$

for $k = 1, 2, \ldots$*. The optimal capacity process* Y^{o} *is given by* $Y_t^{\mathrm{o}} = Y_t^{(n)}$ *for* $t < \tau_n$ *and* $n \ge 1$*.*

Proof. In view of Lemma 8, we have to show only that the process Y^{o} satisfies (55)–(59) in the verification theorem, Theorem 5. To this end, we first make the

following comments on the construction of Y^{o}. If $y^* = \infty$, then the boundary F does not exist, and $Y^{\mathrm{o}} = Y^{(1)}$ is all we need because it reflects the joint process $(X, Y^{(1)})$ along the boundary G in the positive y-direction. On the other hand, if $y^* < \infty$, then the boundary F becomes part of the picture and we need to define Y^{o} in a recursive way. If the initial condition (x, y) is in the interior of the "disinvestment" region $\mathcal{D}$, then the process $Y^{(1)}$ has a jump of size $-(y - \Phi(x))$ at time 0, which instantaneously repositions the joint process $(X, Y^{(1)})$ in the closure of the "wait" region $\mathcal{W}$. Similarly, if the initial condition (x, y) is in the interior of the "investment" region $\mathcal{I}$, then the process $Y^{(1)}$ has a jump of size $G^{[-1]}(x) - y$ at time 0, which instantaneously repositions $(X, Y^{(1)})$ in the closure of the "wait" region. After time 0, the process $Y^{(1)}$ reflects the joint process $(X, Y^{(1)})$ along the boundary G in the positive y-direction, and $(X, Y^{(1)})$ enters the interior of the "disinvestment" region $\mathcal{D}$ after time τ_1 with positive probability. The process $Y^{(2)}$ is the same as $Y^{(1)}$ up to time τ_1, $Y^{(2)}_{\tau_1} \equiv Y^{(1)}_{\tau_1} > y^*$, and $X_{\tau_1} = F(Y^{(2)}_{\tau_1})$. Beyond time τ_1, $Y^{(2)}$ reflects the joint process $(X, Y^{(2)})$ along the boundary F in the negative y-direction. As a result, the process $(X, Y^{(2)})$ is kept outside the interior of $\mathcal{I} \cup \mathcal{D}$ at all times up to τ_2, after which time it enters the interior of the "investment" region $\mathcal{I}$ with positive probability. The process $Y^{(3)}$ is the same as $Y^{(2)}$ up to time τ_2 and $X_{\tau_2} = G(Y_{\tau_2})$. After τ_2, $Y^{(3)}$ reflects $(X, Y^{(3)})$ along the boundary G in the positive y-direction. It follows that the process $(X, Y^{(3)})$ does not enter the interior of $\mathcal{I} \cup \mathcal{D}$ up to time τ_3. Iterating this construction, which ensures that $Y^{(n)}_t = Y^{(n+1)}_t$, for all $t \in [0, \tau_{n+1}]$ and $n \geq 1$, and observing that $\lim_{n\to\infty} \tau_n = \infty$, we can see that Y^{o}_t is defined for all $t \geq 0$ and that (55) is satisfied. Also, if $\xi^{\mathrm{o}+}$ and $\xi^{\mathrm{o}-}$ are the increasing processes providing the minimal decomposition of Y^{o} into $Y^{\mathrm{o}} = y + \xi^{\mathrm{o}+} - \xi^{\mathrm{o}-}$, then both (56) and (57) hold.

To proceed further, we note that the construction of Y^{o} implies that

$$Y^{\mathrm{o}}_t \leq y\mathbf{1}_{\{\bar{X}_t \leq G(y)\}} + G^{[-1]}(\bar{X}_t)\mathbf{1}_{\{\bar{X}_t > G(y)\}}, \tag{115}$$

where $\bar{X}_t = \sup_{s\leq t} X_s$. Combining this inequality with the definition (108) of G and the estimates in (91) and (106), we can see that

$$Y^{\mathrm{o}}_t \leq y\mathbf{1}_{\{\bar{X}_t \leq G(y)\}} + C_4 \bar{X}_t^{\alpha/(1-\beta)}\mathbf{1}_{\{\bar{X}_t > G(y)\}} \tag{116}$$

and

$$\xi^{\mathrm{o}+}_t \leq C_4 \bar{X}_t^{\alpha/(1-\beta)}. \tag{117}$$

Now, we can use (116), the observation that

$$G(Y^{\mathrm{o}}_t) \leq G(y)\mathbf{1}_{\{\bar{X}_t \leq G(y)\}} + \bar{X}_t\mathbf{1}_{\{\bar{X}_t > G(y)\}},$$

which follows immediately from (115), to see that, e.g.,

$$\begin{aligned} G^\alpha\left(Y^{\mathrm{o}}_t\right)\left(Y^{\mathrm{o}}_t\right)^\beta &\leq G^\alpha(y)y^\beta\mathbf{1}_{\{\bar{X}_t \leq G(y)\}} + C_4^\beta \bar{X}_t^{\alpha/(1-\beta)}\mathbf{1}_{\{\bar{X}_t > G(y)\}} \\ &\leq G^\alpha(y)y^\beta + C_4^\beta \bar{X}_t^{\alpha/(1-\beta)}. \end{aligned}$$

In view of this and similar calculations involving the other terms, as well as the estimate (114) and the fact that $\alpha < \frac{\alpha}{1-\beta} < n$ (see Assumption 2), we can conclude that (116)–(117) imply that the estimates (58)–(59) hold true, and the proof is complete. □

6. Examples. We can illustrate our main results by means of the special cases that we now consider.

COROLLARY 10. *Suppose that h is given by* (27) *in Example* 1, *and $K^+, K^+ + K^- > 0$. If $\frac{\alpha}{1-\beta} < n$, then $v < \infty$, while if $\frac{\alpha}{1-\beta} > n > \alpha$, then $v \equiv \infty$, where n is the positive solution of* (10). *In the former case, the following hold true:*

(a) *If $K^- \geq 0$, then $y^* = \infty$,*

$$G(y) = \left[-\frac{rK^+(\alpha - m)}{m\beta}\right]^{1/\alpha} y^{(1-\beta)/\alpha}, \tag{118}$$

and the optimal strategy can be depicted by Figure 1.

(b) *If $K^- < 0$, then $y^* = 0$ and*

$$\lim_{y \downarrow 0} F(y) = \lim_{y \downarrow 0} G(y) = 0, \tag{119}$$

and the optimal strategy can be depicted by Figure 4.

Proof. As we have observed in Example 1, Assumptions 1 and 2 are satisfied and $v < \infty$ if and only if $\frac{\alpha}{1-\beta} < n$. Also, if $\frac{\alpha}{1-\beta} > n > \alpha$, then we have proved in Lemma 3 that $v \equiv \infty$.

The condition distinguishing the two cases follows from a simple inspection of (102), while showing (118) involves elementary calculations. To see (119), we observe that the system of equations (100)–(101), which specifies F and G, is equivalent to

$$\frac{\beta}{\alpha - m} y^{-(1-\beta)} \left[G^{\alpha-m}(y) - F^{\alpha-m}(y)\right] = -\frac{r}{m}\left[K^+ G^{-m}(y) + K^- F^{-m}(y)\right], \tag{120}$$

$$\frac{\beta}{n - \alpha} y^{-(1-\beta)} \left[G^{\alpha-n}(y) - F^{\alpha-n}(y)\right] = \frac{r}{n}\left[K^+ G^{-n}(y) + K^- F^{-n}(y)\right]. \tag{121}$$

Since $m < 0 < \alpha, 1 - \beta$ and F, G are increasing, the right-hand side of (120) remains bounded as $y \downarrow 0$, and $\lim_{y \downarrow 0} y^{-(1-\beta)} = \infty$. It follows that (120) cannot be true unless (119) is satisfied, and the proof is complete. □

Remark 3. In the context of the special case considered in Corollary 10, it is worth noting that the solution w to the HJB equation (49)–(50) that we have constructed following intuition based on economical considerations is finite for all $\alpha \in]0, n[$ and $\beta \in]0, 1[$. Had we adopted a *formal* approach, this observation would have suggested the adoption of the capacity expansion strategy that keeps the process (X, Y) inside the "wait" region $\mathcal{W}$ that is determined by the functions F and G provided by the unique solution to the associated free-boundary problem. However, such a formal approach would have led us to wrong conclusions because

$$w(x, y) < \infty = v(x, y), \quad \text{for all } (x, y) \in \mathcal{S},$$

if $\frac{\alpha}{1-\beta} > n$.

Remark 4. In the special case of Corollary 10 arising when $\alpha = 1 - \beta$ and $K^- < 0$, we can verify that (120) and (121) are satisfied by the functions

$$F(y) = \kappa y \quad \text{and} \quad G(y) = \nu y, \quad \text{for } y \geq 0,$$

where κ and ν are constants satisfying the system of algebraic equations

$$\frac{1 - \alpha}{\alpha - m}\left[\nu^{\alpha-m} - \kappa^{\alpha-m}\right] = -\frac{r}{m}\left[K^+ \nu^{-m} + K^- \kappa^{-m}\right], \tag{122}$$

$$\frac{1 - \alpha}{n - \alpha}\left[\nu^{-(n-\alpha)} - \kappa^{-(n-\alpha)}\right] = \frac{r}{n}\left[K^+ \nu^{-n} + K^- \kappa^{-n}\right]. \tag{123}$$

Abel and Eberly [1] considered this special case with $r > b$, which satisfies our assumptions thanks to the equivalence $r > b \Leftrightarrow n > 1$, and have proved that the system of equations (122)–(123) has a unique solution such that $0 < \kappa < \nu$.

The following special case follows from our general results and (29).

COROLLARY 11. *Suppose that* $K^+, -K^-, K^+ + K^- > 0$, *consider the running payoff function* h *given by* (28) *in Example* 2, *and assume that the associated parameters satisfy* (29). *The following cases hold true:*

(a) *If* $-rK^- \in \left]\left(\beta\eta^\alpha\zeta^{-(1-\beta)} - K\right) \vee 0, rK^+\right[$, *then* $y^* = 0$, $0 < \lim_{y\downarrow 0} F(y) < \lim_{y\downarrow 0} G(y)$, *and the optimal strategy can be depicted by Figure* 3.

(b) *If* $\beta\eta^\alpha\zeta^{-(1-\beta)} > K$ *and* $-rK^- \in \left]0, \beta\eta^\alpha\zeta^{-(1-\beta)} - K\right[$, *then*

$$y^* = \left(\frac{\beta\eta^\alpha}{K - rK^-}\right)^{\frac{1}{1-\beta}} - \zeta > 0,$$

$\lim_{y\downarrow y^*} F(y) = 0$, $\lim_{y\downarrow 0} G(y) > 0$, *and the optimal strategy can be depicted by Figure* 2.

We conclude with the following example that does not satisfy the requirements imposed on the problem data by Assumptions 1 and 2.

Example 3. Suppose that the running payoff function h is given by $h(x, y) = (x + \eta)^\alpha y^\beta$, for some constants $\eta > 0$ and $\alpha, \beta \in]0, 1[$, such that $\frac{\alpha}{1-\beta} < n$. Using the same arguments as the ones in Example 2, we can check that Assumption 1 and (23), (24), and (26) in Assumption 2 all hold true. However, this payoff function does not satisfy the upper bound required by (25) in Assumption 2. Furthermore, if we assume that $K^+, -K^-, K^+ + K^- > 0$, then we can check that the points y_* and y^* defined as in Lemmas 6 and 7 are given by

$$0 < y_* = \left(\frac{\beta\eta^\alpha}{rK^+}\right)^{\frac{1}{1-\beta}} < \left(\frac{\beta\eta^\alpha}{-rK^-}\right)^{\frac{1}{1-\beta}} = y^*.$$

It follows that, at least formally, this example provides a case in which a strategy such as the one depicted by Figure 5 is optimal.

Appendix. Proof of selected results.

Proof of Lemma 6. Suppose that (20) in Assumption 1 is satisfied. Fix any $y \geq 0$, and suppose that $\inf_{x>0} H(x, y) - rK^+ \geq 0$. In this case, $H(x, y) - rK^+ > 0$, for all $x > 0$, because $H(\cdot, y)$ is a strictly increasing function. This implies that $q(x, y) > 0$, for all $x > 0$, and, therefore, the equation $q(x, y) = 0$ has no solution $x > 0$.

Now, fix any $y \geq 0$, and assume that $\inf_{x>0} H(x, y) < rK^+$. Recalling the assumption that $H(\cdot, y)$ is strictly increasing, we define

$$x^\dagger = x^\dagger(y) := \inf\left\{x > 0 : \ H(x, y) - rK^+ > 0\right\} > 0,$$

and we observe that

$$\frac{\partial}{\partial x} q(x, y) = x^{-m-1}\left[H(x, y) - rK^+\right] \begin{cases} < 0, & \text{for all } x \in]0, x^\dagger[, \\ > 0 & \text{for all } x > x^\dagger. \end{cases} \tag{124}$$

Combining the fact that $q(\cdot, y)$ is strictly decreasing in $]0, x^\dagger[$ and strictly increasing in $]x^\dagger, \infty[$, with $q(0, y) = 0$, we can see that $q(x, y) < 0$ for all $x \leq x^\dagger$. In particular, $q(x^\dagger, y) < 0$. Therefore, if $q(x, y) = 0$ has a solution $x > 0$, then this must satisfy $x > x^\dagger$. Also, given that it exists, this solution is unique because $q(\cdot, y)$ is strictly

increasing in $]x^{\dagger}, \infty[$. To prove that the required solution indeed exists, it suffices to show that $\lim_{x\to\infty} q(x,y) = \infty$. The assumption that $\lim_{x\to\infty} H(x,y) = \infty$ implies that, given any constant $M > 0$, there exists $\gamma > x^{\dagger}$ such that $H(x,y) - rK^{+} \geq M$ for all $x \geq \gamma$. However, given any such choice of these constants, we calculate

$$\begin{aligned}\lim_{x\to\infty} q(x,y) &= \lim_{x\to\infty} \left[q(\gamma,y) + \int_{\gamma}^{x} s^{-m-1} \left[H(s,y) - rK^{+} \right] ds \right] \\ &\geq \lim_{x\to\infty} \left[q(\gamma,y) + \frac{M}{m}\gamma^{-m} - \frac{M}{m}x^{-m} \right] = \infty.\end{aligned}$$

If (21) in Assumption 1 also holds and the point $\tilde{y}_*$ defined as in (89) is finite, then $\inf_{x>0} H(x,y) < rK^{+}$ for all $y > \tilde{y}_*$. It follows that (86) uniquely defines a continuous function $\tilde{G} :]\tilde{y}_*, \infty[\to]0, \infty[$. Moreover, the arguments above regarding the solvability of $q(x,y) = 0$ imply (90).

To see that $\tilde{G}$ is C^1 and strictly increasing, we differentiate $q(\tilde{G}(y), y) = 0$ with respect to y to obtain

$$\tilde{G}'(y) = -\tilde{G}^{m+1}(y) \left[H(\tilde{G}(y), y) - rK^{+} \right]^{-1} \int_{0}^{\tilde{G}(y)} s^{-m-1} H_y(s,y)\, ds > 0 \tag{125}$$

for all $y > \tilde{y}_*$. The inequality here follows thanks to (90) and (21) in Assumption 1.

Now, suppose that (25) in Assumption 2 also holds, and observe that this implies that

$$\inf_{x>0} H(x,y) < rK^{+} \quad \text{for all } y > 0.$$

However, this inequality implies that $\tilde{y}_* = 0$. Finally, with regard to (25) in Assumption 2 and (124) above, we calculate

$$\frac{\partial}{\partial x} q(x,y) \leq x^{-m-1} \left[\beta C x^{\alpha} y^{-(1-\beta)} - r\vartheta \right].$$

Combining this inequality with $q(0,y) = 0$, we can see that, given any $y > 0$, $\tilde{G}(y)$ is greater than or equal to the strictly positive solution of the equation

$$\int_{0}^{z} s^{-m-1} \left[\beta C s^{\alpha} y^{-(1-\beta)} - r\vartheta \right] ds = 0,$$

which yields

$$\tilde{G}(y) \geq \left(-\frac{r\vartheta(\alpha - m)}{\beta C m} \right)^{\frac{1}{\alpha}} y^{\frac{1-\beta}{\alpha}} \quad \text{for all } y > 0.$$

However, this implies (91). □

Proof of Lemma 7. Suppose that Assumption 1 holds. We develop the proof in a number of steps.

Step 1. To study the solvability of the system of equations (98) and (99), we first prove that (98) uniquely defines a mapping $L : (\mathbb{R}_{+} \setminus \{0\})^2 \to]0, \infty[$ such that

$$f(x_1, L(x_1,y), y) = 0 \quad \text{and} \quad L(x_1,y) > x_1. \tag{126}$$

To this end, fix any $x_1 > 0$, $y > 0$, and observe that

$$f(x_1, x_1, y) = -\frac{1}{m} r \left(K^{+} + K^{-} \right) x_1^{-m} > 0. \tag{127}$$

Given $M > 0$, observe that the assumption that $\lim_{x\to\infty} H(x,y) = \infty$, for all $y > 0$, implies that there exists a constant $\gamma > x_1$ such that $H(x,y) - rK^+ \geq M$ for all $x \geq \gamma$. For such a choice of parameters, since $m < 0$, we calculate

$$\begin{aligned}
\lim_{x_2\to\infty} f(x_1,x_2,y) &= \lim_{x_2\to\infty}\Big[-\int_{x_1}^{\gamma} s^{-m-1}\left[H(s,y)-rK^+\right]ds \\
&\quad -\int_{\gamma}^{x_2} s^{-m-1}\left[H(s,y)-rK^+\right]ds - \frac{r}{m}\left(K^+ + K^-\right)x_1^{-m}\Big] \\
&\leq \lim_{x_2\to\infty}\left[f(x_1,\gamma,y) - M\int_{\gamma}^{x_2} s^{-m-1}\,ds\right] \\
&= \lim_{x_2\to\infty}\left[f(x_1,\gamma,y) - \frac{M}{m}\gamma^{-m} + \frac{M}{m}x_2^{-m}\right] \\
&= -\infty. \qquad (128)
\end{aligned}$$

Also, it is straightforward to calculate

$$\frac{\partial f}{\partial x_2}(x_1,x_2,y) = -x_2^{-m-1}\left[H(x_2,y)-rK^+\right]\begin{cases} >0, & \text{for all } x_2 \in]0,x^\dagger[, \\ <0, & \text{for all } x_2 > x^\dagger, \end{cases} \qquad (129)$$

where

$$x^\dagger = x^\dagger(y) := \inf\left\{x > 0 :\ H(x,y) - rK^+ > 0\right\}.$$

Combining the fact that $f(x_1,\cdot,y)$ is strictly increasing in the interval $[x_1, x^\dagger[$, if $x_1 < x^\dagger$, and strictly decreasing in the interval $]x^\dagger \vee x_1, \infty[$, with (128) and (127), we can conclude that the equation $f(x_1,x_2,y) = 0$ has a unique solution $x_2 = L(x_1,y)$ which satisfies (126) as well as

$$H(L(x_1,y),y) - rK^+ > 0. \qquad (130)$$

For future reference, we also note that differentiation of $f(x_1,L(x_1,y),y) = 0$ with respect to x_1 yields

$$\frac{\partial}{\partial x_1}L(x_1,y) = \frac{x_1^{-m-1}\left[H(x_1,y)+rK^-\right]}{L^{-m-1}(x_1,y)\left[H(L(x_1,y),y)-rK^+\right]}, \qquad (131)$$

while differentiation of $f(x_1,L(x_1,y),y) = 0$ with respect to y gives

$$\frac{\partial}{\partial y}L(x_1,y) = -L^{m+1}(x_1,y)\left[H(L(x_1,y),y)-rK^+\right]^{-1}\int_{x_1}^{L(x_1,y)} s^{-m-1}H_y(s,y)\,ds. \qquad (132)$$

Step 2. To prove that the system of equations (98) and (99) has a unique solution (x_1,x_2) such that $0 < x_1 < x_2$ we have to show that there exists a unique $x_1 > 0$ such that $g(x_1,L(x_1,y),y) = 0$. To this end, we first observe that the calculation

$$g(x_1,L(x_1,y),y) = \int_{x_1}^{L(x_1,y)} s^{-n-1}\left[H(s,y)-rK^+\right] + \frac{1}{n}r\left(K^+ + K^-\right)x_1^{-n}$$

and the assumptions $\lim_{x\to\infty} H(x,y) = \infty$, $K^+ + K^- > 0$ imply that

(133) there exists a constant $N > 0$ such that $g(x_1,L(x_1,y),y) > 0$ for all $x_1 \geq N$.

Now, with regard to (131), we calculate

$$\frac{\partial}{\partial x_1} g(x_1, L(x_1,y), y) = x_1^{-m-1} \left[L^{m-n}(x_1,y) - x_1^{m-n}\right] \left[H(x_1,y) + rK^-\right]. \tag{134}$$

Since $L(x_1,y) > x_1$ and $m < n$, $L^{m-n}(x_1,y) - x_1^{m-n} < 0$. It follows that if $\inf_{x>0} H(x,y) \geq -rK^-$, then $g(\cdot, L(\cdot,y), y)$ is decreasing, which, combined with (133), implies that the equation $g(x_1, L(x_1,y), y) = 0$ cannot have a solution $x_1 > 0$. Therefore, we must have $\inf_{x>0} H(x,y) < -rK^-$. Assuming that this condition holds, we recall that $H(\cdot,y)$ is strictly increasing, we define

$$x^\ddagger = x^\ddagger(y) := \inf\left\{x > 0 : \ H(x,y) + rK^- > 0\right\},$$

and we observe that

$$g(\cdot, L(\cdot,y), y) \text{ is strictly increasing in }]0, x^\ddagger[\text{ and strictly decreasing in }]x^\ddagger, \infty[. \tag{135}$$

Furthermore, under this condition, there exist $\varepsilon > 0$ and $\delta < x^\ddagger$ such that $H(x_1,y) + rK^- \leq -\varepsilon$ for all $x_1 \leq \delta$. For such a choice of parameters, we calculate

$$\begin{aligned} &\lim_{x_1 \downarrow 0} \int_{x_1}^\infty s^{-n-1} \left[H(s,y) + rK^-\right] ds \\ &\qquad \leq \lim_{x_1 \downarrow 0} \left[\frac{\varepsilon}{n}\delta^{-n} - \frac{\varepsilon}{n} x_1^{-n} + \int_\delta^\infty s^{-n-1} \left[H(s,y) + rK^-\right] ds\right] \\ &\qquad = -\infty. \end{aligned} \tag{136}$$

In view of this, (130), and the assumption that $H(\cdot,y)$ is increasing,

$$\begin{aligned} &\lim_{x_1 \downarrow 0} g(x_1, L(x_1,y), y) \\ &\qquad = \lim_{x_1 \downarrow 0} \left[\int_{x_1}^\infty s^{-n-1} \left[H(s,y) + rK^-\right] ds - \int_{L(x_1,y)}^\infty s^{-n-1} \left[H(s,y) - rK^+\right] ds\right] \\ &\qquad \leq \lim_{x_1 \downarrow 0} \int_{x_1}^\infty s^{-n-1} \left[H(s,y) + rK^-\right] ds \\ &\qquad = -\infty. \end{aligned} \tag{137}$$

However, combining (133) with (135) and (137), we can see that the equation $g(x_1, L(x_1,y), y) = 0$ has a unique solution $x_1 > 0$, which also satisfies

$$H(x_1,y) + rK^- < 0. \tag{138}$$

Step 3. Summarizing the analysis above, under the assumption that the point $\bar{y}^*$ defined as in (102) is finite, the system of equations (98) and (99) uniquely defines two continuous functions $\bar{F}, \bar{G} :]\bar{y}^*, \infty[\to]0, \infty[$ that satisfy $\bar{F}(y) < \bar{G}(y)$, for all $y > \bar{y}^*$, as well as (105). Also, (103)–(104) follow from a simple continuity argument combining the definition of $\bar{y}^*$ and (138).

Step 4. Now, assuming that $\bar{y}^* < \infty$, we consider any point $y > \bar{y}^*$. Differentiating the equation $g(\bar{F}(y), L(\bar{F}(y),y), y) = 0$ with respect to y, using (131), and observing

that $\bar{G}(y) = L(\bar{F}(y), y)$, we calculate

$$\begin{aligned}\bar{F}'(y) = & -\bar{F}^{m+1}(y)\bar{G}^{-n}\left[\bar{G}^{-(n-m)}(y) - \bar{F}^{-(n-m)}(y)\right]^{-1}\left[H(\bar{F}(y), y) + rK^{-}\right]^{-1} \\ & \times \int_{\bar{F}(y)}^{\bar{G}(y)} \left[\left(\frac{\bar{G}(y)}{s}\right)^{n} - \left(\frac{\bar{G}(y)}{s}\right)^{m}\right] \frac{1}{s} H_y(s, y)\, ds > 0, \end{aligned} \tag{139}$$

the inequality following thanks to assumption (21), the first inequality in (105), and the fact that $m < 0 < n$. Also, differentiating the equation $f(\bar{F}(y), L(\bar{F}(y), y), y) = 0$ with respect to y, and using (132) and (139), we calculate

$$\begin{aligned}\bar{G}'(y) = & -\bar{F}^{-n}(y)\bar{G}^{m+1}\left[\bar{G}^{-(n-m)}(y) - \bar{F}^{-(n-m)}(y)\right]^{-1}\left[H(\bar{G}(y), y) - rK^{+}\right]^{-1} \\ & \times \int_{\bar{F}(y)}^{\bar{G}(y)} \left[\left(\frac{\bar{F}(y)}{s}\right)^{n} - \left(\frac{\bar{F}(y)}{s}\right)^{m}\right] \frac{1}{s} H_y(s, y)\, ds > 0, \end{aligned}$$

the inequality following thanks to (105) and (21). However, these calculations show that $\bar{F}$ and $\bar{G}$ are both C^1 and strictly increasing.

Step 5. Finally, suppose that (25) in Assumption 2 is also true. With reference to the equation $f(\bar{F}(y), \bar{G}(y), y) = 0$, we calculate

$$\begin{aligned} 0 = & -\int_{\bar{F}(y)}^{\bar{G}(y)} s^{-m-1}\left[H(s, y) - rK^{+}\right] ds - \frac{1}{m} r\left(K^{+} + K^{-}\right)\bar{F}^{-m}(y) \\ \geq & -\left[\frac{\beta C}{\alpha - m}\bar{G}^{\alpha-m}(y) y^{-(1-\beta)} + \frac{r\vartheta}{m}\bar{G}^{-m}(y)\right] \\ & + \left[\frac{\beta C}{\alpha - m}\bar{F}^{\alpha-m}(y) y^{-(1-\beta)} - \frac{1}{m} r\left(K^{+} + K^{-} - \vartheta\right)\bar{F}^{-m}(y)\right]. \end{aligned}$$

Since $\vartheta < K^{+} + K^{-}$ by assumption, the second term on the right-hand side of this expression is strictly positive. Therefore, we must have

$$\frac{\beta C}{\alpha - m}\bar{G}^{\alpha-m}(y) y^{-(1-\beta)} + \frac{r\vartheta}{m}\bar{G}^{-m}(y) > 0.$$

This inequality can be true only if $\bar{G}(y)$ is strictly greater than the unique strictly positive solution of the equation

$$\frac{\beta C}{\alpha - m} z^{\alpha-m} y^{-(1-\beta)} + \frac{r\vartheta}{m} z^{-m} = 0,$$

which yields

$$\bar{G}(y) \geq \left(-\frac{r\vartheta(\alpha - m)}{\beta C m}\right)^{\frac{1}{\alpha}} y^{\frac{1-\beta}{\alpha}} \quad \text{for all } y > \bar{y}^{*}.$$

However, this implies (106).

Proof of Lemma 8. We develop the proof along a series of steps.

Step 1. We first prove (114). Consider (109), and note that the upper bound in (25) in Assumption 2 implies that

$$0 < A(y) \leq \frac{\beta C}{\sigma^2 (n - m)(n - \alpha)} \int_{y}^{\infty} u^{-(1-\beta)} G^{-(n-\alpha)}(u)\, du. \tag{140}$$

Recalling the inequalities $\alpha < \frac{\alpha}{1-\beta} < n$, we fix any $\varepsilon_0 > 0$ such that

$$\varepsilon_0 < n - \frac{\alpha}{1-\beta} < n - \alpha.$$

Using the fact that G is increasing and the estimate provided by (91) and (106), we calculate

$$\begin{aligned}\int_y^\infty u^{-(1-\beta)} G^{-(n-\alpha)}(u)\,du &\le G^{-\varepsilon_0}(y) \int_y^\infty u^{-(1-\beta)} G^{-(n-\alpha-\varepsilon_0)}(u)\,du \\ &\le \frac{\alpha C_4^{(1-\beta)(n-\alpha-\varepsilon_0)/\alpha}}{(1-\beta)(n-\varepsilon_0)-\alpha} G^{-\varepsilon_0}(y) y^{1-\frac{(1-\beta)(n-\varepsilon_0)}{\alpha}},\end{aligned}$$

which implies that

$$\int_y^\infty u^{-(1-\beta)} G^{-(n-\alpha)}(u)\,du \le \frac{\alpha C_4^{(1-\beta)(n-\alpha-\varepsilon_0)/\alpha}}{(1-\beta)(n-\varepsilon_0)-\alpha} G^{-\varepsilon_0}(y) \quad \text{for all } y \ge 1. \tag{141}$$

Also, the fact that G is increasing implies that

$$\begin{aligned}G^n(y) \int_y^1 u^{-(1-\beta)} G^{-(n-\alpha)}(u)\,du &\le G^\alpha(y) \int_y^1 u^{-(1-\beta)}\,du \\ &\le \frac{1}{\beta} G^\alpha(1) \quad \text{for all } y < 1.\end{aligned} \tag{142}$$

However, (140)–(142) imply that

$$\begin{aligned}A(y)x^n &\le A(y)G^n(y) \\ &\le \frac{\beta C}{\sigma^2(n-m)(n-\alpha)} \left[\frac{\alpha C_4^{(1-\beta)(n-\alpha-\varepsilon_0)/\alpha}}{(1-\beta)(n-\varepsilon_0)-\alpha} G^{n-\varepsilon_0}(y) \mathbf{1}_{\{y\ge 1\}} \right. \\ &\quad \left. + \left(\frac{\alpha C_4^{(1-\beta)(n-\alpha-\varepsilon_0)/\alpha}}{(1-\beta)(n-\varepsilon_0)-\alpha} G^{n-\varepsilon_0}(1) + \frac{1}{\beta} G^\alpha(1) \right) \mathbf{1}_{\{y<1\}} \right] \\ &= C_{51} \left(1 + G^{n-\varepsilon_0}(y)\right), \quad \text{for all } y \ge 0 \text{ and } x \le G(y),\end{aligned} \tag{143}$$

where $C_{51} > 0$ is a constant.

If $y^* < \infty$, then (110), the assumption that $K^+ + K^- > 0$, the lower bound in (25) in Assumption 2, and the fact that F is increasing imply that, given any $y > y^*$,

$$\begin{aligned}B(y) &\le -\frac{C + rK^+}{\sigma^2 m(n-m)} \int_{y^*}^y F^{-m}(u)\,du \\ &\le -\frac{C + rK^+}{\sigma^2 m(n-m)} y F^{-m}(y).\end{aligned}$$

In light of this calculation and the fact that $m < 0$, we can see that

$$\sup_{x \in [F(y), G(y)]} B(y)x^m \le B(y)F^m(y) \le C_{52} y, \quad \text{for all } y > y^*, \tag{144}$$

where $C_{52} > 0$ is a constant. Since R is increasing in x (see (26) in Assumption 2 and (16)), the upper bound in Lemma 2 implies that

$$\begin{aligned}\sup_{x\leq G(y)} R(x,y) &\leq R(G(y),y)\\ &\leq C_1\left(1+y+G^{n-\vartheta}(y)+G^{\alpha}(y)y^{\beta}\right) \quad \text{for all } y\geq 0.\end{aligned}$$

However, combining this estimate with (143) and (144), we can see that w satisfies

$$w(x,y)\leq C_{53}\left(1+y+G^{n-\varepsilon_0\wedge\vartheta}(y)+G^{\alpha}(y)y^{\beta}\right), \quad \text{for all } (x,y)\in\overline{\mathcal{W}}, \tag{145}$$

for some constant $C_{53} > 0$. With regard to the structure of w provided by (112)–(113), this inequality and the estimates provided by (91) and (106) imply that

$$\begin{aligned}w(x,y) &\leq w(x,G^{[-1]}(x))+K^{+}y\\ &\leq C_{53}\left(1+G^{[-1]}(x)+x^{n-\varepsilon_0\wedge\vartheta}+x^{\alpha}\left[G^{[-1]}(x)\right]^{\beta}\right)+K^{+}y\\ &\leq C_{54}\left(1+y+x^{n-\varepsilon_0\wedge\vartheta}+x^{\alpha/(1-\beta)}\right), \quad \text{for } (x,y)\in\mathcal{I}, \qquad (146)\end{aligned}$$

for some constant $C_{54} > 0$. Also, since $\Phi(x) \leq y$, for all $(x,y) \in \mathcal{D}$, and G is increasing,

$$\begin{aligned}w(x,y) &\leq w(x,\Phi(x))+|K^{-}|y\\ &\leq C_{53}\left(1+\Phi(x)+G^{n-\varepsilon_0\wedge\vartheta}(\Phi(x))+G^{\alpha}(\Phi(x))\Phi^{\beta}(x)\right)+|K^{-}|y\\ &\leq C_{55}\left(1+y+G^{n-\varepsilon_0\wedge\vartheta}(y)+G^{\alpha}(y)y^{\beta}\right), \quad \text{for } (x,y)\in\mathcal{D}, \qquad (147)\end{aligned}$$

where $C_{55} > 0$ is a constant. However, in view of the assumption $\frac{\alpha}{1-\beta} < n$, if we choose any

$$\varepsilon_4 \in \left]0, \varepsilon_0\wedge\vartheta\wedge\left(n-\frac{\alpha}{1-\beta}\right)\right[\quad \text{and} \quad C_5 \geq C_{53}\vee C_{54}\vee C_{55},$$

then we can see that (145)–(147) imply (114).

Step 2. To show that w satisfies (54), we first observe that the positivity of A, B and the lower bound in Lemma 2 imply that

$$w(x,y)\geq -C_1(1+y) \quad \text{for all } (x,y)\in\overline{\mathcal{W}}. \tag{148}$$

This estimate and the definition of w in $\mathcal{I}$, provided by (112)–(113), imply that

$$\begin{aligned}w(x,y) &\geq -(C_1+K^{+})G^{[-1]}(x)-C_1\\ &\geq -(C_1+K^{+})C_4x^{\alpha/(1-\beta)}-C_1, \quad \text{for all } (x,y)\in\mathcal{I}, \qquad (149)\end{aligned}$$

the second inequality following thanks to (91) and (106). Also, if $y^{*} < \infty$, then (148) and the definition of w in $\mathcal{D}$, given by (113), imply that

$$\begin{aligned}w(x,y) &\geq -C_1(1+\Phi(x))-|K^{-}|\max\{y,\Phi(x)\}\\ &\geq -(C_1+|K^{-}|)y-C_1. \qquad (150)\end{aligned}$$

However, (148)–(150) establish (54).

Step 3. With reference to the construction of w, we will show that w is C^2 if we prove that w_x, w_{xx}, and w_{yy} are continuous along the free boundaries F and G. To this end, we calculate

$$\begin{aligned} w_x(x,y) &= w_x\left(x, G^{[-1]}(x)\right) + \left[w_y\left(x, G^{[-1]}(x)\right) - K^+\right] \frac{dG^{[-1]}(x)}{dx} \\ &= w_x\left(x, G^{[-1]}(x)\right), \quad \text{for } (x,y) \in \mathcal{I}, \end{aligned} \tag{151}$$

and

$$\begin{aligned} w_{xx}(x,y) &= w_{xx}\left(x, G^{[-1]}(x)\right) + w_{xy}\left(x, G^{[-1]}(x)\right) \frac{dG^{[-1]}(x)}{dx} \\ &= w_{xx}\left(x, G^{[-1]}(x)\right), \quad \text{for } (x,y) \in \mathcal{I}, \end{aligned} \tag{152}$$

the second equalities following thanks to (80) that have been among the requirements leading to the equations specifying the function G. However, these calculations and the structure of w provided by (112)–(113) show that w_x and w_{xx} are continuous along G.

Now, if $y^* > 0$ and $y \in [0, y^*] \cap \mathbb{R}$, we can use (79) and (88) to calculate

$$\begin{aligned} \lim_{x \uparrow G(y)} w_{yy}(x,y) &= A''(y)G^n(y) + R_{yy}(G(y), y) \\ &= \frac{G^{-1}(y)}{\sigma^2(n-m)}\left[G'(y)\left[H(G(y),y) - rK^+\right]\right. \\ &\quad \left. + G^{m+1}(y)\int_0^{G(y)} s^{-m-1} H_y(s,y)\, ds\right] \\ &= 0, \end{aligned} \tag{153}$$

the last equality following thanks to (125). Also, if $y^* < \infty$ and $y > y^*$, we can use (79), (95), and (97) to calculate

$$\begin{aligned} \lim_{x \uparrow G(y)} w_{yy}(x,y) &= A''(y)G^n(y) + B''(y)G^m(y) + R_{yy}(G(y), y) \\ &= 0. \end{aligned} \tag{154}$$

However, combining (153) and (154) with the fact that $w_{yy}(x,y) = 0$, for $(x,y) \in \mathcal{I}$, we conclude that w_{yy} is continuous along G.

Showing that w_x, w_{xx}, and w_{yy} are continuous along F involves similar arguments.

Step 4. By construction, we will prove that w satisfies the HJB equation (49)–(50) if we show that

$$\sigma^2 x^2 w_{xx}(x,y) + bx w_x(x,y) - rw(x,y) + h(x,y) \le 0, \quad \text{for } (x,y) \in \mathcal{I}, \tag{155}$$

$$w_y(x,y) + K^- \ge 0, \quad \text{for } (x,y) \in \mathcal{I},\ y > 0, \tag{156}$$

$$w_y(x,y) - K^+ \le 0, \quad \text{for } (x,y) \in \mathcal{W}, \tag{157}$$

$$w_y(x,y) + K^- \ge 0, \quad \text{for } (x,y) \in \mathcal{W},\ y > 0, \tag{158}$$

and, if $\mathcal{D} \neq \emptyset$,

$$\sigma^2 x^2 w_{xx}(x,y) + bx w_x(x,y) - rw(x,y) + h(x,y) \le 0, \quad \text{for } (x,y) \in \mathcal{D}, \tag{159}$$

$$w_y(x,y) - K^+ \le 0 \quad \text{for } (x,y) \in \mathcal{D}. \tag{160}$$

It is straightforward to see that either of (156) or (160) is equivalent to $K^+ + K^- \geq 0$, which is true by assumption. Recalling that $H \equiv h_y$, we can easily verify that, since $y \leq G^{[-1]}(x)$, for all $(x, y) \in \mathcal{I}$, (151) and (152) imply that (155) is equivalent to

$$\int_y^{G^{[-1]}(x)} \left[H(x,u) - rK^+\right] du \geq 0 \quad \text{for } (x,y) \in \mathcal{I}.$$

However, this inequality follows immediately from the assumption that $H(x,\cdot)$ is strictly decreasing, for all x, and (90) together with the second inequality in (105). Similarly, we can show that if $\mathcal{D} \neq \emptyset$, then (159) is equivalent to

$$\int_{\Phi(x)}^{y} \left[H(x,u) + rK^-\right] du \leq 0, \quad \text{for } (x,y) \in \mathcal{D},$$

where Φ is defined by (111). However, we can see that this inequality is true once we combine the first inequality in (105) with the assumption that $H(x,\cdot)$ is strictly decreasing, for all x, and the assumption that $H(\cdot, 0)$ is strictly increasing.

Now, suppose that $y^* < \infty$, and fix any $y > y^*$. Since $w_y(F(y), y) = -K^-$ and $w_y(G(y), y) = K^+$, we will prove that both (157) and (158) are satisfied if we show that

$$w_{yx}(x,y) \geq 0 \quad \text{for all } x \in]F(y), G(y)[. \tag{161}$$

To this end, we consider the transformation of the independent variable $x > 0$ provided by $z = \ln x$, and we write $w(x,y) = u(\ln x, y)$ for some function $u = u(z,y)$. It follows that (161) is true if and only if

$$u_{yz}(z,y) \geq 0 \quad \text{for all } z \in]\ln F(y), \ln G(y)[. \tag{162}$$

Now, since $w = w(x,y)$ satisfies (77) for $x \in]F(y), G(y)[$, u_y satisfies

$$\sigma^2 u_{yzz}(z,y) + \left(b - \sigma^2\right) u_{yz}(z,y) - r u_y(z,y) + H(e^z, y) = 0 \text{ for } z \in]\ln F(y), \ln G(y)[.$$

Recalling that H_x is continuous and $H_x(\cdot, y) \geq 0$ (see Assumption 1), we can differentiate this equation with respect to z to obtain

$$\begin{aligned}\sigma^2 (u_{yz})_{zz}(z,y) + \left(b - \sigma^2\right)(u_{yz})_z(z,y) - r u_{yz}(z,y) &= -e^z H_x(e^z, y) \\ &\leq 0 \text{ for } z \in]\ln F(y), \ln G(y)[.\end{aligned}$$

This inequality and the maximum principle imply that $u_{yz}(\cdot, y)$ does not have a negative minimum in the interval $]\ln F(y), \ln G(y)[$, and so

$$\begin{aligned}\inf_{z \in]\ln F(y), \ln G(y)[} u_{yz}(z,y) &\geq \min_{z = \ln F(y), \ln G(y)} 0 \wedge u_{yz}(z,y) \\ &= \min_{z = F(y), G(y)} 0 \wedge w_{yx}(x,y) \\ &= 0.\end{aligned}$$

However, this calculation implies (162).

To proceed further, fix any $y \in [0, y^*] \cap \mathbb{R}$. Using the definition of R in (79), the expression for $A'(y)$ provided by (88), and the fact that $G(y)$ satisfies (86), we can

see that if we define $\overline{u}(x,y) = w_y(x,y) - K^+$, then

$$\overline{u}_x(x,y) = \frac{1}{\sigma^2(n-m)}\left[-mx^{m-1}\int_x^{G(y)} s^{-m-1}\left[H(s,y) - rK^+\right]ds + nx^{n-1}\int_x^{G(y)} s^{-n-1}\left[H(s,y) - rK^+\right]ds\right] \quad \text{for } x \in]0, G(y)[.$$

This calculation and the assumption that $H(\cdot, y)$ is strictly increasing imply that $\overline{u}_x(x,y) = w_{yx}(x,y) > 0$, for all $x \in [x^\dagger(y), G(y)[$, where $x^\dagger(y) \in]0, G(y)[$ is the unique point such that $H\left(x^\dagger(y), y\right) - rK^+ = 0$ (see Lemma 6). This observation and the boundary condition $w_y\left(G(y), y\right) = K^+$ imply that

$$w_y(x,y) - K^+ < 0 \quad \text{for all } x \in [x^\dagger(y), G(y)[. \tag{163}$$

Furthermore, since

$$\sigma^2 x^2 \overline{u}_{xx}(x,y) + bx\overline{u}_x(x,y) - r\overline{u}(x,y) = -\left[H(x,y) - rK^+\right] \geq 0, \quad \text{for } x \in]0, x^\dagger(y)[,$$

the maximum principle implies that the function $x \mapsto \overline{u}(x,y) = w_y(x,y) - K^+$ has no positive maximum in the interval $]0, x^\dagger(y)[$, and so

$$\sup_{x \in]0, x^\dagger(y)[}\left[w_y(x,y) - K^+\right] \leq \max_{x=0, x^\dagger(y)} 0 \vee \left[w_y(x,y) - K^+\right] = 0, \tag{164}$$

the equality following thanks to (163) and the fact that

$$\lim_{x \downarrow 0} w_y(x,y) = \lim_{x \downarrow 0} R_y(x,y) = \lim_{x \downarrow 0} \frac{H(x,y)}{r} \in [-K^-, K^+[. \tag{165}$$

The second equality here holds true because of (17), while the inclusion follows from the context (see Lemmas 6 and 7). However, (163) and (164) establish (157). Finally, if we define $\underline{u}(x,y) = w_y(x,y) + K^-$, then (165) and the assumption that $H(\cdot, y)$ is increasing imply that

$$\sigma^2 x^2 \underline{u}_{xx}(x,y) + bx\underline{u}_x(x,y) - r\underline{u}(x,y) = -\left[H(x,y) + rK^-\right] \leq 0 \text{ for all } x \in]0, G(y)[.$$

This calculation and the maximum principle imply that the function $x \mapsto \underline{u}(x,y) = w_y(x,y) + K^-$ has no negative minimum inside $]0, G(y)[$, and so

$$\inf_{x \in]0, G(y)[}\left[w_y(x,y) + K^-\right] = \min_{x=0, G(y)} 0 \wedge \left[w_y(x,y) + K^-\right],$$

which, combined with (165) and the boundary condition $w_y(G(y), y) + K^- = K^+ + K^- > 0$, proves (158), and the proof is complete. □

Acknowledgments. We would like to acknowledge the stimulating environment of the Isaac Newton Institute for Mathematical Sciences in Cambridge, where this research was completed while we were participants in the Developments in Quantitative Finance programme. We are grateful for several fruitful discussions with other participants in the programme. We would also like to thank two anonymous referees, whose comments led to an improvement of the paper.

REFERENCES

[1] A. B. Abel and J. C. Eberly, *Optimal investment with costly reversibility*, Rev. Econom. Stud., 63 (1996), pp. 581–593.

[2] P. Bank, *Optimal control under a dynamic fuel constraint*, SIAM J. Control Optim., 44 (2005), pp. 1529–1541.

[3] S. Bentolila and G. Bertola, *Firing costs and labour demand: How bad is eurosclerosis?*, Rev. Econom. Stud., 57 (1990), pp. 381-402.

[4] M. B. Chiarolla and U. G. Haussmann, *Explicit solution of a stochastic, irreversible investment problem and its moving threshold*, Math. Oper. Res., 30 (2005), pp. 91–108.

[5] M. H. A. Davis, *Markov Models and Optimization*, Chapman and Hall, London, 1993.

[6] M. H. A. Davis, M. A. H. Dempster, S. P. Sethi, and D. Vermes, *Optimal capacity expansion under uncertainty*, Adv. in Appl. Probab., 19 (1987), pp. 156–176.

[7] X. Guo and H. Pham, *Optimal partially reversible investment with entry decisions and general production function*, Stochastic Process. Appl., 115 (2005), pp. 705–736.

[8] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.

[9] T. S. Knudsen, B. Meister, and M. Zervos, *Valuation of investments in real assets with implications for the stock prices*, SIAM J. Control Optim., 36 (1998), pp. 2082–2102.

[10] T. Ø. Kobila, *A class of solvable stochastic investment problems involving singular controls*, Stoch. Stoch. Rep., 43 (1993), pp. 29–63.

[11] A. Øksendal, *Irreversible investment problems*, Finance Stoch., 4 (2000), pp. 223–250.

[12] H. Wang, *Capacity expansion with exponential jump diffusion processes*, Stoch. Stoch. Rep., 75 (2003), pp. 259–274.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 877–899

EXACT CONTROLLABILITY OF A NONLINEAR KORTEWEG–DE VRIES EQUATION ON A CRITICAL SPATIAL DOMAIN*

EDUARDO CERPA†

Abstract. We consider the boundary controllability problem for a nonlinear Korteweg–de Vries equation with the Dirichlet boundary condition. We study this problem for a spatial domain with a critical length for which the linearized control system is not controllable. In order to deal with the nonlinearity, we use a power series expansion of second order. We prove that the nonlinear term gives the local exact controllability around the origin provided that the time of control is large enough.

Key words. controllability, Korteweg–de Vries equation, power series expansion

AMS subject classifications. 93B05, 35Q53

DOI. 10.1137/06065369X

1. Introduction.

Let $L > 0$ be fixed. Let us consider the following Korteweg–de Vries (KdV) control system with the Dirichlet boundary condition

$$\begin{cases} \partial_t y + \partial_x y + \partial_x^3 y + y\partial_x y = 0, \\ y(t,0) = y(t,L) = 0, \\ \partial_x y(t,L) = u(t), \end{cases} \tag{1.1}$$

where the state is $y(t,\cdot) : [0,L] \to \mathbb{R}$ and the control is $u(t) \in \mathbb{R}$. This is a well-known example of a nonlinear dispersive partial differential equation. This equation has been introduced by Korteweg and de Vries in [14] to describe approximately long waves in water of relatively shallow depth. A very good book to understand both physical motivation and deduction of the KdV equation is the book by Whitham [23].

We are concerned with the exact controllability properties of (1.1). In [17] Rosier has proved that this control system is locally exactly controllable around the origin provided that the length of the spatial domain is not critical. This was done using multiplier techniques and the Hilbert Uniqueness Method (HUM) method introduced by Lions (see [15]).

THEOREM 1.1 (see [17, Theorem 1.3]). *Let $T > 0$, and assume that*

$$L \notin N := \left\{ 2\pi \sqrt{\frac{k^2 + kl + l^2}{3}};\ k, l \in \mathbb{N}^* \right\}. \tag{1.2}$$

Then there exists $r > 0$ such that, for every $(y_0, y_T) \in L^2(0,L)^2$ with $\|y_0\|_{L^2(0,L)} < r$ and $\|y_T\|_{L^2(0,L)} < r$, there exist $u \in L^2(0,T)$ and

$$y \in C([0,T], L^2(0,L)) \cap L^2(0,T,H^1(0,L))$$

satisfying (1.1), *$y(0,\cdot) = y_0$, and $y(T,\cdot) = y_T$.*

Moreover, Rosier proved that the linearized control system of (1.1) around the origin, which is given by

$$\begin{cases} \partial_t y + \partial_x y + \partial_x^3 y = 0, \\ y(t,0) = y(t,L) = 0, \\ \partial_x y(t,L) = u(t), \end{cases} \tag{1.3}$$

*Received by the editors March 5, 2006; accepted for publication (in revised form) March 12, 2007; published electronically June 12, 2007.
http://www.siam.org/journals/sicon/46-3/65369.html

†Laboratoire de Mathématiques CNRS-UMR 8628, Université Paris-Sud, Bât. 425, 91405 Orsay Cedex, France (eduardo.cerpa@math.u-psud.fr).

is not controllable if $L \in N$. Indeed, there exists a finite-dimensional subspace of $L^2(0,L)$, denoted by M, which is unreachable for the linear system. More precisely, for every nonzero state $\psi \in M$, for every $u \in L^2(0,T)$, and for every $y \in C([0,T],L^2(0,L)) \cap L^2(0,T,H^1(0,L))$ satisfying (1.3) and $y(0,\cdot) = 0$, one has $y(T,\cdot) \neq \psi$.

Remark 1.2. If one is allowed to use more than one boundary control input, there is no critical spatial domain, and the exact controllability holds for any $L > 0$. More precisely, let us consider the nonlinear control system

$$\begin{cases} \partial_t y + \partial_x y + \partial_x^3 y + y\partial_x y = 0, \\ y(t,0) = u_1(t), \quad y(t,L) = u_2(t), \quad \partial_x y(t,L) = u_3(t), \end{cases} \tag{1.4}$$

where the controls are $u_1(t), u_2(t)$, and $u_3(t)$. As has been pointed out by Rosier in [17], for every $L > 0$ the system (1.4) with $u_1 \equiv 0$ is locally exactly controllable in $L^2(0,L)$ around the origin. Moreover, using all three control inputs, Zhang proved in [24] that for every $L > 0$ the system (1.4) is exactly controllable in the space $H^s(0,L)$ for any $s \geq 0$ in a neighborhood of a given smooth solution of the KdV equation.

Recently, Coron and Crépeau in [8] have proved Theorem 1.1 for the critical lengths $L = 2k\pi$, with $k \in \mathbb{N}^*$ satisfying

$$\nexists (m,n) \in \mathbb{N}^* \times \mathbb{N}^*, \quad \text{with} \quad m^2 + mn + n^2 = 3k^2 \quad \text{and} \quad m \neq n. \tag{1.5}$$

For these values of L, the subspace M of missed directions is one-dimensional and is generated by the function $f(x) = 1 - \cos(x)$. Their method consists, first, in moving along this direction by performing a power series expansion of the solution and then in using a fixed point theorem.

Remark 1.3. The condition (1.5) has been communicated to the author by Coron and Crépeau. They pointed out that if it is not satisfied, then the dimension of the missed directions subspace is higher than one, and the proof given in [8] does not work anymore.

In this paper, we follow the method of Coron and Crépeau to investigate the case of critical lengths for which the subspace M is two-dimensional. The set of lengths for which it holds is denoted by N'. We will see in section 2 that N' contains an infinite number of lengths.

This paper is organized as follows. First, in section 2, we study the linearized control system (1.3), and we provide a complete description of the space M in terms of the length L of the spatial domain $(0,L)$. Then, in section 3, we prove in the case $L \in N'$ that the nonlinear term $y\partial_x y$ allows us to reach all of the missed directions provided that the time of control is large enough. We give an explicit expression of the minimal time required by our method. Finally, in section 4, we get the local exact controllability by means of a fixed point theorem; i.e., we prove our main result.

THEOREM 1.4. *Let $L \in N'$. There exists $T_M > 0$ such that for any $T > T_M$ there exist $C > 0$ and $r > 0$ such that for every $(y_0, y_T) \in L^2(0,L)^2$ with $\|y_0\|_{L^2(0,L)} < r$ and $\|y_T\|_{L^2(0,L)} < r$ there exist $u \in L^2(0,T)$ with*

$$\|u\|_{L^2(0,T)} \leq C(\|y_0\|_{L^2(0,L)} + \|y_T\|_{L^2(0,L)})^{1/2} \tag{1.6}$$

and

$$y \in C([0,T],L^2(0,L)) \cap L^2(0,T,H^1(0,L))$$

satisfying (1.1), $y(0,\cdot) = y_0$, *and* $y(T,\cdot) = y_T$.

Remark 1.5. The power 1/2 in the estimate (1.6) comes, as we will see, from performing a power series expansion of second order to deal with the nonlinearity. The same estimate holds with power 1/3 for the critical lengths studied in [8] (third-order expansion) and with power 1 for the noncritical lengths studied in [17] (first-order expansion).

Remark 1.6. In order to complete the study of the exact controllability of system (1.1), it is necessary to investigate the case where the dimension of the space M is bigger than 2. An approach would be to use the exact controllability of the nonlinear equation around nontrivial stationary solutions proved by Crépeau (in [10] for the domains $(0, 2\pi k)$ and in [11] for any other domain $(0, L)$) and then to apply the method introduced in [5] (see also [1, 2]), that is, the return method (see [3, 4]), together with quasi-static deformations (see also [9]). With such a method, one should obtain the exact controllability of (1.1) for a large time. However, it seems that the minimal time required with this approach is far from being optimal.

Remark 1.7. In Theorem 1.4, we get the local controllability for (1.1) provided that the time of control is large enough. However, we may wonder if this condition on the time is really necessary. This is an interesting open problem since one knows that even if the speed of propagation of the KdV equation is infinite, it may exist a minimal time of control. This is, for example, the case of a nonlinear control system for the Schrödinger equation studied by Beauchard and Coron in [2]. They proved the local controllability of this system along the ground state trajectory for a large time. More recently, Coron proved in [6] and [7, Theorem 9.8] that this local controllability does not hold in small time, even if the Schrödinger equation has an infinite speed of propagation.

Remark 1.8. In [1, 2], there appear Schrödinger linear control systems which are not controllable. One could apply the method used in this paper to prove the local controllability of the corresponding nonlinear control systems. The main difficulty is that in those cases the subspace of missed directions for the linear system is not finite-dimensional.

Remark 1.9. Concerning the stabilization of the KdV equation, some results in the case of periodic boundary conditions can be found in [13] (damping distributed all along the domain), [20] (damping distributed with localized support), and [19] (boundary damping). In the case of the Dirichlet boundary condition, exponential decay of the solution has been obtained in [16] by adding a localized damping term (see also [18] for a generalization of this result). However, the decay rate is unknown. A natural open problem is to design for the control system (1.1) (or the linearized one (1.3)) stabilizing feedback laws which give us an explicit decay rate. This kind of result, even with a prescribed arbitrarily large decay rate, has been obtained in [12, 22] for a general class of second-order (in time) systems including the wave equation and platelike systems. It uses the fact that these systems are time-reversible. This is not the case of the control system (1.1).

2. Linearized control system. We first recall some properties proved by Rosier in [17]. Let $L > 0$ and $T > 0$. In order to study the following linear KdV equation:

$$\begin{cases} \partial_t y + \partial_x y + \partial_x^3 y = f, \\ y(t,0) = y(t,L) = 0, \\ \partial_x y(t,L) = u(t), \\ y(0,\cdot) = y_0, \end{cases} \tag{2.1}$$

we define the space $\mathcal{B} := C([0,T],L^2(0,L)) \cap L^2(0,T,H^1(0,L))$ endowed with the norm

$$\|y\|_{\mathcal{B}} = \max_{t\in[0,T]} \|y(t)\|_{L^2(0,L)} + \left(\int_0^T \|y(t)\|^2_{H^1(0,L)} dt\right)^{1/2}.$$

Let A denote the operator $Aw = -w' - w'''$ on the domain $D(A) \subset L^2(0,L)$ defined by

$$D(A) := \left\{w \in H^3(0,L);\, w(0) = w(L) = w'(L) = 0\right\}.$$

One can see that both A and its adjoint A^* are closed and dissipative. Hence A generates a strongly continuous semigroup of contractions. Using this fact and the multiplier method, Rosier proved the following existence and uniqueness result.

PROPOSITION 2.1 (see [17, Propositions 3.2 and 3.7]). *There exist unique continuous linear maps* Ψ *and* δ

$$\begin{array}{rcl} \Psi : L^2(0,L)\times L^2(0,T)\times L^1(0,T,L^2(0,L)) & \longrightarrow & \mathcal{B}, \\ (y_0,u,f) & \longmapsto & \Psi(y_0,u,f), \\ \delta : L^2(0,L)\times L^2(0,T)\times L^1(0,T,L^2(0,L)) & \longrightarrow & L^2(0,T), \\ (y_0,u,f) & \longmapsto & \delta(y_0,u,f), \end{array}$$

such that, for $y_0 \in D(A)$, $u \in C^2([0,T])$, *with* $u(0) = 0$, *and* $f \in C^1([0,T],L^2(0,L))$, *then* $\Psi(y_0,u,f)$ *is the unique classical solution of* (2.1) *and*

$$\delta(y_0,u,f) = \partial_x \Psi(y_0,u,f)(\cdot,0).$$

The function $\Psi(y_0,u,f)$ is called the mild solution or simply the solution of (2.1) in the context of this paper.

Now we focus our attention on the domains of critical length. In particular, we describe the space M of unreachable states for the linear control system (1.3). Let $L \in N$. There exists a finite number of pairs $\{(k_j,l_j)\}_{j=1}^n \subset \mathbb{N}^*\times\mathbb{N}^*$, with $k_j \geq l_j$, such that

$$L = 2\pi\sqrt{\frac{k_j^2 + k_j l_j + l_j^2}{3}}. \tag{2.2}$$

From the work of Rosier in [17], we know that for each $j \in \{1,\dots,n\}$ there exist two nonzero real-valued functions $\varphi_1^j = \varphi_1^j(x)$ and $\varphi_2^j = \varphi_2^j(x)$ such that $\varphi^j := \varphi_1^j + i\varphi_2^j$ is a solution of

$$\left\{\begin{array}{l} -ip(k_j,l_j)\varphi^j + \varphi^{j\prime} + \varphi^{j\prime\prime\prime} = 0, \\ \varphi^j(0) = \varphi^j(L) = 0, \\ \varphi^{j\prime}(0) = \varphi^{j\prime}(L) = 0, \end{array}\right. \tag{2.3}$$

where, for $(k,l) \in \mathbb{N}^*\times\mathbb{N}^*$, $p(k,l)$ is defined by

$$p(k,l) := \frac{(2k+l)(k-l)(2l+k)}{3\sqrt{3}(k^2+kl+l^2)^{3/2}}.$$

Easy computations lead to

$$\begin{aligned} \varphi_1^j &= C\left(\cos(\gamma_1^j x) - \frac{\gamma_1^j - \gamma_3^j}{\gamma_2^j - \gamma_3^j}\cos(\gamma_2^j x) + \frac{\gamma_1^j - \gamma_2^j}{\gamma_2^j - \gamma_3^j}\cos(\gamma_3^j x)\right), \\ \varphi_2^j &= C\left(\sin(\gamma_1^j x) - \frac{\gamma_1^j - \gamma_3^j}{\gamma_2^j - \gamma_3^j}\sin(\gamma_2^j x) + \frac{\gamma_1^j - \gamma_2^j}{\gamma_2^j - \gamma_3^j}\sin(\gamma_3^j x)\right), \end{aligned} \tag{2.4}$$

where C is a constant and the numbers γ_m^j, with $m = 1, 2, 3$, are the three roots of $x^3 - x + p(k_j, l_j) = 0$. One can easily verify that these roots are given by

$$\gamma_1^j = -\frac{2\pi}{L}\left(\frac{2k_j + l_j}{3}\right), \quad \gamma_2^j = \gamma_1^j + \frac{2\pi k_j}{L}, \quad \gamma_3^j = \gamma_2^j + \frac{2\pi l_j}{L}. \tag{2.5}$$

Moreover, by choosing the constant C, we can assume that

$$\|\varphi_1^j\|_{L^2(0,L)} = \|\varphi_2^j\|_{L^2(0,L)} = 1.$$

Roughly speaking, the functions φ_1^j and φ_2^j for $j = 1, \dots, n$ are unreachable states for the linear KdV control system (1.3) since the following functions:

$$y_1(t,x) = \mathrm{Re}(e^{-ip(k_j,l_j)t}\varphi^j(x)) \quad \text{and} \quad y_2(t,x) = \mathrm{Im}(e^{-ip(k_j,l_j)t}\varphi^j(x))$$

are solutions of (1.3) with $u(t) \equiv 0$, but they do not satisfy the next observability inequality leading to the exact controllability

$$\|y(0,x)\|_{L^2(0,L)} \le C\|\partial_x y(t,0)\|_{L^2(0,T)}.$$

Let us define the following subspaces of $L^2(0,L)$:

$$M := \langle\{\varphi_1^1, \varphi_2^1, \dots, \varphi_1^n, \varphi_2^n\}\rangle \quad \text{and} \quad H := M^\perp.$$

Remark 2.2. If $p(k_j, l_j) = 0$ for some $j \in \{1, \dots, n\}$, then $\varphi_1^j = \varphi_2^j = 1 - \cos(x)$. It occurs when $k_j = l_j$, i.e., if $L = 2\pi k_j$. If k_j satisfies the condition (1.5), then the space M is one-dimensional. This is the case treated in [8]. It corresponds, for example, to the length $L = 2\pi$.

Remark 2.3. If $p(k_j, l_j) \neq 0$, it is easy to see that $\varphi_1^j \perp \varphi_2^j$. Moreover, for distinct $j_1, j_2 \in \{1, \dots, n\}$, $\varphi_m^{j1} \perp \varphi_s^{j2}$ for $m, s = 1, 2$. Let us give some examples. The pair $(2, 1)$ defines a critical length for which the space M is two-dimensional. The pair $(11, 8)$ defines a critical length for which the space M is four-dimensional since the pairs $(11, 8)$ and $(16, 1)$ define the same critical length.

At this point, we can state the following controllability result which follows directly from the work of Rosier in [17, Propositions 3.3 and 3.9].

THEOREM 2.4. *Let $T > 0$. For every $(y_0, y_T) \in H \times H$, there exist $u \in L^2(0,T)$, and $y \in \mathcal{B}$ satisfying* (1.3)*, $y(0,\cdot) = y_0$, and $y(T,\cdot) = y_T$.*

Now let us define the set N' by

$$N' := \left\{ 2\pi\sqrt{\frac{k^2 + kl + l^2}{3}}; (k,l) \in \mathbb{N}^* \times \mathbb{N}^* \text{ satisfying } k > l \text{ and } (2.7) \right\} \tag{2.6}$$

$$\forall m, n \in \mathbb{N}^*\backslash\{k\},\ k^2 + kl + l^2 \neq m^2 + mn + n^2. \tag{2.7}$$

It is easy to see that N' is the set of critical lengths for which the space of unreachable states is two-dimensional. Indeed, let $L \in N'$; from (2.7) there exists a unique pair $(k_1, l_1) := (k, l)$ satisfying (2.2), and since $k_1 > l_1$, $p(k_1, l_1) > 0$, and therefore the functions φ_1^1, φ_2^1 are orthogonal.

Let us follow the proof of Proposition 8.3 in [7] in order to see that N' contains an infinite number of elements. Let $q \ge 1$ be an integer satisfying

$$\forall m, n \in \mathbb{N}^*\backslash\{q\}, \quad m^2 + mn + n^2 \neq 7q^2. \tag{2.8}$$

Let us consider the critical length L_q defined by the pair $(2q, q)$, that is,

$$L_q := 2\pi\sqrt{\frac{(2q)^2 + 2q^2 + q^2}{3}} = 2\pi q\sqrt{\frac{7}{3}}.$$

From (2.8), it is easy to see that $L_q \in N'$. One can verify that (2.8) holds for $q = 1, 2, 3$, and therefore $L_1, L_2, L_3 \in N'$. Moreover, the following lemma says that the set N' contains an infinite number of lengths L_q.

LEMMA 2.5. *There are infinitely many positive integers q satisfying* (2.8).

Proof. Let $q > 3$ be a prime integer which does not satisfy (2.8), that is, such that

$$\exists m, n \in \mathbb{N}^* \backslash \{q\}, \quad m^2 + mn + n^2 = 7q^2. \tag{2.9}$$

From (2.9) one gets

$$-3mn = (m-n)^2 \pmod q, \quad mn = (m+n)^2 \pmod q. \tag{2.10}$$

It is easy to see that $m + n \neq 0 \pmod q$, and consequently from (2.10) we have

$$-3 = \left((m+n)^{-1}(m-n)\right)^2 \pmod q; \tag{2.11}$$

that is, -3 is a square modulo q. Let us introduce the Legendre symbol, where s is a prime and $x \in \mathbb{Z}$ is an integer not divisible by s:

$$\left(\frac{x}{s}\right) := \begin{cases} 1 & \text{if } x \text{ is a square modulo } s, \\ -1 & \text{if } x \text{ is not a square modulo } s. \end{cases}$$

We have the quadratic reciprocity law due to Gauss for every prime integer $z > 2$, $s > 2$ (see [21, Chapter 3])

$$\left(\frac{s}{z}\right) = \left(\frac{z}{s}\right)(-1)^{\epsilon(z)\epsilon(s)}, \tag{2.12}$$

where

$$\epsilon(z) = \begin{cases} 0 & \text{if } z = 1 \pmod 4, \\ 1 & \text{if } z = -1 \pmod 4. \end{cases}$$

From [21, Chapter 3], we also have that for every x, y coprime to s

$$\left(\frac{xy}{s}\right) = \left(\frac{x}{s}\right)\left(\frac{y}{s}\right) \tag{2.13}$$

and for every $s > 2$ prime integer

$$(-1)^{\epsilon(s)} = \left(\frac{-1}{s}\right). \tag{2.14}$$

Using (2.12), (2.14), (2.13), and (2.11) with $s = q$, $z = 3$, and since $\epsilon(3) = 1$, one obtains

$$\left(\frac{q}{3}\right) = \left(\frac{3}{q}\right)(-1)^{\epsilon(q)} = \left(\frac{3}{q}\right)\left(\frac{-1}{q}\right) = \left(\frac{-3}{q}\right) = 1;$$

that is, $q = 1 \pmod 3$.

Hence, if $q > 3$ is a prime integer such that $q = 2 \pmod 3$, then q satisfies (2.8). As there are two possible nonzero congruences modulo 3, the Dirichlet density theorem (see [21, Chapter 4]) says that (2.8) holds on a set of prime integers of density 1/2. In particular, there are infinitely many positive integers q satisfying (2.8). □

From now on and until the end of this paper, we consider $L \in N'$. From (2.7), for each $L \in N'$ we can define a unique

$$p := \frac{(2k+l)(k-l)(2l+k)}{3\sqrt{3}(k^2+kl+l^2)^{3/2}},$$

and the space M is then defined by

$$M := \langle \varphi_1, \varphi_2 \rangle = \{\alpha\varphi_1 + \beta\varphi_2 \,;\, \alpha, \beta \in \mathbb{R}\},$$

where φ_1 and φ_2 are given by (2.4) with γ_m^j replaced by γ_m, where γ_1, γ_2, and γ_3 are the three roots of $x^3 - x + p = 0$. From (2.3) we also have that φ_1 and φ_2 satisfy

$$\begin{cases} \varphi_1' + \varphi_1''' = -p\,\varphi_2, \\ \varphi_1(0) = \varphi_1(L) = 0, \\ \varphi_1'(0) = \varphi_1'(L) = 0, \end{cases} \tag{2.15}$$

and

$$\begin{cases} \varphi_2' + \varphi_2''' = p\,\varphi_1, \\ \varphi_2(0) = \varphi_2(L) = 0, \\ \varphi_2'(0) = \varphi_2'(L) = 0. \end{cases} \tag{2.16}$$

Now we investigate the evolution of the projection on the subspace M of a solution of (1.3). Let us consider $(y, u) \in \mathcal{B} \times L^2(0,T)$ satisfying (1.3). Let us multiply (2.15) by y and integrate on $[0, L]$. Using integrations by parts we get

$$\frac{d}{dt}\left(\int_0^L y(t,x)\varphi_1(x)dx\right) = -p\int_0^L y(t,x)\varphi_2(x)dx. \tag{2.17}$$

Similarly, multiplying (2.16) by y, we get

$$\frac{d}{dt}\left(\int_0^L y(t,x)\varphi_2(x)dx\right) = p\int_0^L y(t,x)\varphi_1(x)dx. \tag{2.18}$$

Hence, from (2.17) and (2.18), we obtain

$$\int_0^L y(t,x)\varphi_1(x)\,dx = \int_0^L y(0,x)(\cos(p\,t)\varphi_1(x) - \sin(p\,t)\varphi_2(x))\,dx, \tag{2.19}$$

$$\int_0^L y(t,x)\varphi_2(x)\,dx = \int_0^L y(0,x)(\sin(p\,t)\varphi_1(x) + \cos(p\,t)\varphi_2(x))\,dx. \tag{2.20}$$

From (2.19) and (2.20), we see that the projection on M of $y(t,\cdot)$, denoted by $P_M(y(t,\cdot))$, only turns in this two-dimensional subspace and therefore conserves its $L^2(0,L)$-norm. The period of this rotation is $2\pi/p$. Furthermore, we see that if the initial condition $y(0,\cdot)$ lies in H, the solution does too for every time t. Combining this rotation with Theorem 2.4, we obtain the following proposition.

PROPOSITION 2.6. *Let $y_0, y_1 \in L^2(0,L)$ be such that*

$$\|P_M(y_0)\|_{L^2(0,L)} = \|P_M(y_1)\|_{L^2(0,L)}.$$

Then there exists $t^* \leq \frac{2\pi}{p}$ and $u \in L^2(0,t^*)$ such that the solution $y = y(t,x)$ of (1.3), with $y(0,\cdot) = y_0$, satisfies $y(t^*,\cdot) = y_1$.

Proof. Let $y_M = y_M(t,x)$ be the solution of (1.3), with $y_M(0,\cdot) = P_M(y_0)$ and without control ($u \equiv 0$). We know that there exists a time $0 < t^* \leq \frac{2\pi}{p}$ such that $y_M(t^*,\cdot) = P_M(y_1)$. On the other hand, from Theorem 2.4 there exists a control $u_H \in L^2(0,t^*)$ such that the corresponding solution $y_H = y_H(t,x)$ of (1.3) satisfies

$$y_H(0,\cdot) = P_H(y_0) \in H \quad \text{and} \quad y_H(t^*,\cdot) = P_H(y_1).$$

Then $y(t,x) := y_H(t,x) + y_M(t,x)$ satisfies (1.3), with $u = u_H$, $y(0,\cdot) = y_0$, and $y(t^*,\cdot) = y_1$, which ends the proof of this proposition. ∎

3. Motion in the missed directions. Let us first explain the general idea of the method. Let $y = y(t,x)$ be a solution of (1.1) with control $u = u(t)$. We consider a power series expansion of (y,u) with the same scaling on the state and on the control

$$\begin{aligned} y &= \epsilon y_1 + \epsilon^2 y_2 + \epsilon^3 y_3 \ldots, \\ u &= \epsilon u_1 + \epsilon^2 u_2 + \epsilon^3 u_3 \ldots. \end{aligned}$$

In this way, we see that the nonlinear term is given by

$$y\partial_x y = \epsilon^2 y_1 \partial_x y_1 + \epsilon^3 y_1 \partial_x y_2 + \epsilon^3 y_2 \partial_x y_1 + (\text{higher terms}),$$

and therefore, for a small ϵ, we have the expansion of second order $y \approx \epsilon y_1 + \epsilon^2 y_2$, where y_1 and y_2 are given by

$$\begin{cases} \partial_t y_1 + \partial_x y_1 + \partial_x^3 y_1 = 0, \\ y_1(t,0) = y_1(t,L) = 0, \\ \partial_x y_1(t,L) = u_1(t), \end{cases}$$

and

$$\begin{cases} \partial_t y_2 + \partial_x y_2 + \partial_x^3 y_2 = -y_1 \partial_x y_1, \\ y_2(t,0) = y_2(t,L) = 0, \\ \partial_x y_2(t,L) = u_2(t), \end{cases}$$

respectively. The strategy consists first in proving that the expansion to the second order of $y = y(t,x)$, i.e., $\epsilon y_1 + \epsilon^2 y_2$, can reach all of the missed directions and then in using a fixed point argument to prove that it is sufficient to get Theorem 1.4. This is a classical approach to study the local controllability of a finite-dimensional control system, and it has been applied in [8] to prove the local exact controllability around the origin of the control system (1.1) for some critical domains.

Now we see that we can "enter" into the subspace M. More precisely, the result we prove is the following one.

PROPOSITION 3.1. *Let $T > 0$. There exists $(u,v) \in L^2(0,T)^2$ such that if $\alpha = \alpha(t,x)$ and $\beta = \beta(t,x)$ are the solutions of*

$$\begin{cases} \partial_t \alpha + \partial_x \alpha + \partial_x^3 \alpha = 0, \\ \alpha(t,0) = \alpha(t,L) = 0, \\ \partial_x \alpha(t,L) = u(t), \\ \alpha(0,\cdot) = 0, \end{cases} \tag{3.1}$$

and

$$\begin{cases} \partial_t\beta + \partial_x\beta + \partial_x^3\beta = -\alpha\partial_x\alpha, \\ \beta(t,0) = \beta(t,L) = 0, \\ \partial_x\beta(t,L) = v(t), \\ \beta(0,\cdot) = 0, \end{cases} \tag{3.2}$$

then

$$\alpha(T,\cdot) = 0 \quad \textit{and} \quad \beta(T,\cdot) \in M\backslash\{0\}.$$

Proof. In order to study the trajectory $\beta = \beta(t,x)$, we set $\beta = \beta^u + \beta^v$, where $\beta^u = \beta^u(t,x)$ and $\beta^v = \beta^v(t,x)$ are the solutions of

$$\begin{cases} \partial_t\beta^u + \partial_x\beta^u + \partial_x^3\beta^u = -\alpha\partial_x\alpha, \\ \beta^u(t,0) = \beta^u(t,L) = 0, \\ \partial_x\beta^u(t,L) = 0, \\ \beta^u(0,\cdot) = 0, \end{cases} \tag{3.3}$$

and

$$\begin{cases} \partial_t\beta^v + \partial_x\beta^v + \partial_x^3\beta^v = 0, \\ \beta^v(t,0) = \beta^v(t,L) = 0, \\ \partial_x\beta^v(t,L) = v(t), \\ \beta^v(0,\cdot) = 0, \end{cases} \tag{3.4}$$

respectively. If $u \in L^2(0,T)$ is given, by Theorem 2.4 one can find $v \in L^2(0,T)$ such that

$$\beta^v(T,\cdot) = -P_H(\beta^u(T,\cdot))$$

and thus $\beta(T,\cdot) = P_M(\beta^u(T,\cdot))$. From this fact, one sees that the proof of Proposition 3.1 can be reduced to prove

$$\exists u \in L^2(0,T) \quad \text{such that} \quad \alpha(T,\cdot) = 0 \quad \text{and} \quad P_M(\beta^u(T,\cdot)) \neq 0. \tag{3.5}$$

Let $u \in L^2(0,T)$. Let us multiply (3.3) by φ_1 and integrate the resulting equality on $[0,L]$. Then, using integration by parts, (2.15), and the boundary and initial conditions in (3.3), one gets

$$\frac{d}{dt}\left(\int_0^L \beta^u(t,x)\varphi_1(x)dx\right) = -p\int_0^L \beta^u(t,x)\varphi_2(x)dx + \frac{1}{2}\int_0^L \alpha^2(t,x)\varphi_1'(x)dx.$$

In a similar way, if we now multiply (3.3) by φ_2, we get

$$\frac{d}{dt}\left(\int_0^L \beta^u(t,x)\varphi_2(x)dx\right) = p\int_0^L \beta^u(t,x)\varphi_1(x)dx + \frac{1}{2}\int_0^L \alpha^2(t,x)\varphi_2'(x)dx.$$

If we call

$$\eta_k(t) := \int_0^L \beta^u(t,x)\varphi_k(x)dx \quad \text{for } k = 1,2,$$

we can write the system

$$\left\{ \begin{array}{l} \begin{pmatrix} \dot\eta_1(t) \\ \dot\eta_2(t) \end{pmatrix} = \begin{pmatrix} 0 & -p \\ p & 0 \end{pmatrix} \begin{pmatrix} \eta_1(t) \\ \eta_2(t) \end{pmatrix} + \begin{pmatrix} \frac{1}{2}\int_0^L \alpha^2(t,x)\varphi_1'(x)dx \\ \frac{1}{2}\int_0^L \alpha^2(t,x)\varphi_2'(x)dx \end{pmatrix}, \\ \eta_1(0) = 0, \quad \eta_2(0) = 0. \end{array} \right. \tag{3.6}$$

The solution of (3.6) is given by

$$\begin{pmatrix} \eta_1(t) \\ \eta_2(t) \end{pmatrix} = \begin{pmatrix} \cos(p\,t) & -\sin(p\,t) \\ \sin(p\,t) & \cos(p\,t) \end{pmatrix} \begin{pmatrix} I_1(t) \\ I_2(t) \end{pmatrix},$$

where

$$I_1(t) := \frac{1}{2}\int_0^t \int_0^L \alpha^2(s,x)(\cos(ps)\varphi_1'(x) + \sin(ps)\varphi_2'(x))dx\,ds,$$
$$I_2(t) := \frac{1}{2}\int_0^t \int_0^L \alpha^2(s,x)(-\sin(ps)\varphi_1'(x) + \cos(ps)\varphi_2'(x))dx\,ds.$$

If we work with complex numbers calling $\varphi := \varphi_1 + i\varphi_2$, we get

$$\eta_1(t) + i\eta_2(t) = \frac{1}{2}e^{ip\,t}\int_0^t \int_0^L e^{-ips}\alpha^2(s,x)\varphi'(x)dx\,ds.$$

Now let us assume that (3.5) fails to be true; i.e., let us suppose that

$$\forall\, u \in L^2(0,T), \quad \eta_1(T) = \eta_2(T) = 0 \quad \text{or} \quad \alpha(T,\cdot) \neq 0. \tag{3.7}$$

If we define

$$U_{ad} := \left\{ u \in L^2(0,T)\,;\ \text{the solution } \alpha \text{ of (3.1) satisfies } \alpha(T,\cdot) = 0 \right\},$$

then condition (3.7) implies that

$$\forall\, u \in U_{ad}, \quad \int_0^T \int_0^L e^{-ips}\alpha^2(s,x)\varphi'(x)dx\,ds = 0. \tag{3.8}$$

Let $\alpha_1 = \alpha_1(t,x)$ and $\alpha_2 = \alpha_2(t,x)$ be two solutions of (3.1) such that

$$\alpha_1(T,\cdot) = \alpha_2(T,\cdot) = 0.$$

Now, for $(\rho_1,\rho_2) \in \mathbb{R}^2$, let $\alpha := \rho_1\alpha_1 + \rho_2\alpha_2$ and $u := \alpha_x(\cdot,L)$. By linearity, we see that $\alpha = \alpha(t,x)$ is a solution of (3.1) and $u \in U_{ad}$. Consequently, (3.8) implies that, for every $(\rho_1,\rho_2) \in \mathbb{R}^2$,

$$\rho_1^2 \int_0^T \int_0^L e^{-ips}\alpha_1^2(s,x)\varphi'(x)dx\,ds + 2\rho_1\rho_2 \int_0^T \int_0^L e^{-ips}\alpha_1(s,x)\alpha_2(s)\varphi'(x)dx\,ds + \rho_2^2 \int_0^T \int_0^L e^{-ips}\alpha_2^2(s,x)\varphi'(x)dx\,ds = 0.$$

Looking at the coefficient of $\rho_1\rho_2$, we get

$$\int_0^T \int_0^L e^{-ips}\alpha_1(s,x)\alpha_2(s,x)\varphi'(x)dx\,ds = 0. \tag{3.9}$$

Let t_1, t_2 be such that $0 < t_1 < t_2 < T$. We choose the trajectories $\alpha_1 = \alpha_1(t,x)$ and $\alpha_2 = \alpha_2(t,x)$ such that

$$\alpha_2 \text{ is not identically equal to } 0, \tag{3.10}$$

$$\alpha_2(t,x)|_{([0,t_1]\cup[t_2,T])\times[0,L]} = 0 \quad \text{and} \quad \alpha_1(t,x)|_{[t_1,t_2]\times[0,L]} = \mathrm{Re}(e^{\lambda t} y_\lambda(x)), \tag{3.11}$$

where $\lambda \in \mathbb{C}\backslash\{\pm ip\}$ and $y_\lambda = y_\lambda(x)$ is a complex-valued function which satisfies

$$\begin{cases} \lambda y_\lambda + y_\lambda' + y_\lambda''' = 0, \\ y_\lambda(0) = y_\lambda(L) = 0. \end{cases} \tag{3.12}$$

If $\lambda \neq \pm ip$, one can see that $\mathrm{Re}(y_\lambda), \mathrm{Im}(y_\lambda) \in H$, and then by Theorem 2.4 there exists such a trajectory $\alpha_1 = \alpha_1(t,x)$.

Let us introduce the operator $\tilde{A}w = -w' - w'''$ on the domain $D(\tilde{A}) \subset L^2(0,L)$ defined by

$$D(\tilde{A}) := \left\{ w \in H^3(0,L);\, w(0) = w(L) = 0, w'(0) = w'(L) \right\}.$$

It is not difficult to see that $i\tilde{A}$ is a self-adjoint operator on $L^2(0,L)$ with compact resolvent. Hence, the spectrum $\sigma(\tilde{A})$ of $\tilde{A}$ consists only of eigenvalues. Furthermore, the spectrum is a discrete subset of $i\mathbb{R}$.

If we take λ such that $(-ip+\lambda) \notin \sigma(\tilde{A})$, the operator $(\tilde{A} - (-ip+\lambda)I)$ is invertible, and thus, there exists a unique complex-valued function $\phi_\lambda = \phi_\lambda(x)$ solution of

$$\begin{cases} (-ip+\lambda)\phi_\lambda + \phi_\lambda' + \phi_\lambda''' = y_\lambda \varphi', \\ \phi_\lambda(0) = \phi_\lambda(L) = 0, \\ \phi_\lambda'(0) = \phi_\lambda'(L). \end{cases} \tag{3.13}$$

We multiply (3.13) by $\alpha_2(t,x)e^{(-ip+\lambda)t}$, integrate on $[0,L]$, and use integrations by parts together with (3.1), and the boundary and initial conditions in (3.13) to get

$$e^{-ip\,t}\int_0^L e^{\lambda t} y_\lambda \alpha_2(t,x)\varphi'(x)dx = \frac{d}{dt}\left(\int_0^L e^{(-ip+\lambda)t}\phi_\lambda(x)\alpha_2(t,x)dx\right) - e^{(-ip+\lambda)t}\phi_\lambda'(L)\partial_x\alpha_2(t,x)\Big|_{x=0}^{L}.$$

Then let us integrate this equality on $[0,T]$ and use the fact that $\alpha_2(0,\cdot) = 0$ and $\alpha_2(T,\cdot) = 0$. We obtain

$$\int_0^T\int_0^L e^{-ip\,t}e^{\lambda t} y_\lambda \alpha_2(t,x)\varphi'(x)dx\,dt = -\phi_\lambda'(L)\int_0^T e^{(-ip+\lambda)t}\left(\partial_x\alpha_2(t,L) - \partial_x\alpha_2(t,0)\right)dt. \tag{3.14}$$

On the other hand, by (3.9) and (3.11), it follows that

$$\int_0^T\int_0^L e^{-ip\,t}\mathrm{Re}(e^{\lambda t} y_\lambda)\alpha_2(t,x)\varphi'(x)dx\,dt = 0, \tag{3.15}$$

and, since one can also take a trajectory $\tilde{\alpha}_1 = \tilde{\alpha}_1(t,x)$ such that

$$\tilde{\alpha}_1(t,x)|_{[t_1,t_2]\times[0,L]} = \mathrm{Im}(e^{\lambda t} y_\lambda(x)),$$

one deduces from (3.9) that

$$\int_0^T \int_0^L e^{-ip\,t}\mathrm{Im}(e^{\lambda t} y_\lambda)\alpha_2(t,x)\varphi'(x)dx\,dt = 0. \tag{3.16}$$

Therefore, from (3.15) and (3.16), one gets

$$\int_0^T \int_0^L e^{-ip\,t} e^{\lambda t} y_\lambda \alpha_2(t,x)\varphi'(x)dx\,dt = 0,$$

and consequently from (3.14), for every $\lambda \neq \pm ip$ such that $(-ip+\lambda) \notin \sigma(\tilde{A})$, one has

$$\phi'_\lambda(L)\int_0^T e^{(-ip+\lambda)t}\left(\partial_x\alpha_2(t,L) - \partial_x\alpha_2(t,0)\right)dt = 0. \tag{3.17}$$

Let $a \in \mathbb{R}\backslash[-1/\sqrt{3}, 1/\sqrt{3}]$. We take $\lambda = 2ai(4a^2-1)$. Let

$$y_\lambda(x) = Ce^{(-\sqrt{3a^2-1}-ai)x} + (1-C)e^{(\sqrt{3a^2-1}-ai)x} - e^{2aix}, \tag{3.18}$$

where

$$C = \frac{e^{2aiL} - e^{(\sqrt{3a^2-1}-ai)L}}{e^{(-\sqrt{3a^2-1}-ai)L} - e^{(\sqrt{3a^2-1}-ai)L}}.$$

One easily checks that such a $y_\lambda = y_\lambda(x)$ satisfies (3.12) and $y_\lambda \neq 0$. Let us define

$$\Sigma := \left\{ a \in \mathbb{R}\backslash[-1/\sqrt{3}, 1/\sqrt{3}]\,;\, \lambda \notin \sigma(\tilde{A}), (\lambda - ip) \notin \sigma(\tilde{A}) \right\},$$

where $\lambda = 2ai(4a^2-1)$. Then the function $S : \Sigma \to \mathbb{C}$, $S(a) = \phi'_\lambda(L)$ is continuous. Now we use the fact that S is not identically equal to the function 0 (the proof of this statement will be given in Lemma 3.6 at the end of this section). Then there exist $\widehat{a} \in \Sigma$ and $\epsilon > 0$ such that, for every $a \in \Sigma$ with $|a - \widehat{a}| < \epsilon$, $S(a) \neq 0$. From (3.17) one gets

$$\forall\, a \in \Sigma, \quad |a-\widehat{a}| < \epsilon, \quad \int_0^T e^{(-p+2a(4a^2-1))i\,t}\left(\partial_x\alpha_2(t,L) - \partial_x\alpha_2(t,0)\right)dt = 0,$$

and since the function $\beta \in \mathbb{C} \mapsto \int_0^T e^{\beta t}\left(\partial_x\alpha_2(t,L) - \partial_x\alpha_2(t,0)\right)dt \in \mathbb{C}$ is holomorphic, it follows that

$$\forall\, \beta \in \mathbb{C}, \quad \int_0^T e^{\beta t}\big(\partial_x\alpha_2(t,L) - \partial_x\alpha_2(t,0)\big)dt = 0,$$

which implies that $\partial_x\alpha_2(t,0) - \partial_x\alpha_2(t,L) = 0$ for every t. In summary, one has that $\alpha_2 = \alpha_2(t,x)$ satisfies

$$\begin{cases} \partial_t\alpha_2 + \partial_x\alpha_2 + \partial_x^3\alpha_2 = 0, \\ \alpha_2(t,0) = \alpha_2(t,L) = 0, \\ \partial_x\alpha_2(t,0) = \partial_x\alpha_2(t,L), \\ \alpha_2(0,\cdot) = 0, \\ \alpha_2(T,\cdot) = 0. \end{cases} \tag{3.19}$$

If we multiply (3.19) by α_2, integrate on $[0, L]$, and use integration by parts together with the boundary conditions, we obtain that

$$\frac{d}{dt}\int_0^L |\alpha_2(t,x)|^2 dx = 0,$$

which, together with $\alpha_2(0,\cdot) = 0$, implies that

$$\alpha_2(t,x) = 0 \quad \forall\, x \in [0,L],\ \forall\, t \in [0,T]. \tag{3.20}$$

But this is in contradiction with (3.10). Thus, we have proved (3.5) and therefore Proposition 3.1. □

From now on, for each $T_c > 0$, we denote by $(u_c, v_c) \in L^2(0,T)^2$ the controls given by Proposition 3.1 and by (α_c, β_c) the corresponding trajectories. Let us define $\tilde{\varphi}_1 := \beta_c(T_c, \cdot)$. Let us notice that, by scaling the controls, we can assume that $\|\tilde{\varphi}_1\|_{L^2(0,L)} = 1$. We will prove now that in any time $T > \pi/p$, we can reach all of the states lying in M.

PROPOSITION 3.2. *Let $T > \pi/p$. Let $\psi \in M$. There exists $(u,v) \in L^2(0,T)^2$ such that if $\alpha = \alpha(t,x)$ and $\beta = \beta(t,x)$ are the solutions of* (3.1) *and* (3.2)*, then*

$$\alpha(T,\cdot) = 0 \quad \textit{and} \quad \beta(T,\cdot) = \psi.$$

Proof. Let $\hat{T} > 0$ be such that $T = (\pi/p) + \hat{T}$. Let T_c be such that $0 < T_c < \hat{T}$. Let $T_a := T - T_c$. If we take in (3.1) and (3.2) the controls

$$(u^1, v^1)(t) = \begin{cases} (0,0) & \text{if } t \in (0, T_a), \\ (u_c(t - T_a), v_c(t - T_a)) & \text{if } t \in (T_a, T), \end{cases}$$

we obtain that $\beta^1(T,\cdot) = \tilde{\varphi}_1$, where $\beta^1 = \beta^1(t,x)$ is the corresponding solution of (3.2). Now we use the rotation showed in section 2 (see, in particular, (2.19) and (2.20)) in order to reach other states lying in M. Let us define $\tilde{\varphi}_2 := \beta^2(T,\cdot)$, where $\beta^2 = \beta^2(t,x)$ is defined by the controls

$$(u^2, v^2)(t) = \begin{cases} (0,0) & \text{if } t \in (0, T_a - \frac{\pi}{2p}), \\ (u_c(t - T_a + \frac{\pi}{2p}), v_c(t - T_a + \frac{\pi}{2p})) & \text{if } t \in (T_a - \frac{\pi}{2p}, T - \frac{\pi}{2p}), \\ (0,0) & \text{if } t \in (T - \frac{\pi}{2p}, T). \end{cases}$$

In a similar way, the controls

$$(u^3, v^3)(t) = \begin{cases} (0,0) & \text{if } t \in (0, T_a - \frac{\pi}{p}), \\ (u_c(t - T_a + \frac{\pi}{p}), v_c(t - T_a + \frac{\pi}{p})) & \text{if } t \in (T_a - \frac{\pi}{p}, T - \frac{\pi}{p}), \\ (0,0) & \text{if } t \in (T - \frac{\pi}{p}, T) \end{cases}$$

allow us to define $\tilde{\varphi}_3 := \beta^3(T,\cdot)$. Notice that $\tilde{\varphi}_3 = -\tilde{\varphi}_1$.

Let T_θ be such that $0 < T_\theta < \min\{\pi/(2p), \hat{T} - T_c\}$, and let $T_b := (\pi/p) + T_\theta$. Let us define $\tilde{\varphi}_4 := \beta^4(T,\cdot)$, where $\beta^4 = \beta^4(t,x)$ is the solution of (3.2), with

$$(u^4, v^4)(t) = \begin{cases} (0,0) & \text{if } t \in (0, T_a - T_b), \\ (u_c(t - T_a + T_b), v_c(t - T_a + T_b)) & \text{if } t \in (T_a - T_b, T - T_b), \\ (0,0) & \text{if } t \in (T - T_b, T). \end{cases}$$

We have thus proved that we can reach the missed directions $\{\tilde{\varphi}_k\}_{k=1}^4$. Let us now define the cones

$$\begin{aligned} M_1 &:= \{d_1\tilde{\varphi}_1 + d_2\tilde{\varphi}_2; d_1 > 0, d_2 \geq 0\},\\ M_2 &:= \{d_1\tilde{\varphi}_2 + d_2\tilde{\varphi}_3; d_1 > 0, d_2 \geq 0\},\\ M_3 &:= \{d_1\tilde{\varphi}_3 + d_2\tilde{\varphi}_4; d_1 > 0, d_2 \geq 0\},\\ M_4 &:= \{d_1\tilde{\varphi}_4 + d_2\tilde{\varphi}_1; d_1 > 0, d_2 \geq 0\}. \end{aligned}$$

By construction of these cones, one has that $M = \cup_{k=1}^4 M_k$.

Remark 3.3. It is easy to see that if one chooses T_c, T_θ such that $T_c < T_\theta$, then the supports of the trajectories $\alpha^k = \alpha^k(t,x)$ for $k = 1, \dots, 4$ are disjoint.

For each $w = (w_1, w_2) \in \mathbb{R}^2$, let us define

$$\rho_w := \sqrt{w_1^2 + w_2^2} \quad \text{and} \quad z_w := (w_1\varphi_1 + w_2\varphi_2)/\rho_w \in M.$$

We have that $z_w \in M_i$ for some $i \in \{1, \dots, 4\}$, and hence there exist $d_{1w} > 0$ and $d_{2w} \geq 0$ such that $z_w = d_{1w}\tilde{\varphi}_i + d_{2w}\tilde{\varphi}_{i+1}$. If we take the control

$$(u_w, v_w) = (d_{1w}^{1/2}u^i + d_{2w}^{1/2}u^{i+1}, d_{1w}v^i + d_{2w}v^{i+1})$$

and use the fact that the trajectories α^k for $k = 1, \dots, 4$ are disjoints, then we see that the corresponding solution $\beta_w = \beta_w(t,x)$ of (3.2) satisfies $\beta_w(T, \cdot) = z_w$.

Finally, let $\psi \in M$. With $R := \|\psi\|_{L^2(0,L)}$ we can write $\psi = Rz_w$ for a $(w_1, w_2) \in \mathbb{R}^2$ such that $w_1^2 + w_2^2 = 1$. It is easy to see that the control $(u, v) = (R^{1/2}u_w, Rv_w)$ allows us to reach the state ψ, and so the proof of this proposition is ended. □

Remark 3.4. The proof of Proposition 3.2 is the only part which needs a time large enough. Hence, Theorem 1.4 holds for $T_M := \pi/p$.

Remark 3.5. In [8] an expansion to the second order is not sufficient, and the authors must go to the third order to enter into the subspace of missed directions. Since in their case this subspace is one-dimensional and since they use an odd order expansion, one can reach all of the missed states with a scaling argument. Our case is different. We can also enter into the subspace of missed directions in any time, but, in order to reach all of these states, our method needs a time large enough.

It remains to prove the following lemma to complete the proof of Proposition 3.1.

LEMMA 3.6. *The function S is not identically equal to* 0.

Proof. Let $a \in \Sigma$ and $\lambda = 2ai(4a^2 - 1)$. Let $\mu \in \mathbb{C}$, and let $y_\mu = y_\mu(x)$ be a solution of

$$\begin{cases} \mu y_\mu + y'_\mu + y'''_\mu = 0,\\ y_\mu(0) = y_\mu(L) = 0. \end{cases}$$

We multiply (3.13) by y_μ and integrate by parts on $[0, L]$. Thus, we get

$$(\lambda - ip + \mu)\int_0^L \phi_\lambda y_\mu dx - \phi'_\lambda(L)\big(y'_\mu(L) - y'_\mu(0)\big) = \int_0^L y_\lambda \varphi' y_\mu dx. \tag{3.21}$$

From now on, we set $\mu = \mu(a) := -\lambda + ip$. With this choice we obtain from (3.21)

$$-S(a)\big(y'_\mu(L) - y'_\mu(0)\big) = \int_0^L y_\lambda \varphi' y_\mu dx.$$

Therefore, if we prove that the function

$$a \in \Sigma \longrightarrow J(a) := \int_0^L y_\lambda \varphi' y_\mu dx \in \mathbb{C}$$

is not identically equal to 0, the proof of this lemma is ended. Let $b \in \mathbb{R}$ be such that $\mu = 2bi(4b^2 - 1)$. We take the function y_μ given by

$$y_\mu(x) = De^{(-\sqrt{3b^2-1}-bi)x} + (1-D)e^{(\sqrt{3b^2-1}-bi)x} - e^{2bix}, \tag{3.22}$$

where

$$D = \frac{e^{2biL} - e^{(\sqrt{3b^2-1}-bi)L}}{e^{(-\sqrt{3b^2-1}-bi)L} - e^{(\sqrt{3b^2-1}-bi)L}}.$$

In the next computations, we use the fact that $e^{i\gamma_1 L} = e^{i\gamma_2 L} = e^{i\gamma_3 L}$ (see (2.5)) and the following formula:

$$\int_0^L e^{(v+iw)x}\varphi' = \frac{\big(1+\gamma_1^2 - 2p/\gamma_1\big)\big(1 - e^{(v+iw+i\gamma_1)L}\big)(vi-w)}{(vi-w)^3 - (vi-w) + p}, \tag{3.23}$$

which holds if $v + iw \neq -i\gamma_m$ for $m = 1, 2, 3$.

We want to show that as $a \to \infty$, the following expression diverges, which is in contradiction with the fact that $J(a) \equiv 0$:

$$R(a) := \frac{(e^{(-\sqrt{3a^2-1}-ai)L} - e^{(\sqrt{3a^2-1}-ai)L})(e^{(-\sqrt{3b^2-1}-bi)L} - e^{(\sqrt{3b^2-1}-bi)L})}{1+\gamma_1^2 - 2p/\gamma_1} J(a).$$

In fact, by using (3.23), one computes explicitly $J(a)$, and thus one sees that, as a tends to infinity, the dominant term of $R(a)$ is given by

$$\begin{aligned}
Z(a) := e^{(\sqrt{3a^2-1}+\sqrt{3b^2-1})L}\Big\{ & \frac{(e^{(-ai-bi)L} - e^{(ai+bi+\gamma_1 i)L})(-2a-2b)}{(-2a-2b)^3 - (-2a-2b) + p} \\
& + \frac{e^{(-ai-bi)L}(-i\sqrt{3a^2-1} - i\sqrt{3b^2-1} + a + b)}{(-i\sqrt{3a^2-1} - i\sqrt{3b^2-1} + a + b)^3 - (-i\sqrt{3a^2-1} - i\sqrt{3b^2-1} + a + b) + p} \\
& - \frac{e^{(ai+bi+\gamma_1 i)L}(i\sqrt{3a^2-1} + i\sqrt{3b^2-1} + a + b)}{(i\sqrt{3a^2-1} + i\sqrt{3b^2-1} + a + b)^3 - (i\sqrt{3a^2-1} + i\sqrt{3b^2-1} + a + b) + p} \\
& + \frac{e^{(ai+bi+\gamma_1 i)L}(i\sqrt{3a^2-1} + a - 2b)}{(i\sqrt{3a^2-1} + a - 2b)^3 - (i\sqrt{3a^2-1} + a - 2b) + p} \\
& - \frac{e^{(-ai-bi)L}(-i\sqrt{3b^2-1} - 2a + b)}{(-i\sqrt{3b^2-1} - 2a + b)^3 - (-i\sqrt{3b^2-1} - 2a + b) + p} \\
& + \frac{e^{(ai+bi+\gamma_1 i)L}(i\sqrt{3b^2-1} - 2a + b)}{(i\sqrt{3b^2-1} - 2a + b)^3 - (i\sqrt{3b^2-1} - 2a + b) + p} \\
& - \frac{e^{(-ai-bi)L}(-i\sqrt{3a^2-1} + a - 2b)}{(-i\sqrt{3a^2-1} + a - 2b)^3 - (-i\sqrt{3a^2-1} + a - 2b) + p}\Big\}.
\end{aligned}$$

Using that as $a \to \infty$, $b \to -\infty$ and $a+b \sim -p/(24a^2)$, we obtain the following asymptotical expression for the right-hand factor of $Z(a)$:

$$\frac{-(e^{\frac{p}{24a^2}iL} - e^{-\frac{p}{24a^2}iL+i\gamma_1 L})}{12a^2} \sim \begin{cases} -\frac{(1-e^{i\gamma_1 L})}{12a^2} & \text{if } e^{i\gamma_1 L} \neq 1, \\ -\frac{ipL}{144a^4} & \text{if } e^{i\gamma_1 L} = 1. \end{cases}$$

One can see that in both cases $Z(a)$ diverges as $a \to \infty$, and therefore $R(a)$ does, which implies that $J(a)$ is not identically equal to 0. It ends the proof of this lemma. □

4. Proof of Theorem 1.4.

4.1. Existence and uniqueness results. Let us recall the existence property proved by Coron and Crépeau in [8] for the following nonlinear KdV equation:

$$\begin{cases} \partial_t y + \partial_x y + \partial_x^3 y + y\partial_x y = f, \\ y(t,0) = y(t,L) = 0, \\ \partial_x y(t,L) = u(t), \\ y(0,\cdot) = y_0. \end{cases} \tag{4.1}$$

Proposition 4.1 (see [8, Proposition 14]). *Let $L > 0$ and $T > 0$. There exist $\epsilon > 0$ and $C > 0$ such that, for every $f \in L^1(0,T,L^2(0,L))$, $u \in L^2(0,T)$, and $y_0 \in L^2(0,L)$ such that*

$$\|f\|_{L^1(0,T,L^2(0,L))} + \|u\|_{L^2(0,T)} + \|y_0\|_{L^2(0,L)} \leq \epsilon,$$

there exists at least one solution of (4.1) which satisfies

$$\|y\|_{\mathcal{B}} \leq C(\|f\|_{L^1(0,T,L^2(0,L))} + \|u\|_{L^2(0,T)} + \|y_0\|_{L^2(0,L)}). \tag{4.2}$$

For the uniqueness, one has the following.

Proposition 4.2 (see [8, Proposition 15]). *Let $T > 0$, and let $L > 0$. There exists $C > 0$ such that for every $(y_{01}, y_{02}) \in L^2(0,L)^2$, $(u_1, u_2) \in L^2(0,T)^2$, and $(f_1, f_2) \in L^1(0,T,L^2(0,L))^2$ for which there exist solutions $y_1 = y_1(t,x)$ and $y_2 = y_2(t,x)$ of (4.1), one has the following estimates:*

$$\int_0^T \int_0^L |\partial_x y_1(t,x) - \partial_x y_2(t,x)|^2 dxdt \leq e^{C(1+\|y_1\|^2_{L^2(0,T,H^1(0,L))} + \|y_2\|^2_{L^2(0,T,H^1(0,L))})} \cdot \left(\|u_1 - u_2\|^2_{L^2(0,T)} + \|f_1 - f_2\|^2_{L^1(0,T,L^2(0,L))} + \|y_{01} - y_{02}\|^2_{L^2(0,L)} \right),$$

$$\int_0^L |y_1(t,x) - y_2(t,x)|^2 dx \leq e^{C(1+\|y_1\|^2_{L^2(0,T,H^1(0,L))} + \|y_2\|^2_{L^2(0,T,H^1(0,L))})} \cdot \left(\|u_1 - u_2\|^2_{L^2(0,T)} + \|f_1 - f_2\|^2_{L^1(0,T,L^2(0,L))} + \|y_{01} - y_{02}\|^2_{L^2(0,L)} \right)$$

for every $t \in [0,T]$.

4.2. Settings and a technical lemma. Until the end of this paper, we adopt some important notations. Let us denote, for $D > 0$ and $R > 0$,

$$B_R^D := \left\{ \xi \in L^2(0,D) \,;\, \|\xi\|_{L^2(0,D)} \leq R \right\}$$

and recall that for each $w = (w_1, w_2) \in \mathbb{R}^2$, we write $\rho_w := \sqrt{w_1^2 + w_2^2}$ and $z_w := (w_1\varphi_1 + w_2\varphi_2)/\rho_w$. We also write $(u_w, v_w) \in L^2(0,T)$ the controls defined in section 3 in order to drive the solutions $\beta_w = \beta_w(t,x)$ from zero at $t = 0$ to z_w at $t = T$.

By the work of Rosier in [17], we know that for each $y_0 \in L^2(0,L)$ there exists a continuous linear affine map (it is a consequence of applying the HUM method to prove Theorem 2.4)

$$\Gamma_0 : h \in H \subset L^2(0,L) \longmapsto \Gamma_0(h) \in L^2(0,T)$$

such that the solution of

$$\begin{cases} \partial_t y + \partial_x y + \partial_x^3 y = 0, \\ y(t,0) = y(t,L) = 0, \\ \partial_x y(t,L) = \Gamma_0(h), \\ y(0,\cdot) = P_H(y_0) \end{cases}$$

satisfies $y(T,\cdot) = h$. Moreover, there exist constants $D_1, D_2 > 0$ such that

$$\forall y_0 \in L^2(0,L),\ \forall h \in H, \quad \|\Gamma_0(h)\|_{L^2(0,T)} \le D_1(\|h\|_{L^2(0,L)} + \|y_0\|_{L^2(0,L)}), \tag{4.3}$$

$$\forall y_0 \in L^2(0,L),\ \forall h, g \in H, \quad \|\Gamma_0(h) - \Gamma_0(g)\|_{L^2(0,T)} \le D_2\|h - g\|_{L^2(0,L)}. \tag{4.4}$$

Let $y_0 \in L^2(0,L)$ be such that $\|y_0\|_{L^2(0,L)} < r$, $r > 0$ to be chosen later. Let us define the functions G and F

$$\begin{aligned} G : L^2(0,L) &\longrightarrow L^2(0,T), \\ z = P_H(z) + w_1\varphi_1 + w_2\varphi_2 &\mapsto G(z) = \Gamma_0(P_H(z)) + \rho_w^{1/2}u_w + \rho_w v_w, \end{aligned}$$

$$\begin{aligned} F : B_{\epsilon_1}^T \cap L^2(0,T) &\longrightarrow L^2(0,L), \\ u &\longmapsto F(u) = y(T,\cdot), \end{aligned}$$

where $y = y(t,x)$ is the solution of

$$\begin{cases} \partial_t y + \partial_x y + \partial_x^3 y + y\partial_x y = 0, \\ y(t,0) = y(t,L) = 0, \\ \partial_x y(t,L) = u(t), \\ y(0,\cdot) = y_0, \end{cases} \tag{4.5}$$

and ϵ_1 is small enough so that the function F is well defined. It holds if $\epsilon_1 + r < \epsilon$, where ϵ is given by Proposition 4.1. The map F is even continuous according to Proposition 4.2. Let $y_T \in L^2(0,L)$ be such that $\|y_T\| < r$. Let Λ_{y_0,y_T} denote the map

$$\begin{aligned} \Lambda_{y_0,y_T} : B_{\epsilon_2}^L \cap L^2(0,L) &\longrightarrow L^2(0,L), \\ z &\longmapsto \Lambda_{y_0,y_T}(z) = z + y_T - F \circ G(z), \end{aligned}$$

where ϵ_2 is small enough so that Λ_{y_0,y_T} is well defined (ϵ_2 exists according to Proposition 4.1 and to the continuity of G).

Let us notice that if we find a fixed point $\tilde{z} \in L^2(0,L)$ of the map Λ_{y_0,y_T}, then we will have $F \circ G(\tilde{z}) = y_T$, and this means that the control $u := G(\tilde{z}) \in L^2(0,T)$ drives the solution of (4.5) from y_0 at $t = 0$ to y_T at $t = T$.

Let us assert the following technical result which will be needed to study the map Λ_{y_0,y_T}.

LEMMA 4.3. *There exist $\epsilon_3 > 0$ and $C_3 > 0$ such that for every $z, y_0 \in B^L_{\epsilon_3}$ the following estimate holds:*

$$\|z - F(G(z))\|_{L^2(0,L)} \leq C_3(\|y_0\|_{L^2(0,L)} + \|z\|^{3/2}_{L^2(0,L)}).$$

Proof. Let $z, y_0 \in L^2(0,L)$. Let $w = (w_1, w_2) \in \mathbb{R}^2$ be such that $z = P_H(z) + \rho_w z_w$. Let $y = y(t,x)$ be a solution of

$$\begin{cases} \partial_t y + \partial_x y + \partial_x^3 y + y\partial_x y = 0, \\ y(t,0) = y(t,L) = 0, \\ \partial_x y(t,L) = G(z), \\ y(0,\cdot) = y_0. \end{cases} \tag{4.6}$$

From (4.3) and since $\rho_w \leq \|z\|_{L^2(0,L)}$, one deduces that if $\|z\|_{L^2(0,L)}$ is smaller than 1 (and therefore $\|z\|_{L^2(0,L)} \leq \|z\|^{1/2}_{L^2(0,L)}$), then there exists a constant D_3 such that

$$\|G(z)\|_{L^2(0,T)} \leq D_3(\|y_0\|_{L^2(0,L)} + \|z\|^{1/2}_{L^2(0,L)}). \tag{4.7}$$

Remark 4.4. Let us notice that the controls u_w, v_w in the definition of the map G drive the solution β_w from the origin at $t = 0$ to the state z_w at $t = T$, with $\|z_w\|_{L^2(0,L)} = 1$, and therefore they are uniformly bounded.

By using (4.2) and (4.7), one can find $\epsilon_2, C_2 > 0$ such that for every $z, y_0 \in B^L_{\epsilon_2}$ the unique solution of (4.6) satisfies

$$\|y\|_{\mathcal{B}} \leq C_2(\|y_0\|_{L^2(0,L)} + \|z\|^{1/2}_{L^2(0,L)}). \tag{4.8}$$

Let $\tilde{y} = \tilde{y}(t,x)$, $\alpha_w = \alpha_w(t,x)$, $\beta_w = \beta_w(t,x)$, and $\beta^0 = \beta^0(t,x)$ be the solutions of

$$\begin{cases} \partial_t \tilde{y} + \partial_x \tilde{y} + \partial_x^3 \tilde{y} = 0, \\ \tilde{y}(t,0) = \tilde{y}(t,L) = 0, \\ \partial_x \tilde{y}(t,L) = \Gamma_0(P_H(z)), \\ \tilde{y}(0,\cdot) = P_H(y_0), \end{cases}$$

$$\begin{cases} \partial_t \alpha_w + \partial_x \alpha_w + \partial_x^3 \alpha_w = 0, \\ \alpha_w(t,0) = \alpha_w(t,L) = 0, \\ \partial_x \alpha_w(t,L) = u_w(t), \\ \alpha_w(0,\cdot) = 0, \end{cases}$$

$$\begin{cases} \partial_t \beta_w + \partial_x \beta_w + \partial_x^3 \beta_w = -\alpha_w \partial_x \alpha_w, \\ \beta_w(t,0) = \beta_w(t,L) = 0, \\ \partial_x \beta_w(t,L) = v_w(t), \\ \beta_w(0,\cdot) = 0, \end{cases}$$

$$\begin{cases} \partial_t \beta^0 + \partial_x \beta^0 + \partial_x^3 \beta^0 = 0, \\ \beta^0(t,0) = \beta^0(t,L) = 0, \\ \partial_x \beta^0(t,L) = 0, \\ \beta^0(0,\cdot) = P_M(y_0), \end{cases}$$

respectively. Let us define

$$\phi := y - \tilde{y} - \rho_w^{1/2}\alpha_w - \rho_w \beta_w - \beta^0.$$

We have that $\phi = \phi(t,x)$ satisfies

$$\begin{cases} \partial_t\phi + \partial_x\phi + \partial_x^3\phi + \phi\partial_x\phi = -\partial_x(\phi a) - b, \\ \phi(t,0) = \phi(t,L) = 0, \\ \partial_x\phi(t,L) = 0, \\ \phi(0) = 0, \end{cases}$$

where

$$\begin{aligned} a &:= \tilde{y} + \rho_w^{1/2}\alpha_w + \rho_w\beta_w + \beta^0, \\ b &:= \tilde{y}\partial_x\tilde{y} + \partial_x(\tilde{y}(\rho_w^{1/2}\alpha_w + \rho_w\beta_w + \beta^0)) + \rho_w^{3/2}\partial_x(\alpha_w\beta_w) \\ &\quad + \rho_w^2\beta_w\partial_x(\beta_w) + \rho_w^{1/2}\partial_x(\alpha_w\beta^0) + \rho_w\partial_x(\beta_w\beta^0) + \beta^0\partial_x\beta^0. \end{aligned}$$

It is easy to see that there exists $C_4 > 0$ such that for every $z, y_0 \in B^L_{\epsilon_2}$

$$\|\phi\|_{\mathcal{B}} \le C_4(\|y_0\|_{L^2(0,L)} + \|z\|^{1/2}_{L^2(0,L)}), \tag{4.9}$$

$$\|a\|_{\mathcal{B}} \le C_4(\|y_0\|_{L^2(0,L)} + \|z\|^{1/2}_{L^2(0,L)}), \tag{4.10}$$

$$\|b\|_{L^1(0,T,L^2(0,L))} \le C_4(\|y_0\|_{L^2(0,L)} + \|z\|^{3/2}_{L^2(0,L)}). \tag{4.11}$$

One can also prove that there exists $C_5 > 0$ such that for every $f, g \in \mathcal{B}$

$$\|\partial_x(fg)\|_{L^1(0,T,L^2(0,L))} \le C_5\|f\|_{\mathcal{B}}\|g\|_{\mathcal{B}}. \tag{4.12}$$

By (4.2), (4.11), and (4.12), there exists $C_6 > 0$ such that

$$\|\phi\|^2_{\mathcal{B}} \le C_6(\|\phi\|^2_{\mathcal{B}}\|a\|^2_{\mathcal{B}} + \|y_0\|^2_{L^2(0,L)} + \|z\|^3_{L^2(0,L)});$$

i.e., one has

$$\|\phi\|^2_{\mathcal{B}}(1 - C_6\|a\|^2_{\mathcal{B}}) \le C_6(\|y_0\|^2_{L^2(0,L)} + \|z\|^3_{L^2(0,L)}),$$

which, together with (4.10), implies the existence of ϵ_3 and C_7 such that for every $z, y_0 \in B^L_{\epsilon_3}$

$$\|\phi\|_{\mathcal{B}} \le C_7(\|y_0\|_{L^2(0,L)} + \|z\|^{3/2}_{L^2(0,L)}). \tag{4.13}$$

Finally, from (4.13) and using that $\|\beta^0(0)\|_{L^2(0,L)} = \|\beta^0(T)\|_{L^2(0,L)}$ (β^0 turns only in the subspace M), one obtains with $C_3 := C_7 + 1$

$$\begin{aligned} \|z - F\circ G(z)\|_{L^2(0,L)} &\le \|z - F\circ G(z) + \beta^0(T)\|_{L^2(0,L)} + \|\beta^0(T)\|_{L^2(0,L)} \\ &= \|\phi(T)\|_{L^2(0,L)} + \|\beta^0(0)\|_{L^2(0,L)} \\ &\le \|\phi\|_{\mathcal{B}} + \|y_0\|_{L^2(0,L)} \\ &\le C_7(\|y_0\|_{L^2(0,L)} + \|z\|^{3/2}_{L^2(0,L)}) + \|y_0\|_{L^2(0,L)} \\ &\le C_3(\|y_0\|_{L^2(0,L)} + \|z\|^{3/2}_{L^2(0,L)}), \end{aligned}$$

which ends the proof of Lemma 4.3. ∎

4.3. Fixed point in the subspace H. For $w = (w_1, w_2) \in \mathbb{R}^2$ fixed, let us study the map $\Pi := P_H \circ \Lambda_{y_0,y_T}(\cdot + \rho_w z_w)$ on the subspace H (recall that $\rho_w z_w = w_1\varphi_1 + w_2\varphi_2$):

$$\begin{array}{rcl} \Pi : H & \longrightarrow & H, \\ h & \longmapsto & \Pi(h) = h + P_H(y_T) - P_H(F \circ G(h + \rho_w z_w)). \end{array}$$

In order to prove the existence of a fixed point of the map Π, we will apply the Banach fixed point theorem to the restriction of Π to the closed ball $B_R^L \cap H$, with $R > 0$ small enough. By using Lemma 4.3 we see that

$$\begin{array}{rcl} \|\Pi(h)\|_{L^2(0,L)} & \leq & \|y_T\|_{L^2(0,L)} + \|h + \rho_w z_w - F \circ G(h + \rho_w z_w)\|_{L^2(0,L)} \\ & \leq & \|y_T\|_{L^2(0,L)} + C_3(\|y_0\|_{L^2(0,L)} + \|h + \rho_w z_w\|^{3/2}_{L^2(0,L)}) \\ & \leq & C_3'(\|y_0\|_{L^2(0,L)} + \|y_T\|_{L^2(0,L)} + \rho_w^{3/2}) + C_3\|h\|^{3/2}_{L^2(0,L)} \\ & \leq & C_3'(2r + \rho_w^{3/2}) + C_3\|h\|^{3/2}_{L^2(0,L)}, \end{array}$$

where $C_3' := C_3 + 1$. Hence, if we choose R such that $R^{3/2} \leq \frac{R}{2C_3}$ and r, ρ_w such that

$$C_3'(2r + \rho_w^{3/2}) \leq \frac{R}{2},$$

then it follows that

$$\|\Pi(h)\|_{L^2(0,L)} \leq R \quad \text{and so} \quad \Pi(B_R^L \cap H) \subset (B_R^L \cap H).$$

It remains to prove that the map Π is a contraction. Let $g, h \in B_R^L \cap H$. Let $y = y(t,x)$, $q = q(t,x)$, $\tilde{y} = \tilde{y}(t,x)$, and $\tilde{q} = \tilde{q}(t,x)$ be the solutions of the following problems:

$$\left\{ \begin{array}{l} \partial_t y + \partial_x y + \partial_x^3 y + y\partial_x y = 0, \\ y(t,0) = y(t,L) = 0, \\ \partial_x y(t,L) = G(g + \rho_w z_w), \\ y(0,\cdot) = y_0, \end{array} \right.$$

$$\left\{ \begin{array}{l} \partial_t q + \partial_x q + \partial_x^3 q + q\partial_x q = 0, \\ q(t,0) = q(t,L) = 0, \\ \partial_x q(t,L) = G(h + \rho_w z_w), \\ q(0,\cdot) = y_0, \end{array} \right.$$

$$\left\{ \begin{array}{l} \partial_t \tilde{y} + \partial_x \tilde{y} + \partial_x^3 \tilde{y} = 0, \\ \tilde{y}(t,0) = \tilde{y}(t,L) = 0, \\ \partial_x \tilde{y}(t,L) = \Gamma_0(g), \\ \tilde{y}(0,\cdot) = P_H(y_0), \end{array} \right.$$

$$\left\{ \begin{array}{l} \partial_t \tilde{q} + \partial_x \tilde{q} + \partial_x^3 \tilde{q} = 0, \\ \tilde{q}(t,0) = \tilde{q}(t,L) = 0, \\ \partial_x \tilde{q}(t,L) = \Gamma_0(h), \\ \tilde{q}(0,\cdot) = P_H(y_0), \end{array} \right.$$

repsectively. Let us define $\phi := y - \tilde{y}$, $\psi := q - \tilde{q}$, and $\gamma := \phi - \psi$. One sees that γ satisfies

$$\left\{ \begin{array}{l} \partial_t \gamma + \partial_x \gamma + \partial_x^3 \gamma + \gamma\partial_x \gamma = -\partial_x(\gamma a) - b, \\ \gamma(t,0) = \gamma(t,L) = 0, \\ \partial_x \gamma(t,L) = 0, \\ \gamma(0) = 0, \end{array} \right. \tag{4.14}$$

where

$$a := \tilde{y} + \psi \quad \text{and} \quad b := \partial_x \left(q(\tilde{y} - \tilde{q})\right) + (\tilde{y} - \tilde{q})\partial_x(\tilde{y} - \tilde{q}).$$

It is easy to see that there exists a constant $C_8 > 0$ such that

$$\|b\|_{L^1(0,T,L^2(0,L))} \leq C_8 \left(\|q\|_{\mathcal{B}} + \|\tilde{y}\|_{\mathcal{B}} + \|\tilde{q}\|_{\mathcal{B}}\right) \|\tilde{y} - \tilde{q}\|_{\mathcal{B}}, \tag{4.15}$$

$$\|\partial_x(a\gamma)\|_{L^1(0,T,L^2(0,L))} \leq C_8 \left(\|q\|_{\mathcal{B}} + \|\tilde{y}\|_{\mathcal{B}} + \|\tilde{q}\|_{\mathcal{B}}\right) \|\gamma\|_{\mathcal{B}}. \tag{4.16}$$

By using Proposition 4.2, (4.15), and (4.16) we get the existence of $C_9 > 0$ such that

$$\|\gamma\|_{\mathcal{B}}^2 \leq C_9(\|q\|_{\mathcal{B}} + \|\tilde{y}\|_{\mathcal{B}} + \|\tilde{q}\|_{\mathcal{B}})^2(\|\tilde{y} - \tilde{q}\|_{\mathcal{B}}^2 + \|\gamma\|_{\mathcal{B}}^2). \tag{4.17}$$

In addition, since $w := \tilde{y} - \tilde{q}$ satisfies the following linear equation:

$$\begin{cases} \partial_t w + \partial_x w + \partial_x^3 w = 0, \\ w(t,0) = w(t,L) = 0, \\ \partial_x w(t,L) = \Gamma_0(g) - \Gamma_0(h), \\ w(0,\cdot) = 0, \end{cases}$$

there exists $C_{10} > 0$ such that

$$\|\tilde{y} - \tilde{q}\|_{\mathcal{B}} \leq C_{10}\|\Gamma_0(g) - \Gamma_0(h)\|_{L^2(0,T)},$$

and so, from (4.4), one gets

$$\|\tilde{y} - \tilde{q}\|_{\mathcal{B}} \leq C_{10}D_2\|g - h\|_{L^2(0,L)}. \tag{4.18}$$

Moreover, it is easy to see that there exists a constant $C_{11} > 0$ such that

$$\|q\|_{\mathcal{B}} + \|\tilde{q}\|_{\mathcal{B}} + \|\tilde{y}\|_{\mathcal{B}} \leq C_{11}(\|y_0\|_{L^2(0,L)} + \|h\|_{L^2(0,L)} + \|g\|_{L^2(0,L)} + \rho_w^{1/2}). \tag{4.19}$$

Thus, using (4.17)–(4.19) we see that if R, ρ_w, r are small enough, it follows that

$$\|\gamma\|_{\mathcal{B}} \leq \tfrac{1}{2}\|g - h\|_{L^2(0,L)}.$$

Therefore, we have

$$\begin{array}{rcl} \|\Pi(g) - \Pi(h)\|_{L^2(0,L)} & \leq & \|g - F \circ G(g + \rho_w z_w) - h + F \circ G(h + \rho_w z_w)\|_{L^2(0,L)} \\ & = & \|\gamma(T)\|_{L^2(0,L)} \leq \|\gamma\|_{\mathcal{B}} \\ & \leq & \frac{1}{2}\|g - h\|_{L^2(0,L)}, \end{array}$$

which implies the existence of a unique fixed point $h(y_0, y_T, w_1, w_2) \in B_R^L \cap H$ of the map $\Pi|_{B_R^L \cap H}$. Moreover, the more precise proposition follows.

PROPOSITION 4.5. *There exist $R_0 > 0$, $D > 1$ such that for every $0 < R < R_0$, for every $y_0, y_T \in B_{R/D}^L$, $(w_1, w_2) \in \mathbb{R}^2$, with $\rho_w < R/D$, there exists a unique $h(y_0, y_T, w_1, w_2) \in B_R^L \cap H$ fixed point of the map $\Pi|_{B_R^L \cap H}$.*

4.4. Fixed point in the subspace M. We now apply the Brouwer fixed point theorem to the restriction of the map

$$\begin{array}{rcl} \tau : M & \longrightarrow & M, \\ w_1\varphi_1 + w_2\varphi_2 & \to & P_M(\rho_w z_w + y_T - F \circ G(\rho_w z_w + h(y_0, y_T, w_1, w_2))) \end{array}$$

to the closed ball $B^L_{\hat{R}} \cap M$, with $\hat{R}$ small enough. Using Lemma 4.3, the continuity (in a neighborhood of $0 \in (L^2(0,L))^2 \times \mathbb{R}^2$) of the map $(y_0, y_T, w_1, w_2) \longmapsto h(y_0, y_T, w_1, w_2)$ and choosing r small enough, we get the existence of a radius $\hat{R} > 0$ such that $\tau(B^L_{\hat{R}} \cap M) \subset B^L_{\hat{R}} \cap M$. This inclusion and the continuity of the map τ allow us to apply the Brouwer fixed point theorem. Therefore, there exists $(\tilde{w}_1, \tilde{w}_2) \in \mathbb{R}^2$, with $\tilde{w}_1^2 + \tilde{w}_2^2 \leq \hat{R}^2$, such that $\tilde{h} := h(y_0, y_T, \tilde{w}_1, \tilde{w}_2)$ satisfies

$$P_M(y_T - F \circ G(\tilde{h} + \tilde{w}_1\varphi_1 + \tilde{w}_2\varphi_2)) = 0. \tag{4.20}$$

Using the fact that

$$\Pi(\tilde{h}) = P_H(\tilde{h} + y_T - F \circ G(\tilde{h} + \tilde{w}_1\varphi_1 + \tilde{w}_2\varphi_2)) = \tilde{h},$$

we obtain

$$P_H(y_T - F \circ G(\tilde{h} + \tilde{w}_1\varphi_1 + \tilde{w}_2\varphi_2)) = 0,$$

which together with (4.20) implies that

$$y_T = F \circ G(\tilde{h} + \tilde{w}_1\varphi_1 + \tilde{w}_2\varphi_2),$$

which ends the proof of Theorem 1.4. Let us remark that from our proof it follows that if r is chosen small enough, one can take $\hat{R} := rD$, where $D > 0$ is given by Proposition 4.5. By using this proposition one obtains the estimate (1.6).

Acknowledgments. I thank Jean-Michel Coron for having attracted my attention to this problem, for his constant support, and for fruitful discussions. I also thank Emmanuelle Crépeau for interesting remarks.

REFERENCES

[1] K. Beauchard, *Local controllability of a* 1D *Schrödinger equation*, J. Math. Pures Appl., 84 (2005), pp. 851–956.

[2] K. Beauchard and J.-M. Coron, *Controllability of a quantum particle in a moving potential well*, J. Funct. Anal., 232 (2006), pp. 328–389.

[3] J.-M. Coron, *Global asymptotic stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.

[4] J.-M. Coron, *On the controllability of* 2-D *incompressible perfect fluids*, J. Math. Pures Appl., 75 (1996), pp. 155–188.

[5] J.-M. Coron, *Local controllability of a* 1-D *tank containing a fluid modeled by the shallow water equations*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 513–554.

[6] J.-M. Coron, *On the small time local controllability of a quantum particle in a moving one-dimensional infinite square potential well*, C. R. Math. Acad. Sci. Paris, 342 (2006), pp. 103–108.

[7] J.-M. Coron, *Control and nonlinearity*, in Math. Surveys Monogr., American Mathematical Society, Providence, RI, 2007.

[8] J.-M. Coron and E. Crépeau, *Exact boundary controllability of a nonlinear KdV equation with critical lengths*, J. Eur. Math. Soc., 6 (2004), pp. 367–398.

[9] J.-M. Coron and E. Trélat, *Global steady-state controllability of one-dimensional semilinear heat equations*, SIAM J. Control Optim., 43 (2004), pp. 549–569.

[10] E. Crépeau, *Exact controllability of the Korteweg-de Vries equation around a non-trivial stationary solution*, Internat. J. Control, 74 (2001), pp. 1096–1106.

[11] E. Crépeau, *Contrôlabilité exacte d'équations dispersives issues de la mécanique*, Ph.D. thesis, Laboratoire de Mathématiques, Université Paris-Sud, Paris, 2002.

[12] V. Komornik, *Rapid boundary stabilization of linear distributed systems*, SIAM J. Control Optim., 35 (1997), pp. 1591–1613.

[13] V. Komornik, D. L. Russell, and B.-Y. Zhang, *Stabilisation de l'équation Korteweg-de Vries*, C. R. Math. Acad. Sci. Paris, 312 (1991), pp. 841–843.
[14] D. J. Korteweg and G. de Vries, *On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves*, Philos. Mag., 39 (1895), pp. 422–443.
[15] J. L. Lions, *Contrôlabilité exacte, perurbations et stabilization de systèmes distribués*, Masson, Paris, 1988.
[16] G. Perla Menzala, C. F. Vasconcellos, and E. Zuazua, *Stabilization of the Korteweg-de Vries equation with localized damping*, Quart. Appl. Math., Vol. LX, No. 1 (2002), pp. 111–129.
[17] L. Rosier, *Exact boundary controllability for the Korteweg-de Vries equation on a bounded domain*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 33–55.
[18] L. Rosier and B.-Y Zhang, *Global Stabilization of the generalized Korteweg-de Vries equation posed on a finite domain*, SIAM J. Control Optim., 45 (2006), pp. 927–956.
[19] D. L. Russell and B.-Y Zhang, *Smoothing and decay properties of the Korteweg-de Vries equation on a periodic domain with point dissipation*, J. Math. Anal. Appl., 190 (1995), pp. 449–488.
[20] D. L. Russell and B.-Y Zhang, *Exact controllability and stabilizability of the Korteweg-de Vries equation*, Trans. Amer. Math. Soc., 348 (1996), pp. 3643–3672.
[21] J.-P. Serre, *A Course in Arithmetic*, Springer-Verlag, New York, 1973.
[22] J. M. Urquiza, *Rapid exponential feedback stabilization with unbounded control operators*, SIAM J. Control Optim., 43 (2005), pp. 2233–2244.
[23] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.
[24] B.-Y. Zhang, *Exact boundary controllability of the Korteweg-de Vries equation*, SIAM J. Control Optim., 37 (1999), pp. 543–565.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 900–929

SWITCHING GAMES OF STOCHASTIC DIFFERENTIAL SYSTEMS*

SHANJIAN TANG† AND SHUI-HUNG HOU‡

Abstract. A two-player, zero-sum, switching game is formulated for general stochastic differential systems and is studied using a combined dynamic programming and viscosity solution approach. The existence of the game value is proved. For the proof of the related dynamic programming principle (DDP) for the lower and upper value functions, the measurability problem, of the same kind as mentioned in the paper of Fleming and Souganidis, is also encountered, and we are able to get around it via a delicate adaptation of their technique. Moreover, the traditional direct method to prove the time continuity of lower and upper value functions also gives rise to a serious measurability problem. To get around the new difficulty, a subtle dynamic programming argument is developed to obtain the time continuity, which in return is used to derive the DDP for random intermediate times from the DDP with deterministic intermediate times.

Key words. stochastic differential games, dynamic programming inequalities, switching strategies, value function, viscosity solution

AMS subject classifications. 49N70, 49L25, 60H30, 49L20, 90C39, 93E20

DOI. 10.1137/050642204

1. Introduction.

Consider a differential game of the following stochastic differential system on Wiener space $(\Omega, \mathcal{F}, P)$:

$$\begin{cases} dy(t) = f(t, y(t), a(t), b(t))\,dt + g(t, y(t), a(t), b(t))\,dw(t), \quad t \in (s, 1], \\ y(s) = x \end{cases} \tag{1.1}$$

with the cost functional

$$\begin{aligned} J_{s,x}(a(\cdot), b(\cdot)) = E_{sx}\Bigg\{ \int_s^1 f^0(t, y(t), a(t), b(t))\,dt + h(y(1)) \\ + \sum_{i \geq 1} k(\theta_i, a_{i-1}, a_i) - \sum_{j \geq 1} l(\tau_j, b_{j-1}, b_j) \Bigg\}. \end{aligned} \tag{1.2}$$

Here f, g, f^0, and h are given maps; $w(\cdot)$ is the coordinate process in Ω, and its natural filtration is denoted by $\mathcal{F}_t$. The subscript sx of the expectation operator E indicates that the underlying mathematical expectation is taken under the condition that the underlying system state process $y(\cdot)$ takes the value x at time s. The first player chooses the control a from a given finite set A to minimize the payoff (1.2), and each

*Received by the editors October 9, 2005; accepted for publication (in revised form) March 21, 2007; published electronically June 12, 2007. This work was partially supported by the Natural Science Foundation of China under grant 10325101 (Distinguished Youth Foundation, entitled with "Control Theory of Stochastic Systems"), by the Chang Jiang Scholars Program, by the Science Foundation of Chinese Ministry of Education under grant 20030246004 (entitled with "Optimal Control Theory and Their Applications for Stochastic Systems"), and by the Croucher Foundation.
http://www.siam.org/journals/sicon/46-3/64220.html

†Department of Finance and Control Sciences, School of Mathematical Sciences, Fudan University, and Key Laboratory of Mathematics for Nonlinear Sciences (Fudan University), Ministry of Education, Shanghai 200433, China (sjtang@fudan.edu.cn).

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong (mahough@polyu.edu.hk).

of his instantaneous actions is related with one positive cost k, while the second player chooses the control b from a given finite set B to maximize the payoff (1.2), and each of his instantaneous actions is associated with the other positive cost l.

For differential switching games, a key point in connecting value functions with the corresponding Isaacs' equations is to prove the following fact: It is the best way for a player to keep his underlying switching position for some time of a positive length, whenever he is not on his switching set. In the deterministic case, such an assertion is easy to understand from the following almost obvious fact: If he is not on his switching set, a player will keep away from the set for some deterministic time interval of a positive length, as the system state evolves continuously. See Yong [10] for details. In the stochastic case, the situation becomes complicated due to the nature of diffusion: Even if he is not on his switching set, it is possible for a player to arrive at his switching set in an arbitrarily short time. That is, if he is not on his switching set, although the system state still evolves continuously, a player can keep away from the switching set only for some *random* (rather than deterministic, in general) time interval, *almost surely* (rather than uniformly, in general) of a positive length. Then the intuition of the dynamic programming principle for the underlying switching game suggests that if he is not on his switching set, the optimal action of a player has to vary with different events, even within a very short deterministic time period. We show in section 3 by using arguments quite different from the deterministic case that, whenever he is not on his switching set, a player's best action is to keep his underlying switching position, *before he escapes from a sufficiently small ball centered at the current state*, within some deterministic time interval of a positive length.

It has been widely recognized that the dynamic programming method is both easy and efficient for the study of deterministic optimal control and differential games within the framework of viscosity solutions. The general nonsmooth feature of inf-sup functions is no longer a difficulty in view of the notion of viscosity solutions. However, applications of dynamic programming to optimal stochastic controls and stochastic differential games still encounter difficulties; the reader is referred to Bertsekas and Shreve [1] and Fleming and Souganidis [6] for detail. It was noticed by Fleming and Souganidis [6], in the study of classical stochastic differential games, that the conventional proof of the dynamic programming principle for the lower and upper value functions encounters a serious measurability issue. In this paper, we observe that the traditional direct approach to show the time continuity of lower and upper value functions also gives rise to a serious measurability problem. The difficulty is circulated using a dynamic programming argument.

In this paper, the coefficients of differential games are allowed to grow linearly, and a powerful simple test function is given to prove the uniqueness of unbounded viscosity solutions for the associated Isaacs' system of variational inequalities.

The rest of our paper is organized as follows. Section 2 is devoted to the formulation of our stochastic switching game, the definitions of some restrictive class of admissible controls, and strategies to be used in the following sections. Several useful dynamic programming results and the time continuous properties are established in section 3. The existence of the value is proved for our game in section 4.

There are some related papers which remain to be mentioned. For the optimal switching problem, the reader is referred to Capuzzo Dolcetta and Evans in [2] in the deterministic case, and to Evans and Friedman [4], Tang and Yong [7] and the references therein in the stochastic case. For the switching game, the reader is referred to Yong [10] in the deterministic case with the dynamic programming approach and the notion of viscosity solution, and to Yamada [8, 9] in the stochastic and infinite

time-horizon case with an analytical approach rather than the dynamic programming approach.

2. Preliminaries. Let $A = \{1, 2, \dots, m\}, B = \{1, 2, \dots, n\}$, and X be a finite-dimensional Euclidean space. Let $f : [0,1] \times X \times A \times B \to X, g : [0,1] \times X \times A \times B \to X \times W, f^0 : [0,1] \times X \times A \times B \to \mathbb{R}, k : [0,1] \times A \times A \to \mathbb{R}_+ \equiv [0, \infty)$, and $l : [0,1] \times B \times B \to \mathbb{R}_+$ be continuous functions satisfying the following hypotheses.

Hypothesis 1. There exists a constant $L > 0$ such that for all $x, \widehat{x} \in X, t \in [0,1], a \in A$, and $b \in B$,

$$\begin{aligned}
&|f(t,x,a,b) - f(t,\widehat{x},a,b)| + |g(t,x,a,b) - g(t,\widehat{x},a,b)| \le L|x - \widehat{x}|,\\
&|f(t,x,a,b)| + |g(t,x,a,b)| \le L(1 + |x|),\\
&|f^0(t,x,a,b) - f^0(t,\widehat{x},a,b)| + |h(x) - h(\widehat{x})| \le L|x - \widehat{x}|,\\
&|f^0(t,0,a,b)| + |h(0)| \le L.
\end{aligned}$$

Hypothesis 2. For all $a, \widehat{a}, \widetilde{a} \in A, a \ne \widehat{a} \ne \widetilde{a}$, and $0 \le s \le t \le 1$,

$$\begin{aligned}
&k(t,a,\widetilde{a}) < k(t,a,\widehat{a}) + k(t,\widehat{a},\widetilde{a}),\\
&k(t,a,\widehat{a}) > 0, k(t,a,a) = 0,\\
&k(t,a,\widetilde{a}) \le k(s,a,\widetilde{a}).
\end{aligned}$$

Hypothesis 3. For all $b, \widehat{b}, \widetilde{b} \in B, b \ne \widehat{b} \ne \widetilde{b}$, and $0 \le s \le t \le 1$,

$$\begin{aligned}
&l(t,b,\widetilde{b}) < l(t,b,\widehat{b}) + l(t,\widehat{b},\widetilde{b}),\\
&l(t,b,\widehat{b}) > 0, l(t,b,b) = 0,\\
&l(t,b,\widetilde{b}) \le l(s,b,\widetilde{b}).
\end{aligned}$$

For $s, \hat{s} \in [0,1]$ such that $s < \hat{s}$, let

$$\Omega_{s,\hat{s}} = \{\omega \in C([s,\hat{s}]; \mathbb{R}^d) : \omega(s) = 0\}. \tag{2.1}$$

Denote by $\mathcal{F}_{s,\hat{s}}$ the topological σ-field of $\Omega_{s,\hat{s}}$ and consider the Wiener space $(\Omega_{s,\hat{s}}, \mathcal{F}_{s,\hat{s}}, P_{s,\hat{s}})$. Let

$$\Omega_s = \Omega_{s,1}, \quad P_s = P_{s,1}, \quad \mathcal{F}_s = \mathcal{F}_{s,1}, \tag{2.2}$$

and

$$\begin{cases}
\omega_1 = \omega|_{[s,\hat{s}]},\\
\omega_2 = (\omega - \omega(\hat{s}))|_{[\hat{s},1]},\\
\Pi\omega = (\omega_1, \omega_2).
\end{cases} \tag{2.3}$$

We see that the map $\Pi : \Omega_s \to \Omega_{s,\hat{s}} \times \Omega_{\hat{s}}$ induces an identification

$$\Omega_s = \Omega_{s,\hat{s}} \times \Omega_{\hat{s}}. \tag{2.4}$$

Moreover, the inverse of Π is defined in an evident way: $\Omega_s = \Pi^{-1}(\Omega_{s,\hat{s}}, \Omega_{\hat{s}})$. Finally, we have

$$P_s = P_{s,\hat{s}} \otimes P_{\hat{s}}.$$

Define

$$w(r,\omega)=\omega(r),\quad (\omega,r)\in\Omega_s\times[s,1]. \tag{2.5}$$

Then $\{w(r), r\in[s,1]\}$ is a standard Wiener process.

DEFINITION 2.1. *An admissible switching process for player* I *(resp.,* II*) on* $[s,1]$ *with initial value* a_0 *(resp.,* b_0*) is defined to be a pair of sequences* $\{a_i,\theta_i\}_{i\geq 0}$ *(resp.,* $\{b_i,\tau_i\}_{i\geq 0}$*) such that each* θ_i *(resp.,* τ_i*) is an* $\mathcal{F}_{s,\cdot}$*-stopping time, with*

$$s=\theta_0\leq\theta_1\leq\cdots\leq 1\quad \text{a.s.}$$
$$(\text{resp., } s=\tau_0\leq\tau_1\leq\cdots\leq 1\quad \text{a.s.}),$$

each a_i *(resp.,* b_i*) is* $\mathcal{F}_{s,\theta_i}$*- (resp.,* $\mathcal{F}_{s,\tau_i}$*-) measurable with values in* A *(resp.,* B*), and*

$$E\sum_{i\geq 1}k(\theta_i,a_{i-1},a_i)<\infty\quad\left(\text{resp., } E\sum_{j\geq 1}l(\tau_j,b_{j-1},b_j)<\infty\right).$$

Denote by $\mathcal{A}^a[s,\hat{s}]$ *(resp.,* $\mathcal{B}^b[s,\hat{s}]$*) the totality of the admissible switchings for player* I *(resp.,* II*) on* $[s,\hat{s}]$ *with the initial value* a *(resp.,* b*).*

We shall identify $\{a_i,\theta_i\}_{i\geq 0}\in\mathcal{A}^a[s,1]$ with

$$a(r)=\sum_{i\geq 1}a_{i-1}\chi_{[\theta_{i-1},\theta_i)}(r),\quad r\in[s,1].$$

Note that in the case of $\theta_1=\theta_2$ the term $a_1\chi_{[\theta_1,\theta_2)}(r)$ will be void, but we still keep it in the above expression. This is due to the fact that the sequence $\{a_i,\theta_i\}$ with or without (a_1,θ_1) represents two different switching controls and their costs are different. A similar identification will also be used for $\{b_i,\tau_i\}\in\mathcal{B}^b[t,1]$.

Following Elliott and Kalton [3] and Fleming and Souganidis [6], we define in the switching game an admissible strategy as follows.

DEFINITION 2.2. *For* $s\in[0,1]$ *and* $a\in A$ *(resp.,* $b\in B$*), an admissible strategy* $\alpha^{a,s}$ *(resp.,* $\beta^{b,s}$*) with the initial value* a *(resp.,* b*) for player* I *(resp.,* II*) on* $[s,1]$ *is a mapping* $\alpha^{a,s}:\cup_{b\in B}\mathcal{B}^b[s,1]\to\mathcal{A}^a[s,1]$ *(resp.,* $\beta^{b,s}:\cup_{a\in A}\mathcal{A}^a[s,1]\to\mathcal{B}^b[s,1]$*) such that*

$$b(r)=\widehat{b}(r)\quad(\text{resp., } a(r)=\widehat{a}(r))\quad\text{a.s. }\forall r\in[s,\hat{s}]$$

implies

$$\alpha^{a,t}[b(\cdot)](r)=\alpha^{a,t}[\widehat{b}(\cdot)](r)\quad(\text{resp., } \beta^{b,t}[a(\cdot)](r)=\beta^{b,t}[\widehat{a}(\cdot)](r))$$

for $r\in[s,\hat{s}]$.

We denote all admissible strategies with the initial value a (resp., b) for player I (resp., II) on $[s,1]$ by $\Gamma^a[s,1]$ (resp., $\Delta^b[s,1]$). We adopt the convention that

$$\begin{aligned}&\mathcal{A}^a[1,1]=a,\quad \Gamma^a[1,1]=a,\\ &\mathcal{B}^b[1,1]=b,\quad \Delta^b[1,1]=b.\end{aligned} \tag{2.6}$$

Set for $(s,x)\in[0,1]\times X$,

(2.7)

$$V_{a,b}(s,x)=\inf_{\alpha\in\Gamma^a[s,1]}\sup_{b(\cdot)\in\mathcal{B}^b[s,1]}J_{s,x}(\alpha(b(\cdot)),b(\cdot)),\quad V(s,x)=(V_{a,b}(s,x))_{a\in A,b\in B};$$

$$U^{a,b}(s,x)=\sup_{\beta\in\Delta^b[s,1]}\inf_{a(\cdot)\in\mathcal{A}^a[s,1]}J_{s,x}(a(\cdot),\beta(a(\cdot))),\quad U(s,x)=(U^{a,b}(s,x))_{a\in A,b\in B}.$$

The matrix-valued functions V and U are called the lower and the upper value function, respectively. If $V = U$, we say that the above stochastic switching game has a value. Our aim is to study the existence of the value of our stochastic switching game. U and V should satisfy the dynamic programming principle. However, the conventional proof leads to serious technical problems related to measurability issues, which have been noticed by Fleming and Souganidis [6] in the study of classical stochastic differential games . To circumvent these problems, we borrow the techniques of Fleming and Souganidis [6] and introduce in the following the concepts of restrictive class of admissible strategies, π-admissible switching processes, and π-admissible strategies.

Consider $s \in [0,1]$, $\hat{s} \in (s,1)$, and $b(\cdot) \in \mathcal{B}^b[s,1]$. For $P_{s,\hat{s}}$-a.s. $\omega_1 \in \Omega_{s,\hat{s}}$, the map $b(\omega_1) : [\hat{s},1] \times \Omega_{\hat{s}} \to B$ defined by

$$b(\omega_1)(\omega_2)(r) = b(\omega_1,\omega_2)(r), \quad r \in [\hat{s},1],$$

is an admissible control for player II, i.e., $b(\omega_1) \in \mathcal{B}^{b(\hat{s})}[\hat{s},1]$.

DEFINITION 2.3. *Let $\alpha \in \Gamma^a[\hat{s},1]$. If for $\forall s \in (0,\hat{s})$ and $\forall b(\cdot) \in \mathcal{B}^b[s,1], b \in B$, the map $(\tau,\omega) \mapsto \alpha[b(\omega_1)](\omega_2)(\tau)$ is $\mathcal{B}([\hat{s},\tau]) \otimes \mathcal{F}_{s,\tau}$-measurable for every $\tau \in [\hat{s},1]$, then α is called an r-strategy with initial value a for player* I *on $[s,1]$. The set of r-strategies with initial value a on $[s,1]$ of player* I *is denoted by $\Gamma_1^a[s,1]$.*

Similarly, we define r-strategies with initial value $b \in B$ on $[s,1]$ for player II and denote their collection by $\Delta_1^b[s,1]$.

Set

(2.8)
$$V_{a,b}^1(s,x) = \inf_{\alpha \in \Gamma_1^a[s,1]} \sup_{b(\cdot) \in \mathcal{B}^b[s,1]} J_{s,x}(\alpha(b(\cdot)), b(\cdot)), \quad V^1(s,x) = (V_{a,b}^1(s,x))_{a \in A, b \in B};$$

$$U_1^{a,b}(s,x) = \sup_{\beta \in \Delta_1^b[s,1]} \inf_{a(\cdot) \in \mathcal{A}^a[s,1]} J_{s,x}(a(\cdot), \beta(a(\cdot))), \quad U_1(s,x) = (U_1^{a,b}(s,x))_{a \in A, b \in B}.$$

The matrix-valued functions V^1 and U_1 are called the r-lower and the r-upper value function, respectively.

Let $\pi_s = \{s = t_0 < t_1 < \cdots < t_M = 1\}$ be a partition of $[s,1]$, and denote by $||\pi_s|| = \max_{1 \le i \le M}(t_i - t_{i-1})$ its mesh. The notions of π-admissible switching processes and π-admissible strategies are then defined as follows.

DEFINITION 2.4. *Let $a(\cdot) = \{a_i, \theta_i\}_{i \ge 0} \in \mathcal{A}^a[s,1]$. If each θ_i is a π_s-valued stopping time, then it is called a π-admissible switching process with initial value $a \in A$ on $[s,1]$ for player* I. *The set of π-admissible switching processes with initial value $a \in A$ on $[s,1]$ for player* I *is denoted by $\mathcal{A}_\pi^a[s,1]$. The π-admissible switching processes with initial value $b \in B$ on $[s,1]$ for player* II *are defined in a similar way, and their collection is denoted by $\mathcal{B}_\pi^b[s,1]$.*

DEFINITION 2.5. *$\alpha \in \Gamma^a[s,1]$ is called a π-admissible strategy with initial value $a \in A$ on $[s,1]$ for player* I, *if it satisfies the following properties:* (1) *$\forall b(\cdot) \in \mathcal{B}^b[s,1], b \in B$, $\alpha[b(\cdot)] \in \mathcal{A}_\pi^a[s,1]$.* (2) *Fix $b \in B$. If $s \in [t_{i_0}, t_{i_0+1})$, then $\alpha[b_1(\cdot)]|_{[s,t_{i_0+1})} = \alpha[b_2(\cdot)]|_{[s,t_{i_0+1})}$ $\forall\, b_1(\cdot), b_2(\cdot) \in \mathcal{B}^b[s,1]$.* (3) *If $b(\cdot) = \bar{b}(\cdot)$ on $[s,t_i)$, then $\alpha[b(\cdot)](t_i) = \alpha[\bar{b}(\cdot)](t_i), P_s$-a.s. for $i \in \{i_0+1, \ldots, M\}$. The collection of π-admissible strategies with initial value $a \in A$ on $[s,1]$ for player I is denoted by $\Gamma_\pi^a[s,1]$. The π-admissible strategies with initial value $b \in B$ on $[s,1]$ for player* II *are defined in a similar way, and their collection is denoted by $\Delta_\pi^b[s,1]$.*

It is crucial, in our case of the switching game, that $\alpha[b(\cdot)](t_i)$ in Definition 2.5 is required to be independent of $b(t_i)$ for $\alpha \in \Gamma_\pi^a[s,1]$ and $i = i_0+1, \ldots, M$. Definition 2.5 differs from Fleming and Souganidis' in that $\alpha[b(\cdot)]|_{[s,t_{i_0+1})}$ may depend on $b(s-)$

even if $\alpha \in \Gamma_\pi^a[s,1]$, and this is due to the fact that the initial position of a player is crucial in our switching game.

According to the definitions of $V_{a,b}^1(s,x), U_1^{a,b}(s,x), V_{a,b}(s,x)$, and $U^{a,b}(s,x)$, we have immediately the following two relations:

$$V_{a,b}(s,x) \le V_{a,b}^1(s,x) \text{ and } U_1^{a,b}(s,x) \le U^{a,b}(s,x). \tag{2.9}$$

Next, let us introduce some operators. For any $m \times n$ matrix-valued function $W(\cdot,\cdot) = (W^{a,b}(\cdot,\cdot))$ defined on $[0,1] \times X$, we define for $(a,b,s,x) \in A \times B \times [0,1] \times X$

$$\begin{aligned} M^{a,b}[W](s,x) &= \min_{\hat{a} \ne a} \{W^{\hat{a},b}(s,x) + k(s,a,\hat{a})\}, \\ M_{a,b}[W](s,x) &= \max_{\hat{b} \ne b} \{W^{a,\hat{b}}(s,x) - l(s,b,\hat{b})\}. \end{aligned} \tag{2.10}$$

The two operators are called obstacle operators. According to the definitions, for any $(a,b,s,x) \in A \times B \times [0,1] \times X$, the following are true:

$$\begin{aligned} M_{a,b}[V](s,x) \le V_{a,b}(s,x) \le M^{a,b}[V](s,x), \\ M_{a,b}[V^1](s,x) \le V_{a,b}^1(s,x) \le M^{a,b}[V^1](s,x), \\ M_{a,b}[U](s,x) \le U^{a,b}(s,x) \le M^{a,b}[U](s,x), \\ M_{a,b}[U_1](s,x) \le U_1^{a,b}(s,x) \le M^{a,b}[U_1](s,x). \end{aligned} \tag{2.11}$$

Before closing this section, we state without proof the following result on the continuity in the space variable of the costs and the value functions.

PROPOSITION 2.1. (1) *For any $a(\cdot) \in \mathcal{A}^a[s,1], b(\cdot) \in \mathcal{B}^b[s,1], \alpha \in \Gamma^a[s,1]$, and $\beta \in \Delta^b[s,1]$, the functions $J_{sx}(\alpha[b(\cdot)], b(\cdot))$ and $J_{sx}(a(\cdot), \beta[a(\cdot)])$, $(s,x) \in [0,T] \times X$, grow linearly, are Lipschitz continuous in the space variable x, and are Hölder-continuous in the time variable s, uniformly with respect to the other variable s and x, respectively, and uniformly as well with respect to the four parameters: $\alpha, a(\cdot), \beta$, and $b(\cdot)$.*

(2) *The functions V, V^1, U, and U_1 grow linearly and are Lipschitz continuous in the space variable x, uniformly with respect to the time variable s.*

The time continuity of value functions turns out to be a measurability issue and will be considered in the next section.

3. Dynamic programming and time continuity of various value functions. In this section, we use the Bellman dynamic programming principle to study the time continuity and the dynamics of various value functions related to our game.

PROPOSITION 3.1. (1) *The lower value function $V^1(\cdot,\cdot)$ satisfies the following suboptimality condition: For any $(a,b) \in A \times B$, $x \in X$, and $0 \le s < \hat{s} \le 1$,*

(3.1a)

$$\begin{aligned} V_{a,b}^1(s,x) \le \inf_{\alpha \in \Gamma_1^a[s,1]} \sup_{b(\cdot) \in \mathcal{B}^b[s,1]} E_{sx} \Bigg\{ & \int_s^{\hat{s}} f^0(r, y(r), \alpha[b(\cdot)](r), b(r))\, dr \\ & + \sum_{\theta_i \le \hat{s}} k(\theta_i, a_{i-1}, a_i) - \sum_{\tau_j \le \hat{s}} l(\tau_j, b_{j-1}, b_j) + V_{\alpha[b(\cdot)](\hat{s}), b(\hat{s})}^1(\hat{s}, y(\hat{s})) \Bigg\}, \end{aligned}$$

where $\{a_i, \theta_i\}$ and $\{b_j, \tau_j\}$ are associated with $\alpha[b(\cdot)]$ and $b(\cdot)$, respectively, and $\alpha[b(\cdot)](\hat{s}) = \alpha[b(\cdot)](\hat{s}+0), b(\hat{s}) = b(\hat{s}+0)$.

(2) *The upper value function $U_1(\cdot,\cdot)$ satisfies the following superoptimality condition: For any $(a,b) \in A \times B$, $x \in X$, and $0 \le s < \hat{s} \le 1$,*

(3.1b)

$$U_1^{a,b}(s,x) \geq \sup_{\beta\in\Delta_1^b[s,1]}\inf_{a(\cdot)\in\mathcal{A}^a[s,1]} E_{sx}\Big\{\int_s^{\hat{s}} f^0(r,y(r),a(\cdot),\beta[a(\cdot)](r))\,dr + \sum_{\theta_i\leq\hat{s}} k(\theta_i,a_{i-1},a_i) - \sum_{\tau_j\leq\hat{s}} l(\tau_j,b_{j-1},b_j) + U_1^{a(\hat{s}),\beta[a(\cdot)](\hat{s})}(\hat{s},y(\hat{s}))\Big\},$$

where $\{a_i,\theta_i\}$ *and* $\{b_j,\tau_j\}$ *are associated with* $a(\cdot)$ *and* $\beta[a(\cdot)]$*, respectively, and* $\beta[a(\cdot)](\hat{s}) = \beta[a(\cdot)](\hat{s}+0), b(\hat{s}) = b(\hat{s}+0)$.

Proof of Proposition 3.1. We prove only the inequality (3.1a); the inequality (3.1b) can be proved in the same manner.

Let (s,x,a,b) be fixed, and let $W_{a,b}(s,x)$ be the right-hand side of (3.1a). Then, $\forall\varepsilon>0$, there exists $\alpha\in\Gamma_1^a[s,1]$ such that

(3.2)

$$W_{a,b}(s,x) \geq E_{sx}\Big\{\int_s^{\hat{s}} f^0(r,y(r),\alpha[b(\cdot)](r),b(r))\,dr + \sum_{\theta_i\leq\hat{s}} k(\theta_i,a_{i-1},a_i) - \sum_{\tau_j\leq\hat{s}} l(\tau_j,b_{j-1},b_j) + V^1_{\alpha[b(\cdot)](\hat{s}),b(\hat{s})}(\hat{s},y(\hat{s}))\Big\} - \varepsilon$$

for every $b(\cdot)\in\mathcal{B}^b[s,1]$. Also, for each $\hat{a}\in A,\hat{b}\in B,\xi\in X$,

$$V^1_{\hat{a},\hat{b}}(\hat{s},\xi) = \inf_{\alpha\in\Gamma_1^{\hat{a}}[\hat{s},1]}\sup_{b(\cdot)\in\mathcal{B}^{\hat{b}}[\hat{s},1]} J_{\hat{s}\xi}(\alpha[b(\cdot)],b(\cdot)); \tag{3.3}$$

thus there exists $\alpha_\xi^{\hat{a},\hat{b}}\in\Gamma_1^{\hat{a}}[\hat{s},1]$ for which

$$V^1_{\hat{a},\hat{b}}(\hat{s},\xi) \geq \sup_{b(\cdot)\in\mathcal{B}^{\hat{b}}[\hat{s},1]} J_{\hat{s}\xi}(\alpha_\xi^{\hat{a},\hat{b}}[b(\cdot)],b(\cdot)) - \varepsilon. \tag{3.4}$$

Next let $\{A_i : i=1,2,\dots\}$ be a partition of X by Borel sets, and choose $\xi_i\in A_i (i=1,2,\dots)$. If the diameter of the A_i's is sufficiently small, then for $i=1,2,\dots$ and $w\in A_i$,

(3.5)

$$|J_{\hat{s}w}(\alpha[b(\cdot)],b(\cdot)) - J_{\hat{s}\xi_i}(\alpha[b(\cdot)],b(\cdot))| < \varepsilon \quad \text{for any } b(\cdot)\in\mathcal{B}^{\hat{b}}[\hat{s},1] \text{ and } \alpha\in\Gamma_1^{\hat{a}}[\hat{s},1],$$

and

$$|V^1_{\hat{a},\hat{b}}(\hat{s},w) - V^1_{\hat{a},\hat{b}}(\hat{s},\xi_i)| < \varepsilon. \tag{3.6}$$

Now we use the strategies α and $\alpha_{\xi_i}^{\hat{a},\hat{b}}, i=1,\dots,\hat{a}\in A,\hat{b}\in B$, to construct a new admissible strategy $\widetilde{\alpha}\in\Gamma_1^a[s,1]$ as follows: For $(r,\omega)\in[s,1]\times\Omega_s$ and $b(\cdot)\in\mathcal{B}^b[s,1]$, we define

(3.7)

$$\alpha[b(\cdot)](\omega)(r) = \begin{cases} \alpha[b(\cdot)](\omega)(r) & \text{if } r<\hat{s},\\ \sum_{i=1}^\infty\sum_{\hat{a}\in A,\hat{b}\in B}\chi_{\{b(\hat{s})=\hat{b},\alpha[b(\cdot)](\hat{s})=\hat{a}\}}\chi_{A_i}(y_{sx}(\hat{s}))\alpha_{\xi_i}^{\hat{a},\hat{b}}[b(\omega_1)](\omega_2)(r) & \text{if } r\geq\hat{s},\end{cases}$$

where $\omega=(\omega_1,\omega_2)\in\Omega_{t,\hat{s}}\times\Omega_{\hat{s}}$ and $b(\omega_1)(\cdot)\in\mathcal{B}^{b(\hat{s})}[\hat{s},1]$ is given by $b(\omega_1)(\omega_2)(r) = b(\omega_1,\omega_2)(r)$.

Consequently for any $b(\cdot) \in \mathcal{B}^b[s,1]$, using (3.2), (3.4), and (3.6), we obtain

$$
\begin{aligned}
W_{a,b}(s,x) \geq & E_{sx}\Bigg\{\int_s^{\hat{s}} f^0(r,y(r),\alpha[b(\cdot)](r),b(r))\,dr \\
& + \sum_{\theta_i \leq \hat{s}} k(\theta_i, a_{i-1}, a_i) - \sum_{\tau_j \leq \hat{s}} l(\tau_j, b_{j-1}, b_j) \\
& + \sum_{i=1}^{\infty} \sum_{\hat{a}\in A, \hat{b}\in B} \chi_{\{\alpha[b(\cdot)](\hat{s})=\hat{a}, b(\hat{s})=\hat{b}\}} \chi_{A_i}(y_{s,x}(\hat{s})) V^1_{\hat{a},\hat{b}}(\hat{s}, y(\hat{s}))\Bigg\} - \varepsilon \\
\geq & E_{sx}\Bigg\{\int_s^{\hat{s}} f^0(r,y(r),\alpha[b(\cdot)](r),b(r))\,dr \\
& + \sum_{\theta_i \leq \hat{s}} k(\theta_i, a_{i-1}, a_i) - \sum_{\tau_j \leq \hat{s}} l(\tau_j, b_{j-1}, b_j) \\
& + \sum_{i=1}^{\infty} \sum_{\hat{a}\in A, \hat{b}\in B} \chi_{\{\alpha[b(\cdot)](\hat{s})=\hat{a}, b(\hat{s})=\hat{b}\}} \chi_{A_i}(y_{s,x}(\hat{s})) V^1_{\hat{a},\hat{b}}(\hat{s}, \xi_i)\Bigg\} - 2\varepsilon.
\end{aligned} \tag{3.8}
$$

On the other hand, for $y_{sx}(\hat{s}) \in A_i, i = 1,2,\ldots$ and $\forall\, b(\cdot) \in \mathcal{B}^{\hat{b}}[\hat{s},1]$, we derive from (3.4) and (3.5) that

$$
V^1_{\hat{a},\hat{b}}(\hat{s},\xi_i) \geq J_{\hat{s}\xi_i}(\alpha^{\hat{a},\hat{b}}_{\xi_i}[b(\cdot)], b(\cdot)) - \varepsilon \geq J_{\hat{s}y_{sx}(\hat{s})}(\alpha^{\hat{a},\hat{b}}_{\xi_i}[b(\cdot)], b(\cdot)) - 2\varepsilon. \tag{3.9}
$$

Combining the above inequalities, we have

$$
\begin{aligned}
W_{a,b}(s,x) \geq & E_{sx}\Bigg\{\int_s^{\hat{s}} f^0(r,y(r),\alpha[b(\cdot)](r),b(r))\,dr \\
& + \sum_{\theta_i \leq \hat{s}} k(\theta_i, a_{i-1}, a_i) - \sum_{\tau_j \leq \hat{s}} l(\tau_j, b_{j-1}, b_j) \\
& + \sum_{i=1}^{\infty} \chi_{A_i}(y_{s,x}(\hat{s})) E_{\hat{s}y_{sx}(\hat{s})}\Bigg\{\int_{\hat{s}}^1 f^0(r,y(r),\widetilde{\alpha}[b(\cdot)](r),b(r))\,dr \\
& + \sum_{\theta_i > \hat{s}} k(\theta_i, a_{i-1}, a_i) - \sum_{\tau_j > \hat{s}} l(\tau_j, b_{j-1}, b_j) + h(y(1))\Bigg\}\Bigg\} - 4\varepsilon.
\end{aligned} \tag{3.10}
$$

Therefore,

$$
W_{a,b}(s,x) \geq \; J_{sx}(\widetilde{\alpha}, b(\cdot)) - 4\varepsilon,
$$

which in turn implies

$$
W_{a,b}(s,x) \geq \; V^1_{a,b}(s,x) - 4\varepsilon,
$$

and the result now follows. □

From Proposition 3.1, we can obtain the following time continuity of V^1 and U_1.

PROPOSITION 3.2. *There exists $L > 0$ such that for any $a \in A, b \in B, x \in X$, and $s,t \in [0,1]$*

$$
\begin{aligned}
|V^1_{a,b}(s,x) - V^1_{a,b}(t,x)| \; &\leq L(1+|x|)\sqrt{|s-t|}, \\
|U_1^{a,b}(s,x) - U_1^{a,b}(t,x)| \; &\leq L(1+|x|)\sqrt{|s-t|}.
\end{aligned} \tag{3.11}
$$

Proof of Proposition 3.2. We prove only the $\frac{1}{2}$-Hölder continuity of the r-lower value function V^1 in the time variable the $\frac{1}{2}$-Hölder continuity of the r-upper value function U_1 in the time variable can be proved in the same way.

Suppose that $s < t$. First, we prove the following:

$$V^1_{a,b}(s,x) - V^1_{a,b}(t,x) \le L(1+|x|)\sqrt{t-s}. \tag{3.12}$$

From Proposition 3.1 and Hypothesis 3, we derive

$$\begin{aligned}
&V^1_{a,b}(s,x)\\
&\le \sup_{b(\cdot)\in\mathcal{B}^b[s,1]} E_{sx}\left\{\int_s^t f^0(r,y(r),a,b(r))\,dr - \sum_{\tau_j\le t} l(\tau_j,b_{j-1},b_j) + V^1_{a,b(t)}(t,y(t))\right\}\\
&\le \sup_{b(\cdot)\in\mathcal{B}^b[s,1]} E_{sx}\left\{\int_s^t f^0(r,y(r),a,b(r))\,dr - \sum_{\tau_j\le t} l(t,b_{j-1},b_j) + V^1_{a,b(t)}(t,y(t))\right\}\\
&\le \sup_{b(\cdot)\in\mathcal{B}^b[s,1]} E_{sx}\left\{\int_s^t f^0(r,y(r),a,b(r))\,dr - l(t,b,b(t)) + V^1_{a,b(t)}(t,y(t))\right\}\\
&\le \sup_{b(\cdot)\in\mathcal{B}^b[s,1]} E_{sx}\left\{\int_s^t f^0(r,y(r),a,b(r))\,dr + V^1_{a,b}(t,y(t))\right\}.
\end{aligned} \tag{3.13}$$

Note that in the last step, we have used the relation (2.11). We then have

$$\begin{aligned}
&V^1_{a,b}(s,x) - V^1_{a,b}(t,x)\\
&\le \sup_{b(\cdot)\in\mathcal{B}^b[s,1]} E_{sx}\left\{\int_s^t f^0(r,y(r),a,b(r))\,dr + V^1_{a,b}(t,y(t)) - V^1_{a,b}(t,x)\right\},
\end{aligned} \tag{3.14}$$

which proves (3.12) by the uniformly Lipschitz continuity of $V^1_{a,b}(t,x)$ in x and the following estimate:

$$E|y_{sx}(t) - x| \le L(1+|x|)\sqrt{t-s}.$$

Second, we prove the following:

$$V^1_{a,b}(s,x) - V^1_{a,b}(t,x) \ge -L(1+|x|)\sqrt{t-s}. \tag{3.15}$$

In fact, for any $\widehat{b}(\cdot) \in \mathcal{B}^b[t,1]$ and $\alpha \in \Gamma^a_1[s,1]$, we define

$$b(r) = \begin{cases} b, & r \in [s,t),\\ \widehat{b}(r), & r \in [t,1], \end{cases}$$

and

$$\begin{aligned}
&\widehat{\alpha}(\omega_1)[\widehat{b}(\cdot)](r) = \alpha[b(\cdot)](\omega_1)(r), \quad r \in [t,1],\\
&\widehat{\alpha}(\omega_1)[\widehat{b}(\cdot)](t-0) = a.
\end{aligned} \tag{3.16}$$

Then we see that $b(\cdot) \in \mathcal{B}^b[s,1]$ and $\widehat{\alpha}(\omega_1) \in \Gamma_1^a[s,1]$, a.s. It follows that

(3.17)
$$\begin{aligned} &J_{sx}(\alpha[b(\cdot)], b(\cdot)) \\ &\geq E_{sx}\Big\{\int_s^t f^0(r, y(r), \alpha[b(\cdot)](r), b(r))\,dr + l(t,a,a(t)) + J_{tx}(\alpha(\omega_1)[b(\cdot)], \widehat{b}(\cdot)) \\ &\quad + \int_t^1 [f^0(r, y(r), \widehat{\alpha}(\omega_1)[\widehat{b}(\cdot)](r), \widehat{b}(r)) - f^0(r, y(r), \widehat{\alpha}(\omega_1)[\widehat{b}(\cdot)](r), b(r))]\,dr\Big\}. \end{aligned}$$

Here we have used Hypothesis 2. Then we see that

(3.18)
$$\begin{aligned} &\sup_{b(\cdot)\in\mathcal{B}^b[s,1]} J_{sx}(\alpha[b(\cdot)], b(\cdot)) \\ &\geq \sup_{b(\cdot)\in\mathcal{B}^b[s,1]} E_{sx}[V^1_{\alpha[b(\cdot)](t),b}(t,x) + l(t,a,\alpha[b(\cdot)](t))] - L(1+|x|)\sqrt{t-s} \\ &\geq V^1_{a,b}(t,x) - L(1+|x|)|s-t|^{1/2}, \end{aligned}$$

which implies (3.15). □

Remark 3.1. It is still true to replace in Proposition 3.1 the deterministic time $\hat{s} \in (s,1]$ with a stopping time τ which takes its values in $(s,1]$. In fact, in this version of Proposition 3.2, it is sufficient to note that, for any $(x,a,b) \in X \times A \times B$, the two random variables $V^1_{a,b}(\tau,x)$ and $U_1^{a,b}(\tau,x)$ may be sufficiently approximated by

$$\sum_{i=0}^{N-1} V^1_{a,b}(t_i,x)\chi_{[t_i,t_{i+1})}(\tau) \quad \text{and} \quad \sum_{i=0}^{N-1} U_1^{a,b}(t_i,x)\chi_{[t_i,t_{i+1})}(\tau),$$

respectively, by letting N be sufficiently large. Here we have used the following notation:

$$t_i = \frac{i(1-s)}{N}, \quad i = 0,1,\ldots,N.$$

For $(s,x,\delta) \in [0,1] \times X \times (0,\infty)$ and $(a(\cdot),b(\cdot)) \in \mathcal{A}^a[s,1] \times \mathcal{B}^b[s,1]$, define

$$\tau^\delta_{s,x}(a(\cdot),b(\cdot)) := \inf\Big\{t \in [s,T] : |y_{s,x}^{a(\cdot),b(\cdot)}(t) - x| \geq \delta\Big\} \wedge T,$$

where $y_{s,x}^{a(\cdot),b(\cdot)}$ is the solution of the system (1.1) corresponding to $(a(\cdot),b(\cdot)) \in \mathcal{A}^a[s,1] \times \mathcal{B}^b[s,1]$, which will occasionally be abbreviated as y_{sx} or simply y to simplify the notation. It is easy to see that $\tau^\delta_{s,x}(a(\cdot),b(\cdot))$ is a stopping time for any triplet $(s,x,\delta) \in [0,1] \times X \times [0,\infty)$ and any pair $(a(\cdot),b(\cdot)) \in \mathcal{A}^a[s,1] \times \mathcal{B}^b[s,1]$. To simplify the notation, we shall simply write τ^δ for $\tau^\delta_{s,x}(a(\cdot),b(\cdot))$ when the dependence on $(s,x,a(\cdot),b(\cdot))$ is not confused from the context. We also have

$$|y_{s,x}^{a(\cdot),b(\cdot)}(t \wedge \tau^d) - x| \leq \delta$$

for $s \leq t \leq 1, a(\cdot) \in \mathcal{A}^a[s,1]$, and $b(\cdot) \in \mathcal{B}^b[s,1]$. Moreover, we have

$$\lim_{\hat{s}\to s+} \sup_{a(\cdot)\in\mathcal{A}^a[s,1], b(\cdot)\in\mathcal{B}^b[s,1]} \frac{P(\{\hat{s} \geq \tau^\delta\})}{\hat{s}-s} = 0.$$

In fact, we have

$$P(\{\tau^\delta \le \hat{s}\}) = P\left(\left\{\sup_{s\le t\le \hat{s}} |y_{s,x}^{a(\cdot),b(\cdot)}(t) - x| \ge \delta\right\}\right) \le \sum_{i=1}^{N} P_{s,x}^{\delta,i}(a(\cdot),b(\cdot)),$$

where N is the dimension of the state space X, for $i = 1,\dots,N$, e_i is the unit vector of X whose ith component is one, and

$$P_{s,x}^{\delta,i}(a(\cdot),b(\cdot)) := P\left(\left\{\sup_{s\le t\le \hat{s}} \langle e_i, y_{s,x}^{a(\cdot),b(\cdot)}(t) - x\rangle \ge \delta N^{-\frac{1}{2}}\right\}\right).$$

Define

$$f_{s,x}^{\delta} := \sup\{|f(t,y,a,b)| : (t,a,b) \in [s,1]\times A\times B, |y-x|\le\delta\}$$

and

$$g_{s,x}^{\delta} := \sup\{|g(t,y,a,b)| : (t,a,b) \in [s,1]\times A\times B, |y-x|\le\delta\}.$$

For $\theta \in X$, from Itô's formula, it follows that the process

$$\begin{aligned} &Z_{s,x}^{\delta,\theta}(t;a(\cdot),b(\cdot)) \\ &:= \exp\left\{\left\langle \theta, y_{s,x}(t\wedge\tau^\delta) - x - \int_s^{t\wedge\tau^\delta} f(r,y_{s,x}(r),a(r),b(r))\,dr\right\rangle \right. \\ &\quad \left. -\frac{1}{2}\int_s^{t\wedge\tau^\delta} |g^*(r,y_{s,x}(r),a(r),b(r))e_i|^2\,dr\right\}, \quad t\in[s,1], \end{aligned}$$

is a continuous martingale, and $E[Z_{s,x}^{\delta,\theta}(t;a(\cdot),b(\cdot))] = 1$ for any $(t,a(\cdot),b(\cdot),\theta) \in [s,T]\times\mathcal{A}^a[s,1]\times\mathcal{B}^b[s,1]\times X$. Therefore, using Doob's inequality, we have for $h := \hat{s}-s, \lambda>0, \delta_0\ge\delta>0, a(\cdot)\in\mathcal{A}^a[s,1]$, and $b(\cdot)\in\mathcal{B}^b[s,1]$

$$\begin{aligned} &P_{s,x}^{\delta,i}(a(\cdot),b(\cdot)) \\ &\le P\left(\left\{\sup_{s\le t\le\hat{s}} Z_{s,x}^{\delta,\lambda e_i}(t;a(\cdot),b(\cdot)) \ge \exp\left[\lambda(\delta N^{-\frac{1}{2}} - hf_{s,x}^{\delta_0}) - \frac{1}{2}\lambda^2 h|g_{s,x}^{\delta_0}|^2\right]\right\}\right) \\ &\le \exp\left[-\lambda(\delta N^{-\frac{1}{2}} - hf_{s,x}^{\delta_0}) + \frac{1}{2}\lambda^2 h|g_{s,x}^{\delta_0}|^2\right]. \end{aligned}$$

As h is sufficiently small, take

$$\lambda = \frac{\delta N^{-\frac{1}{2}} - hf_{s,x}^{\delta_0}}{h|g_{s,x}^{\delta_0}|^2},$$

and we further have

$$P_{s,x}^{\delta,i}(a(\cdot),b(\cdot)) \le \exp\left\{\frac{-|\delta N^{-\frac{1}{2}} - hf_{s,x}^{\delta_0}|^2}{2h|g_{s,x}^{\delta_0}|^2}\right\}.$$

Hence,

$$\begin{aligned}
&\lim_{\hat{s}\to s+} \sup_{a(\cdot)\in\mathcal{A}^a[s,1],b(\cdot)\in\mathcal{B}^b[s,1]} \frac{P(\{\tau^\delta \le \hat{s}\})}{\hat{s}-s} \\
&\le \lim_{\hat{s}\to s+} \sup_{a(\cdot)\in\mathcal{A}^a[s,1],b(\cdot)\in\mathcal{B}^b[s,1]} h^{-1}\sum_{i=1}^{N} P^{\delta,i}_{s,x}(a(\cdot),b(\cdot)) \\
&\le \lim_{h\to 0} Nh^{-1}\exp\left\{\frac{-|\delta N^{-\frac{1}{2}} - hf^{\delta_0}_{s,x}|^2}{2h|g^{\delta_0}_{s,x}|^2}\right\} = 0.
\end{aligned}$$

The desired result then follows.

PROPOSITION 3.3. (1) *The r-lower value function $V^1(\cdot,\cdot)$ satisfies the following: Suppose at $(a,b,s,x)\in A\times B\times[0,1]\times X$*

$$V^1_{a,b}(s,x) > M_{a,b}[V^1](s,x). \tag{3.19a}$$

Then there exist a deterministic time $s_0 > s$ and a sufficiently small number $\delta_0 > 0$, such that for all $\hat{s}\in[s,s_0]$ and $\delta\in(0,\delta_0]$,

$$V^1_{a,b}(s,x) \le E_{sx}\left\{\int_s^{\hat{s}\wedge\tau^\delta} f^0(r,y^{a,b}(r),a,b)\,dr + V^1_{a,b}(\hat{s}\wedge\tau^\delta, y^{a,b}(\hat{s}\wedge\tau^\delta))\right\}. \tag{3.19b}$$

Here we have abbreviated $\tau^\delta_{s,x}(a,b)$ as τ^δ.

(2) *The r-upper value function $U_1(\cdot,\cdot)$ satisfies the following: Suppose at $(a,b,s,x)\in A\times B\times[0,1]\times X$*

$$U^{a,b}_1(s,x) < M^{a,b}[U_1](s,x). \tag{3.20a}$$

Then there exist a deterministic time $s_0 > s$ and a sufficiently small number $\delta_0 > 0$, such that for all $\hat{s}\in[s,s_0]$ and $\delta\in(0,\delta_0]$,

$$U^{a,b}_1(s,x) \ge E_{sx}\left\{\int_s^{\hat{s}\wedge\tau^\delta} f^0(r,y^{a,b}(r),a,b)\,dr + U^{a,b}_1(\hat{s}\wedge\tau^\delta, y^{a,b}(\hat{s}\wedge\tau^\delta))\right\}. \tag{3.20b}$$

Here we have abbreviated $\tau^\delta_{s,x}(a,b)$ as τ^δ.

Remark 3.2. Proposition 3.3 can be viewed as a stochastic version of Theorem 3.2 by Yong [10]. However, it is by no means trivial and is of stochastic nature in its formulation. The upper limits of the integrals in (3.19b) and (3.20b) are more complicated than the deterministic counterparts: The former are a deterministic time $\hat{s} > s$ which is sufficiently close to the initial time s, stopped by the first time of the system state process $y^{a,b}_{s,x}$ (steered by both players I and II with constant actions $a\in A$ and $b\in B$, respectively) escaping from a sufficiently small ball centered at the initial state x, while the latter are simply a deterministic time $\hat{s} > s$ which is sufficiently close to the initial time s. Obviously, both coincide. Our proof below is quite different from the deterministic case and is of stochastic nature; it includes a delicate analysis.

Proof of Proposition 3.3. We prove only statement (1); the proof of statement (2) is similar.

If statement (1) were not true, then there would exist sequences $\hat{s}\to s, \delta\to 0+$, and $\varepsilon\to 0+$ such that

$$V^1_{a,b}(s,x) - \varepsilon > E_{sx}\left\{\int_s^{\hat{s}\wedge\tau^\delta} f^0(s,y(s;a,b),a,b)\,ds + V^1_{a,b}(\hat{s}\wedge\tau^\delta, y(\hat{s}\wedge\tau^\delta;a,b))\right\}. \tag{3.21}$$

On the other hand, using Proposition 3.2 and the idea exposed in Remark 3.1, we can show the following analogy to Proposition 3.1 (1):

$$\begin{aligned} V^1_{a,b}(s,x) \le \inf_{\alpha\in\Gamma^a_1[s,1]} \sup_{b(\cdot)\in\mathcal{B}^b[s,1]} E_{sx}\Big\{ & \int_s^{\hat{s}\wedge\tau^\delta_{a(\cdot),b(\cdot)}} f^0(r,y(r),\alpha[b(\cdot)](r),b(r))\,dr \\ & + \sum_{\theta_i\le\hat{s}\wedge\tau^\delta_{a(\cdot),b(\cdot)}} k(\theta_i,a_{i-1},a_i) - \sum_{\tau_j\le\hat{s}\wedge\tau^\delta_{a(\cdot),b(\cdot)}} l(\tau_j,b_{j-1},b_j) \\ & + V^1_{\alpha[b(\cdot)](\hat{s}),b(\hat{s})}(\hat{s}\wedge\tau^\delta_{a(\cdot),b(\cdot)}, y(\hat{s}\wedge\tau^\delta_{a(\cdot),b(\cdot)}))\Big\}, \end{aligned}$$

where $\{a_i,\theta_i\}$ and $\{b_j,\tau_j\}$ are associated with $\alpha[b(\cdot)]$ and $b(\cdot)$, respectively; $\alpha[b(\cdot)](\hat{s}) = \alpha[b(\cdot)](\hat{s}+0), b(\hat{s}) = b(\hat{s}+0)$, and $\tau^\delta_{a(\cdot),b(\cdot)} := \tau^\delta_{s,x}(a(\cdot),b(\cdot))$. Therefore, we have

$$\begin{aligned} V^1_{a,b}(s,x) \le \sup_{b(\cdot)\in\mathcal{B}^{b,s}} E_{sx}\Big\{ & \int_s^{\hat{s}\wedge\tau^\delta_{a,b(\cdot)}} f^0(r,y(r),a,b(r))\,dr \\ & - \sum_{\tau_j\le\hat{s}\wedge\tau^\delta_{a,b(\cdot)}} l(\tau_j,b_{j-1},b_j) + V^1_{a,b(\hat{s})}(\hat{s}\wedge\tau^\delta_{a,b(\cdot)}, y(\hat{s}\wedge\tau^\delta_{a,b(\cdot)}))\Big\}. \end{aligned} \tag{3.22}$$

Furthermore, by definition, we conclude that there exists $b^\varepsilon(\cdot)\in\mathcal{B}^b[s,1]$ such that

$$\begin{aligned} V^1_{a,b}(s,x) - \varepsilon \le E_{sx}\Big\{ & \int_s^{\hat{s}\wedge\tau^\delta_{a,b^\varepsilon(\cdot)}} f^0(r,y(r),a,b^\varepsilon(r))\,dr \\ & - \sum_{\tau^\varepsilon_j\le\hat{s}\wedge\tau^\delta_{a,b^\varepsilon(\cdot)}} l(\tau^\varepsilon_j,b^\varepsilon_{j-1},b^\varepsilon_j) + V^1_{a,b^\varepsilon(\hat{s})}(\hat{s}\wedge\tau^\delta_{a,b^\varepsilon(\cdot)}, y(\hat{s}\wedge\tau^\delta_{a,b^\varepsilon(\cdot)}))\Big\}, \end{aligned} \tag{3.23}$$

where $\{\tau^\varepsilon_j,b^\varepsilon_j\} = b^\varepsilon(\cdot)$.

Set

$$\begin{aligned} B_1 &:= \{\omega : b^\varepsilon(r\wedge\tau^\delta_{a,b^\varepsilon(\cdot)},\omega) \ne b \text{ for some } r\in[s,\hat{s}\wedge\tau^\delta_{a,b^\varepsilon(\cdot)}]\}, \\ B_1^c &:= \Omega\backslash B_1 = \{\omega : b^\varepsilon(r\wedge\tau^\delta_{a,b^\varepsilon(\cdot)},\omega) = b\ \forall\, r\in[s,\hat{s}\wedge\tau^\delta_{a,b^\varepsilon(\cdot)}]\}. \end{aligned}$$

Note that B_1 and B_1^c depend on $(\delta,\hat{s})$. Then the two inequalities (3.21) and (3.23) yield

$$E[\chi_{B_1}] > 0 \text{ for sufficiently small positive } \delta \text{ and } \hat{s}. \tag{3.24}$$

Combining (3.21) and (3.23), we have

$$\text{(I)} + \text{(II)} + \text{(III)} > 0, \tag{3.25}$$

where

(3.26)

$$\begin{aligned}
\text{(I)} &= E_{sx}\left\{\int_s^{\hat{s}\wedge\tau^{\delta}_{a,b^{\varepsilon}(\cdot)}} f^0(r,y(r;a,b^{\varepsilon}(\cdot)),a,b^{\varepsilon}(r))\,dr - \int_s^{\hat{s}\wedge\tau^{\delta}_{a,b}} f^0(r,y(r;a,b),a,b)\,dr\right\}\\
&\le C((\hat{s}-s)(1+|x|+\delta))E\chi_{B_1},\\
\text{(II)} &= E_{sx}\left\{-\sum_{\tau^{\varepsilon}_j\le\hat{s}\wedge\tau^{\delta}_{a,b^{\varepsilon}(\cdot)}} l(\tau^{\varepsilon}_j,b^{\varepsilon}_{j-1},b^{\varepsilon}_j)\right\} \le -E_{sx}\left\{\sum_{\tau^{\varepsilon}_j\le\hat{s}\wedge\tau^{\delta}_{a,b^{\varepsilon}(\cdot)}} l(\hat{s}\wedge\tau^{\delta}_{a,b^{\varepsilon}(\cdot)},b^{\varepsilon}_{j-1},b^{\varepsilon}_j)\right\},\\
\text{(III)} &= E_{sx}[V^1_{a,b^{\varepsilon}(\hat{s})}(\hat{s}\wedge\tau^{\delta}_{a,b^{\varepsilon}(\cdot)},y(\hat{s}\wedge\tau^{\delta}_{a,b^{\varepsilon}(\cdot)};a,b^{\varepsilon}(\cdot))) - V^1_{a,b}(\hat{s}\wedge\tau^{\delta}_{a,b},y(\hat{s}\wedge\tau^{\delta}_{a,b};a,b))]\\
&= E_{sx}\Big\{[V^1_{a,b^{\varepsilon}(\hat{s})}(\hat{s}\wedge\tau^{\delta}_{a,b^{\varepsilon}(\cdot)},y(\hat{s}\wedge\tau^{\delta}_{a,b^{\varepsilon}(\cdot)};a,b^{\varepsilon}(\cdot)))\\
&\qquad - V^1_{a,b}(\hat{s}\wedge\tau^{\delta}_{a,b},y(\hat{s}\wedge\tau^{\delta}_{a,b};a,b))]\chi_{B_1}\Big\}.
\end{aligned}$$

Hence, noting Propositions 2.1 and 3.2, we have

$$\begin{aligned}
0 &\le \text{(I)} + \text{(II)} + \text{(III)}\\
&\le \{M_{a,b}[V^1](s,x) - V^1_{a,b}(s,x) + C[\sqrt{\hat{s}-s}(1+|x|)+\delta]\}E\chi_{B_1}
\end{aligned}\tag{3.27}$$

for some positive constant C, which implies that

$$M_{a,b}[V^1](s,x) - V^1_{a,b}(s,x) \ge -C\sqrt{\hat{s}-s}(1+|x|)+\delta. \tag{3.28}$$

Letting $\delta\to 0+$ and $\varepsilon\to 0+$, we have

$$M_{a,b}[V^1](s,x) \ge V^1_{a,b}(s,x),$$

which contradicts (3.19a). □

Note that the time continuity of V^1 and U_1 given by Proposition 3.2 is used in the proof of Proposition 3.3.

Denote by $C^{0,1}(X,\mathbb{R}^{m\times n})$ the totality of $\mathbb{R}^{m\times n}$-valued uniformly Lipschitz continuous functions on X. For $\varphi(\cdot)=(\varphi_{a,b}(\cdot))_{a\in A,b\in B}\in C^{0,1}(X,\mathbb{R}^{m\times n})$, define

(3.29a)

$$\begin{aligned}
F_{a,b}(s,\hat{s})\varphi(x) = \sup_{\hat{b}\in B}\inf_{a(\cdot)\in\mathcal{A}^a[s,\hat{s}]} E_{sx}\Big\{&\varphi_{a(\hat{s}-),\hat{b}}(y(\hat{s})) + \int_s^{\hat{s}} f^0(r,y(r),a(r),\hat{b})\,dr\\
&+\sum_{s\le\theta_i<\hat{s}} k(\theta_i,a_{i-1},a_i) - l(s,b,\hat{b})\Big\},\quad (a,b,x)\in A\times B\times X;\\
F(s,\hat{s})\varphi = (F_{a,b}(s,\hat{s})\varphi)_{a\in A,b\in B}&
\end{aligned}$$

and

(3.29b)

$$\begin{aligned}
G^{a,b}(s,\hat{s})\varphi(x) = \inf_{\hat{a}\in A}\sup_{b(\cdot)\in\mathcal{B}^b[s,\hat{s}]} E_{sx}\Big\{&\varphi_{\hat{a},b(\hat{s}-)}(y(\hat{s})) + \int_s^{\hat{s}} f^0(r,y(r),\hat{a},b(r))\,dr\\
&+k(s,a,\hat{a}) - \sum_{s\le\tau_j<\hat{s}} l(\tau_j,a_{j-1},a_j)\Big\},\quad (a,b,x)\in A\times B\times X;\\
G(s,\hat{s})\varphi = (G^{a,b}(s,\hat{s})\varphi)_{a\in A,b\in B}&.
\end{aligned}$$

It is easily shown that $F(s,\hat{s})$ and $G(s,\hat{s})$ are self-mappings on $C^{0,1}(X,\mathbb{R}^{m\times n})$. Therefore, the function $V^\pi : [0,1]\times X \to C^{0,1}(X,\mathbb{R}^{m\times n})$ given by

(3.30a)
$$V^\pi(1,x) = (V^\pi_{a,b}(1,x))_{a\in A, b\in B}, \quad V^\pi_{a,b}(1,x) \equiv h(x) \text{ for } (a,b)\in A\times B, \text{ with } x\in X;$$
$$V^\pi(s,x) = F(s,t_{i_0+1}) \prod_{i=i_0+2}^{M} F(t_{i-1},t_i)h(x), \quad x\in X, \quad \text{if } s\in[t_{i_0},t_{i_0+1}),$$

is well defined. Let $V^\pi_{a,b}(s,x)$ be the (a,b)-component of the matrix $V^\pi(s,x)$. Similarly, define $U_\pi = (U^{a,b}_\pi)_{a\in A,b\in B} : [0,1]\times X \to C^{0,1}(X,\mathbb{R}^{m\times n})$ as follows:

(3.30b)
$$U_\pi(1,x) = (U^{a,b}_\pi(1,x))_{a\in A,b\in B}, \quad U^{a,b}_\pi(1,x) \equiv h(x) \text{ for } (a,b)\in A\times B, \text{ with } x\in X;$$
$$U_\pi(s,x) = G(s,t_{i_0+1}) \prod_{i=i_0+2}^{M} G(t_{i-1},t_i)h(x), \ x\in X, \quad \text{if } s\in[t_{i_0},t_{i_0+1}).$$

We have the following.

PROPOSITION 3.4. *For $(a,b,s,x)\in A\times B\times[0,1]\times X$ and $\hat{s}\in\pi\cap[s,1]$, we have*

$$\begin{aligned} V^\pi_{a,b}(s,x) = \inf_{\alpha\in\Gamma^a[s,1]} \sup_{b(\cdot)\in\mathcal{B}^b_\pi[s,1]} E_{sx}\Bigg\{ &\int_s^{\hat{s}} f^0(r,y(r),\alpha[b(\cdot)](r),b(r))\,dr \\ &+ \sum_{s\le\theta_i<\hat{s}} k(\theta_i,a_{i-1},a_i) - \sum_{s\le\tau_j<\hat{s}} l(\tau_j,b_{j-1},b_j) \\ &+ V^\pi_{\alpha[b(\cdot)](\hat{s}-),b(\hat{s}-)}(\hat{s},y(\hat{s}))\Bigg\}, \end{aligned} \tag{3.31a}$$

where $b(\cdot)=\{b_j,\tau_j\}$ and $a(\cdot)=\alpha[b(\cdot)]=\{a_i,\theta_i\}$, and

$$\begin{aligned} U^{a,b}_\pi(s,x) = \sup_{\beta\in\Delta^b[s,1]} \inf_{a(\cdot)\in\mathcal{A}^a_\pi[s,1]} E_{sx}\Bigg\{ &\int_s^{\hat{s}} f^0(r,y(r),a(r),\beta[a(\cdot)](r))\,dr \\ &+ \sum_{s\le\theta_i<\hat{s}} k(\theta_i,a_{i-1},a_i) - \sum_{s\le\tau_j<\hat{s}} l(\tau_j,b_{j-1},b_j) \\ &+ U^{a(\hat{s}-),\beta[a(\cdot)](\hat{s}-)}_\pi(\hat{s},y(\hat{s}))\Bigg\}, \end{aligned} \tag{3.31b}$$

where $a(\cdot)=\{a_i,\theta_i\}$ and $b(\cdot)=\beta[a(\cdot)]=\{b_j,\tau_j\}$.

Proof of Proposition 3.4. We prove only (3.31a) here; the proof of (3.31b) is identical and therefore will be omitted.

For $a(\cdot)\in\mathcal{A}^a[s,1]$ and $b(\cdot)\in\mathcal{B}^b[s,1]$, set

$$\begin{aligned} \widehat{J}_{sx}(a(\cdot),b(\cdot)) = E_{sx}\Bigg\{ &\int_s^{\hat{s}} f^0(r,y(r),a(r),b(r))\,dr \\ &+ \sum_{s\le\theta_i<\hat{s}} k(\theta_i,a_{i-1},a_i) - \sum_{s\le t_j<\hat{s}} l(t_j,b_{j-1},b_j) \\ &+ V^\pi_{a(\hat{s}-),b(\hat{s}-)}(\hat{s},y(\hat{s}))\Bigg\}. \end{aligned} \tag{3.32}$$

The desired result can be derived from the following assertion: For $(a,b,s,x)\in A\times B\times[0,1]\times X$ and $\forall\varepsilon>0$, there exist $\beta_\varepsilon\in\Delta^b_\pi[s,1]$ and $\alpha_\varepsilon\in\Gamma^a[s,1]$ such that

$$V^\pi_{a,b}(s,x) \ge \widehat{J}_{sx}(\alpha_\varepsilon[b(\cdot)],b(\cdot)) - \varepsilon \quad \forall b(\cdot)\in\mathcal{B}^b_\pi[s,1] \tag{3.33}$$

and

$$V^{\pi}_{a,b}(s,x) \le \widehat{J}_{sx}(a(\cdot),\beta_\varepsilon[a(\cdot)]) + \varepsilon \quad \forall a(\cdot) \in \Gamma^a[s,1]. \tag{3.34}$$

In fact, the inequality (3.33) implies (3.31a) with the equality sign replaced by "$\ge$." On the other hand, for any $\alpha \in \Gamma^a[s,1]$, the pair of strategies of $\beta_\varepsilon \in \Delta^b_\pi[s,1]$ and $\alpha \in \Gamma^a[s,1]$ define a pair of switching processes $a_\varepsilon(\cdot) \in \mathcal{A}^a[s,1]$ and $b_\varepsilon \in \mathcal{B}^b_\pi[s,1]$ such that

$$\widehat{J}_{sx}(a_\varepsilon(\cdot),\beta_\varepsilon) = \widehat{J}_{sx}(\alpha, b_\varepsilon(\cdot)), \tag{3.35}$$

and this gives the other inequality in (3.31a). We invite the reader to see Fleming and Souganidis [6] for the details of the proof.

We conclude the proof by establishing (3.33) and (3.34). For $\varphi \in C^{0,1}(X,\mathbb{R}^{m\times n})$, define

$$\begin{aligned}\psi_{a,b}(s,x,\hat{s},\varphi,\widetilde{b}) = \inf_{a(\cdot)\in\mathcal{A}^a[s,\hat{s})} E_{sx}\Bigg\{ & \varphi_{a(\hat{s}-),\widetilde{b}}(y(\hat{s})) + \int_s^{\hat{s}} f^0(r,y(r),a(r),\widetilde{b})\,dr \\ & + \sum_{s\le\theta_i<\hat{s}} k(\theta_i,a_{i-1},a_i) - l(s,b,\widetilde{b})\Bigg\}.\end{aligned} \tag{3.36}$$

Here $y_{sx}(\cdot)$ is the solution of (1.1) with $b(r) = \widetilde{b}, r \in [s,\hat{s}]$.

$$F_{a,b}(s,\hat{s})\varphi(x) = \sup_{\widetilde{b}\in B} \psi_{a,b}(s,x,\hat{s},\varphi,\widetilde{b}). \tag{3.37}$$

If $s \in [t_{i_0}, t_{i_0+1})$ for $i_0 \in \{0,1,\dots,M-1\}$, let $D_M = h, D_j = F(t_j,t_{j+1})D_{j+1}$, for $j = i_0+1,\dots,M-1$, and $D_{s,i_0} = F(t,t_{i_0+1})D_{i_0+1}$. Thus,

$$D_{s,i_0}(x) = V^{\pi}_{a,b}(s,x), \tag{3.38}$$

and, in particular,

$$D_{i_0}(x) = V^{\pi}_{a,b}(t_{i_0},x). \tag{3.39}$$

We partition X into Borel sets $\{A_i : i = 1,2,\dots\}$ of diameter less than δ, where δ is to be specified later, and we choose $x_i \in A_i$. Given $\gamma > 0$, we can choose δ small enough and $\widetilde{b}^a_{i(j-1)} \in B$ for $i = 1,2,\dots$ and $j = i_0+1,\dots,M$ such that

$$\psi_{a,b}(t_{j-1},x_i,t_j,D_j,\widetilde{b}^a_{i(j-1)}) > F_{a,b}(t_{j-1},t_j)D_j(x_i) - \gamma, \tag{3.40a}$$

and thus

$$\begin{aligned}E_{t_{j-1}x_i}\Bigg\{ & D_j^{a(t_j-),\widetilde{b}^a_{i(j-1)}}(y(t_j)) + \int_{t_{j-1}}^{t_j} f^0(r,y(r),a(r),\widetilde{b}^a_{i(j-1)})\,dr \\ & + \sum_{t_{j-1}\le\theta_i<t_j} k(\theta_i,a_{i-1},a_i) - l(t_{j-1},b,\widetilde{b})\Bigg\} > D^{a,b}_{j-1}(x_i) - \gamma \quad \forall a(\cdot) \in \mathcal{A}^a[t_{j-1},t_j].\end{aligned} \tag{3.40b}$$

We also choose $a^{\tilde{b}}_{i(j-1)}(\cdot) \in \mathcal{A}^a(t_{j-1}, t_j)$ such that, for $a(\cdot) = a^{\tilde{b}}_{ij}(\cdot)$ and $b(r) = \widetilde{b}, r \in (t_{j-1}, t_j]$,

$$\begin{aligned} & E_{t_{j-1}x_i}\Big\{ D_j^{a(t_j-),\tilde{b}}(y(t_j; a^{\tilde{b}}_{i(j-1)}(\cdot), \widetilde{b})) + \int_{t_{j-1}}^{t_j} f^0(r, y(r), a^{\tilde{b}}_{i(j-1)}(r), \widetilde{b})\, dr \\ & \qquad + \sum_{t_{j-1} \le \theta_i < t_j} k(\theta_i, a_{i-1}, a_i) - l(t_{j-1}, b, \widetilde{b}) \Big\} \\ & < \quad \psi_{a,b}(t_{j-1}, x_i, t_j, D_j, \widetilde{b}) + \gamma = D_j^{a,b}(x_i) + \gamma, \end{aligned} \tag{3.41}$$

where for $j = i_0 + 1$ we replace t_{i_0} by s. Here $y_{t_j x_i}(\cdot; a^{\tilde{b}}_{ij}(\cdot), \widetilde{b})$ is the solution of (1.1) with the initial data (t_{j-1}, x_i) and on the switchings $a(\cdot) = a^{\tilde{b}}_{ij}(\cdot)$ and $b(\cdot) \equiv \widetilde{b}$.

We need to introduce more notations. As before, we identify $\omega \in \Omega_s$ with the pair $(\omega_{1j}, \omega_{2j})$ for $j = i_0 + 2, \dots, M$, where $\omega_{1j} = \omega|_{[s,t_{j-1}]}$ and $\omega_{2j} = \omega - \omega_{t_{j-1}}|_{[t_{j-1},1]}$. With this identification, the Wiener measure P_s on Ω_s can be regarded as the product measure $P_{1j} \otimes P_{2j}$ of the two probability measures P_{1j} and P_{2j}, which are defined on the two measure spaces $(\Omega_{s,t_{j-1}}, \mathcal{F}_{s,t_{j-1}})$ and $(\Omega_{t_{j-1}}, \mathcal{F}_{t_{j-1}})$, respectively. In view of this identification, we will be writing

$$E^{P_{2j}} \equiv E_{t_{j-1}x_i}. \tag{3.42}$$

The strategies $\alpha_\varepsilon \in \Gamma^a[s,1]$ and $\beta_\varepsilon \in \Delta^b_\pi[s,1]$ are defined as follows. Let $(a, b, s, x) \in A \times B \times [0,1) \times X$ be fixed. For $a(\cdot) \in \mathcal{A}^a[s,1]$, we define

$$\begin{cases} \beta_\varepsilon[a(\cdot)](r) = b\chi_{[s,s)} + \chi_{[s,t_{i_0+1})}(r) \displaystyle\sum_{i=1}^{\infty} \widetilde{b}^a_{ii_0}\chi_{A_i}(x) \\ \qquad\qquad + \displaystyle\sum_{j=i_0+1}^{M-1} \chi_{[t_j,t_{j+1})}(r) \sum_{\substack{\tilde{a}\in A \\ i=1}}^{\infty} \widetilde{b}^{\tilde{a}}_{ij}\chi_{A_i}(y_{sx}(t_j))\chi_{\{a(t_j-)=\tilde{a}\}}, r \in [s,1), \\ \beta_\varepsilon[a(\cdot)](1) = \beta_\varepsilon[a(\cdot)](1-), \end{cases} \tag{3.43}$$

where the random variable $y_{sx}(\cdot)$ is defined successively on intervals $[s, t_{i_0+1}]$, $[t_j, t_{j+1}]$, $j = i_0 + 1, \dots, M-1$, as the solution to (1.1) with $b(r) = \beta_\varepsilon[a(\cdot)](r)$. Note that $\forall a(\cdot) \in \mathcal{A}^a[s,1]$ and $r \in (s,1)$, $\beta_\varepsilon[a(\cdot)](r)$ depends only on $a(\cdot)|_{[s,r)}$ and is independent of $a(r)$. For $b(\cdot) \in \mathcal{B}^b[s,1]$, we define

$$\begin{cases} \alpha_\varepsilon[b(\cdot)](r) = \chi_{[s,s)} \displaystyle\sum_{i=1}^{\infty} \widetilde{a}^b_{ii_0}\chi_{A_i}(x) + \chi_{[s,t_{i_0+1})}(r) \sum_{\tilde{b}\in B, i=1}^{\infty} \widetilde{a}^{\tilde{b}}_{ii_0}(r)\chi_{A_i}(x)\chi_{\{b(s)=\tilde{b}\}} \\ \qquad\qquad + \displaystyle\sum_{j=i_0+1}^{M-1} \chi_{[t_j,t_{j+1})}(r) \sum_{\tilde{b}\in B, i=1}^{\infty} \widetilde{a}^{\tilde{b}}_{ij}(r)\chi_{A_i}(y_{sx}(t_j))\chi_{\{b(r)=\tilde{b}\}}, r \in [s,1); \\ \alpha_\varepsilon[b(\cdot)](1) = \alpha_\varepsilon[b(\cdot)](1-), \end{cases} \tag{3.44}$$

where again $y_{sx}(\cdot)$ is defined successively on intervals $[s, t_{i_0+1}]$, $[t_j, t_{j+1}]$, $j = i_0 + 1, \dots, M-1$, as the solution to (1.1) with $a(r) = \alpha_\varepsilon[b(\cdot)](r)$. Note that for any $b(\cdot) \in \mathcal{B}^b[s,1]$ and $r \in [s,1]$, $\alpha_\varepsilon[b(\cdot)](r)$ depends on $b(r)$.

For either $a(\cdot) \in \mathcal{A}^a[s,1]$ and $b(\cdot) = \beta_\varepsilon[a(\cdot)]$ or $b(\cdot) \in \mathcal{B}^b_\pi[s,1]$ and $a(\cdot) = \alpha_\varepsilon[b(\cdot)]$, we have

$$
\begin{aligned}
&V^\pi_{a,b}(s,x) - \widehat{J}_{sx}(a(\cdot),b(\cdot)) \\
&= E^{P_s} \sum_{t_{i_0+1} \le t_j \le \hat{s}} \Big\{ D_{j-1}^{a(t_{j-1}-),b(t_{j-1}-)}(y_{sx}(t_{j-1})) \\
&\quad + E^{P_s}\Big[-\int_{t_{j-1}}^{t_j} f^0(r, y_{sx}(r), a(r), b(r))\, dr - \sum_{t_{j-1} \le \theta_i < t_j} k(\theta_i, a_{i-1}, a_i) \\
&\quad + \sum_{t_{j-1} \le \tau_j < t_j} l(\tau_j, b_{j-1}, b_j) - D_j^{a(t_j-),b(t_j-)}(y_{sx}(t_j)) \;\Big|\; \mathcal{F}_{s,t_{j-1}} \Big] \Big\}.
\end{aligned} \tag{3.45}
$$

To obtain (3.33) and (3.34), it suffices to show that the following statements hold:

$$
\begin{aligned}
&D_{j-1}^{a(t_{j-1}-),b(t_{j-1}-)}(y_{sx}(t_{j-1})) \;\ge \\
&E^{P_s}\Big[\int_{t_{j-1}}^{t_j} f^0(r, y_{sx}(r), a(r), b(r))\, dr + \sum_{t_{j-1} \le \theta_i < t_j} k(\theta_i, a_{i-1}, a_i) \\
&- \sum_{t_{j-1} \le \tau_j < t_j} l(\tau_j, b_{j-1}, b_j) + D_j^{a(t_j-),b(t_j-)}(y_{sx}(t_j)) \;\Big|\; \mathcal{F}_{s,t_{j-1}} \Big] - \varepsilon(t_j - t_{j-1}), \\
&\qquad P_s\text{-a.s.} \quad \forall\, b(\cdot) \in \mathcal{B}^b_\pi[s,1] \text{ and } a(\cdot) = \alpha_\varepsilon[b(\cdot)]
\end{aligned} \tag{3.46}
$$

and

$$
\begin{aligned}
&D_{j-1}^{a(t_{j-1}-),b(t_{j-1}-)}(y_{sx}(t_{j-1})) \\
&\le E^{P_s}\Big[\int_{t_{j-1}}^{t_j} f^0(r, y_{sx}(r), a(r), b(r))\, dr + \sum_{t_{j-1} \le \theta_i < t_j} k(\theta_i, a_{i-1}, a_i) \\
&- \sum_{t_{j-1} \le \tau_j < t_j} l(\tau_j, b_{j-1}, b_j) + D_j^{a(t_j-),b(t_j-)}(y_{sx}(t_j)) \;\Big|\; \mathcal{F}_{s,t_{j-1}} \Big] + \varepsilon(t_j - t_{j-1}), \\
&\qquad P_s\text{-a.s.} \quad \forall\, a(\cdot) \in \mathcal{A}^a[s,1] \text{ and } b(\cdot) = \beta_\varepsilon[a(\cdot)].
\end{aligned} \tag{3.47}
$$

They can be derived from (3.41) and (3.40b), separately. □

It is easy to see that $V^\pi_{a,b}(s,x)$ and $U^{a,b}_\pi(s,x)$ grow in a linear way in the state variable $x \in X$, uniformly with respect to $(a,b,s) \in A \times B \times [0,T]$ and the partition π. Analogous to the first part of the proof of Proposition 3.2, we can also show some time continuity of V^π and U_π. These properties are summarized into the following.

PROPOSITION 3.5. *There is a positive constant L such that for $s \in [0,1], t \in \pi \cap [s,1]$, $a \in A, b \in B, x \in X$, and $y \in X$ we have*

$$
\begin{aligned}
&|V^\pi_{a,b}(s,x)| + |U^{a,b}_\pi(s,x)| \le L(1+|x|), \\
&|V^\pi_{a,b}(s,x) - V^\pi_{a,b}(s,y)| + |U^{a,b}_\pi(s,x) - U^{a,b}_\pi(s,y)| \le L|x-y|, \\
&|V^\pi_{a,b}(s,x) - V^\pi_{a,b}(t,x)| + |U^{a,b}_\pi(s,x) - U^{a,b}_\pi(t,x)| \le L(1+|x|)\sqrt{t-s}.
\end{aligned} \tag{3.48}
$$

Next we assume that $\pi_i = \{j/2^i\}_{j=0}^{2^i}, i = 0,1,2,\ldots$. Then $||\pi_i|| = 1/2^i$, and, from the definition of V^{π_i} and $U_{\pi_i}, i = 0,1,2,\ldots$, and Proposition 3.4, we have

$$
V^{\pi_0}_{a,b} \le V^{\pi_1}_{a,b} \le \cdots \le V^{\pi_i}_{a,b} \le \cdots \le V_{a,b}, \quad (a,b) \in A \times B \tag{3.49a}
$$

and

$$U^{a,b}_{\pi_0} \geq U^{a,b}_{\pi_1} \geq \cdots \geq U^{a,b}_{\pi_i} \geq \cdots \geq U^{a,b}, \quad (a,b) \in A \times B. \tag{3.49b}$$

PROPOSITION 3.6. *For $(a,b) \in A \times B$, let $v_{a,b} = \lim_{i\to\infty} V^{\pi_i}_{a,b}$ and $v := (v_{a,b})_{a\in A, b\in B}$. Then, for $s,t \in [0,1]$ and $x,y \in X$, we have*

$$\begin{aligned} v_{a,b}(s,x) &\leq V_{a,b}(s,x), \\ |v_{a,b}(s,x)| &\leq L(1+|x|), \\ |v_{a,b}(s,x) - v_{a,b}(t,x)| &\leq L(1+|x|)\sqrt{|t-s|}, \\ |v_{a,b}(s,x) - v_{a,b}(s,y)| &\leq L|x-y|. \end{aligned} \tag{3.50a}$$

Similarly, let $u^{a,b} = \lim_{i\to\infty} U^{a,b}_{\pi_i}$ and $u := (u^{a,b})_{a\in A, b\in B}$. Then, for $s,t \in [0,1]$ and $x,y \in X$,

$$\begin{aligned} u^{a,b}(s,x) &\geq U^{a,b}(s,x), \\ |u^{a,b}(s,x)| &\leq L(1+|x|), \\ |u^{a,b}(s,x) - u^{a,b}(t,x)| &\leq L(1+|x|)\sqrt{|t-s|}, \\ |u^{a,b}(s,x) - u^{a,b}(s,y)| &\leq L|x-y|. \end{aligned} \tag{3.50b}$$

Proof of Proposition 3.6. First, we prove (3.50a). Assume, without loss of generality, that $s < t$. Let $\{t_i\}_{i=1}^{\infty} \subset \cup_{i=0}^{\infty}\pi_i \cap [t,1]$ and $\lim_{i\to\infty} t_i = t$. Then we have from Proposition 3.5 that

$$\begin{aligned} &|v_{a,b}(s,x) - v_{a,b}(t,x)| \\ &\leq |v_{a,b}(s,x) - v_{a,b}(t_i,x)| + |v_{a,b}(t_i,x) - v_{a,b}(t,x)| \\ &\leq L(1+|x|)(\sqrt{t_i - s} + \sqrt{t_i - t}), \quad i = 1,2,\ldots. \end{aligned} \tag{3.51}$$

This concludes the $\frac{1}{2}$-Hölder continuity in the time variable of $v_{a,b}$. Its linear growth and uniform Lipschitz continuity in the space variable x is straightforward.

In an identical way, we can show (3.50b). ∎

Passing to the limit $\|\pi\| \to 0$ in Proposition 3.4, we obtain that two functions v and u satisfy the following dynamic programming principle.

PROPOSITION 3.7. *For $(a,b,s,x) \in A \times B \times [0,1] \times X$ and $\hat{s} \in \cup_{i=0}^{\infty}\pi_i \cap [s,1]$, we have*

$$\begin{aligned} v_{a,b}(s,x) = \lim_{i\to\infty} \inf_{\alpha\in\Gamma^a[s,1]} \sup_{b(\cdot)\in\mathcal{B}^b_{\pi_i}[s,1]} E_{sx}\Bigg\{ &\int_s^{\hat{s}} f^0(r,y(r),\alpha[b(\cdot)](r),b(r))\,dr \\ &+ \sum_{s\leq\theta_j<\hat{s}} k(\theta_j,a_{j-1},a_j) - \sum_{s\leq\tau_j<\hat{s}} l(\tau_j,b_{j-1},b_j) \\ &+ v_{\alpha[b(\cdot)](\hat{s}-),b(\hat{s}-)}(\hat{s},y(\hat{s}))\Bigg\}, \end{aligned} \tag{3.52a}$$

where $b(\cdot) = \{b_j,\tau_j\}$ and $a(\cdot) = \alpha[b(\cdot)] = \{a_j,\theta_j\}$, and

$$\begin{aligned} u^{a,b}(s,x) = \lim_{i\to\infty} \sup_{\beta\in\Delta^b[s,1]} \inf_{a(\cdot)\in\mathcal{A}^a_{\pi_i}[s,1]} E_{sx}\Bigg\{ &\int_s^{\hat{s}} f^0(r,y(r),a(r),\beta[a(\cdot)](r))\,dr \\ &+ \sum_{s\leq\theta_j<\hat{s}} k(\theta_j,a_{j-1},a_j) - \sum_{s\leq\tau_j<\hat{s}} l(\tau_j,b_{j-1},b_j) \\ &+ u^{a(\hat{s}-),\beta[a(\cdot)](\hat{s}-)}(\hat{s},y(\hat{s}))\Bigg\}, \end{aligned} \tag{3.52b}$$

where $a(\cdot) = \{a_j,\theta_j\}$ and $b(\cdot) = \beta[a(\cdot)] = \{b_j,\tau_j\}$.

Proof of Proposition 3.7. We only derive the equality (3.52a) from the equality (3.31a) in Proposition 3.4. The proof of the equality (3.52b) is similar.

Since $\hat{s} \in \cup_{i=0}^{\infty} \pi_i \cap [s,1]$, we have $\hat{s} \in \pi_i \cap [s,1]$ when i is sufficiently large. From Proposition 3.4, we have that, when i is sufficiently large,

$$\begin{aligned} V_{a,b}^{\pi_i}(s,x) = \inf_{\alpha \in \Gamma^a[s,1]} \sup_{b(\cdot) \in \mathcal{B}_{\pi_i}^b[s,1]} E_{sx} \Big\{ & \int_s^{\hat{s}} f^0(r, y(r), \alpha[b(\cdot)](r), b(r))\, dr \\ & + \sum_{s \le \theta_j < \hat{s}} k(\theta_j, a_{j-1}, a_j) - \sum_{s \le \tau_j < \hat{s}} l(\tau_j, b_{j-1}, b_j) \\ & + V_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}^{\pi_i}(\hat{s}, y(\hat{s})) \Big\}, \end{aligned} \tag{3.53}$$

where $b(\cdot) = \{b_j, \tau_j\}$ and $a(\cdot) = \alpha[b(\cdot)] = \{a_j, \theta_j\}$.

Set, for any $C > 0$,

$$O_C(x) := \{y : |y - x| \le C\}, \quad O_C^c(x) := \{y : |y - x| > C\}.$$

It is easy to see from Propositions 3.5 and 3.6 that

$$\begin{aligned} & E_{sx} |V_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}^{\pi_i}(\hat{s}, y(\hat{s})) - v_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}(\hat{s}, y(\hat{s}))| \\ & \le E_{sx} \{ \chi_{O_C^c(x)}(y(\hat{s})) |V_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}^{\pi_i}(\hat{s}, y(\hat{s})) - v_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}(\hat{s}, y(\hat{s}))| \} \\ & \quad + E_{sx} \{ \chi_{O_C(x)}(y(\hat{s})) |V_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}^{\pi_i}(\hat{s}, y(\hat{s})) - v_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}(\hat{s}, y(\hat{s}))| \} \\ & \le \{ E_{sx} |V_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}^{\pi_i}(\hat{s}, y(\hat{s})) - v_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}(\hat{s}, y(\hat{s}))|^2 \}^{\frac{1}{2}} \{ P(O_C^c(x)) \}^{\frac{1}{2}} \\ & \quad + E_{sx} \{ \chi_{O_C(x)}(y(\hat{s})) |V_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}^{\pi_i}(\hat{s}, y(\hat{s})) - v_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}(\hat{s}, y(\hat{s}))| \}. \end{aligned} \tag{3.54}$$

From Propositions 3.5 and 3.6, we have that, for any given positive constant C,

$$\lim_{i \to \infty} V^{\pi_i}(\hat{s}, y) = v(\hat{s}, y) \quad \text{uniformly in } y \in O_C(x),$$

which implies that

$$\lim_{i \to \infty} E_{sx} \{ \chi_{O_C(x)}(y(\hat{s})) |V_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}^{\pi_i}(\hat{s}, y(\hat{s})) - v_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}(\hat{s}, y(\hat{s}))| \} = 0. \tag{3.55}$$

Moreover,

$$E_{sx} |V_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}^{\pi_i}(\hat{s}, y(\hat{s})) - v_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}(\hat{s}, y(\hat{s}))|^2 \le L(1 + E_{sx}|y(\hat{s})|^2). \tag{3.56}$$

Since

$$\begin{aligned} & E_{sx}|y(\hat{s})|^2 \le L(1 + |x|^2), \\ & P(O_C^c(x)) \le C^{-2} E_{sx}|y(\hat{s}) - x|^2 \le LC^{-2}(1 + |x|^2), \end{aligned} \tag{3.57}$$

we see that

$$0 \le \overline{\lim_{i \to \infty}} \sup_{\alpha, b(\cdot)} E_{sx} |V_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}^{\pi_i}(\hat{s}, y(\hat{s})) - v_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}(\hat{s}, y(\hat{s}))| \le LC^{-2}(1 + |x|^2)$$

for an arbitrary sufficiently large positive number C, and therefore

$$\lim_{i \to \infty} \sup_{\alpha, b(\cdot)} E_{sx} |V_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}^{\pi_i}(\hat{s}, y(\hat{s})) - v_{\alpha[b(\cdot)](\hat{s}-), b(\hat{s}-)}(\hat{s}, y(\hat{s}))| = 0. \tag{3.58}$$

The last equality implies (3.52a) immediately. □

Note that

$$v_{a,b}(1,x) = u^{a,b}(1,x) = h(x), \quad (a,b,x) \in A \times B \times X. \tag{3.59}$$

It follows from Proposition 3.7 that, for $(a,b,s,x) \in A \times B \times [0,1] \times X$,

$$\begin{aligned} v_{a,b}(s,x) = \lim_{i\to\infty} \inf_{\alpha\in\Gamma^a[s,1]} \sup_{b(\cdot)\in\mathcal{B}^b_{\pi_i}[s,1]} E_{sx}\Big\{ & \int_s^1 f^0(r,y(r),\alpha[b(\cdot)](r),b(r))\,dr \\ & + \sum_{s\le\theta_j<1} k(\theta_j,a_{j-1},a_j) - \sum_{s\le\tau_j<1} l(\tau_j,b_{j-1},b_j) + h(y(1))\Big\}, \end{aligned} \tag{3.60a}$$

where $b(\cdot) = \{b_j,\tau_j\}$ and $a(\cdot) = \alpha[b(\cdot)] = \{a_j,\theta_j\}$, and

$$\begin{aligned} u^{a,b}(s,x) = \lim_{i\to\infty} \sup_{\beta\in\Delta^b[s,1]} \inf_{a(\cdot)\in\mathcal{A}^a_{\pi_i}[s,1]} E_{sx}\Big\{ & \int_s^1 f^0(r,y(r),a(r),\beta[a(\cdot)](r))\,dr \\ & + \sum_{s\le\theta_j<1} k(\theta_j,a_{j-1},a_j) - \sum_{s\le\tau_j<1} l(\tau_j,b_{j-1},b_j) + h(y(1))\Big\}, \end{aligned} \tag{3.60b}$$

where $a(\cdot) = \{a_j,\theta_j\}$ and $b(\cdot) = \beta[a(\cdot)] = \{b_j,\tau_j\}$. From the above two formulas, we have

$$M_{a,b}[v](s,x) \le v_{a,b}(s,x) \le M^{a,b}[v](s,x), \quad (a,b,s,x) \in A \times B \times [0,1] \times X, \tag{3.61a}$$

and

$$M_{a,b}[u](s,x) \le u^{a,b}(s,x) \le M^{a,b}[u](s,x), \quad (a,b,s,x) \in A \times B \times [0,1] \times X. \tag{3.61b}$$

In view of the time continuity given by Proposition 3.6, the deterministic time $\hat{s}$ may be replaced in Proposition 3.7 with an arbitrary stopping time which takes values in $[s,1]$. That is, we have the following.

PROPOSITION 3.8. *For $(a,b,s,x) \in A \times B \times [0,1] \times X$ and any stopping time τ which take values in $[s,1]$, we have*

$$\begin{aligned} v_{a,b}(s,x) = \lim_{i\to\infty} \inf_{\alpha\in\Gamma^a[s,1]} \sup_{b(\cdot)\in\mathcal{B}^b_{\pi_i}[s,1]} E_{sx}\Big\{ & \int_s^\tau f^0(r,y(r),\alpha[b(\cdot)](r),b(r))\,dr \\ & + \sum_{s\le\theta_j<\tau} k(\theta_j,a_{j-1},a_j) - \sum_{s\le\tau_j<\tau} l(\tau_j,b_{j-1},b_j) \\ & + v_{\alpha[b(\cdot)](\tau-),b(\tau-)}(\tau,y(\tau))\Big\}, \end{aligned} \tag{3.62a}$$

where $b(\cdot) = \{b_j,\tau_j\}$ and $a(\cdot) = \alpha[b(\cdot)] = \{a_j,\theta_j\}$, and

$$\begin{aligned} u^{a,b}(s,x) = \lim_{i\to\infty} \sup_{\beta\in\Delta^b[s,1]} \inf_{a(\cdot)\in\mathcal{A}^a_{\pi_i}[s,1]} E_{sx}\Big\{ & \int_s^\tau f^0(r,y(r),a(r),\beta[a(\cdot)](r))\,dr \\ & + \sum_{s\le\theta_j<\tau} k(\theta_j,a_{j-1},a_j) - \sum_{s\le\tau_j<\tau} l(\tau_j,b_{j-1},b_j) \\ & + u^{a(\tau-),\beta[a(\cdot)](\tau-)}(\tau,y(\tau))\Big\}, \end{aligned} \tag{3.62b}$$

where $a(\cdot) = \{a_j,\theta_j\}$ and $b(\cdot) = \beta[a(\cdot)] = \{b_j,\tau_j\}$.

Proceeding similarly as in the proof of Proposition 3.3, we derive from Proposition 3.8 the following.

PROPOSITION 3.9. (1) *The lower value function* $v(\cdot,\cdot) := (v_{a,b})_{a\in A, b\in B}$ *satisfies the following: Suppose at* $(a,b,s,x) \in A\times B\times[0,1]\times X$,

$$v_{a,b}(s,x) > M_{a,b}[v](s,x) \quad (\text{resp., } v_{a,b}(s,x) < M^{a,b}[v](s,x)). \tag{3.63a}$$

Then there exist a deterministic time $s_0 > s$ *and a sufficiently small number* $\delta_0 > 0$, *such that for all* $\hat{s} \in [s, s_0]$ *and* $\delta \in (0, \delta_0]$,

(3.63b)

$$v_{a,b}(s,x) \leq \ (\text{resp., } \geq)\ E_{sx}\left\{\int_s^{\hat{s}\wedge\tau^\delta} f^0(r, y^{a,b}(r), a, b)\, dr + v_{a,b}(\hat{s}\wedge\tau^\delta, y^{a,b}(\hat{s}\wedge\tau^\delta))\right\}.$$

Here we have abbreviated $\tau^\delta_{s,x}(a,b)$ *as* τ^δ.

(2) *The upper value function* $u(\cdot,\cdot) := (u^{a,b})_{a\in A, b\in B}$ *satisfies the following: Suppose at* $(a,b,s,x) \in A\times B\times[0,1]\times X$,

$$u^{a,b}(s,x) < M^{a,b}[u](s,x) \quad (\text{resp., } u^{a,b}(s,x) > M_{a,b}[u](s,x)). \tag{3.64a}$$

Then there exist a deterministic time $s_0 > s$ *and a sufficiently small number* $\delta_0 > 0$, *such that for all* $\hat{s} \in [s, s_0]$ *and* $\delta \in (0, \delta_0]$,

(3.64b)

$$u^{a,b}(s,x) \geq \ (\text{resp., } \leq)\ E_{sx}\left\{\int_s^{\hat{s}\wedge\tau^\delta} f^0(r, y^{a,b}(r), a, b)\, dr + u^{a,b}(\hat{s}\wedge\tau^\delta, y^{a,b}(\hat{s}\wedge\tau^\delta))\right\}.$$

Here we have abbreviated $\tau^\delta_{s,x}(a,b)$ *as* τ^δ.

4. Viscosity solutions, uniqueness result, dynamic programming equations, and existence of the game value. In this section, we shall introduce the generalized notion of viscosity solution for our Isaacs' system of variational inequalities. The value functions defined in sections 2 and 3 turn out to be its viscosity sub- or supersolutions. We then prove the uniqueness of the viscosity solution and establish the existence of the value of our stochastic switching game.

Define for $(a,b,t,x,q,Q) \in A\times B\times[0,1]\times X\times X\times\mathcal{S}$,

$$H^{a,b}(t,x,q,Q) := \ f^0(t,x,a,b) + < q, f(t,x,a,b) > + \frac{1}{2}\mathrm{tr}(Qgg^*(t,x,a,b)). \tag{4.1}$$

Here $\mathcal{S}$ is the set of all real symmetric transformations in X. Let $C^{1,2}([0,1)\times X)$ be the set of all continuous functions which are continuously differentiable in t and twice continuously differentiable in x.

Associated with our stochastic switching game is the following Isaacs' system of quasi-variational inequalities where W is to be solved:

(1) For $(a,b,t,x) \in A\times B\times[0,1]\times X$,

$$M_{a,b}[W](t,x) \leq W_{a,b}(t,x) \leq M^{a,b}[W](t,x); \tag{4.2}$$

(2) for $(a,b,t,x) \in A\times B\times[0,1]\times X$ such that $W_{a,b}(t,x) > M_{a,b}[W](t,x)$,

$$\frac{\partial}{\partial t}W_{a,b}(t,x) + H^{a,b}(t,x,\nabla W_{a,b}(t,x), \nabla^2 W_{a,b}(t,x)) \geq 0; \tag{4.3}$$

(3) for $(a,b,t,x) \in A \times B \times [0,1] \times X$ such that $W_{a,b}(t,x) < M^{a,b}[W](t,x)$,

$$\frac{\partial}{\partial t} W_{a,b}(t,x) + H^{a,b}(t,x,\nabla W_{a,b}(t,x), \nabla^2 W_{a,b}(t,x)) \le 0; \tag{4.4}$$

(4) the terminal condition

$$W_{a,b}(1,x) = h(x), \quad (a,b,x) \in A \times B \times X. \tag{4.5}$$

DEFINITION 4.1. *An $\mathbb{R}^{m\times n}$-valued continuous function $W = (W_{a,b})_{a\in A, b\in B}$ on $[0,T]\times X$ is called a viscosity sub- (resp., super-) solution of* (4.2)–(4.5) *if it satisfies* (4.2) *and* (4.5)*, and moreover, for any $\varphi(\cdot,\cdot) \in C^{1,2}([0,1)\times X)$ and $(a,b) \in A \times B$, whenever $W_{a,b}(\cdot,\cdot) - \varphi(\cdot,\cdot)$ attains a local maximum (resp., minimum) at $(t_0,x_0) \in [0,1)\times X$ and*

$$W_{a,b}(t_0,x_0) > M_{a,b}[W](t_0,x_0) \quad (\text{resp., } W_{a,b}(t_0,x_0) < M^{a,b}[W](t_0,x_0)),$$

we have

$$\frac{\partial}{\partial t}\varphi(t_0,x_0) + H^{a,b}(t_0,x_0,\nabla\varphi(t_0,x_0),\nabla^2\varphi(t_0,x_0)) \ge 0$$
$$\left(\textit{resp.,}\ \frac{\partial}{\partial t}\varphi(t_0,x_0) + H^{a,b}(t_0,x_0,\nabla\varphi(t_0,x_0),\nabla^2\varphi(t_0,x_0)) \le 0\right).$$

An $\mathbb{R}^{m\times n}$-valued function $W = (W_{a,b})_{a\in A,b\in B}$ on $[0,T]\times X$ is called a viscosity solution of (4.2)–(4.5) *if it is both a viscosity sub- and supersolution of* (4.2)–(4.5).

Propositions 3.3 and 3.9 imply the following result.

PROPOSITION 4.1. (1) *The r-lower and r-upper value functions V^1 and U_1 are viscosity sub- and supersolutions of* (4.2)–(4.5), respectively.

(2) The functions $v = (v_{a,b})$ and $u = (u^{a,b})$ defined in Proposition 3.6 *are viscosity solutions of* (4.2)–(4.5).

Proof of Proposition 4.1. We now prove that the r-lower value function V^1 is a viscosity subsolution of (4.2)–(4.5). From the definition, it follows that $V^1_{a,b} = h$ for $(a,b) \in A \times B$. In view of (2.11), we see that V^1 satisfies (4.2).

Consider $\varphi(\cdot,\cdot) \in C^{1,2}([0,1)\times X)$ and $(a,b) \in A\times B$. Assume that $V^1_{a,b}(\cdot,\cdot) - \varphi(\cdot,\cdot)$ attains a local maximum at $(s,x) \in [0,1)\times X$ and

$$V^1_{a,b}(s,x) > M_{a,b}[V^1](s,x).$$

From Proposition 3.3, we see that there exist a deterministic time $s_0 > s$ and a sufficiently small number $\delta_0 > 0$, such that for all $\hat{s} \in (s,s_0]$ and $\delta \in (0,\delta_0]$, we have

$$V^1_{a,b}(s,x) \le E_{sx}\left\{\int_s^{\hat{s}\wedge\tau^\delta} f^0(r,y^{a,b}(r),a,b)\,dr + V^1_{a,b}(\hat{s}\wedge\tau^\delta, y^{a,b}(\hat{s}\wedge\tau^\delta))\right\}.$$

Here we have abbreviated $\tau^\delta_{s,x}(a,b)$ as τ^δ. For sufficiently small $\hat{s} \in [s,s_0]$ and $\delta \in (0,\delta_0]$, we have

$$V^1_{a,b}(\hat{s}\wedge\tau^\delta, y^{a,b}(\hat{s}\wedge\tau^\delta)) - \varphi(\hat{s}\wedge\tau^\delta, y^{a,b}(\hat{s}\wedge\tau^\delta)) \le V^1(s,x) - \varphi(s,x). \tag{4.6}$$

Therefore,

$$E_{sx}\left\{\int_s^{\hat{s}\wedge\tau^\delta} f^0(r,y^{a,b}(r),a,b)\,dr + \varphi(\hat{s}\wedge\tau^\delta, y^{a,b}(\hat{s}\wedge\tau^\delta)) - \varphi(s,x)\right\} \ge 0.$$

From Itô's formula, we conclude that

$$E_{sx}\int_s^{\hat{s}\wedge\tau^\delta}\left[\frac{\partial}{\partial t}\varphi(r,y^{a,b}(r))+H^{a,b}(r,y^{a,b}(r),\nabla\varphi(r,y^{a,b}(r)),\nabla^2\varphi(r,y^{a,b}(r)))\right]dr\geq 0.$$

Noting (see the arguments following Remark 3.1) that

$$\lim_{\hat{s}\to s+}\frac{P(\{\tau^\delta\leq\hat{s}\})}{\hat{s}-s}=0,$$

we have

$$\begin{aligned}0&\leq\lim_{\hat{s}\to s+}(\hat{s}-s)^{-1}E_{sx}\left\{\int_s^{\hat{s}\wedge\tau^\delta}\left[\frac{\partial}{\partial t}\varphi(r,y^{a,b}(r))+H^{a,b}(\cdots)\right]dr\right\}\\&=\lim_{\hat{s}\to s+}(\hat{s}-s)^{-1}E_{sx}\left\{\chi_{\{\hat{s}\leq\tau^\delta\}}\int_s^{\hat{s}}\left[\frac{\partial}{\partial t}\varphi(r,y^{a,b}(r))+H^{a,b}(\cdots)\right]dr\right\}\\&=\frac{\partial}{\partial t}\varphi(s,x)+H^{a,b}(s,x,\nabla\varphi(s,x),\nabla^2\varphi(t,x)).\end{aligned}$$

Concluding the above, we see that V^1 is a viscosity subsolution.

Noting (3.59), (3.61a), and (3.61b), we can prove all other assertions in Proposition 4.1 in an identical way. □

Let us introduce the following sets, which are adopted from Evans and Ishii [5]. For function $v:[0,1]\times X\to[-\infty,+\infty]$ and $(s,z)\in[0,1)\times X$, define

$$\begin{aligned}\wp^{2,+}v(s,z):=\Big\{&(p,q,Q)\in\mathbb{R}\times X\times\mathcal{S}:v(t,x)\\&\leq v(s,z)+p(t-s)+\langle q,x-z\rangle+\tfrac{1}{2}\langle Q(x-z),x-z\rangle\\&\quad+o(|t-s|+|x-z|^2)\text{ as }[0,1]\times X\ni(t,x)\to(s,z)\Big\},\end{aligned}\tag{4.7}$$

$$\begin{aligned}\bar{\wp}^{2,+}v(s,z):=\Big\{&(p,q,Q)\in\mathbb{R}\times X\times\mathcal{S}:\exists(t_i,x_i)\in[0,1]\times X,\\&(p_i,q_i,Q_i)\in\wp^{2,+}v(t_i,x_i),\\&(t_i,x_i,v(t_i,x_i),p_i,q_i,Q_i)\to(s,z,v(s,z),p,q,Q)\Big\}.\end{aligned}\tag{4.8}$$

Define for $(s,z)\in[0,1)\times X$

$$\wp^{2,-}v(s,z)=-\wp^{2,+}(-v)(s,z)\text{ and }\bar{\wp}^{2,-}v(s,z)=-\bar{\wp}^{2,+}(-v)(s,z).\tag{4.9}$$

The following result is standard.

PROPOSITION 4.2. *An $\mathbb{R}^{m\times n}$-valued function $W=(W_{a,b})_{a\in A,b\in B}$ on $[0,T]\times X$ is a viscosity sub- (resp., super-) solution of* (4.2)–(4.5) *if and only if it satisfies* (4.2) *and* (4.5)*, and moreover, for any $(t,x,a,b)\in[0,1)\times X\times A\times B$, whenever $(p,q,Q)\in\bar{\wp}^{2,+}W_{a,b}(t,x)$ (resp., $\bar{\wp}^{2,-}W_{a,b}(t,x)$) and*

$$W_{a,b}(t_0,x_0)>M_{a,b}[W](t_0,x_0)\quad(\text{resp., }W_{a,b}(t_0,x_0)<M^{a,b}[W](t_0,x_0)),$$

we have

$$p+H^{a,b}(t,x,q,Q)\geq 0\text{ (resp., }p+H^{a,b}(t,x,q,Q)\leq 0).\tag{4.10}$$

Now let us make a further assumption that will play an important role in the proof of the uniqueness result.

Hypothsis 4. For any loop $\{a_i, b_i\}_{i=1}^{j+1} \subset A \times B$, with the properties that

$$\begin{aligned} &j \le mn, \quad a_{j+1} = a_1, \quad b_{j+1} = b_1, \\ &\text{and either} \quad a_{i+1} = a_i \quad \text{or} \quad b_{i+1} = b_i \quad \forall\, 1 \le i \le j, \end{aligned} \tag{4.11}$$

we have

$$\sum_{i=1}^{j} k(s, a_i, a_{i+1}) - \sum_{i=1}^{j} l(s, b_i, b_{i+1}) \neq 0 \quad \forall s \in [0,1]. \tag{4.12}$$

THEOREM 4.1. *Assume Hypotheses* 1–4. *If* W *and* $\widehat{W}$ *are continuous viscosity sub- and supersolutions of* (4.2)–(4.5), *respectively, and satisfy for* $(t, x, y, a, b) \in [0,1] \times X \times X \times A \times B$ *the following:*

$$\begin{aligned} &|W_{a,b}(t,x)| + |\widehat{W}_{a,b}(t,x)| \le C(1 + |x|), \\ &|W_{a,b}(t,x) - W_{a,b}(t,y)| + |\widehat{W}_{a,b}(t,x) - \widehat{W}_{a,b}(t,y)| \le C|x - y|, \end{aligned} \tag{4.13}$$

then

$$W_{a,b}(t,x) \le \widehat{W}_{a,b}(t,x) \quad \forall (t,x,a,b) \in [0,1] \times X \times A \times B. \tag{4.14}$$

Proof of Theorem 4.1. We prove the theorem by contradiction. So suppose that $\exists (\bar{a}, \bar{b}, \bar{t}, \bar{x}) \in A \times B \times (0,1) \times X$ such that

$$W_{\bar{a},\bar{b}}(\bar{t}, \bar{x}) - \widehat{W}_{\bar{a},\bar{b}}(\bar{t}, \bar{x}) = \eta > 0. \tag{4.15}$$

Consider the following test function:

$$\psi(t,x,y) = \frac{|x-y|^2}{2\varepsilon} + \alpha e^{-\beta t}(1 + |x|^2 + |y|^2), \quad (t,x,y) \in [0,1] \times X \times X, \tag{4.16}$$

with parameters $\alpha > 0$ and $\beta > 0$. We choose a sufficiently small $\alpha > 0$ such that it does not depend on the parameter $\beta > 0$ and that it satisfies the following:

$$\psi(\bar{t}, \bar{x}, \bar{x}) < \frac{\eta}{2} \quad \forall \beta > 0. \tag{4.17}$$

Now consider the function

$$\Psi^{a,b}(t,x,y) = W_{a,b}(t,x) - \widehat{W}_{a,b}(t,y) - \psi(t,x,y), \quad (a,b,t,x,y) \in A \times B \times [0,1] \times X \times X. \tag{4.18}$$

From (4.13), (4.15), and (4.17), we see that there is a point $(a_0, b_0, t_0, x_0, y_0) \in A \times B \times [0,1] \times X \times X$ such that

$$\Psi^{a_0,b_0}(t_0, x_0, y_0) = \max_{\substack{a \in A \\ b \in B}} \sup_{t,x,y} \Psi^{a,b}(t,x,y) \ge \Psi^{\bar{a},\bar{b}}(\bar{t}, \bar{x}, \bar{x}) \ge \frac{\eta}{2}. \tag{4.19}$$

At this stage, we have the following two conclusions.

Conclusion 1. Following the arguments of Yamada [9, pp. 424–425], we can show the following assertion: Without loss of generality, we may assume that

$$M_{a_0,b_0}[W](t_0, x_0) < W_{a_0,b_0}(t_0, x_0), \quad M^{a_0,b_0}[\widehat{W}](t_0, y_0) > \widehat{W}_{a_0,b_0}(t_0, y_0). \tag{4.20}$$

Otherwise, we have

$$M_{a_0,b_0}[W](t_0,x_0)=W_{a_0,b_0}(t_0,x_0)\quad\text{or}\quad M^{a_0,b_0}[\widehat{W}](t_0,y_0)=\widehat{W}_{a_0,b_0}(t_0,y_0). \tag{4.21}$$

Consequently, there is $b_1\in B$ or $a_1\in A$ such that

$$W_{a_0,b_0}(t_0,x_0)=W_{a_0,b_1}(t_0,x_0)-l(t_0,b_0,b_1) \tag{4.22}$$

or

$$\widehat{W}_{a_0,b_0}(t_0,y_0)=\widehat{W}_{a_1,b_0}(t_0,y_0)+k(t_0,a_0,a_1). \tag{4.23}$$

On the other hand, from (4.19), we have

$$\Psi^{a_0,b_0}(t_0,x_0,y_0)\geq\Psi^{a_1,b_0}(t_0,x_0,y_0), \tag{4.24}$$

which implies immediately

$$W_{a_0,b_0}(t_0,x_0)-\widehat{W}_{a_0,b_0}(t_0,y_0)\geq W_{a_1,b_0}(t_0,x_0)-\widehat{W}_{a_1,b_0}(t_0,y_0). \tag{4.25}$$

Therefore, we have

$$\begin{aligned}0\geq&W_{a_0,b_0}(t_0,x_0)-W_{a_1,b_0}(t_0,x_0)-k(t_0,a_0,a_1)\\ \geq&\widehat{W}_{a_0,b_0}(t_0,y_0)-\widehat{W}_{a_1,b_0}(t_0,y_0)-k(t_0,a_0,a_1),\end{aligned} \tag{4.26}$$

which shows the following:

$$W_{a_0,b_0}(t_0,x_0)=W_{a_1,b_0}(t_0,x_0)+k(t_0,a_0,a_1) \tag{4.27}$$

if (4.23) is true. In summary, there is $b_1\in B$ or $a_1\in A$ such that either (4.22) or (4.27) is true. Moreover, we have

$$W_{a_0,b_0}(t_0,x_0)-\widehat{W}_{a_0,b_0}(t_0,y_0)=W_{a_1,b_0}(t_0,x_0)-\widehat{W}_{a_1,b_0}(t_0,y_0), \tag{4.28}$$

from which it follows that

$$\Psi^{a_0,b_0}(t_0,x_0,y_0)=\Psi^{a_1,b_0}(t_0,x_0,y_0)=\max_{\substack{a\in A\\ b\in B}}\sup_{t,x,y}\Psi^{a,b}(t,x,y). \tag{4.29}$$

Symmetrically, we have

$$\Psi^{a_0,b_0}(t_0,x_0,y_0)=\Psi^{a_0,b_1}(t_0,x_0,y_0)=\max_{\substack{a\in A\\ b\in B}}\sup_{t,x,y}\Psi^{a,b}(t,x,y). \tag{4.30}$$

Set $(\tilde{a}_1,\tilde{b}_1):=(a_0,b_0)$. We can repeat the above argument to start from the pair of parameters $(\tilde{a}_2,\tilde{b}_2)$—which is $(a_1,\tilde{b}_1)$ or $(\tilde{a}_1,b_1)$—to find a new pair of parameters (a_2,b_2) such that either

$$W_{\tilde{a}_2,\tilde{b}_2}(t_0,x_0)=W_{\tilde{a}_2,b_2}(t_0,x_0)-l(t_0,\tilde{b}_2,b_2) \tag{4.31}$$

or

$$W_{\tilde{a}_2,\tilde{b}_2}(t_0,x_0)=W_{a_2,\tilde{b}_2}(t_0,x_0)+k(t_0,\tilde{a}_2,a_2) \tag{4.32}$$

is true. Moreover,

$$\Psi^{\tilde{a}_2,b_2}(t_0,x_0,y_0)=\Psi^{a_2,\tilde{b}_2}(t_0,x_0,y_0)=\max_{\substack{a\in A\\ b\in B}}\sup_{t,x,y}\Psi^{a,b}(t,x,y). \tag{4.33}$$

Then we can continue the procedure until we find a loop $\{\tilde{a}_i,\tilde{b}_i\}_{i=1}^{j+1}$ which satisfies the properties (4.11). Summing up (4.31)–(4.32) for the loop, we get

$$\sum_{i=1}^{j}k(s,\tilde{a}_i,\tilde{a}_{i+1})-\sum_{i=1}^{j}l(s,\tilde{b}_i,\tilde{b}_{i+1})=0. \tag{4.34}$$

Then we get a contradiction to Hypothesis 4.

Conclusion 2. On the maximum point (a_0,b_0,t_0,x_0,y_0), we have the following properties:

(i) There is a constant $C_{\alpha,\beta}$, which depends on positive α,β such that $|x_0|+|y_0|\le C_{\alpha,\beta}$;

(ii) from (4.13) and the following inequality:

$$2\Psi^{a_0,b_0}(t_0,x_0,y_0)\ge\Psi^{a_0,b_0}(t_0,x_0,x_0)+\Psi^{a_0,b_0}(t_0,y_0,y_0),$$

we obtain $|x_0-y_0|\le\varepsilon C_{\alpha,\beta}$. Hence, $|x_0-y_0|\to 0$ as $\varepsilon\to 0$, while keeping α and β fixed;

(iii) since $\Psi^{a_0,b_0}(1,x_0,y_0)\le h(x_0)-h(y_0)\le C|x_0-y_0|$, we conclude from (4.19) that $t_0\in[0,1)$ whenever $\varepsilon>0$ is sufficiently small.

A simple computation gives rise to the following:

$$\begin{aligned}
\partial_t\psi(t,x,y)&=-\beta\alpha e^{-\beta t}(1+|x|^2+|y|^2),\\
\partial_x\psi(t,x,y)&=\frac{(x-y)}{\varepsilon}+2\alpha e^{-\beta t}x,\\
\partial_y\psi(t,x,y)&=\frac{(y-x)}{\varepsilon}+2\alpha e^{-\beta t}y,\\
\partial^2_{(x,y)}\psi(t,x,y)&=\frac{1}{\varepsilon}\begin{pmatrix}I&-I\\-I&I\end{pmatrix}+2\alpha e^{-\beta t}\begin{pmatrix}I&0\\0&I\end{pmatrix}.
\end{aligned} \tag{4.35}$$

Then, applying Theorem 9 of Evans and Ishii [5] to the function

$$W_{a_0,b_0}(t,x)+(-\widehat{W}_{a_0,b_0})(t,y)-\psi(t,x,y)$$

at the point (t_0,x_0,y_0), we can find $p_1,p_2\in\mathbb{R}$ and $Q_1,Q_2\in\mathcal{S}$ such that

$$\begin{aligned}
&(p_1,\partial_x\psi(t_0,x_0,y_0),Q_1)\in\bar{\wp}^{2,+}W_{a_0,b_0}(t_0,x_0),\\
&(p_2,\partial_y\psi(t_0,x_0,y_0),Q_2)\in\bar{\wp}^{2,+}(-\widehat{W}_{a_0,b_0})(t_0,y_0),\\
&p_1+p_2=\partial_t\psi(t_0,x_0,y_0),\\
&\begin{pmatrix}Q_1&0\\0&Q_2\end{pmatrix}\le\partial^2_{(x,y)}\psi(t_0,x_0,y_0)+\varepsilon\Big(\partial^2_{(x,y)}\psi(t_0,x_0,y_0)\Big)^2.
\end{aligned} \tag{4.36}$$

By the definitions of viscosity sub- and supersolutions, and in view of (4.20), we have

$$\begin{aligned}
p_1+H^{a,b}\left(t_0,x_0,\frac{(x_0-y_0)}{\varepsilon}+2\alpha e^{-\beta t_0}x_0,Q_1\right)&\ge 0,\\
-p_2+H^{a,b}\left(t_0,y_0,-\frac{(y_0-x_0)}{\varepsilon}-2\alpha e^{-\beta t_0}y_0,-Q_2\right)&\le 0.
\end{aligned} \tag{4.37}$$

Thus, we have (see (4.36))

$$\begin{aligned}
&\beta\alpha e^{-\beta t_0}(1+|x_0|^2+|y_0|^2)\\
&\le H^{a,b}\left(t_0,x_0,\frac{(x_0-y_0)}{\varepsilon}+2\alpha e^{-\beta t_0}x_0,Q_1\right)\\
&\quad - H^{a,b}\left(t_0,y_0,-\frac{(y_0-x_0)}{\varepsilon}-2\alpha e^{-\beta t_0}y_0,-Q_2\right)\\
&\le \frac{1}{2}\mathrm{tr}[(g^*Q_1g)(t_0,x_0,a_0,b_0)+(g^*Q_2g)(t_0,y_0,a_0,b_0)]\\
&\quad + \left[\left\langle\frac{x_0-y_0}{\varepsilon},f(t_0,x_0,a_0,b_0)-f(t_0,y_0,a_0,b_0)\right\rangle\right.\\
&\quad \left.+2\alpha e^{-\beta t_0}\Big(\langle x_0,f(t_0,x_0,a_0,b_0)\rangle+\langle y_0,f(t_0,y_0,a_0,b_0)\rangle\Big)\right]\\
&\quad + f^0(t_0,x_0,a_0,b_0)-f^0(t_0,y_0,a_0,b_0).
\end{aligned}$$

Set

$$\begin{aligned}
\text{(I)} &:= \frac{1}{2}\mathrm{tr}[(g^*Q_1g)(t_0,x_0,a_0,b_0)+(g^*Q_2g)(t_0,y_0,a_0,b_0)]\\
\text{(II)} &:= \left\langle\frac{x_0-y_0}{\varepsilon},f(t_0,x_0,a_0,b_0)-f(t_0,y_0,a_0,b_0)\right\rangle\\
&\quad + 2\alpha e^{-\beta t_0}\left(\langle x_0,f(t_0,x_0,a_0,b_0)\rangle+\langle y_0,f(t_0,y_0,a_0,b_0)\rangle\right)\\
\text{(III)} &:= f^0(t_0,x_0,a_0,b_0)-f^0(t_0,y_0,a_0,b_0).
\end{aligned}$$

We now estimate (I), (II), and (III) separately. It is immediate that

$$\begin{aligned}
&\partial^2_{(x,y)}\psi(t_0,x_0,y_0)+\varepsilon\Big(\partial^2_{(x,y)}\psi(t_0,x_0,y_0)\Big)^2\\
&\le \frac{3}{\varepsilon}\begin{pmatrix} I & -I\\ -I & I\end{pmatrix}+4\alpha e^{-\beta t_0}\begin{pmatrix} I & -I\\ -I & I\end{pmatrix}\\
&\quad +(4\varepsilon\alpha^2e^{-2\beta t_0}+2\alpha e^{-\beta t_0})\begin{pmatrix} I & 0\\ 0 & I\end{pmatrix}.
\end{aligned}\tag{4.38}$$

Then we have

$$\begin{aligned}
\text{(I)} &\le \frac{C}{\varepsilon}|x_0-y_0|^2+C\alpha e^{-\beta t_0}|x_0-y_0|^2\\
&\quad + C(2\varepsilon\alpha^2e^{-2\beta t_0}+\alpha e^{-\beta t_0})(1+|x_0|^2+|y_0|^2),\\
\text{(II)} &\le \frac{C}{\varepsilon}|x_0-y_0|^2+C\alpha e^{-\beta t_0}(1+|x_0|^2+|y_0|^2),\\
\text{(III)} &\le C|x_0-y_0|.
\end{aligned}\tag{4.39}$$

Hence, we have

$$\begin{aligned}
&\beta\alpha e^{-\beta t_0}(1+|x_0|^2+|y_0|^2)\\
&\le \frac{C}{\varepsilon}|x_0-y_0|^2+C|x_0-y_0|+C\alpha e^{-\beta t_0}(|x_0-y_0|^2+1+|x_0|^2+|y_0|^2)\\
&\quad + C(2\varepsilon\alpha^2e^{-2\beta t_0}+\alpha e^{-\beta t_0})(1+|x_0|^2+|y_0|^2).
\end{aligned}\tag{4.40}$$

Letting $\varepsilon \to 0$, we get

$$(4.41)\quad \beta\alpha e^{-\beta t}(1+2|x|^2) \le C\alpha e^{-\beta t}(1+2|x|^2) + C(2\varepsilon\alpha^2 e^{-2\beta t} + \alpha e^{-\beta t})(1+2|x|^2)$$

for some $(t,x) \in [0,1] \times X$, which immediately implies $\beta \le C + C\alpha$. Since we can choose β to be sufficiently large so that $\beta > C + \alpha C$, we arrive at a contradiction. Hence, (4.14) is proved. □

Remark 4.1. Note that the stochastic nature leads to the corresponding Isaacs' system of variational inequalities involving a second-order differential operator, and thus the proof of the uniqueness of the viscosity solutions necessarily involves the computation of the second-order differentials of the chosen test function, say, ψ in our proof. Due to this feature, the test function used by Yong [10] does not seem to carry over to our case. Here we use a different test function. It is both simpler and more powerful in proving the uniqueness of unbounded viscosity solutions, as is shown in the above proof.

THEOREM 4.2. *Let Hypotheses* 1–4 *be satisfied. Then our stochastic differential switching game described by* (1.1) *and* (1.2) *has a value. The function* $V^1 = v = V = U = u = U^1$ *is the unique viscosity solution of* (4.2)–(4.5).

Proof of Theorem 4.2. From Proposition 4.1, we see that V^1 is a viscosity subsolution and v is a viscosity supersolution, while u is a viscosity subsolution and U^1 is a viscosity supersolution. From Theorem 4.1, it follows immediately that

$$V^1_{a,b}(t,x) \le v_{a,b}(t,x) \text{ and } u^{a,b}(t,x) \le U_1^{a,b}(t,x), \quad (t,x,a,b) \in [0,1] \times X \times A \times B.$$

In view of Proposition 3.6, we have

$$V^1_{a,b} \le v_{a,b} \le V_{a,b}, \quad U_1^{a,b} \ge u^{a,b} \ge U^{a,b}, \quad (a,b) \in A \times B.$$

Combining these inequalities with (2.9), we have

$$V^1_{a,b} = v_{a,b} = V_{a,b}, \quad U_1^{a,b} = u^{a,b} = U^{a,b}, \quad (a,b) \in A \times B.$$

In short form, we have

$$V^1 = v = V, \quad U_1 = u = U.$$

From Proposition 4.1, we also know that u and v are two viscosity solutions of (4.2)–(4.5). By Theorem 4.1, we have $u = v$.

Concluding the above, we have

$$V^1 = v = V = U_1 = u = U.$$

Therefore, our stochastic differential switching game described by (1.1) and (1.2) has a value, and the function $V^1 = v = V = U = u = U^1$ is the unique viscosity solution of (4.2)–(4.5). □

Acknowledgments. The first author thanks the Department of Applied Mathematics, The Hong Kong Polytechnic University for their hospitality during his recent visit to Hong Kong. He is also grateful to both anonymous referees for the very careful reading of the original manuscript and their helpful detailed comments. In particular, a referee pointed out an error in our original arguments.

REFERENCES

[1] D. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[2] I. Capuzzo Dolcetta and L. C. Evans, *Optimal switching for ordinary differential equations*, SIAM J. Control Optim., 22 (1984), pp. 143–161.

[3] R. J. Elliott and N. J. Kalton, *The Existence of Value in Differential Games*, Mem. Amer. Math. Soc. 126, American Mathematical Society, Providence, RI, 1972.

[4] L. C. Evans and A. Friedman, *Optimal stochastic switching and the Dirichlet problem for the Bellman equation*, Trans. Amer. Math. Soc., 253 (1979), pp. 365–389.

[5] L. C. Evans and H. Ishii, *The maximum principle for semicontinuous functions*, Differential Integral Equations, 3 (1990), pp. 1001–1014.

[6] W. H. Fleming and P. E. Souganidis, *On the existence of value functions of two-player, zero-sum stochastic differential games*, Indiana Univ. Math. J., 38 (1989), pp. 293–314.

[7] S. Tang and J. Yong, *Finite horizon stochastic optimal switching and impulse controls with a viscosity solution approach*, Stoch. Stoch. Rep., 45 (1993), pp. 145–176.

[8] N. Yamada, *A system of elliptic variational inequalities associated with a stochastic switching game*, Hiroshima Math. J., 13 (1983), pp. 109–132.

[9] N. Yamada, *Viscosity solutions for a system of elliptic variational inequalities with bilateral obstacles*, Funkcial. Ekvac., 30 (1987), pp. 417–425.

[10] J. Yong, *A zero-sum differential game in a finite duration with switching strategies*, SIAM J. Control Optim., 28 (1990), pp. 1234–1250.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 930–951

A BIRKHOFF CONTRACTION FORMULA WITH APPLICATIONS TO RICCATI EQUATIONS*

JIMMIE LAWSON† AND YONGDO LIM‡

Abstract. In this paper we show that the symplectic Hamiltonian operators on a Hilbert space give rise to linear fractional transformations on the open convex cone of positive definite operators that contract a natural invariant Finsler metric, the Thompson or part metric, on the convex cone. More precisely, the constants of contraction for the Hamiltonian operators satisfy the classical Birkhoff formula: the Lipschitz constant for the corresponding linear fractional transformations on the cone of positive definite operators is equal to the hyperbolic tangent of one fourth the diameter of the image. By means of the close connections between Hamilitonian operators and Riccati equations, this result and the associated machinery are applied to obtain convergence results for discrete algebraic Riccati equations and Riccati differential equations.

Key words. Riccati equation, Birkhoff formula, contraction, symplectic group, control theory, Lie semigroup, Hamiltonian operator, positive definite operator

AMS subject classifications. 49N10, 93B03, 22E15

DOI. 10.1137/050637637

1. Introduction.

Connections between linear control theory, the Riccati equation, and the symplectic group are well known; see, for example, Hermann [13], Shayman [22], Jurdjevic [14, Chapter 8], and [23], and the references cited in those sources. In this paper we focus particularly on connections to the symplectic subsemigroup, which consists of those symplectic transformations that are sometimes called Hamiltonian. In [15] we studied in some detail this subsemigroup of symplectic operators in the infinite dimensional setting and its close connection to Riccati differential equations arising in linear control systems. The canonical triple factorization of symplectic Hamiltonian operators and their action via linear fractional transformation on the open convex cone $\mathcal{P}_0$ of positive definite operators on a Hilbert space have played key roles in the study of Riccati equations via Lie semigroup theory. In this paper we study the contraction property of symplectic Hamiltonian operators acting on the convex cone $\mathcal{P}_0$ for the natural *invariant* Finsler metric (Thompson's part metric), and apply it to finite- and infinite dimensional discrete algebraic Riccati equations and Riccati differential equations.

One of our main results is the Birkhoff theorem (section 5) for symplectic Hamiltonian operators with respect to Thompson's metric $p(X, Y)$: each symplectic Hamiltonian $g = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$, regarded as the self map on $\mathcal{P}_0$ given by the linear fractional transformation

$$g.X = (AX + B)(CX + D)^{-1},$$

*Received by the editors August 5, 2005; accepted for publication (in revised form) December 28, 2006; published electronically June 19, 2007.
http://www.siam.org/journals/sicon/46-3/63763.html
†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 (lawson@math.lsu.edu).
‡Department of Mathematics, Kyungpook National University, Taegu 702-701, Korea (ylim@knu.ac.kr). The work of this author was supported by grant R01-2006-000-10211-0 from the Basic Research Program of the Korea Science and Engineering Foundation.

satisfies the contraction formula

$$\sup_{\substack{X,Y\in\mathcal{P}_0\\ X\neq Y}} \frac{p(g(X),g(Y))}{p(X,Y)} = \tanh\left(\frac{\operatorname{diam}(g)}{4}\right),$$

where $\operatorname{diam}(g)$ denotes the diameter of the image $g(\mathcal{P}_0)$ for the Thompson's metric. The diameter is completely determined by $\operatorname{diam}(g) = p(BD^{-1}, AC^{-1})$ when both BD^{-1} and AC^{-1} are positive definite; otherwise $\operatorname{diam}(g) = \infty$ (Theorem 5.8). This beautiful and important formula had its origin with Birkhoff [4] for Möbius transformations with positive entries with respect to the Riemannian metric $p(a,b) = |\log a - \log b|$ on the positive reals. Liverani and Wojtkowski [18] and Lim [16] have generalized it to fractional transformations on the symmetric cone of positive definite matrices and on symmetric cones arising from Euclidean Jordan algebras with respect to the invariant Finsler metric associated with the spectral norm. In the linear setting, the Birkhoff formula for positive linear maps on Banach spaces for Hilbert's projective (pseudo)metric [5] is well known, with many applications in analysis [4], [8], [17]; see also [20], [21] and the references therein. It has also found applications in control theory, primarily in filtering theory; see, e.g., [3], [7].

In the connections between linear control theory, the Riccati equation, and symplectic Hamiltonians, the contraction property of symplectic Hamiltonians with explicitly given contraction coefficient is applied to the iterative method of solution for discrete algebraic Riccati equations,

$$X = A^*XA - A^*XB(R + B^*XB)^{-1}B^*XA + H,$$

and to the asymptotic behavior of solutions of the Riccati differential equation,

$$\dot{K}(t) = R(t) + A(t)K(t) + K(t)A^*(t) - K(t)S(t)K(t),$$

on an arbitrary Hilbert space. Bougerol [6] has proved that symplectic Hamiltonian matrices are contractions for the standard Riemannian metric on the symmetric space of positive definite matrices and given applications to Kalman filtering theory (cf. [12], [9]). However, in the Riemannian metric case, there is no explicit formula for the contraction coefficient of Hamiltonian matrices. In section 7, we prove that under the invertiblity condition of A and $BR^{-1}B^*$, the discrete Riccati equation has a unique positive solution X_∞ approached by any iteration $X_n \in \mathcal{P}_0$ with the rate of convergence determined by the *computable* Birkhoff constant with respect to Thompson's part metric. Using the *best estimation* given by the Birkhoff constant on the Lie wedge of the symplectic Hamiltonian semigroup studied in section 6, we prove in section 8 that, under the uniform boundedness condition $S^{1/2}(t)R(t)S^{1/2}(t) \geq \mu I$, the solution $K(t)$ of the Riccati differential equation with $K(0) \in \mathcal{P}_0$ is exponentially attracting for Thompson's part metric. These results, obtained mainly from the Birkhoff formula and the invariant Finsler metric, provide new techniques for study of Riccati equations, even for the finite dimensional case, where illustrative numerical experiments can be calculated.

2. Symplectic Hamiltonian operators. In this section we review some basic material on the algebraic structure of the symplectic Lie group and the associated symplectic Hamiltonian semigroup from [15].

Let E be a real Hilbert space with inner product $\langle\cdot,\cdot\rangle\colon E\times E\to\mathbb{R}$, and let $V_E = E\oplus E$. We denote members of V_E by column vectors $\left[\begin{smallmatrix} x\\ y\end{smallmatrix}\right]$, where $x, y \in E$. The

standard symplectic form Q on V_E is defined by

$$Q\left(\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}\right) := \langle x_1, y_2\rangle - \langle y_1, x_2\rangle.$$

We denote by $\mathrm{End}(V_E)$ (resp., $\mathrm{End}(E)$) the set of bounded linear operators on V_E (resp., E), and by $\mathrm{GL}(V_E)$ (resp., $\mathrm{GL}(E)$) those that are invertible. We shall always assume that the topology is generated by the operator norm. For a bounded linear transformation A on E, let A^* denote the unique linear operator such that $\langle Ax, y\rangle = \langle x, A^*y\rangle$ for all x, y in E. We call A^* the *adjoint* of A. We say that A is *symmetric* if $A^* = A$. A bounded symmetric operator A on E is *positive semidefinite* if $\langle x, Ax\rangle \geq 0$ for all $x \in E$. We denote by $\mathcal{P}$ (resp., $\mathcal{P}_0$) all positive semidefinite (resp., positive semidefinite invertible) bounded operators on E.

For (V_E, Q) a standard sympletic space, the symplectic Lie group is defined by

$$\mathrm{Sp}(V_E) := \{M \in \mathrm{GL}(V_E) \colon \forall x, y \in V_E,\ Q(Mx, My) = Q(x, y)\}$$

and has the following characterizations.

Proposition 2.1 (see Proposition 2.5 of [15]). *Let $M \in GL(V_E)$. The following are equivalent:*

1. *$M \in Sp(V_E)$; i.e., M preserves $Q(\cdot, \cdot)$.*
2. *$M^*JM = J$, where $J = \left[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right] \in End(V_E)$.*
3. *If M has block matrix form $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$, then*
   (a) *A^*C, B^*D are symmetric;*
   (b) *$A^*D - C^*B = I$.*

Members of $\mathrm{Sp}(V_E)$ viewed as linear operators on V_E are called *linear sympletic maps.*

Recall that the symplectic Lie algebra $\mathfrak{sp}(V_E)$ consists of all $X \in \mathrm{End}(V_E)$ such that $\exp(tX) \in \mathrm{Sp}(V_E)$ for all $t \in \mathbb{R}$.

Proposition 2.2. *Let $X \in End(V_E)$. The following are equivalent:*

1. *$X \in \mathfrak{sp}(V_E)$.*
2. *$X^*J + JX = 0$.*
3. *If $X = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$, then*
   (a) *B and C are symmetric;*
   (b) *$D = -A^*$.*

We consider four subsets of $\mathrm{Sp}(V_E)$:

$$\mathcal{S} = \left\{\begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathrm{Sp}(V_E) : D \text{ is invertible, } B^*D \in \mathcal{P},\ CD^* \in \mathcal{P}\right\},$$

$$\mathcal{S}_1 = \left\{\begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathrm{Sp}(V_E) : D \text{ is invertible, } B^*D \in \mathcal{P}_0,\ CD^* \in \mathcal{P}\right\},$$

$$\mathcal{S}_2 = \left\{\begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathrm{Sp}(V_E) : D \text{ is invertible, } B^*D \in \mathcal{P},\ CD^* \in \mathcal{P}_0\right\},$$

$$\mathcal{S}_0 = \mathcal{S}_1 \cap \mathcal{S}_2.$$

We define

$$\Gamma^U = \left\{\begin{bmatrix} I & B \\ 0 & I \end{bmatrix} : B \in \mathcal{P}\right\}, \qquad \Gamma_0^U = \left\{\begin{bmatrix} I & B \\ 0 & I \end{bmatrix} : B \in \mathcal{P}_0\right\},$$

$$\Gamma^L = \left\{ \begin{bmatrix} I & 0 \\ C & I \end{bmatrix} : C \in \mathcal{P} \right\}, \qquad \Gamma_0^L = \left\{ \begin{bmatrix} I & 0 \\ C & I \end{bmatrix} : C \in \mathcal{P}_0 \right\}.$$

We further define a group H of block diagonal matrices by

$$H = \left\{ \begin{bmatrix} A^* & 0 \\ 0 & A^{-1} \end{bmatrix} : A \in \mathrm{GL}(E) \right\}.$$

THEOREM 2.3. *We have that $\mathcal{S}$ is a subsemigroup of $Sp(V_E)$ and $\mathcal{S}\mathcal{S}_i\mathcal{S} \subseteq \mathcal{S}_i$ for $i = 0, 1, 2$; i.e., $\mathcal{S}_i$ is a semigroup ideal. We alternatively have that $\mathcal{S} = \Gamma^U H \Gamma^L$, $\mathcal{S}_1 = \Gamma_0^U H \Gamma^L$, $\mathcal{S}_2 = \Gamma^U H \Gamma_0^L$, and $\mathcal{S}_0 = \Gamma_0^U H \Gamma_0^L$. Furthermore these "triple decompositions" are unique: the multiplication mapping from $\Gamma^U \times H \times \Gamma^L$ to $\mathcal{S}$ is a homeomorphism.*

Proof. The proof follows from Theorem 6.7 and [15, Lemmas 6.4, 6.5]. See also [6], [9]. □

The unique triple factorization of a symplectic Hamiltonian $M = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix} \right] \in \mathcal{S}$ is given by

$$M = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} (D^{-1})^* & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix}.$$

This factorization occurs more generally for any member $M \in \mathrm{Sp}(V_E)$ with invertible $(2,2)$-entry. The semigroup $\mathcal{S}$ of the preceding theorem is called the *sympletic semigroup*, and members of $\mathcal{S}$ are sometimes called *Hamiltonian* operators of $\mathrm{Sp}(V_E)$.

3. Fractional transformations and compressions. In this section we show that Hamiltonians arise exactly as compressions of the open convex cone of positive definite operators under the canonical fractional transformation action.

We consider the lower block triangular subgroup $\mathbf{P}$ of $\mathrm{Sp}(V_E)$ given by

$$\mathbf{P} := \left\{ \begin{bmatrix} A & 0 \\ C & D \end{bmatrix} \in \mathrm{Sp}(V_E) : A, C, D \in \mathrm{End}(E) \right\}.$$

We note from Proposition 2.1 that such a lower triangular block matrix is in $\mathrm{Sp}(V_E)$ if and only if $A^* = D^{-1}$ and $A^*C = D^{-1}C$ is symmetric. We denote by $\mathcal{M}$ the homogeneous space

$$\mathcal{M} := \mathrm{Sp}(V_E)/\mathbf{P}.$$

In the finite dimensional setting, $\mathbf{P}$ is a parabolic subgroup and the homogeneous space is a flag manifold of $\mathrm{Sp}(V_E)$. The set $\mathrm{Sym}(E)$ of symmetric operators in $\mathrm{End}(E)$ is embedded into $\mathcal{M}$ as a dense open subset (see Lemma 9.2 of [15]):

$$X \mapsto \left[\begin{smallmatrix} I & X \\ 0 & I \end{smallmatrix} \right] \mathbf{P} \in \mathcal{M}.$$

If $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathrm{Sp}(V_E)$ and $X \in \mathrm{Sym}(E)$ such that $CX + D$ is invertible, then

$$M \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} \mathbf{P} = \begin{bmatrix} I & (AX + B)(CX + D)^{-1} \\ 0 & I \end{bmatrix} \mathbf{P}.$$

This defines the (partial) action by fractional transformations of $\mathrm{Sp}(V_E)$ on $\mathrm{Sym}(E) \subseteq \mathcal{M}$:

$$M.X = (AX + B)(CX + D)^{-1} \quad \text{if } (CX + D)^{-1} \text{ exists.} \tag{3.1}$$

For $X, Y \in \mathrm{Sym}(E)$, we define

$$X < Y :\Longleftrightarrow Y - X \in \mathcal{P}_0,$$
$$X \leq Y :\Longleftrightarrow Y - X \in \mathcal{P}.$$

The order $\leq$ is sometimes called the *Loewner order*. For $X \leq Y$ (resp., $X < Y$) we define the order intervals

$$[X, Y] = \{Z \in \mathrm{Sym}(E) : X \leq Z \leq Y\},$$
$$(X, Y) = \{Z \in \mathrm{Sym}(E) : X < Z < Y\},$$

respectively.

PROPOSITION 3.1 (see Propositions 9.6 and 9.7 of [15]). *The sets* $\{(-(1/n)A, (1/n)A) : n \in \mathbb{N}\}$ *form a basis of open sets at* 0 *in* $Sym(E)$ *for any* $A \in \mathcal{P}_0$. *For an element* $A \in Sym(E)$, *the following are equivalent:*

1. $A \in \mathcal{P}$;
2. $A + X$ *is invertible for all* $X \in \mathcal{P}_0$;
3. $A + rI$ *is invertible for all* $r > 0$.

PROPOSITION 3.2 (see Propositions 9.6 and 9.9 of [15]). *Each order interval* $[A, B] = \{X \in Sym(E) : A \leq X \leq B\}$ *for* $A \leq B$ *is closed in* $\mathcal{M}$, *the interior of* $[A, B]$ *is equal to* (A, B), *and the closure* $\overline{\mathcal{P}}$ *of* $\mathcal{P}$ *in* $\mathcal{M}$ *has interior* $\mathcal{P}_0$.

Let us call a member of $\mathrm{Sp}(V_E)$ a *compression* if it carries $\mathcal{P}_0$ into itself under the action of fractional transformation (3.1).

LEMMA 3.3. *If* $M = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in Sp(V_E)$ *is a compression and the image of* $0_E \in \mathrm{Sym}(E) \subseteq \mathcal{M}$ *under* M *is in* $\mathcal{P}$, *then* M *belongs to the sympletic semigroup* $\mathcal{S}$.

Proof. The image of 0_E under M is BD^{-1}. This means that D is invertible, and hence M has a triple decomposition in $\mathrm{Sp}(V_E)$ of the form

$$M = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} (D^{-1})^* & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix}.$$

Since $0_E \in \mathcal{M}$ corresponds to $\mathbf{P}$ in $\mathrm{Sp}(V_E)/\mathbf{P}$, we conclude that the last two factors of M applied to it return 0_E. Thus, by (3.1),

$$M.0_E = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix}.0 = BD^{-1}.$$

Since the latter is in $\mathcal{P}$ by hypothesis, we conclude that BD^{-1} is positive semidefinite, and hence that the first factor of M is in $\mathcal{S}$. The second factor is trivially in $\mathcal{S}$.

Let $X \in \mathcal{P}_0$. Then

$$\begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix}.X = \begin{bmatrix} D^* & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix} M.X,$$

where the right-hand side must be in $\mathrm{Sym}(E) \subseteq \mathcal{M}$. It follows that $D^{-1}CX + I$ is invertible for all $X \in \mathcal{P}_0$. Since X is invertible, $(D^{-1}CX + I)X^{-1} = D^{-1}C + X^{-1}$ is invertible for all $X \in \mathcal{P}_0$. It then follows from Proposition 3.1 that $D^{-1}C$ is in $\mathcal{P}$. Thus the third factor of M is also in $\mathcal{S}$. ∎

THEOREM 3.4. *Let* $M \in Sp(V_E)$. *The following are equivalent:*

1. $M.\mathcal{P}_0 \subseteq \overline{\mathcal{P}}$;
2. M *is a compression;*

3. $M \in \mathcal{S}$.

Proof. If $M.\mathcal{P}_0 \subseteq \overline{\mathcal{P}}$, then since M is a homeomorphism, it must carry $\mathcal{P}_0$ into $\text{int}\overline{\mathcal{P}}$, which by Proposition 3.2 is $\mathcal{P}_0$. Thus M is a compression. The converse is immediate. Hence items 1 and 2 are equivalent.

It turns out that elements of $\mathcal{S}$ carry $\mathcal{P}_0$ into itself (Proposition 7.1 of [15]). Conversely suppose that $M.\mathcal{P}_0 \subseteq \mathcal{P}_0$. Define $M_n = M \circ t_n$, where t_n has matrix representation $\left[\begin{smallmatrix} I & (1/n)I \\ 0 & I \end{smallmatrix}\right]$. Since $t_n.0 = (1/n)I$, we conclude that $M_n.0 \in \mathcal{P}_0$. Hence by the preceding lemma, $M_n \in \mathcal{S}$.

Let $A := M.I$. By hypothesis we may write the result in this form with $A \in \mathcal{P}_0$. Since A is in the open order interval $(0, 2A)$, we have for n large enough that $M_n.I \in (0, 2A)$. Since M_n is order-preserving (Proposition 3.5), we have

$$0 \leq M_n.0 \leq M_n.I \leq 2A.$$

Since the interval $[0, 2A]$ is closed in $\mathcal{M}$ (Proposition 3.2), we conclude that

$$M.0 = \lim_n M_n.0 \in [0, 2A].$$

We can now apply the preceding lemma to M to conclude that $M \in \mathcal{S}$. □

PROPOSITION 3.5 (see Proposition 9.10 of [15]). *Members of the symplectic semigroup $\mathcal{S}$ satisfy the following monotonicity properties:*

1. *For $g \in \mathcal{S}$ and $X, Y \in \mathcal{P}_0$, $X \leq Y$ if and only if $g(X) \leq g(Y)$.*
2. *For $g \in \mathcal{S}$ and $X, Y \in \mathcal{P}$, $X \leq Y$ implies $g(X) \leq g(Y)$.*

4. Hamiltonian operators and the standard sector. There is an alternative context in which Hamiltonian operators arise naturally. We consider the quadratic form $\mathcal{Q}$ on the symplectic space (V_E, Q),

$$\mathcal{Q}(w) = \langle x, y \rangle, \qquad w = \begin{bmatrix} x \\ y \end{bmatrix} \in V_E,$$

and the *standard sector* of the symplectic space (V_E, Q), which is defined by

$$\mathcal{C} = \{w \in V_E : \mathcal{Q}(w) \geq 0\}.$$

By $\mathcal{C}^\circ$ we denote the interior of $\mathcal{C}$:

$$\mathcal{C}^\circ = \{w \in V_E : \mathcal{Q}(w) > 0\}.$$

By continuity $g(\mathcal{C}^\circ) \subset \mathcal{C}^\circ$ for any $\mathcal{Q}$-monotone g. Each member of H, the subgroup of block diagonals in $\text{Sp}(V_E)$, acts as an $\mathcal{Q}$-isometry.

The following is immediate from the triple decompositions of $\mathcal{S}$ and $\mathcal{S}_0$ (Theorem 2.3) and the preservation of (strict) $\mathcal{Q}$-monotonicity under composition.

THEOREM 4.1. *Each member of $\mathcal{S}$ (resp., $\mathcal{S}_0$) is $\mathcal{Q}$-monotone (resp., strictly $\mathcal{Q}$-monotone). Furthermore, each member of $\mathcal{S}$ (resp., $\mathcal{S}_0$) increases (resp., strictly increases) the quadratic form.*

Remark. The quadratic form $\mathcal{Q}$ and the associated sector $\mathcal{C}$ define two natural closed subsemigroups in $\text{Sp}(V_E)$ containing the symplectic semigroup $\mathcal{S}$: the subsemigroup of (strictly) monotone maps and the subsemigroup of (strictly) symplectic maps (strictly) increasing the quadratic form. It turns out that these are all the same in the finite dimensional case [19].

We derive an explicit relationship between the action of symplectic Hamiltonians on the sector $\mathcal{C}^\circ$ and the Möbius action of fractional transformation on the positive definite cone $\mathcal{P}_0$.

LEMMA 4.2. *Let* $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] \in \mathcal{C}^\circ$. *Then there exists a positive definite operator* P *on* E *such that* $y = Px$. *In particular,*

$$\mathcal{C}^\circ = \left\{ \begin{bmatrix} x \\ Px \end{bmatrix} : x \neq 0, P \in \mathcal{P}_0 \right\}.$$

Proof. Let W be the subspace generated by x and y. If x and y are linearly dependent, then $y = \lambda x$ for some $\lambda > 0$, so we may take $P = \lambda I$. Suppose that W is two-dimensional. Then it is enough to construct a positive definite operator A on W sending x into y by observing that $P := \begin{bmatrix} A & 0 \\ 0 & I_{W^\perp} \end{bmatrix}$ is positive definite.

Suppose that $x = (x_1, x_2), y = (y_1, y_2) \in \mathbb{R}^2$ such that $l := x_1y_1 + x_2y_2 > 0$. We will solve the equations $ax_1 + bx_2 = y_1, bx_1 + dx_2 = y_2$ with $a > 0, ad > b^2$.

Case 1. If $x_1 = 0$, then take $b = y_1/x_2, d = y_2/x_2 > 0$, and a (positive) large enough. If $x_2 = 0$, then take $b = y_2/x_1, a = y_1/x_1 > 0$, and d large enough.

Case 2. $x_1 \neq 0$ and $x_2 \neq 0$: If $x_1x_2 > 0$, then take

$$b < \min\left\{ \frac{y_1}{x_2}, \frac{y_1y_2}{l} \right\}, \quad a = \frac{(y_1 - bx_2)}{x_1}, \quad d = \frac{(y_2 - bx_1)}{x_2}.$$

If $x_1x_2 < 0$, then take $b > \max\{y_1/x_2, y_1y_2/l\}$, $a = (y_1 - bx_2)/x_1$, and $d = (y_2 - bx_1)/x_2$. ∎

A *slice* of the sector $\mathcal{C}^\circ$ consists of sets of the form

$$\mathcal{P}_x = \left\{ P_x := \begin{bmatrix} Px \\ x \end{bmatrix} : P \in \mathcal{P}_0 \right\}.$$

The preceding lemma shows that the sector $\mathcal{C}^\circ$ is the disjoint union of slices.

PROPOSITION 4.3. *Let* $g = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathcal{S}$. *Then for* $P > 0$,

$$g(P_x) = (g.P)_{(CP+D)x}.$$

Proof. We calculate that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} Px \\ x \end{bmatrix} = \begin{bmatrix} (AP + B)x \\ (CP + D)x \end{bmatrix} = \begin{bmatrix} (AP + B)(CP + D)^{-1}y \\ y \end{bmatrix} = \begin{bmatrix} (g.P)y \\ y \end{bmatrix},$$

where $y = (CP + D)x$. ∎

5. Contractions and the Birkhoff formula. In this section, we show that each element of $\mathcal{S}$ (resp., $\mathcal{S}_0$) is a contraction (resp., strict contraction) of $\mathcal{P}_0$ for a natural invariant metric on it, with an explicit contraction constant given by the Birkhoff formula.

For $A, B \in \mathcal{P}_0$, we define

$$M(A/B) := \inf\{t > 0 : A \leq tB\},$$
$$m(A/B) := \sup\{t > 0 : tB \leq A\}.$$

Then $M(A/B) = m(B/A)^{-1}$. *Thompson's metric* (sometimes called the *part metric*) on $\mathcal{P}_0$ is defined by

$$p(A, B) = \log(\max\{M(A/B), M(B/A)\});$$

see, e.g., [24], [25], [20].

LEMMA 5.1. *The set $\mathcal{P}_0$ becomes a complete metric space with respect to the metric p, and the metric p induces the topology of $\mathcal{P}_0$.*

Proof. The space $\mathrm{Sym}(E)$ of symmetric operators equipped with the operator norm is a Banach space satisfying that $0 \le A \le B$ implies $\|A\| \le \|B\|$. It follows from Proposition 3.1 that for $A, B \in \mathcal{P}_0$ there exists $t \in \mathbb{R}$ such that $A \le tB$. It then follows from Lemma 3 of [24] that the Thompson metric p is indeed a metric and is complete on $\mathcal{P}_0$, and by Proposition 1.1 of [21] that the Thompson metric induces the same topology. □

LEMMA 5.2. *The metric p is invariant under the block diagonal group H and inversion $j(A) = A^{-1}$.*

Proof. The lemma follows directly from the observations

$$\forall D \in \mathrm{GL}(E), M(D^*AD/D^*BD) = M(A/B), \quad M(A^{-1}/B^{-1}) = M(B/A),$$

where the last equality follows from the fact that inversion on $\mathcal{P}_0$ is order-reversing (cf. Proposition 9.8 of [15]). □

A map $\gamma : [0, 1] \to \mathcal{P}_0$ is said to be a *minimal geodesic* for the metric p if, whenever $0 \le t_1 \le t_2 \le 1$, we have

$$p(\gamma(t_1), \gamma(t_2)) = (t_2 - t_1)p(\gamma(0), \gamma(1)).$$

PROPOSITION 5.3 (see Proposition 1.10 of [20]). *Let $A, B \in \mathcal{P}_0$. Then*

$$\gamma(t) = A^{1/2}(A^{-1/2}BA^{-1/2})^t A^{1/2}$$

is a minimal geodesic curve from A to B with respect to p.

For $X \in \mathrm{Sym}(E)$, the *order unit norm* for the order unit I is given by

$$\|X\| = \inf\{t > 0 : -tI \le X \le tI\}.$$

LEMMA 5.4. *On $\mathrm{Sym}(E)$ we have the following:*

1. *The order unit norm agrees with the operator norm on $\mathrm{Sym}(E)$.*
2. *For $X \in \mathcal{P}_0$, $m(X/I) = \|X^{-1}\|^{-1}$.*
3. *The map $X \mapsto m(X/I)$ is continuous on $\mathcal{P}$.*

Proof. Part 1. Let us temporarily denote the order unit norm by $\|X\|_{or}$. Then

$$|\langle x, Xx\rangle| \le \|x\|(\|X\|\,\|x\|) = \|X\|\langle x, Ix\rangle = \langle x, \|X\|Ix\rangle$$

implies that $\|X\|_{or} \le \|X\|$. For $\|x\| = 1$ and $X \ge 0$, we have

$$\|X^{1/2}x\|^2 = \langle X^{1/2}x, X^{1/2}x\rangle = \langle x, Xx\rangle \le \langle x, \|X\|_{or}Ix\rangle = \|X\|_{or}.$$

It follows that $\|X\| \le \|X^{1/2}\|^2 \le \|X\|_{or}$. We then have for arbitrary symmetric X

$$\|X\|^2 = \|X^*X\| = \|X^2\| = \|X^2\|_{or} \le \|X\|_{or}^2,$$

since $-tI \le X \le tI$ implies that $t^2I - X^2 = (tI + X)^{1/2}(tI - X)(tI + X)^{1/2} \ge 0$.

2. For $X \in \mathcal{P}_0$, we have directly that

$$\begin{aligned} m(X/I) &= \sup\{t > 0 : tI \le X\} = \sup\{t > 0 : (1/t)I \ge X^{-1}\} \\ &= \sup\{(1/s) > 0 : X^{-1} \le sI\} = \|X^{-1}\|^{-1}. \end{aligned}$$

3. It follows from part 2 that the function $X \mapsto m(X/I)$ is continuous on $\mathcal{P}_0$. For $X \in \mathcal{P}$, let $0 \le X_n \to X$. Then for $\varepsilon > 0$, $X_n + \varepsilon I \to X + \varepsilon I$ in $\mathcal{P}_0$, which in turn implies $m(X_n + \varepsilon I/I) \to m(X + \varepsilon I/I)$. Since $m(A + \varepsilon I/I) = m(A/I) + \varepsilon$ for $A \in \mathcal{P}$, the desired conclusion follows. □

Remarks. (1) The map $X \mapsto m(X/I)$ on $\mathcal{P}_0$ is one of special interest; it agrees with the smallest eigenvalue function in the finite dimensional case.

(2) For $X \in \mathcal{P}_0$, $||X|| = M(X/I)$, and hence $p(I, X) = \max\log\{||X||, ||X^{-1}||\}$. Thus for $X, Y \in \mathcal{P}_0$,

$$p(X, Y) = p(I, X^{-1/2}YX^{-1/2}) = \log\max\{||X^{-1/2}YX^{-1/2}||, ||X^{1/2}Y^{-1}X^{1/2}||\}.$$

Identifying the tangent bundle $T\mathcal{P}_0$ of $\mathcal{P}_0$ with $\mathcal{P}_0 \times \mathrm{Sym}(E)$, we define a Finsler structure on $\mathcal{P}_0$ by

$$|X|_A := ||A^{-1/2}XA^{-1/2}||$$

for $A \in \mathcal{P}_0$, $X \in \mathrm{Sym}(E)$. Then it is easy to see that $|\cdot|_A$ is a norm on the tangent space $\mathrm{Sym}(E)$ at A.

THEOREM 5.5 (see Theorem 1.1 of [21]). *Let $A, B \in \mathcal{P}_0$. Then*

$$p(A, B) = \inf\left\{\int_0^1 |\psi'(t)|_{\psi(t)}dt\right\},$$

where the infimum is taken over all piecewise C^1 maps ψ from $A = \psi(0)$ to $B = \psi(1)$. In particular,

$$p(A, B) = \int_0^1 |\gamma'(t)|_{\gamma(t)}dt,$$

where $\gamma(t) = A^{1/2}(A^{-1/2}BA^{-1/2})^tA^{1/2}$.

For notational convenience, we denote for $A \in \mathcal{P}$ and $D \in \mathrm{GL}(E)$,

$$\begin{aligned} t_A &:= \begin{bmatrix} I & A \\ 0 & I \end{bmatrix} \in \Gamma^U, \\ \tilde{t}_A &:= \begin{bmatrix} I & 0 \\ A & I \end{bmatrix} \in \Gamma^L, \\ h_D &:= \begin{bmatrix} D^* & 0 \\ 0 & D^{-1} \end{bmatrix} \in H. \end{aligned}$$

Then under the action of fractional transformation (3.1),

$$t_A(B) = A + B, \quad h_D(B) = D^*BD, \quad \tilde{t}_A(B) = (A + B^{-1})^{-1} = (jt_Aj)(B)$$

for $B \in \mathcal{P}_0$, where $j(A) = A^{-1}$, the inversion operator on $\mathcal{P}_0$.

PROPOSITION 5.6. *Let $X, Y \in \mathcal{P}_0$ and let $D \in \mathrm{GL}(E)$. Then*

$$t_X \circ h_D \circ \tilde{t}_Y = h_{Y^{-1/2}D} \circ t_{Y^{1/2}(D^{-1})^*XD^{-1}Y^{1/2}} \circ \tilde{t}_I \circ h_{Y^{1/2}}.$$

Proof. The proof is straightforward. □

Set $\infty := \left[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right]\mathbf{P} \in \mathcal{M}$. It is easy to see that for $g \in \mathrm{Sp}(V_E)$, $g \cdot \infty = \infty$ if and only if g is an upper triangular block matrix.

LEMMA 5.7. *Let $0 < A \le B$. If $X, Y \in [A, B]$, then $p(X, Y) \le p(A, B)$.*

Proof. Suppose that $p(X, Y) = \log M(X/Y)$. Since $A \le X \le M(X/A)A \le M(X/A)Y$, we have $M(X/Y) \le M(X/A)$. The fact $A \le X$ implies that $m(X/A) \ge 1$ and hence $M(A/X) = m(X/A)^{-1} \le 1$. Thus $M(X/A) \ge 1$. Therefore

$$p(X, Y) = \log M(X/Y) \le \log M(X/A) = p(A, X). \tag{5.1}$$

Now, $X \le B \le M(B/A)A$ implies that $M(X/A) \le M(B/A)$ and hence by (5.1)

$$p(X, A) = \log M(X/A) \le \log M(B/A) \le p(B, A).$$

Therefore $p(X, Y) \le p(A, B)$. Similarly, we have that $p(X, Y) \le p(A, B)$ when $p(X, Y) = \log M(X/Y)$. □

THEOREM 5.8. *Let $g \in \mathcal{S}_0$. Then $g(\mathcal{P}_0) = (g(0), g(\infty))$, $g(\overline{\mathcal{P}}) = [g(0), g(\infty)] \subseteq \mathcal{P}_0$, and the diameter $\Delta(g)$ of $g(\mathcal{P}_0)$ for the metric p is the distance $p(g(0), g(\infty))$. If $g \in \mathcal{S} \setminus \mathcal{S}_0$, then $\Delta(g) = \infty$.*

Proof. Let $g = t_X \circ h_D \circ \tilde{t}_Y \in \mathcal{S}_0$. By Theorem 2.3, $X, Y \in \mathcal{P}_0$. Then $g(0) = X$ and $g(\infty) = X + D^*Y^{-1}D$. Suppose that $Z \in (X, X + D^*Y^{-1}D)$. Then $Z = X + A = X + D^*Y^{-1}D - B$ for some $A, B \in \mathcal{P}_0$. Note that $A = D^*Y^{-1}D - B$, so $A < D^*Y^{-1}D$. Since the inversion j is order-reversing on $\mathcal{P}_0$ (cf. Proposition 9.8 of [15]), $W := (DA^{-1}D^* - Y)^{-1} \in \mathcal{P}_0$. This implies that $Z = X + A = g(W) \in g(\mathcal{P}_0)$. Conversely, suppose that $Z \in g(\mathcal{P}_0)$. Then $Z = g(W) = X + D^*(Y + W^{-1})^{-1}D$ for some $W \in \mathcal{P}_0$. It is obvious that $X < Z$. Since $W^{-1} \in \mathcal{P}_0$, we have that $Y^{-1} > (Y + W^{-1})^{-1}$. Thus $D^*Y^{-1}D > D^*(Y + W^{-1})^{-1}D$. This implies that

$$g(\infty) - Z = (X + D^*Y^{-1}D) - (X + D^*(Y + W^{-1})^{-1}D) > 0.$$

Therefore $Z \in (g(0), g(\infty))$. So, $g(\mathcal{P}_0) = (g(0), g(\infty))$. The second assertion follows from this, Proposition 3.1, the fact that g acts as a homeomorphism on $\mathcal{M}$, and our computation of $g(0)$ and $g(\infty)$.

That the diameter of $g(\mathcal{P}_0) = (g(0), g(\infty))$ is the Thompson distance $p(g(0), g(\infty))$ follows from the preceding lemma.

Suppose that $g = t_A \circ h_D \circ \tilde{t}_B \in \mathcal{S} \setminus \mathcal{S}_0$. Then by Theorem 2.3 either A or B lies in $\mathcal{P} \setminus \mathcal{P}_0$. Suppose that $A \in \mathcal{P} \setminus \mathcal{P}_0$. Pick $C \in g(\mathcal{P}_0)$ with $C > 0$. Let $Y_n = g(\frac{1}{n}I) \in \mathcal{P}_0$. Then $Y_n \to g(0) = A$. Since $g(0) = A \in \mathcal{P} \setminus \mathcal{P}_0$, for each $k > 0$, there exists $n_k > 0$ such that $Y_{n_k} \notin [\frac{1}{k}C, kC]$, that is, $Y_{n_k} \not\le kC$ or $\frac{1}{k}C \not\le Y_{n_k}$. By definition, $M(C/Y_{n_k}) \ge k$ or $M(Y_{n_k}/C) \ge k$. Therefore $\log k \le p(C, Y_{n_k}) \to \infty$, and hence $\Delta(g) = \infty$. Similarly, if $g(\infty) = B \in \mathcal{P} \setminus \mathcal{P}_0$, then $\Delta(g) = \infty$. □

LEMMA 5.9. *Let $A, X \in \mathcal{P}_0$. Then*

$$|(I + A)^{-1}U(I + A)^{-1}|_{X+(I+A^{-1})^{-1}} \le \left(\sqrt{m(X/I)} + \sqrt{1 + m(X/I)}\right)^{-2} |U|_A$$

for all $U \in \mathrm{Sym}(E)$.

Proof. First, we show that

$$(I + A)X(I + A) + A^2 + A \ge \left(\sqrt{m(X/I)} + \sqrt{1 + m(X/I)}\right)^2 A.$$

It immediately follows from $m(X/I)I \le X$ that

$$m(X/I)(I+A)^2 = m(X/I)(I+A)I(I+A) \le (I+A)X(I+A).$$

We then have

$$\begin{aligned}
&(I+A)X(I+A) + A^2 + A \\
&\ge m(X/I)(I+A)^2 + A^2 + A \\
&= (m(X/I)+1)A^2 + m(X/I)I + (2m(X/I)+1)A \\
&\ge 2\sqrt{m(X/I)(m(X/I)+1)}A + (2m(X/I)+1)A \\
&= (\sqrt{m(X/I)} + \sqrt{1+m(X/I)})^2 A,
\end{aligned}$$

where the second inequality follows from the fact that the square of $\sqrt{m(X/I)+1}A - \sqrt{m(X/I)}I$ is positive semidefinite.

Set $k := (\sqrt{m(X/I)} + \sqrt{1+m(X/I)})^{-2}$. Then $-tI \le kA^{-1/2}UA^{-1/2} \le tI$ for some $t > 0$, or equivalently $(-t/k)A \le U \le (t/k)A$. From the first paragraph, we obtain that

$$-t\big((I+A)X(I+A) + A^2 + A\big) \le \frac{-t}{k}A \le U \le \frac{t}{k}A \le t((I+A)X(I+A) + A^2 + A).$$

Since $(I+A)^{-1}(A^2+A)(I+A)^{-1} = (I+A^{-1})^{-1}$, this implies that

$$-t(X + (I+A^{-1})^{-1}) \le (I+A)^{-1}U(I+A)^{-1} \le t(X + (I+A^{-1})^{-1})$$

and hence

$$-tI \le (X + (I+A^{-1})^{-1})^{-1/2}(I+A)^{-1}U(I+A)^{-1}(X + (I+A^{-1})^{-1})^{-1/2} \le tI.$$

Therefore from the definition of the order unit norm,

$$||(X+(I+A^{-1})^{-1})^{-1/2}(I+A)^{-1}U(I+A)^{-1}(X+(I+A^{-1})^{-1})^{-1/2}|| \le k||A^{-1/2}UA^{-1/2}||,$$

and the lemma follows immediately. □

LEMMA 5.10. *Let $X \in \mathcal{P}_0$ and let $0 < \alpha < \beta$. Then*

$$M(X + \beta I/X + \alpha I) \ge \frac{m(X/I) + \beta}{m(X/I) + \alpha} \ge 1.$$

In particular,

$$M\left(X + \frac{\beta}{\beta+1}I/X + \frac{\alpha}{\alpha+1}\right) \ge \frac{m(X/I) + \frac{\beta}{\beta+1}}{m(X/I) + \frac{\alpha}{\alpha+1}} \ge 1.$$

Proof. If $X + \beta I \le t(X + \alpha I)$ for $t > 0$, then

$$m(X/I) + \beta = m(X + \beta I/I) \le m(t(X+\alpha I)/I) = t(m(X/I) + \alpha). \qquad \square$$

Let us introduce the Lipschitz constant (the least coefficient of contraction) of $g \in \mathcal{S}$,

$$N(g) = \sup_{\substack{A,B \in \mathcal{P}_0 \\ A \ne B}} \frac{p(g(A), g(B))}{p(A,B)}.$$

Note that $N(g_1g_2) \le N(g_1)N(g_2)$.

THEOREM 5.11. *Let $g \in \mathcal{S}$. Then*

$$N(g) = \tanh\left(\frac{\Delta(g)}{4}\right).$$

Proof. Let $g \in \mathcal{S}_0$. Note that $N(g) = N(h \circ g \circ h')$ for any $h, h' \in H$ by the H-invariance of the metric. By Proposition 5.6, we may assume that $g = t_X \circ \tilde{t}_I$ for some $X \in \mathcal{P}_0$. Then $g(0) = X, g(\infty) = X + I$, and hence $\Delta(g) = p(X, X+I) = \log M(X+I/X) = \log(1+M(I/X)) = \log(1+\frac{1}{m(X/I)})$. A straightforward calculation yields

$$\begin{aligned}\tanh\left(\frac{\Delta(g)}{4}\right) &= \tanh\left(\frac{1}{4}\log\left(1+\frac{1}{m(X/I)}\right)\right)\\ &= \frac{(1+\frac{1}{m(X/I)})^{\frac{1}{4}} - (1+\frac{1}{m(X/I)})^{-\frac{1}{4}}}{(1+\frac{1}{m(X/I)})^{\frac{1}{4}} + (1+\frac{1}{m(X/I)})^{-\frac{1}{4}}}\\ &= \left(\sqrt{m(X/I)} + \sqrt{1+m(X/I)}\right)^{-2}. \qquad (5.2)\end{aligned}$$

Furthermore, for the differential of the mapping $g(Y) = X + (I + Y^{-1})^{-1}$, we have

$$\mathrm{d}g(A)(U) = (I+A^{-1})^{-1}(A^{-1}UA^{-1})(I+A^{-1})^{-1} = (I+A)^{-1}U(I+A)^{-1}$$

for $A \in \mathcal{P}_0, U \in \mathrm{Sym}(E)$.

Let $A, B \in \mathcal{P}_0$, and let $\gamma(t) = A^{\frac{1}{2}}(A^{-\frac{1}{2}}BA^{-\frac{1}{2}})^tA^{\frac{1}{2}}$ be the minimal geodesic curve passing from A to B. Then by Lemma 5.9,

$$\begin{aligned}p(g(A), g(B)) &\le \int_0^1 |(g\circ\gamma)'(t)|_{g(\gamma(t))}dt\\ &= \int_0^1 |\mathrm{d}g(\gamma(t))(\gamma'(t))|_{g(\gamma(t))}dt\\ &= \int_0^1 |(I+\gamma(t))^{-1}\gamma'(t)(I+\gamma(t))^{-1}|_{X+(I+\gamma(t)^{-1})^{-1}}dt\\ &\le (\sqrt{m(X/I)} + \sqrt{1+m(X/I)})^{-2}\int_0^1 |\gamma'(t)|_{\gamma(t)}dt\\ &= (\sqrt{m(X/I)} + \sqrt{1+m(X/I)})^{-2}p(A,B),\end{aligned}$$

where in the last equality we have used the fact that the distance $p(A, B)$ is equal to the Finsler length of the geodesic curve $\gamma(t)$. Therefore

$$N(g) \le \left(\sqrt{m(X/I)} + \sqrt{1+m(X/I)}\right)^{-2}.$$

To show that equality holds, it is enough to show that

$$\left(\sqrt{m(X/I)} + \sqrt{1+m(X/I)}\right)^{-2} \le \sup_{\substack{\alpha,\beta\in\mathbb{R}^+\\ \alpha<\beta}} \frac{p(g(\alpha I), g(\beta I))}{p(\alpha I, \beta I)}.$$

By Lemma 5.10, we obtain that

$$\begin{aligned}\sup_{\substack{\alpha,\beta\in\mathbb{R}^+\\ \alpha<\beta}} \frac{p(g(\alpha I), g(\beta I))}{p(\alpha I, \beta I)} &= \sup_{\substack{\alpha,\beta\in\mathbb{R}^+\\ \alpha<\beta}} \frac{p(X+(I+\alpha^{-1}I)^{-1}), X+(I+\beta^{-1}I)^{-1})}{p(\alpha I, \beta I)} \\ &= \sup_{\substack{\alpha,\beta\in\mathbb{R}^+\\ \alpha<\beta}} \frac{\log M\left(X+\frac{\beta}{\beta+1}I/X+\frac{\alpha}{\alpha+1}I\right)}{\log\frac{\beta}{\alpha}} \\ &\geq \sup_{\substack{\alpha,\beta\in\mathbb{R}^+\\ \alpha<\beta}} \frac{\log\frac{m(X/I)+\frac{\beta}{\beta+1}}{m(X/I)+\frac{\alpha}{\alpha+1}}}{\log\frac{\beta}{\alpha}} = \sup_{\substack{\alpha,\beta\in\mathbb{R}^+\\ \alpha<\beta}} \frac{\log\frac{g(\beta)}{g(\alpha)}}{\log\frac{\beta}{\alpha}},\end{aligned}$$

where $g = \left[\begin{smallmatrix} 1+m(X/I) & m(X/I) \\ 1 & 1 \end{smallmatrix}\right] \in \mathrm{SL}(2,\mathbb{R})$ is the usual Möbius transformation on $\mathbb{R}$. By the Birkhoff formula on the positive reals [4],

$$\begin{aligned}\sup_{\substack{\alpha,\beta\in\mathbb{R}^+\\ \alpha<\beta}} \frac{\log\frac{g(\beta)}{g(\alpha)}}{\log\frac{\beta}{\alpha}} &= \tanh\left(\frac{\Delta(g)}{4}\right) \\ &= \tanh\left(\frac{1}{4}\log\left(1+\frac{1}{m(X/I)}\right)\right) \\ &\overset{5.2}{=} \left(\sqrt{m(X/I)}+\sqrt{1+m(X/I)}\right)^{-2}.\end{aligned}$$

This shows that the Birkhoff formula holds for $\mathcal{S}_0$.

It follows from previous result that every member of $\mathcal{S}_0$ is a strict contraction. By definition we have that the operator $g_n := \left[\begin{smallmatrix} I & 0 \\ (1/n)I & I \end{smallmatrix}\right]$ is in $\mathcal{S}_2$, the operator $h_n := \left[\begin{smallmatrix} I & (1/n)I \\ 0 & I \end{smallmatrix}\right]$ is in $\mathcal{S}_1$, and $g_n, h_n \to e$, the identity element of $\mathrm{Sp}(V_E)$. Then for any $g \in \mathcal{S}$, $g_n h_n g \to g$ and $g_n h_n g \in \mathcal{S}_0$ since $\mathcal{S}_1$ and $\mathcal{S}_2$ are ideals by Theorem 2.3 and $\mathcal{S}_0 = \mathcal{S}_1 \cap \mathcal{S}_2$ by definition. It follows from standard continuity arguments and the density of $\mathcal{S}_0$ in $\mathcal{S}$ that all members of $\mathcal{S}$ are contractions.

Define $\sigma : \mathcal{S} \to \mathbb{R}^+ = [0,\infty)$ by

$$\sigma(t_A h_D \tilde{t}_B) = \left(\sqrt{m(Q/I)}+\sqrt{1+m(Q/I)}\right)^{-2}, \quad \text{where } Q = B^{1/2}(D^{-1})^* A D^{-1} B^{1/2}.$$

Then σ is well defined from the unique triple factorization of $\mathcal{S}$ (Theorem 2.3) and is continuous by Lemma 5.4. By Proposition 5.6 and the calculation above, $\sigma(g) = N(g) = \tanh\left(\frac{\Delta(g)}{4}\right)$ for any $g \in \mathcal{S}_0$. Let $g = t_A h_D \tilde{t}_B \in \mathcal{S} \setminus \mathcal{S}_0$. Then either A or B is not invertible; thus $m(B^{1/2}(D^{-1})^* A D^{-1} B^{1/2}/I) = 0$, and hence $\sigma(g) = 1$. By Theorem 5.8, $\tanh(\Delta(g)/4) = 1$. For small positive ϵ, pick $g_\epsilon \in \mathcal{S}_0$ sufficiently close to the identity such that $\sigma(g) - \epsilon \leq \sigma(g_\epsilon g)$. Then

$$\sigma(g) - \epsilon \leq \sigma(g_\epsilon g) = N(g_\epsilon g) \leq N(g_\epsilon) N(g) \leq N(g) \leq 1,$$

which shows that $N(g) = 1 = \sigma(g) = \tanh(\Delta(g)/4)$. Thus the Birkhoff formula holds for $\mathcal{S} \setminus \mathcal{S}_0$, which completes the proof. □

We refer the reader to the references [6] and [12] for applications of contraction results to Riccati transformations and control. There it is shown that the Riccati transformation of linear filtering/control theory is a contraction on the space of positive definite matrices. The metric used there is the standard Riemannian metric on

the symmetric space of positive definite matrices. Since we extend these results to the infinite dimensional case as well, it has been necessary to substitute the Thompson metric for the Riemannian metric. We have sharpened the results in another sense by calculating the constant of contraction, the one given by the Birkhoff formula. These formulas have been derived in the finite dimensional case in [18].

6. The Birkhoff formula on the Lie wedge. For the symplectic semigroup $\mathcal{S}$, which is a closed subsemigroup of the symplectic Lie group $\mathrm{Sp}(V_E)$, the Lie wedge of $\mathcal{S}$,

$$\mathfrak{L}(\mathcal{S}) := \{X \in \mathfrak{g} : \exp(tX) \in S \ \forall \ t \geq 0\},$$

which is the tangent object of $\mathcal{S}$ in the Lie algebra, is explicitly described as follows.

PROPOSITION 6.1 (see Proposition 8.1 of [15]). *The symplectic semigroup $\mathcal{S}$ has Lie wedge*

$$\mathfrak{L}(\mathcal{S}) = \left\{ \begin{bmatrix} A & B \\ C & -A^* \end{bmatrix} : B, C \geq 0 \right\}.$$

Setting

$$\mathfrak{h} = \left\{ \begin{bmatrix} A & 0 \\ 0 & -A^* \end{bmatrix} : A \in \mathrm{End}(V_E) \right\},$$
$$W = \left\{ \begin{bmatrix} 0 & R \\ S & 0 \end{bmatrix} : R, S \geq 0 \right\},$$

we have $\mathfrak{L}(\mathcal{S}) = \mathfrak{h} \oplus W$. In particular, $\mathfrak{h}$ is the Lie subalgebra of the subgroup H of block diagonal matrices.

We recall the Birkhoff constant map

$$N : \mathcal{S} \to [0,1], \quad N(g) = \tanh\Big(\frac{\triangle(g)}{4}\Big) = \sup_{\substack{X,Y>0 \\ X \neq Y}} \frac{p(g(X), g(Y))}{p(X,Y)},$$

and define

$$f : \mathfrak{L}(\mathcal{S}) \to [0,\infty), \quad \begin{bmatrix} A & R \\ S & -A^* \end{bmatrix} \mapsto \sqrt{m(S^{1/2}RS^{1/2}/I)}.$$

Then f is a continuous, homogeneous, Ad_H-invariant function and is an extension of the map $X \mapsto m(X/I)$ on $\mathcal{P}$.

THEOREM 6.2. *We have* $\log \circ N \circ \exp \leq -2f$ *on* $\mathfrak{L}(\mathcal{S})$.

Proof. Let $\left[\begin{smallmatrix} 0 & R \\ S & 0 \end{smallmatrix}\right] \in W^\circ$, the interior of W, i.e., $R, S > 0$. Then

$$\exp \begin{bmatrix} 0 & R \\ S & 0 \end{bmatrix} = \begin{bmatrix} S^{-1/2} & 0 \\ 0 & S^{1/2} \end{bmatrix} \cdot \exp \begin{bmatrix} 0 & S^{1/2}RS^{1/2} \\ I & 0 \end{bmatrix} \cdot \begin{bmatrix} S^{1/2} & 0 \\ 0 & S^{-1/2} \end{bmatrix},$$

and it follows by homogeneity of the Thompson metric that

$$N\left(\exp \begin{bmatrix} 0 & R \\ S & 0 \end{bmatrix} \right) = N\left(\exp \begin{bmatrix} 0 & S^{1/2}RS^{1/2} \\ I & 0 \end{bmatrix} \right).$$

Setting $X = S^{1/2}RS^{1/2}$ and $g = \exp \left[\begin{smallmatrix} 0 & X \\ I & 0 \end{smallmatrix}\right]$, we have

$$g = \begin{bmatrix} \cosh X^{1/2} & X^{1/2} \sinh X^{1/2} \\ X^{-1/2} \sinh X^{1/2} & \cosh X^{1/2} \end{bmatrix}$$

and therefore

$$g(0) = X^{1/2} \sinh X^{1/2} (\cosh X^{1/2})^{-1} = X^{1/2} \tanh X^{1/2}, \qquad g(\infty) = X^{1/2} \coth X^{1/2}.$$

Then

$$\begin{aligned} \triangle(g) &= p(g(0), g(\infty)) \\ &= p(X^{1/2} \tanh X^{1/2}, X^{1/2} \coth X^{1/2}) \\ &= p(I, \coth^2 X^{1/2}) \\ &= \log M(\coth X^{1/2}/I)^2 \\ &\le \log \coth^2 m(X^{1/2}/I), \end{aligned}$$

where the third equality follows from the homogeneity of the metric, the fourth from $\coth X^{1/2} \ge I$, and the last inequality from the fact that for $t > 0$, $tI \le X$ implies $t^n I \le X^n$ and hence $(\exp t)I \le \exp X$ and consequently $(\coth t)I \ge \coth X$. By direct computation we have

$$N(g) = \tanh \left(\frac{\triangle(g)}{4} \right) \le \tanh \left(\frac{1}{2} \log \coth m(X^{1/2}/I) \right) = e^{-2m(X^{1/2}/I)} = e^{-2\sqrt{m(X/I)}}.$$

By continuity of $N(\cdot)$ and $m(\cdot/I)$, the asserted inequality holds for arbitrary members of W.

Finally, the assertion of the inequality on all of $\mathfrak{L}(S)$ follows from the preceding, from the fact that both sides of the inequality reduce to 0 on $\mathfrak{h}$, from the Lie–Trotter product formula, and from the multiplicative property of the Birkhoff constant function: $N(gh) \le N(g)N(h)$. □

Remark. In the finite dimensional case, the inequality in Theorem 6.2 becomes an equality on W: $\log \circ N \circ \exp = -2f$. This follows from the fact that

$$|| \coth X|| = \coth ||X^{-1}||^{-1}, \quad X > 0.$$

Set $R = \left\{ \left[\begin{smallmatrix} 0 & X \\ I & 0 \end{smallmatrix}\right] : X > 0 \right\} \subseteq W^\circ$, the interior of W.

THEOREM 6.3. *We have*

$$\mathcal{S}_0 = \Gamma_0^U H \Gamma_0^L = H(\exp R)H = H(\exp W^\circ).$$

Proof. The first equality follows by Theorem 2.3. We have observed in the proof of Theorem 6.2 that $\exp W^\circ \subseteq H(\exp R)H$ and hence $H \exp W^\circ \subseteq H(\exp R)H$. Since W° is Ad_H-invariant, $(\exp R)H \subseteq H \exp W^\circ$, and therefore $H(\exp R)H \subseteq H \exp W^\circ$, the third equality is proved.

Let $X > 0$. Then

$$\begin{aligned} \exp \begin{bmatrix} 0 & X \\ I & 0 \end{bmatrix} &= \begin{bmatrix} \cosh X^{1/2} & X^{1/2} \sinh X^{1/2} \\ X^{-1/2} \sinh X^{1/2} & \cosh X^{1/2} \end{bmatrix} \\ &= \begin{bmatrix} I & X^{1/2} \tanh X^{1/2} \\ 0 & I \end{bmatrix} \begin{bmatrix} (\cosh X^{1/2})^{-1} & 0 \\ 0 & \cosh X^{1/2} \end{bmatrix} \begin{bmatrix} I & 0 \\ X^{-1/2} \tanh X^{1/2} & I \end{bmatrix} \\ &\in \mathcal{S}_0 = \Gamma_0^U H \Gamma_0^L \end{aligned}$$

because $X^{1/2}\tanh X^{1/2} > 0$ and $X^{-1/2}\tanh X^{1/2} > 0$. The ideal property of $\mathcal{S}_0$ implies that $H(\exp R)H \subseteq \mathcal{S}_0$. However, the explicit triple decomposition and Proposition 5.6 imply that

$$\exp\begin{bmatrix} 0 & X \\ I & 0 \end{bmatrix} \in H \cdot \begin{bmatrix} 0 & (\sinh X^{1/2})^2 \\ I & 0 \end{bmatrix}\begin{bmatrix} I & 0 \\ I & I \end{bmatrix} \cdot H,$$

and thus each element in the right-hand side belongs to $H(\exp R)H$. Suppose that $g \in \mathcal{S}_0 = \Gamma_0^U H \Gamma_0^L$. Then by Proposition 5.6, $g = h_{D_1}\left[\begin{smallmatrix} I & A \\ 0 & I \end{smallmatrix}\right]\left[\begin{smallmatrix} I & 0 \\ I & I \end{smallmatrix}\right]h_{D_2}$ for some $A > 0$ and $D_i \in \mathrm{GL}(E)$. Set $X = [\log(A^{1/2} + (A+I)^{1/2})]^2$. Then $X > 0$, and by direct computation $\sinh X^{1/2} = A^{1/2}$, so that $g \in H(\exp R)H$. □

7. Discrete algebraic Riccati equations. The discrete algebraic Riccati equation (DARE) arises in the context of minimizing a quadratic cost for discrete-time linear time-invariant systems (see, for example, [23, Chapter 8.4]). We consider (DARE) on a Hilbert space E:

$$X = A^*XA - A^*XB(R + B^*XB)^{-1}B^*XA + H, \tag{7.1}$$

where R and H are symmetric and positive definite [10].

It will be convenient to work with a simpler form of (DARE). We begin with the following result.

LEMMA 7.1.

1. *If $A + B$ is invertible, then $A(A+B)^{-1}B = A - A(A+B)^{-1}A$.*
2. *$B(I + B^*XB)^{-1} = (I + BB^*X)^{-1}B$.*

Proof. For the first assertion move the longer term from the right-hand side to the left, factor, and simplify. For the second eliminate the inverses by moving the expressions to the other side of the equation. □

Lemma 7.1 can be used to show that (DARE) is equivalent to

$$X = A^*X(I + GX)^{-1}A + H, \qquad G = BR^{-1}B^*. \tag{7.2}$$

Indeed,

$$\begin{aligned}
& X - XB(R + B^*XB)^{-1}B^*X \\
= \quad & X - XBR^{-1/2}(I + R^{-1/2}B^*XBR^{-1/2})^{-1}R^{-1/2}B^*X \\
\overset{C=BR^{-1/2}}{=} \quad & X - XC(I + C^*XC)^{-1}C^*X \\
\overset{Lemma\ 7.1(2)}{=} \quad & X - X(I + CC^*X)^{-1}CC^*X \\
\overset{G=CC^*}{=} \quad & X - X(I + GX)^{-1}GX \\
\overset{Lemma\ 7.1(1)}{=} \quad & X - X\Big(I - (I + GX)^{-1}\Big) \\
= \quad & X(I + GX)^{-1}.
\end{aligned}$$

THEOREM 7.2. *If A is invertible and $G = BR^{-1}B^*$ is positive definite, then* (DARE) *has a unique positive definite solution X_∞ and the iteration*

$$X_{n+1} = H + A^*X_n(I + GX_n)^{-1}A$$

starting at any point $X_0 \in \mathcal{P}_0$ converges to X_∞ with

$$p(X_\infty, X_n) \le \frac{L^n}{1-L}p(X_1, X_0),$$

where $\Lambda = H^{-1/2}A^*G^{-1/2}$, $L = \tanh\left((1/4)\log\|I + \Lambda\Lambda^*\|\right)$.

Proof. We note that positive definite solutions of (DARE) correspond to positive definite fixed points of the map

$$X \mapsto A^*X(I + GX)^{-1}A + H \tag{7.3}$$

on $\mathcal{P}_0$. Under the fractional transformation, the mapping (7.3) becomes

$$X \mapsto H + A^*X(I + GX)^{-1}A = \begin{bmatrix} I & H \\ 0 & I \end{bmatrix} \begin{bmatrix} A^* & 0 \\ 0 & A^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ G & I \end{bmatrix} .X.$$

The operator of the right-hand side,

$$\begin{bmatrix} I & H \\ 0 & I \end{bmatrix} \begin{bmatrix} A^* & 0 \\ 0 & A^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ G & I \end{bmatrix}, \tag{7.4}$$

belongs to $\mathcal{S}_0$ and hence is a strict contraction for Thompson's metric p by Theorems 5.11 and 5.8. By completeness of the metric, it has a unique fixed point on the positive definite cone $\mathcal{P}_0$, and therefore (DARE) has a unique positive definite solution. Obviously the solution X_∞ is represented as a limit of iteration $X_{n+1} = H + A^*X_n(I + GX_n)^{-1}A$ with initial point in $\mathcal{P}_0$. Set $X_\infty = \lim_{n\to\infty} X_n$, $X_0 > 0$. The p-diameter of the map $X \mapsto H + A^*X(I + GX)^{-1}A$ is computed from Theorem 5.8:

$$\Delta = p(H, H + A^*G^{-1}A) = p(I, I + \Lambda\Lambda^*) = \log||I + \Lambda\Lambda^*||,$$

where $\Lambda =: \Lambda(H, R, A, B) = H^{-1/2}A^*G^{-1/2}$. Then its contraction constant is

$$L := \tanh\left(\frac{\Delta}{4}\right) = \tanh\left(\frac{\log||I + \Lambda\Lambda^*||}{4}\right),$$

and the error bound may be estimated by $p(X_\infty, X_n) \le \frac{L^n}{1-L}\, p(X_1, X_0)$. ∎

Remark. We observe that the unique positive definite solution $S(H, R, A, B)$ in the above theorem depends on the parameters H, R, A, B, where H, R vary over the positive definite operators and A, B over invertible operators on E. This defines the solution map of DARE

$$S : \mathcal{P}_0 \times \mathcal{P}_0 \times \mathrm{GL}(E) \times \mathrm{GL}(E) \to \mathcal{P}_0, \quad (H, R, A, B) \to S(H, R, A, B).$$

In the finite dimensional case it is shown in [2] that the solution map is continuous and extends to the set of singular A. Under the additional condition that A is stable (leaves the unit ball invariant) but without the invertibility condition of $G = BR^{-1}B^*$, (DARE) has a unique positive definite solution (Corollary 5.7 of [11]).

Remark. In [9] (DARE) is called the *standard symplectic form* (SSF) when both H and G are invertible, which is the case of Theorem 7.2; that is, the associated symplectic Hamiltonian (7.4) is in $\mathcal{S}_0$. An efficient numerical method is developed, the so-called *structure-preserving doubling* (SDP) algorithm, which requires the fact that the (positive) powers of the associated symplectic Hamiltonian Z remain in SSF (Theorem 2.1 of [9]). We have already obtained the semigroup property (even ideal property) of $\mathcal{S}_0$ in Theorem 2.3. The SDP algorithm produces the sequence $Z^{2^k}, k = 1, 2, \ldots$, and the rate of convergence is estimated in terms of eigenvalues of the associated symplectic pencil (Theorem 3.1 of [9]; see also [10]).

Remark. The unique positive definite solution of (DARE) or (7.3) lies in the open order interval

$$(H, H + A^*G^{-1}A).$$

Thus it is more effective to begin the iteration method starting at a point in this interval. There are three positive definite operators lying in the interval. The harmonic-geometric-arithmetic inequalities of the positive definite operators $A, B > 0$ (cf. [1]),

$$2(A^{-1} + B^{-1})^{-1} \leq A\#B := A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2} \leq \frac{A+B}{2},$$

imply that the open order interval $(H, H+A^*G^{-1}A)$ contains the harmonic, geometric, and arithmetic means of H and $H + A^*G^{-1}A$:

$$2(H^{-1} + (H + A^*G^{-1}A)^{-1})^{-1}, \quad H\#(H + A^*G^{-1}A), \quad H + \frac{A^*G^{-1}A}{2}.$$

One can also show that $\frac{1}{2}(H + H\#(H + 4A^*G^{-1}A))$ lies in the interval $(H, H + A^*G^{-1}A)$.

8. Stability of Riccati differential equations. We consider the control system given by the *basic group control equation* (BGCE) on $\mathrm{Sp}(V_E)$:

$$\dot{g}(t) = u(t)g(t), \tag{BGCE}$$

where $u : \mathbb{I} \to \mathfrak{sp}(V_E)$, $\mathbb{I}$ a (finite or infinite) subinterval of $\mathbb{R}$, is called a *steering* or *control* function. In the case that E is finite dimensional, we assume that $u(\cdot)$ belongs to the class of measurable functions from $\mathbb{I}$ into $\mathfrak{sp}(V_E)$, which are locally bounded, that is, bounded on every finite subinterval, and in the case of general E we assume that $u(\cdot)$ is a regulated function, that is, a function that on each finite subinterval of its domain is a uniform limit of piecewise constant functions. A solution of (BGCE), called a *trajectory*, is an absolutely continuous function $x(\cdot)$ from $\mathbb{I}$ into G such that the equation (BGCE) holds a.e., where a.e. means on the complement of a set of measure 0 in the finite dimensional setting and the complement of a countable set otherwise. The solution for initial condition $g(0) = \mathrm{id}_{V(E)}$ is called the *fundamental solution* of the basic group control equation and denoted $\Phi(t)$. By right invariance the general solution to (BGCE) with initial condition $g(t_0) = g_0$ is then given by $g(t) = \Phi(t)(\Phi(t_0))^{-1}g_0$.

Proposition 8.1 (see [15, Proposition 8.3]). *Each solution $\Phi(t)$ for $t \geq 0$ of the basic group control equation on $Sp(V_E)$,*

$$\dot{g}(t) = u(t)g(t), \quad g(0) = id_{V(E)}, \quad u(t) \in \mathfrak{L}(\mathcal{S}),$$

is contained in the semigroup $\mathcal{S}$; i.e., the attainable set is contained in $\mathcal{S}$. If $\Phi(s) \in \mathcal{S}_i$ for some s and some $i = 0, 1$, or 2, then $\Phi(t) \in \mathcal{S}_i$ for all $t > s$.

We return to the material on the Lie wedge of the symplectic semigroup at the beginning of section 6. Note that $\mathfrak{L}(S) = \mathfrak{h} \oplus W$ has interior $\mathfrak{h} \oplus W^\circ$ in $\mathfrak{sp}(V_E)$, where

$$W^\circ = \left\{ \begin{bmatrix} 0 & R \\ S & 0 \end{bmatrix} : R, S > 0 \right\}.$$

Since the exponential function is locally a homeomorphism in a neighborhood $N(0)$ of the 0-matrix, we conclude that members of $(\mathfrak{h} + W^\circ) \cap N(0)$ are carried by the exponential map into the interior of $\mathcal{S}$. Since for any $Y \in \mathfrak{sp}(V_E)$, $\exp(Y) = (\exp(1/n)Y)^n$

and $\mathcal{S}$ is a subsemigroup, we conclude that $\exp(\mathfrak{h} + W^\circ)$ is carried into the interior of $\mathcal{S}$. It follows readily from the homeomorphic triple decomposition of Theorem 2.3 that the interior of $\mathcal{S}$ is contained in $\mathcal{S}_0$ (indeed they are equal), so $\exp(\mathfrak{h}+W^\circ) \subseteq \mathcal{S}_0$.

We need the following elementary lemma.

LEMMA 8.2. *Let $\Phi : \mathbb{R}^+ \times X \to X$ be a continuous semiflow of the nonnegative reals on a Hausdorff space X. Set $\phi_t(x) = \Phi(t, x)$. If ϕ_t has exactly one fixed point for each $t = 1/2^n$, $n \in \mathbb{N}$, then the fixed point is a common one for all ϕ_t, $t \in \mathbb{R}^+$.*

Consider on the Hilbert space E the Riccati differential equation

$$\text{(RDE)} \qquad \dot{K}(t) = R(t) + A(t)K(t) + K(t)A^*(t) - K(t)S(t)K(t), \quad K(t_0) = K_0,$$

where the coefficient functions are locally bounded and measurable in the finite dimensional case and regulated otherwise. It was shown in [15, section 5] that the solution of equation (RDE) for the case that $R(t), S(t), K_0 \geq 0$ arises through the fundamental solution of basic control equation (BGCE) acting by fractional transformations on $\mathcal{P} \subseteq \mathcal{M}$:

$$K(t) = \Phi(t)(\Phi(t_0))^{-1}(K_0), \qquad \text{where } u(t) = \begin{bmatrix} A(t) & R(t) \\ S(t) & -A(t)^* \end{bmatrix}.$$

In the case of constant coefficients with $R, S > 0$, then for $t > 0$, $\Phi(t) = \exp(tM)$ lies in $\mathcal{S}_0$ (as we have seen), where $M = \left[\begin{smallmatrix} A & R \\ S & -A^* \end{smallmatrix}\right]$. It follows that for each $t > 0$, $\exp(tM)$ is a strict contraction on $\mathcal{P}_0$ by Theorems 5.8 and 5.11 and hence has a unique fixed point. Hence by the preceding Lemma 8.2 we conclude that there is a common fixed point P^* for all ϕ_t, $t \geq 0$. Hence the vector field, given by (RDE), must have a 0-vector at P^*, i.e., the algebraic Riccati equation (ARE)

$$R + AK + KA^* - KSK = 0, \qquad R, S > 0,$$

must have a unique positive definite solution. (Note that another solution would yield another fixed point for the ϕ_t.) We have thus rederived from our machinery the following familiar result.

PROPOSITION 8.3. *The ARE*

$$R + AK + KA^* - KSK = 0, \qquad R, S > 0,$$

has a unique positive definite solution.

Recall the homogeneous function defined on the Lie wedge $\mathfrak{L}(\mathcal{S})$,

$$f : \mathfrak{L}(\mathcal{S}) \to [0, \infty), \qquad \begin{bmatrix} A & R \\ S & -A^* \end{bmatrix} \mapsto \sqrt{m(S^{1/2}RS^{1/2}/I)}.$$

COROLLARY 8.4. *Let $\Phi(t)$ be the fundamental solution of the basic control equation*

$$\dot{g}(t) = u(t)g(t), \quad g(0) = \mathrm{id}_{V(E)}, \quad u(t) \in \mathfrak{L}(\mathcal{S}).$$

If there exists $\mu > 0$ such that $u(t) \in f^{-1}([\mu, \infty))$ for all $t \geq 0$, then $N(\Phi(t)) \leq e^{-2t\mu}$ for each $t \geq 0$.

Proof. The density of the set of piecewise constant controls yields that $\Phi(t)$ is a limit of finite products of elements of the form

$$\exp(\alpha_1 X_1) \exp(\alpha_2 X_2) \cdots \exp(\alpha_n X_n),$$

where $\sum_{i=1}^{n} \alpha_i = t$ and $\alpha_i \geq 0, X_i \in f^{-1}([\mu, \infty))$ for $i = 1, 2, \ldots, n$. Theorem 6.2 ensures that

$$N(\exp \alpha_i X_i) \leq e^{-2f(\alpha_i X_i)} = e^{-2\alpha_i f(X_i)} \leq e^{-2\alpha_i \mu}$$

and hence

$$N(\exp(\alpha_1 X_1) \exp(\alpha_2 X_2) \cdots \exp(\alpha_n X_n)) \leq e^{-2\alpha_1 \mu} e^{-2\alpha_2 \mu} \cdots e^{-2\alpha_n \mu} = e^{-2t\mu}.$$

By continuity (see the last part of the proof of Theorem 5.11), $N(\Phi(t)) \leq e^{-2t\mu}$. ∎

Example. Let $u(t) = \left[\begin{smallmatrix} A(t) & R(t) \\ S(t) & -A(t)^* \end{smallmatrix}\right] \in \mathfrak{L}(\mathcal{S})$. Then $u(t) \in f^{-1}([\mu, \infty))$ for all $t \geq 0$ if and only if $m(S^{1/2}(t)R(t)S^{1/2}(t)/I) \geq \mu^2$ for all $t \geq 0$, and this includes the case when $S(t)$ is invertible and $R(t) \geq \mu^2 S^{-1}(t)$ for all $t \geq 0$.

The next theorem shows that under general conditions two solutions of the Riccati differential equation (RDE) exponentially converge toward each other.

THEOREM 8.5. *Let $K_1(t), K_2(t)$ be two solutions with initial values $K_1(t_0) = K_1 > 0$ and $K_2(t_0) = K_2 > 0$ of the Riccati differential equation*

$$\dot{K}(t) = R(t) + A(t)K(t) + K(t)A^*(t) - K(t)S(t)K(t), \text{ where } R(t), S(t) \geq 0.$$

If there exists $\mu > 0$ and $t_1 \geq t_0$ such that $m(S(t)^{1/2}R(t)S(t)^{1/2}/I) \geq \mu^2$ for all $t \geq t_1$, then

$$p(K_1(t), K_2(t)) \leq e^{-2(t-t_1)\mu} p(K_1, K_2)$$

for $t \geq t_1$.

Proof. Let $u(t) = \left[\begin{smallmatrix} A(t) & R(t) \\ S(t) & -A(t)^* \end{smallmatrix}\right]$. Since $R(t), S(t) \geq 0$, then $u(t) \in \mathfrak{L}(\mathcal{S}), t \geq t_0$. Let $\Phi(t)$ be the fundamental solution of the basic group control equation

$$\dot{g}(t) = u(t)g(t), g(0) = \mathrm{id}_{V_E}.$$

Then

$$K_1(t) = \Phi(t)\Phi(t_0)^{-1}(K_1) = \Phi(t)\Phi(t_1)^{-1}\Phi(t_1)\Phi(t_0)^{-1}(K_1)$$

and $K_2(t) = \Phi(t)\Phi(t_1)^{-1}\Phi(t_1)\Phi(t_0)^{-1}(K_2)$. Note that $\Psi(t) := \Phi(t+t_1)\Phi(t_1)^{-1}$ is the fundamental solution of the basic group control equation

$$\dot{g}(t) = u(t + t_1)g(t), \qquad g(0) = \mathrm{id}_{V(E)}.$$

By assumption,

$$u(t + t_1) \in f^{-1}([\mu, \infty)) = \left\{ \left[\begin{array}{cc} A & R \\ S & -A^* \end{array}\right] \in \mathcal{L}(S) : m(S^{1/2}RS^{1/2}) \geq \mu^2 \right\},$$

and by the previous corollary $N(\Psi(t)) \leq e^{-2t\mu}$ for all $t \geq 0$. Similarly, $\Phi(t_1)\Phi(t_0)^{-1} \in \mathcal{S}$, a contraction. Therefore for $t \geq t_1$

$$\begin{aligned} p(K_1(t), K_2(t)) &= p(\Psi(t - t_1)\Phi(t_1)\Phi(t_0)^{-1}(K_1), \Psi(t - t_1)\Phi(t_1)\Phi(t_0)^{-1}(K_2)) \\ &\leq e^{-2(t-t_1)\mu} p(\Phi(t_1)\Phi(t_0)^{-1}(K_1), \Phi(t_1)\Phi(t_0)^{-1}(K_2)) \\ &\leq e^{-2(t-t_1)\mu} p(K_1, K_2). \end{aligned}$$ ∎

There has been extensive study of conditions for the existence of a (positive definite) solution K to the ARE

$$R + AK + KA^* - KSK = 0, \quad R, S > 0;$$

see, for example [23, Chapter 8.4]. The preceding allows one to draw results in the converse direction.

COROLLARY 8.6. *If K^* is the unique positive definite solution for the constant coefficient ARE, then all solutions of the corresponding Riccati differential equation $\dot{K} = R + AK + KA^* - KSK$, $R, S > 0$, that enter the space of positive definite operators converge exponentially toward K^*.*

Proof. We consider a trajectory of the given Riccati differential equation that takes on a value $K_0 > 0$ at some time t_0. Then the trajectory satisfies (RDE) with initial condition K_0 at time t_0. Since K^* satisfies ARE, the value of the Riccati differential equation at K^* is 0, and thus the solution through K^* is constant. Since the coefficients R, S are constant and positive definite, the appropriate boundedness condition of the previous theorem for $S^{1/2}RS^{1/2}$ is satisfied. Hence the first trajectory converges exponentially toward the second trajectory with constant value K^*. ∎

REFERENCES

[1] T. ANDO, *Topics on Operator Inequalities*, Lecture Notes of Hokkaido University, Sapporo, Japan, 1978.

[2] W. N. ANDERSON, G. D. KLEINDORFER, P. R. KLEINDORFER, AND M. B. WOODROOFE, *Consistent estimates of the parameters of a linear system*, Ann. Math. Statist., 40 (1969), pp. 2064–2075.

[3] R. ATAR AND O. ZEITOUNI, *Exponential stability for nonlinear filtering,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 33 (1997), pp. 697–725.

[4] G. BIRKHOFF, *Extensions of Jentzch's theorem,* Trans. Amer. Math. Soc., 85 (1957), pp. 219–227.

[5] G. BIRKHOFF, *Lattice Theory*, 3rd ed., Amer. Math. Soc. Colloq. Publ. 25, AMS, Providence, RI, 1967.

[6] P. BOUGEROL, *Kalman filtering with random coefficients and contractions*, SIAM J. Control Optim., 31 (1993), pp. 942–959.

[7] A. BUDHIRAJA AND H. KUSHNER, *Robustness of nonlinear filters over the infinite time interval,* SIAM J. Control Optim., 36 (1998), pp. 1618–1637.

[8] P. J. BUSHELL, *Hilbert's metric and positive contraction mappings in a Banach space,* Arch. Ration. Mech. Anal., 52 (1973), pp. 330–338.

[9] M. CHU, N. BUONO, F. DIELE, T. POLITI, AND S. RAGNI, *On the semigroup of standard symplectic matrices and its applications,* Linear Algebra and Appl., 389 (2004), pp. 215–225.

[10] E. CHU, H. FAN, W. LIN, AND C. WANG, *A structure-preserving doubling algorithm for periodic discrete-time algebraic Riccati equations,* Internat. J. Control, 77 (2004), pp. 767–788.

[11] S. M. EL-SAYED AND A. C. M. RAN, *On an iteration method for solving a class of nonlinear matrix equations,* SIAM J. Matrix Anal. Appl., 23 (2001), pp. 632–645.

[12] S. FAKHFAKH, *Stability of Riccati's equation with random stationary coefficients,* Appl. Math. Optim. 40 (1999), pp. 141–162.

[13] R. HERMANN, *Cartanian Geometry, Nonlinear Waves, and Control Theory*, Part A, Interdisciplinary Math. 20, Math Sci Press, Boorline, MA, 1979.

[14] V. JURDJEVIC, *Geometric Control Theory*, Cambridge Press, Cambridge, UK, 1997.

[15] J. LAWSON AND Y. LIM, *The symplectic semigroup and Riccati differential equations: A case study,* J. Dynam. Control Systems, 12 (2006), pp. 49–77.

[16] Y. LIM, *Birkhoff formula for conformal compressions of symmetric cones,* Amer. J. Math. 125 (2003), pp. 167–182.

[17] C. LIVERANI, *Decay of correlations,* Ann. Math., 142 (1995), pp. 239–301.

[18] C. LIVERANI AND M. WOJTKOWSKI, *Generalization of the Hilbert metric to the space of positive definite matrices,* Pacific J. Math., 166 (1994), pp. 339–356.

[19] C. Liverani and M. Wojtkowski, *Ergodicity in Hamiltonian Systems*, Dynam. Report. Expositions Dynam. Systems (N.S.) 4, Springer, Berlin, 1995, pp. 130–202.
[20] R. D. Nussbaum, *Hilbert's Projective Metric and Iterated Nonlinear Maps,* Mem. Amer. Math. Soc., 75 (1988), number 391.
[21] R. D. Nussbaum, *Finsler structures for the part metric and Hilbert's projective metric and applications to ordinary differential equations,* Differential and Integral Equations, 7 (1994), pp. 1649–1707.
[22] M. A. Shayman, *Phase portrait of the matrix Riccati equation,* SIAM J. Control Optim., 24 (1986), pp. 1–65.
[23] E. Sontag, *Mathematical Control Theory*, 2nd ed., Springer, Berlin, 1998.
[24] A. C. Thompson, *On certain contraction mappings in a partially ordered vector spaces,* Proc. Amer. Math. Soc., 14 (1963), pp. 438–443.
[25] E. Vesentini, *Invariant metrics on cones,* Ann. Scuola Norm. Sup. Pisa Cl. Sci., 3 (1976), pp. 671–696.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 952–982

ERROR ESTIMATES FOR THE NUMERICAL APPROXIMATION OF A DISTRIBUTED CONTROL PROBLEM FOR THE STEADY-STATE NAVIER–STOKES EQUATIONS*

EDUARDO CASAS†, MARIANO MATEOS‡, AND JEAN-PIERRE RAYMOND§

Abstract. We obtain error estimates for the numerical approximation of a distributed control problem governed by the stationary Navier–Stokes equations, with pointwise control constraints. We show that the L^2-norm of the error for the control is of order h^2 if the control set is not discretized, while it is of order h if it is discretized by piecewise constant functions. These error estimates are obtained for local solutions of the control problem, which are nonsingular in the sense that the linearized Navier–Stokes equations around these solutions define some isomorphisms, and which satisfy a second order sufficient optimality condition. We establish a second order necessary optimality condition. The gap between the necessary and sufficient second order optimality conditions is the usual gap known for finite dimensional optimization problems.

Key words. optimal control, stationary Navier–Stokes equations, numerical approximation, error estimates

AMS subject classifications. 65N30, 65N15, 49M05, 49M25

DOI. 10.1137/060649999

1. Introduction.

The goal of this paper is to derive some error estimates for the numerical approximation of a distributed optimal control problem governed by the steady-state Navier–Stokes equations, with pointwise control constraints. More precisely we consider the following problem:

$$\text{(P)} \qquad \inf\Big\{F(\mathbf{u},\mathbf{y}) \mid \mathbf{u}\in U_{ad} \text{ and } (\mathbf{u},\mathbf{y}) \text{ satisfies (1.2)}\Big\},$$

where

$$F(\mathbf{u},\mathbf{y}) = \frac{1}{2}\int_\Omega |\mathbf{y}(x)-\mathbf{y}_d(x)|^2dx + \frac{N}{2}\int_\omega |\mathbf{u}(x)|^2dx, \tag{1.1}$$

$$-\nu\Delta\mathbf{y} + (\mathbf{y}\cdot\nabla)\mathbf{y} + \nabla p = \mathbf{f} + \mathcal{C}\mathbf{u} \text{ in } \Omega, \quad \operatorname{div}\mathbf{y} = 0 \text{ in } \Omega, \quad \mathbf{y} = 0 \text{ on } \Gamma, \tag{1.2}$$

$\mathcal{C}$ is a localization operator, $\omega\subset\Omega$, $N>0$, $\nu>0$, and

$$U_{ad} = \Big\{\mathbf{u}\in L^2(\omega;\mathbb{R}^m) \mid \alpha\le\mathbf{u}(x)\le\beta \text{ for almost every (a.e.) } x\in\omega\Big\}.$$

In this setting, Ω is a bounded open and connected subset in $\mathbb{R}^d$, of class C^2, with $d=2$ or $d=3$, and ω is a nonempty open subset in Ω. We can easily show that

*Received by the editors January 14, 2006; accepted for publication (in revised form) January 9, 2007; published electronically June 28, 2007. The first two authors were partially supported by the Spanish Ministry of Education and Science under projects MTM2005-06817 and "Ingenio Mathematica (i-MATH)" CSD2006-00032 (Consolider Ingenio 2010).

http://www.siam.org/journals/sicon/46-3/64999.html

†Departmento de Matemática Aplicada y Ciencias de la Computación, E.T.S.I. Industriales y de Telecomunicación, Universidad de Cantabria, 39005 Santander, Spain (eduardo.casas@unican.es).

‡Departmento de Matemáticas, E.P. de Gijón, Universidad de Oviedo, Campus de Viesques, 33203 Gijón, Spain (mmateos@uniovi.es).

§Laboratoire MIP, UMR CNRS 5640, Université Paul Sabatier, 31062 Toulouse Cedex 9, France (raymond@mip.ups-tlse.fr).

problem (P) admits at least one solution. On one hand, uniqueness of solution to problem (P) is not necessarily guaranteed even if (1.2) has a unique solution (which is not necessarily the case). On the other hand, we can only hope to obtain error estimates for solutions to problem (P) which are locally unique. Local uniqueness can be proved for solutions satisfying first order and sufficient second order optimality conditions. When first order optimality conditions in qualified form are satisfied by a local solution $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ of problem (P), we have

$$\bar{\mathbf{u}} = \mathrm{Proj}_{[\alpha,\beta]}\left(-\frac{1}{N}\mathcal{C}^*\bar{\mathbf{\Phi}}\right),$$

where $\mathrm{Proj}_{[\alpha,\beta]}$ is a projection operator and $\bar{\mathbf{\Phi}}$ is the adjoint state associated with $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$. Thus, even if $\bar{\mathbf{\Phi}}$ is regular, because of the projection operator $\mathrm{Proj}_{[\alpha,\beta]}$ (due to control constraints), $\bar{\mathbf{u}}$ is only a Lipschitz function.

Assuming that $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ satisfies first order and sufficient second order optimality conditions, we can define a discrete control problem (P_h) by discretizing the state equation (1.2) with a finite element method (here h is the mesh size of the underlying triangulation, and we assume that the family of triangulations is regular; see section 4). We consider two cases, the case where the control set in (P_h) is still U_{ad}, and the case where the control set U_{ad}^h is the set of functions in U_{ad} which are piecewise constant on the elements of the triangulation. We show that there exists $\hat{h}$ such that, for all $0 < h \leq \hat{h}$, the discrete control problem (P_h) admits at least one local solution $\bar{\mathbf{u}}_h$ in a ball $B_\rho(\bar{\mathbf{u}})$. We prove that the corresponding sequences $\{\bar{\mathbf{u}}_h\}_h$ strongly converge to $\bar{\mathbf{u}}$ in L^2 (see Theorem 4.11). When the control set in (P_h) is U_{ad}, we show that

$$\|\bar{\mathbf{u}}_h - \bar{\mathbf{u}}\|_{L^2} \leq Ch^2, \tag{1.3}$$

while if the control set is U_{ad}^h, we prove that

$$\|\bar{\mathbf{u}}_h - \bar{\mathbf{u}}\|_{L^2} \leq Ch \tag{1.4}$$

(see Theorem 4.18). To the best of our knowledge both results are new. For numerical computations it seems easier to solve $(\mathrm{P_h})$ when the control set is discretized, that is, when controls belong to U_{ad}^h. However, it is also possible to solve it without a priori discretizing the control set (see, e.g., [16]).

Before comparing our results with the ones existing in the literature, let us make some comments. Knowing that $\bar{\mathbf{u}}$ is a Lipschitz function, the error estimate (1.4), obtained when the discrete control set is defined with piecewise constant functions, is consistent with estimates obtained by approximating Lipschitz functions by piecewise constant functions. The result obtained in (1.3) is more surprising. Indeed, as we are going to see, this kind of result is already known for problems without control constraints. But in that case the optimal control belongs to H^2, and the error estimate is then directly derived from error estimates for the adjoint state. Here we obtain the same order of error estimate, but with control constraints. As far as we know, this kind of result was not previously known. Moreover, our method is quite general, and it can be used in some other problems, provided that we are able to obtain error estimates for the discrete state and discrete adjoint equations.

Let us come back to the existing results in the literature. For optimal control problems of the steady-state Navier–Stokes equations with a distributed control and a slightly different functional, Gunzburger, Hou, and Svobodny have proved error

estimates similar to (1.3) in the case when there is no control constraints and when the control acts everywhere in Ω (see [13, end of section 5.2]). But for a distributed control localized in Ω, the error estimate is only of order $h^{3/2-\varepsilon}$ (see [13, end of section 5.3]). To prove these estimates they do not assume that the optimal solution $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$, which they want to approximate, satisfies a sufficient second order optimality condition. But they assume that the optimality system satisfied by $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ is regular, in the sense that the corresponding linearized optimality system defines some isomorphism. This approach is the extension—to optimality systems of control problems—of the classical one used in the numerical approximation of the steady-state Navier–Stokes equations; see, e.g., [12]. This method has been used in the literature for other similar problems [17] and for the boundary control of the stationary Navier–Stokes equations [14, 15]. Observe that the estimates are not the same if the boundary of the domain where the control is applied is empty or nonempty [14, Theorem 4.6 and the assumptions in Theorem 3.5]. In any case this method cannot be used for problems with control constraints. Another approach used more recently for problems without control constraints is the one by Deckelnick and Hinze [10], which is based on the Kantorovich convergence theorem of the Newton method. In that case a second order sufficient optimality condition is needed, but the Kantorovich convergence theorem is proved only for systems of equations and not for generalized equations. Thus this method cannot be used for problems with control constraints.

For problems with control constraints the obtention of both optimality conditions and error estimates is more complicated. Indeed even if the nonlinear Navier–Stokes equations are well posed, the linearized ones are not necessarily well posed. Thus in general one can obtain optimality conditions only in nonqualified form, that is, optimality conditions of Fritz–John type. Such optimality conditions for optimal control problems of the stationary Navier–Stokes equations have been obtained by Abergel and Casas [1]; see also Casas [3]. Optimality conditions in qualified form, that is, optimality conditions of Karush–Kuhn–Tucker type, may be obtained either by assuming that data of the problem are small enough with respect to the viscosity parameter ν (see, e.g., Roubiček and Tröltzsch [19], Tröltzsch and Wachsmuth [21], De Los Reyes [18]) or by assuming some qualification condition of the set of feasible controls as in Gunzburger, Hou, and Svobodny [15, condition (2.7)] or in [1].

Here, since we are mainly interested in the numerical approximation of control problem (P), we assume that the local optimal solution $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ we want to approximate is a nonsingular solution, that is, that the linearized Navier–Stokes equations about $\bar{\mathbf{y}}$ define some isomorphism. As already mentioned, this is the classical assumption used in the numerical approximation of the Navier–Stokes equations (see, e.g., [12, p. 297]). Thanks to this assumption we derive a necessary optimality condition of the form

$$J''(\bar{\mathbf{u}})\mathbf{v}^2 \geq 0 \qquad \forall\, \mathbf{v} \in C_{\bar{\mathbf{u}}}, \tag{1.5}$$

where $C_{\bar{\mathbf{u}}}$ is the set of directions belonging to the tangent cone at $\bar{\mathbf{u}}$ to U_{ad} satisfying $J'(\bar{\mathbf{u}})\mathbf{v} = 0$; see Theorem 3.6 and Corollary 3.7 (here $J(\mathbf{u}) = F(\mathbf{u}, \mathbf{y}_{\mathbf{u}})$, where $\mathbf{y}_{\mathbf{u}}$ is the unique solution to (1.2) corresponding to $\mathbf{u}$, when $\mathbf{u}$ belongs to some ball $B_\rho(\bar{\mathbf{u}})$). The weakest sufficient optimality condition we can state is the following:

$$J''(\bar{\mathbf{u}})\mathbf{v}^2 > 0 \qquad \forall\, \mathbf{v} \in C_{\bar{\mathbf{u}}} \text{ such that } \ \mathbf{v} \neq 0. \tag{1.6}$$

Under this condition, and assuming that the first order optimality conditions are in qualified form, we prove that $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ is the unique local solution to (P) in some ball

$B_\rho(\bar{\mathbf{u}})$. (See Theorem 3.8. Notice that we cannot hope to prove such a result without assuming that $\bar{\mathbf{u}}$ satisfies the first order optimality conditions in qualified form and condition (1.6).) This local uniqueness result is essential to carry out some numerical analysis of the control problem. The discrete state equation is stated in section 4. The well posedness of the discrete state equation is performed in Theorem 4.8, and error estimates are obtained in Lemma 4.10. The discrete adjoint equation is studied in section 4.3. Its well posedness and error estimates are proved in Lemmas 4.12 and 4.13. Error estimates for the control problem are obtained in section 4.4.

Let us finally mention that in the case of control problems governed by scalar semilinear elliptic equations, this approach to derive error estimates has been developed by Arada, Casas, and Tröltzsch [2], Casas [4, 5], Casas, Mateos, and Tröltzsch [6], and Casas and Raymond [7].

2. Assumptions and preliminary results. Let us recall that Ω is a bounded open and connected subset in $\mathbb{R}^d$, of class C^2, with $d = 2$ or $d = 3$, and that ω is a nonempty open subset in Ω. We assume that $M : \omega \to \mathbb{R}^{d\times m}$ is a Lipschitz function, with $1 \le m \le d$ ($\mathbb{R}^{d\times m}$ denotes the space of $d \times m$ real matrices). Let us consider the linear operator $\mathcal{C} \in \mathcal{L}(L^2(\omega;\mathbb{R}^m), L^2(\Omega;\mathbb{R}^d))$, defined by $(\mathcal{C}\mathbf{u})(x) = M(x)\mathbf{u}(x)\chi_\omega(x)$, where χ_ω is the characteristic function of ω. In the functional $F : L^2(\Omega;\mathbb{R}^d) \times L^2(\omega;\mathbb{R}^m) \longmapsto \mathbb{R}$, defined in (1.1), we assume that $N > 0$ and $\mathbf{y}_d \in L^{\bar r}(\Omega;\mathbb{R}^d)$, for some $\bar r > d$, are given fixed. For $\mathbf{u} \in L^2(\omega;\mathbb{R}^m)$, we denote by u_j the components of $\mathbf{u}$, that is, $\mathbf{u} = (u_j)_{1\le j\le m}$. For $1 \le j \le m$, let $-\infty \le \alpha_j < \beta_j \le +\infty$ be extended real numbers, and set

$$U_{ad} = \Big\{\mathbf{u} \in L^2(\omega;\mathbb{R}^m) \mid \alpha_j \le u_j(x) \le \beta_j \text{ for a.e. } x \in \omega,\ 1 \le j \le m\Big\}.$$

In the case when $\alpha_j = -\infty$, this means that the corresponding constraint is absent. The same convention is adopted if $\beta_j = \infty$.

In (1.2) we assume that $\nu > 0$ and $\mathbf{f} \in L^{\bar r}(\Omega;\mathbb{R}^d)$.

To study (1.2) we have to introduce some function spaces and operators. Throughout the following we set $\mathbf{H}^1(\Omega) = H^1(\Omega;\mathbb{R}^d)$, $\mathbf{H}^1_0(\Omega) = H^1_0(\Omega;\mathbb{R}^d)$, $\mathbf{H}^{-1}(\Omega) = (\mathbf{H}^1_0(\Omega))'$, $\mathbf{L}^p(\Omega) = L^p(\Omega;\mathbb{R}^d)$, and $\mathbf{W}^{s,p}(\Omega) = W^{s,p}(\Omega;\mathbb{R}^d)$ for $1 \le p \le \infty$ and $s > 0$. We introduce different spaces of divergence-free vector fields:

$$\mathbf{V}^0_n(\Omega) = \Big\{\mathbf{u} \in \mathbf{L}^2(\Omega) \mid \text{ div } \mathbf{u} = 0 \ \text{ in } \Omega,\ \mathbf{u}\cdot\mathbf{n} = 0 \text{ in } H^{-1/2}(\Gamma)\Big\},$$
$$\mathbf{V}^1_0(\Omega) = \mathbf{H}^1_0(\Omega) \cap \mathbf{V}^0_n(\Omega),$$

where $\mathbf{n}$ is the outward unit normal to Γ. The dual space of $\mathbf{V}^1_0(\Omega)$ with respect to the pivot space $\mathbf{V}^0_n(\Omega)$ is denoted by $\mathbf{V}^{-1}(\Omega)$. Thus we have

$$\mathbf{V}^1_0(\Omega) \hookrightarrow \mathbf{V}^0_n(\Omega) \hookrightarrow \mathbf{V}^{-1}(\Omega),$$

with dense and continuous imbeddings. The orthogonal projector from $\mathbf{L}^2(\Omega)$ onto $\mathbf{V}^0_n(\Omega)$ will be denoted by P. The operator P can be extended to a bounded operator from $\mathbf{H}^{-1}(\Omega)$ to $\mathbf{V}^{-1}(\Omega)$. For notational simplicity this extension will still be denoted by P.

Let us consider the bilinear form on $\mathbf{H}^1_0(\Omega)$ defined by

$$a(\mathbf{y},\mathbf{z}) = \nu \int_\Omega \nabla\mathbf{y} : \nabla\mathbf{z}\, dx = \nu \sum_{i,j=1}^{d} \int_\Omega \partial_{x_i} y_j\, \partial_{x_i} z_j,$$

and the trilinear form on $b : \mathbf{L}^4(\Omega) \times \mathbf{H}_0^1(\Omega) \times \mathbf{L}^4(\Omega)$ defined by

$$b(\mathbf{y}, \mathbf{z}, \mathbf{\Phi}) = \int_\Omega (\mathbf{y} \cdot \nabla)\mathbf{z} \cdot \mathbf{\Phi}\, dx.$$

We define $A \in \mathcal{L}(\mathbf{H}_0^1(\Omega), \mathbf{H}^{-1}(\Omega))$ by

$$\big\langle A\mathbf{y}, \mathbf{z}\big\rangle_{\mathbf{H}^{-1}(\Omega),\mathbf{H}_0^1(\Omega)} = a(\mathbf{y}, \mathbf{z}) \qquad \forall\, \mathbf{z},\, \mathbf{y} \in \mathbf{H}_0^1(\Omega),$$

and the nonlinear operator B from $\mathbf{H}_0^1(\Omega)$ to $\mathbf{H}^{-1}(\Omega)$ by

$$\big\langle B(\mathbf{y}), \mathbf{z}\big\rangle_{\mathbf{H}^{-1}(\Omega),\mathbf{H}_0^1(\Omega)} = b(\mathbf{y}, \mathbf{y}, \mathbf{z}) \qquad \forall\, \mathbf{z},\, \mathbf{y} \in \mathbf{H}_0^1(\Omega).$$

Equation (1.2) is equivalent to the variational problem

$$\begin{aligned} &\text{Find } \mathbf{y} \in \mathbf{V}_0^1(\Omega) \quad \text{such that} \\ &a(\mathbf{y}, \mathbf{z}) + b(\mathbf{y}, \mathbf{y}, \mathbf{z}) = (\mathbf{f} + \mathcal{C}\mathbf{u}, \mathbf{z}) \quad \forall\, \mathbf{z} \in \mathbf{V}_0^1(\Omega), \end{aligned} \tag{2.1}$$

or to the weak formulation

$$\mathbf{y} \in \mathbf{V}_0^1(\Omega), \quad \big\langle A\mathbf{y} + B(\mathbf{y}), \mathbf{z}\big\rangle_{\mathbf{H}^{-1}(\Omega),\mathbf{H}_0^1(\Omega)} = \big\langle \mathbf{f} + \mathcal{C}\mathbf{u}, \mathbf{z}\big\rangle_{\mathbf{H}^{-1}(\Omega),\mathbf{H}_0^1(\Omega)} \quad \forall \mathbf{z} \in \mathbf{V}_0^1(\Omega).$$

This last equation is equivalent to

$$\mathbf{y} \in \mathbf{V}_0^1(\Omega), \quad PA\mathbf{y} + PB(\mathbf{y}) = P(\mathbf{f} + \mathcal{C}\mathbf{u}) \quad \text{in } \mathbf{V}^{-1}(\Omega),$$

which we shall simply write in the form

$$\mathbf{y} \in \mathbf{V}_0^1(\Omega), \quad A\mathbf{y} + B(\mathbf{y}) = \mathbf{f} + \mathcal{C}\mathbf{u} \quad \text{in } \mathbf{V}^{-1}(\Omega). \tag{2.2}$$

We know that, for all $\mathbf{u} \in L^2(\omega; \mathbb{R}^m)$, equation (2.1), or equivalently (2.2), admits at least one solution $\mathbf{y} \in \mathbf{V}_0^1(\Omega)$. The pressure appearing in (1.2) is the unique function in

$$L_0^2(\Omega) = \left\{ v \in L^2(\Omega) : \int_\Omega v(x)\, dx = 0 \right\},$$

obeying

$$\nabla p = (I - P)(\mathbf{f} + \mathcal{C}\mathbf{u} + \nu\Delta\mathbf{y} - (\mathbf{y} \cdot \nabla)\mathbf{y}). \tag{2.3}$$

It is a consequence of [12, Chapter 1, Lemma 2.1].

The following properties are well known. For all $\mathbf{y} \in \mathbf{L}^4(\Omega)$ obeying $\operatorname{div} \mathbf{y} = 0$ in Ω, and $\mathbf{z}, \mathbf{w} \in \mathbf{H}_0^1(\Omega)$

$$b(\mathbf{y}, \mathbf{z}, \mathbf{w}) = -b(\mathbf{y}, \mathbf{w}, \mathbf{z}) \quad \text{and} \quad b(\mathbf{y}, \mathbf{z}, \mathbf{z}) = 0. \tag{2.4}$$

The next lemma follows directly from Green's formula.

LEMMA 2.1. *For all $\mathbf{y} \in \mathbf{H}_0^1(\Omega)$, the operators $B'(\mathbf{y}) \in \mathcal{L}(\mathbf{H}_0^1(\Omega), \mathbf{H}^{-1}(\Omega))$ and $B'(\mathbf{y})^* \in \mathcal{L}(\mathbf{H}_0^1(\Omega), \mathbf{H}^{-1}(\Omega))$ satisfy*

$$\langle B'(\mathbf{y})\mathbf{z}, \mathbf{\Phi}\rangle = b(\mathbf{y}, \mathbf{z}, \mathbf{\Phi}) + b(\mathbf{z}, \mathbf{y}, \mathbf{\Phi})$$

and

$$\langle B'(\mathbf{y})^*\mathbf{\Phi}, \mathbf{z}\rangle = \int_\Omega (\nabla \mathbf{y})^T \mathbf{\Phi}\cdot \mathbf{z}\, dx - b(\mathbf{y}, \mathbf{\Phi}, \mathbf{z}) - \int_\Omega (\operatorname{div} \mathbf{y})\, \mathbf{\Phi}\cdot \mathbf{z}\, dx$$

for all $\mathbf{z}$, $\mathbf{\Phi} \in \mathbf{H}_0^1(\Omega)$. *Moreover,* $B'' \in \mathcal{L}(\mathbf{H}_0^1(\Omega)\times \mathbf{H}_0^1(\Omega), \mathbf{H}^{-1}(\Omega))$ *obeys*

$$\langle B''(\mathbf{y}, \mathbf{z}), \mathbf{\Phi}\rangle = b(\mathbf{y}, \mathbf{z}, \mathbf{\Phi}) + b(\mathbf{z}, \mathbf{y}, \mathbf{\Phi}) \quad \forall\, \mathbf{z}, \mathbf{y}, \mathbf{\Phi} \in \mathbf{H}_0^1(\Omega).$$

The following regularity result will be used throughout this paper. It is an immediate consequence of the classical result by Cattabriga [8].

THEOREM 2.2. *There exists a constant* $C > 0$ *such that if* $\mathbf{u} \in L^2(\omega; \mathbb{R}^m)$ *and if* $\mathbf{y} \in \mathbf{V}_0^1(\Omega)$ *is a solution to* (2.2), *then*

$$\|\mathbf{y}\|_{\mathbf{V}_0^1(\Omega)} \le C(\|\mathbf{f}\|_{\mathbf{L}^2(\Omega)} + \|\mathbf{u}\|_{L^2(\omega;\mathbb{R}^m)}).$$

There exists a constant $C_r > 0$ *such that if* $\mathbf{u} \in L^r(\omega; \mathbb{R}^m)$, $\mathbf{f} \in \mathbf{L}^r(\Omega)$ *with* $2 \le r < \infty$ *and* $\mathbf{y} \in \mathbf{V}_0^1(\Omega)$ *is a solution to* (2.2) *and* p *the associated pressure, then* $\mathbf{y} \in \mathbf{W}^{2,r}(\Omega)$, $p \in \mathbf{W}^{1,r}(\Omega)$, *and*

$$\|p\|_{\mathbf{W}^{1,r}(\Omega)} + \|\mathbf{y}\|_{\mathbf{W}^{2,r}(\Omega)} \le C_r(1 + \|\mathbf{f}\|^7_{\mathbf{L}^r(\Omega)} + \|\mathbf{u}\|^7_{L^r(\omega;\mathbb{R}^m)}). \tag{2.5}$$

Proof. The estimate of $\|\mathbf{y}\|_{\mathbf{V}_0^1(\Omega)}$ is classical. Using this estimate, since $d \le 3$, we can write

$$\|\mathbf{y}\|_{\mathbf{L}^6(\Omega)} \le C\|\mathbf{y}\|_{\mathbf{V}_0^1(\Omega)}$$

and

$$\|\mathbf{y}\otimes\mathbf{y}\|_{(\mathbf{L}^3(\Omega))^d} \le C\|\mathbf{y}\|^2_{\mathbf{V}_0^1(\Omega)}.$$

Thus, from estimates for the Stokes equation, we successively deduce

$$\|\mathbf{y}\|_{\mathbf{W}^{1,3}(\Omega)} \le C(\|\mathbf{y}\otimes\mathbf{y}\|_{(\mathbf{L}^3(\Omega))^d} + \|\mathbf{f}\|_{\mathbf{L}^2(\Omega)} + \|\mathbf{u}\|_{L^2(\omega;\mathbb{R}^m)})$$

$$\le C(\|\mathbf{f}\|^2_{\mathbf{L}^2(\Omega)} + \|\mathbf{u}\|^2_{L^2(\omega;\mathbb{R}^m)} + 1)$$

and

$$\|\mathbf{y}\|_{\mathbf{H}^2(\Omega)} \le C(\|\mathbf{y}\|_{\mathbf{W}^{1,3}(\Omega)}\|\mathbf{y}\|_{\mathbf{L}^6(\Omega)} + \|\mathbf{f}\|_{\mathbf{L}^2(\Omega)} + \|\mathbf{u}\|_{L^2(\omega;\mathbb{R}^m)})$$

$$\le C(\|\mathbf{f}\|^3_{\mathbf{L}^2(\Omega)} + \|\mathbf{u}\|^3_{L^2(\omega;\mathbb{R}^m)} + 1).$$

Therefore

$$\|(\mathbf{y}\cdot\nabla)\mathbf{y}\|_{(\mathbf{L}^3(\Omega))^d} \le C\|\mathbf{y}\|_{\mathbf{H}^1(\Omega)}\|\mathbf{y}\|_{\mathbf{H}^2(\Omega)} \le C(\|\mathbf{f}\|^4_{\mathbf{L}^2(\Omega)} + \|\mathbf{u}\|^4_{L^2(\omega;\mathbb{R}^m)} + 1),$$

which yields

$$\|\mathbf{y}\|_{\mathbf{W}^{2,r}(\Omega)} \le C(\|\mathbf{f}\|^4_{\mathbf{L}^r(\Omega)} + \|\mathbf{u}\|^4_{L^r(\omega;\mathbb{R}^m)} + 1)$$

if $2 \le r \le 3$. Next we have

$$\|(\mathbf{y}\cdot\nabla)\mathbf{y}\|_{(\mathbf{L}^r(\Omega))^d} \le C_r\|\mathbf{y}\|_{\mathbf{H}^2(\Omega)}\|\mathbf{y}\|_{\mathbf{W}^{2,3}(\Omega)} \le C(\|\mathbf{f}\|^7_{\mathbf{L}^2(\Omega)} + \|\mathbf{u}\|^7_{L^2(\omega;\mathbb{R}^m)} + 1)$$

if $3 \le r < \infty$, which provides the desired estimate. ∎

It is well known that the solution of (1.2) is unique when ν is large enough with respect to the right-hand side; see, for instance, Temam [20]. Since this is a strong assumption we are interested in the solutions of (1.2) for which the equation is locally unique. These solutions, called nonsingular solutions, are defined below.

DEFINITION 2.3. *A function $\mathbf{y} \in \mathbf{V}_0^1(\Omega)$ is a nonsingular solution of* (1.2)*, or equivalently* (2.2)*, if $P(A + B'(\mathbf{y}))$ is an isomorphism from $\mathbf{V}_0^1(\Omega)$ into $\mathbf{V}^{-1}(\Omega)$. If, moreover, $A\mathbf{y} + B(\mathbf{y}) = \mathbf{f} + \mathcal{C}\mathbf{u}$ in $\mathbf{V}^{-1}(\Omega)$, with $\mathbf{u} \in L^2(\omega;\mathbb{R}^m)$, we will also say that the pair $(\mathbf{u},\mathbf{y})$ is a nonsingular solution of* (1.2).

Remark 2.4. For a nonsingular solution $(\mathbf{u},\mathbf{y})$ of (1.2), the condition $P(A + B'(\mathbf{y})) \in \mathrm{isom}(\mathbf{V}_0^1(\Omega),\mathbf{V}^{-1}(\Omega))$ corresponds to the one stated in [12, Chapter 4, condition (3.4)], which is used to get the error estimates for the approximation of the Navier–Stokes equations.

The following theorem is a straightforward consequence of the implicit function theorem and will be useful in what follows.

THEOREM 2.5. *Let $(\bar{\mathbf{u}},\bar{\mathbf{y}}) \in L^2(\omega;\mathbb{R}^m)\times\mathbf{V}_0^1(\Omega)$ be a nonsingular solution of* (1.2)*; then there exist an open neighborhood $\mathcal{O}(\bar{\mathbf{u}})$ of $\bar{\mathbf{u}}$ in $L^2(\omega;\mathbb{R}^m)$, an open neighborhood $\mathcal{O}(\bar{\mathbf{y}})$ of $\bar{\mathbf{y}}$ in $\mathbf{V}_0^1(\Omega)$, and a mapping G from $\mathcal{O}(\bar{\mathbf{u}})$ to $\mathcal{O}(\bar{\mathbf{y}})$ of class C^∞ such that, for all $\mathbf{u} \in \mathcal{O}(\bar{\mathbf{u}})$, $G(\mathbf{u}) = \mathbf{y}_\mathbf{u}$ is the unique solution in $\mathcal{O}(\bar{\mathbf{y}})$ to* (2.2)*. Moreover, if $\mathbf{z}_\mathbf{v} = G'(\mathbf{u})\mathbf{v} \in \mathbf{V}_0^1(\Omega)$ and $\mathbf{w} = G''(\mathbf{u})\mathbf{v}^2 \in \mathbf{V}_0^1(\Omega)$, then $\mathbf{z}_\mathbf{v}$ and $\mathbf{w}$ satisfy the equations*

$$A\mathbf{z}_\mathbf{v} + B'(\mathbf{y}_\mathbf{u})\mathbf{z}_\mathbf{v} = \mathcal{C}\mathbf{v} \quad \text{in } \mathbf{V}^{-1}(\Omega), \tag{2.6}$$

$$A\mathbf{w} + B'(\mathbf{y}_\mathbf{u})\mathbf{w} + B''(\mathbf{z}_\mathbf{v},\mathbf{z}_\mathbf{v}) = 0 \quad \text{in } \mathbf{V}^{-1}(\Omega), \tag{2.7}$$

and $P(A + B'(\mathbf{y}_\mathbf{u}))$ is an isomorphism from $\mathbf{V}_0^1(\Omega)$ into $\mathbf{V}^{-1}(\Omega)$ for all $\mathbf{u} \in \mathcal{O}(\bar{\mathbf{u}})$.

LEMMA 2.6. *Let $(\bar{\mathbf{u}},\bar{\mathbf{y}})$ be as in Theorem* 2.5*, and let $\bar{p}$ be the associated pressure (the solution of* (2.3) *corresponding to $\bar{\mathbf{y}}$). Let $(\mathbf{u}_k)_k$ be a sequence in $\mathcal{O}(\bar{\mathbf{u}})$ weakly converging to $\bar{\mathbf{u}}$ in $L^2(\omega;\mathbb{R}^m)$. Let $\mathbf{y}_k$ be the solution to* (1.2) *in $\mathcal{O}(\bar{\mathbf{y}})$ corresponding to $\mathbf{u}_k$, and let p_k be the associated pressure. Then $(\mathbf{y}_k)_k$ converges to $\bar{\mathbf{y}}$ in $\mathbf{V}_0^1(\Omega)$, and $(p_k)_k$ converges to $\bar{p}$ in $L_0^2(\Omega)$.*

Proof. The proof is an easy consequence of Theorem 2.2 and of formula (2.3). ∎

3. Analysis of the control problem. The existence of a solution of problem (P) can be obtained by the usual approach of taking a minimizing sequence, which is bounded in $L^2(\omega;\mathbb{R}^m)\times\mathbf{V}_0^1(\Omega)$, and passing to the limit; see, for instance, [18] for a detailed proof. In this section we will derive the first and second order optimality conditions for a local solution $(\bar{\mathbf{u}},\bar{\mathbf{y}})$ in $U_{ad}\times\mathbf{V}_0^1(\Omega)$.

3.1. First order optimality conditions. Let us precisely define local solutions of (P).

DEFINITION 3.1. *We shall say that $(\bar{\mathbf{u}},\bar{\mathbf{y}}) \in U_{ad}\times\mathbf{V}_0^1(\Omega)$ is a local solution of* (P) *if and only if $(\bar{\mathbf{u}},\bar{\mathbf{y}})$ satisfies* (1.2) *and there exist neighborhoods $\mathcal{O}(\bar{\mathbf{u}})$ of $\bar{\mathbf{u}}$ in $L^2(\omega;\mathbb{R}^m)$ and $\mathcal{O}(\bar{\mathbf{y}})$ of $\bar{\mathbf{y}}$ in $\mathbf{V}_0^1(\Omega)$ such that $F(\bar{\mathbf{u}},\bar{\mathbf{y}}) \le F(\mathbf{u},\mathbf{y})$ for all pairs $(\mathbf{u},\mathbf{y}) \in (U_{ad}\cap\mathcal{O}(\bar{\mathbf{u}}))\times\mathcal{O}(\bar{\mathbf{y}})$ satisfying* (1.2).

The following theorem was proved by Abergel and Casas [1] for a slightly different functional, but the proof can be repeated for our problem step by step, just by doing the obvious modifications.

THEOREM 3.2. *Let* $(\bar{\mathbf{u}}, \bar{\mathbf{y}}) \in U_{ad} \times \mathbf{V}_0^1(\Omega)$ *be a local solution of* (P)*; then there exist a real number* $\bar{\lambda}$ *and some elements* $\bar{\mathbf{\Phi}} \in \mathbf{W}^{2,\bar{r}}(\Omega)$ *and* $\bar{\pi}, \bar{p} \in W^{1,\bar{r}}(\Omega)$ *such that*

$$\bar{\lambda} + \|\bar{\mathbf{\Phi}}\|_{\mathbf{V}_0^1(\Omega)} > 0, \tag{3.1}$$

$$-\nu\Delta\bar{\mathbf{y}} + (\bar{\mathbf{y}}\cdot\nabla)\bar{\mathbf{y}} + \nabla\bar{p} = \mathbf{f} + \mathcal{C}\bar{\mathbf{u}} \text{ in } \Omega,\ \operatorname{div}\bar{\mathbf{y}} = 0 \text{ in } \Omega,\ \bar{\mathbf{y}} = 0 \text{ on } \Gamma, \tag{3.2}$$

$$-\nu\Delta\bar{\mathbf{\Phi}} + (\nabla\bar{\mathbf{y}})^T\bar{\mathbf{\Phi}} - (\bar{\mathbf{y}}\cdot\nabla)\bar{\mathbf{\Phi}} + \nabla\bar{\pi} = \bar{\lambda}(\bar{\mathbf{y}} - \mathbf{y}_d) \text{ in } \Omega, \tag{3.3}$$

$$\operatorname{div}\bar{\mathbf{\Phi}} = 0 \text{ in } \Omega,\ \bar{\mathbf{\Phi}} = 0 \text{ on } \Gamma, \tag{3.4}$$

$$\int_\omega (\mathcal{C}^*\bar{\mathbf{\Phi}} + \bar{\lambda}N\bar{\mathbf{u}})\cdot(\mathbf{u} - \bar{\mathbf{u}})\,dx \geq 0\ \forall\,\mathbf{u} \in U_{ad}. \tag{3.5}$$

These conditions for optimality are of Fritz–John type, and we are interested in the cases where $\bar{\lambda}$ can be chosen equal to one. Gunzburger, Hou, and Svobodny [14] introduced an assumption on U_{ad} for the local solution $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$. The control set U_{ad} is said to have the property (C) at $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ if the system

$$-\nu\Delta\mathbf{\Phi} + (\nabla\bar{\mathbf{y}})^T\mathbf{\Phi} - (\mathbf{y}\cdot\nabla)\mathbf{\Phi} + \nabla\pi = \bar{\lambda}(\bar{\mathbf{y}} - \mathbf{y}_d) \text{ in } \Omega,\ \operatorname{div}\mathbf{\Phi} = 0 \text{ in } \Omega,\ \mathbf{\Phi} = 0 \text{ on } \Gamma,$$

admits at least a nonzero solution $(\mathbf{\Phi}, \pi) \in \mathbf{V}_0^1(\Omega) \times L_0^2(\Omega)$, and if for any nonzero solution $(\mathbf{\Phi}, \pi)$ we can find $\mathbf{u} \in U_{ad}$ such that

$$\int_\omega \mathcal{C}^*\mathbf{\Phi}\cdot(\mathbf{u} - \bar{\mathbf{u}})\,dx < 0.$$

It is obvious that if U_{ad} has the property (C) at $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$, then (3.2)–(3.5) hold with $\bar{\lambda} = 1$.

Here we will make a different assumption which will be crucial in what follows, in particular for the numerical analysis. We consider only local solutions $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ of (P) such that $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ is a nonsingular solution of (2.2). In that case we shall say that $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ is a *local nonsingular solution* of (P). For such a local nonsingular solution we can apply Theorem 2.5 and define the control problem

$$\inf\Big\{J(\mathbf{u}) \mid \mathbf{u} \in U_{ad} \cap \mathcal{O}(\bar{\mathbf{u}})\Big\}, \tag{$\mathrm{P}_{\mathcal{O}(\bar{\mathbf{u}})}$}$$

where $J : \mathcal{U} \longmapsto \mathbb{R}$ is given by $J(\mathbf{u}) = F(\mathbf{u}, G(\mathbf{u}))$. Then $\bar{\mathbf{u}}$ is a local solution of $(\mathrm{P}_{\mathcal{O}(\bar{\mathbf{u}})})$. Let us study the differentiability properties of J.

THEOREM 3.3. *Function* J *is of class* C^∞ *in* $\mathcal{O}(\bar{\mathbf{u}})$*, and for every* $\mathbf{u} \in \mathcal{O}(\bar{\mathbf{u}})$ *and* $\mathbf{v} \in L^2(\omega;\mathbb{R}^m)$ *we have*

$$J'(\mathbf{u})\mathbf{v} = \int_\omega (\mathcal{C}^*\mathbf{\Phi}_\mathbf{u} + N\mathbf{u})\cdot\mathbf{v}\,dx, \tag{3.6}$$

$$J''(\mathbf{u})\mathbf{v}^2 = \int_\Omega (|\mathbf{z}_\mathbf{v}|^2 - 2(\mathbf{z}_\mathbf{v}\cdot\nabla)\mathbf{z}_\mathbf{v}\cdot\mathbf{\Phi}_\mathbf{u})dx + N\int_\omega |\mathbf{v}|^2 dx, \tag{3.7}$$

where $\mathbf{z}_\mathbf{v}$ *is the solution of* (2.6) *and* $\mathbf{\Phi}_\mathbf{u} \in \mathbf{V}_0^1(\Omega)$ *satisfies*

$$\begin{cases} -\nu\Delta\mathbf{\Phi}_\mathbf{u} + (\nabla\mathbf{y}_\mathbf{u})^T\mathbf{\Phi}_\mathbf{u} - (\mathbf{y}_\mathbf{u}\cdot\nabla)\mathbf{\Phi}_\mathbf{u} + \nabla\pi_\mathbf{u} = \mathbf{y}_\mathbf{u} - \mathbf{y}_d \text{ in } \Omega, \\ \operatorname{div}\mathbf{\Phi}_\mathbf{u} = 0 \text{ in } \Omega,\ \mathbf{\Phi}_\mathbf{u} = 0 \text{ on } \Gamma. \end{cases} \tag{3.8}$$

The proof follows easily from Theorem 2.5. The only delicate point is the definition of $\mathbf{\Phi}_\mathbf{u}$. Let us remark that (3.8) is equivalent to the equation $A^*\mathbf{\Phi}_\mathbf{u} +$

$B'(\mathbf{y_u})^*\mathbf{\Phi_u} = \mathbf{y_u} - \mathbf{y}_d$ in $\mathbf{V}^{-1}(\Omega)$, and due to Theorem 2.5 the operator $P(A^* + B'(\mathbf{y_u})^*)$ is an isomorphism from $\mathbf{V}_0^1(\Omega)$ into $\mathbf{V}^{-1}(\Omega)$.

By using the previous theorem we get the following result.

THEOREM 3.4. *Let $(\bar{\mathbf{u}}, \bar{\mathbf{y}}) \in U_{ad} \times \mathbf{V}_0^1(\Omega)$ be a local nonsingular solution of* (P)*, and let $\bar{p}$ be the associated pressure; then there exist some elements $\bar{\mathbf{\Phi}} \in \mathbf{V}_0^1(\Omega)$ and $\bar{\pi} \in L_0^2(\Omega)$ such that* (3.2)–(3.5) *hold with $\bar{\lambda} = 1$.*

Proof. It is enough to take into account that $J'(\bar{\mathbf{u}})(\mathbf{u} - \bar{\mathbf{u}}) \geq 0$ for all $\mathbf{u} \in U_{ad}$ and to use (3.6). ∎

Using the first order necessary conditions we can deduce some extra regularity for the optimal control, the state, and the adjoint state.

THEOREM 3.5. *Let $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ be a local nonsingular solution of* (P) *and let $\bar{\mathbf{\Phi}}$ be the adjoint state as defined by* (3.3)–(3.4) *with $\bar{\lambda} = 1$. Then $\bar{\mathbf{y}}, \bar{\mathbf{\Phi}} \in \mathbf{W}^{2,\bar{r}}(\Omega)$, $\bar{p}, \bar{\pi} \in W^{1,\bar{r}}(\Omega)$, and $\bar{\mathbf{u}} \in C^{0,1}(\bar{\omega}; \mathbb{R}^m)$.*

Proof. Taking into account that $\mathcal{C}\bar{\mathbf{u}} \in \mathbf{L}^2(\Omega)$ and the assumption on $\mathbf{f}$, it is enough to apply Theorem 2.2 to deduce that $\bar{\mathbf{y}}$ belongs to $\mathbf{H}^2(\Omega)$ and that $\bar{\mathbf{\Phi}}$ belongs to $\mathbf{W}^{2,\bar{r}}(\Omega)$. On the other hand, $\bar{\mathbf{\Phi}} \in \mathbf{W}^{2,\bar{r}}(\Omega) \subset C^{0,1}(\bar{\Omega}; \mathbb{R}^d)$ because $\bar{r} > d$. Now using the Lipschitz property of the function M defining $\mathcal{C}$ and the representation of the optimal control deduced from (3.5), we obtain

$$\bar{u}_j(x) = \mathrm{Proj}_{[\alpha_j, \beta_j]}\left(-\frac{1}{N}(\mathcal{C}^*\bar{\mathbf{\Phi}})_j(x)\right) \quad \text{for a.e. } x \in \omega, \tag{3.9}$$

which gives the desired regularity for $\bar{\mathbf{u}}$. Now still using Theorem 2.2, we obtain the regularity of $\bar{\mathbf{y}}$. ∎

3.2. Second order optimality conditions. To perform the numerical analysis of the problem as well as the analysis of the algorithms of optimization, second order sufficient conditions are required. These sufficient conditions should be as unrestrictive as possible. One way of measuring this is to compare them with the necessary second order conditions and check if the gap is small. This is the reason why we first introduce the second order necessary conditions.

Second order conditions have to be written for directions $\mathbf{v} \in T_{U_{ad}}(\bar{\mathbf{u}})$ such that $J'(\mathbf{u})\mathbf{v} = 0$, where $T_{U_{ad}}(\bar{\mathbf{u}})$ is the tangent cone at $\bar{\mathbf{u}}$ to U_{ad}. To characterize these directions, we introduce $\bar{\mathbf{d}}(x) = \mathcal{C}^*\bar{\mathbf{\Phi}}(x) + N\bar{\mathbf{u}}(x)$ for $x \in \omega$, and the following conditions:

$$v_j(x) = 0 \text{ if } \bar{d}_j(x) \neq 0, \tag{3.10}$$

$$v_j(x) \geq 0 \text{ if } -\infty < \alpha_j = \bar{u}_j(x) \text{ and } \bar{d}_j(x) = 0, \tag{3.11}$$

$$v_j(x) \leq 0 \text{ if } \bar{u}_j(x) = \beta_j < \infty \text{ and } \bar{d}_j(x) = 0. \tag{3.12}$$

Now we define the cone

$$C_{\bar{\mathbf{u}}} = \Big\{\mathbf{v} \in L^2(\omega; \mathbb{R}^m) \mid \mathbf{v} \text{ satisfies (3.10)–(3.12)}\Big\}.$$

Notice that

$$\begin{aligned} &J'(\bar{\mathbf{u}})\mathbf{v} = \int_\omega \bar{\mathbf{d}}(x) \cdot \mathbf{v}(x)\, dx \;\forall\, \mathbf{v} \in L^2(\omega; \mathbb{R}^m), \\ &\bar{\mathbf{d}}(x) \cdot \mathbf{v}(x) = 0 \text{ for a.e. } x \in \omega \text{ and all } \mathbf{v} \in C_{\bar{\mathbf{u}}}. \end{aligned} \tag{3.13}$$

THEOREM 3.6. *Let $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ be a nonsingular local solution of* (P)*. Then*

$$J''(\bar{\mathbf{u}})\mathbf{v}^2 \geq 0 \quad \forall\, \mathbf{v} \in C_{\bar{\mathbf{u}}}.$$

Proof. We sketch the proof in the case where $-\infty < \alpha_j < \beta_j < \infty$ for all $1 \le j \le m$. The modifications for the other cases are obvious. Take $\mathbf{v} \in C_{\bar{\mathbf{u}}}$, and for $\varepsilon < \min\{(\beta_j - \alpha_j)/2 : \ 1 \le j \le m\}$ define

$$v_{j,\varepsilon}(x) = \begin{cases} 0 & \text{if } \alpha_j < \bar{u}_j(x) < \alpha_j + \varepsilon, \\ 0 & \text{if } \beta_j - \varepsilon < \bar{u}_j(x) < \beta_j, \\ \mathrm{Proj}_{[-\frac{1}{\varepsilon}, \frac{1}{\varepsilon}]}(v_j(x)) & \text{otherwise.} \end{cases}$$

It is clear that $|v_{j,\varepsilon}(x)| \le |v_j(x)|$ and that $v_{j,\varepsilon}(x) \to v_j(x)$ for a.e. $x \in \omega$ as $\varepsilon \to 0$, and hence $\mathbf{v}_\varepsilon \to \mathbf{v}$ in $L^2(\omega; \mathbb{R}^m)$. A simple inspection convinces us that $\mathbf{v}_\varepsilon \in C_{\bar{\mathbf{u}}}$. Let us check that $\bar{\mathbf{u}} + \rho \mathbf{v}_\varepsilon \in U_{ad}$ for every $0 < \rho < \varepsilon^2$. If $\bar{d}_j(x) \ne 0$, then $v_{j,\varepsilon}(x) = 0$. So $\bar{u}_j(x) + \rho v_{j,\varepsilon}(x) = \bar{u}_j(x) \in [\alpha_j, \beta_j]$. For $\bar{d}_j(x) = 0$, we have the following:

(1) If $\bar{u}_j(x) = \alpha_j$, then $v_j(x) \ge 0$ and $v_{j,\varepsilon}(x) \ge 0$. So clearly $\alpha_j \le u_j(x) + \rho v_{j,\varepsilon}(x)$. For the other inequality we write $u_j(x) + \rho v_{j,\varepsilon}(x) \le \alpha_j + \varepsilon^2 \frac{1}{\varepsilon} \le \frac{\alpha_j + \beta_j}{2} < \beta_j$. If $\bar{u}_j(x) = \beta_j$, the argument is completely analogous.

(2) If $\alpha_j < \bar{u}_j(x) < \alpha_j + \varepsilon$, then $\bar{u}_j(x) + \rho v_{j,\varepsilon}(x) = \bar{u}_j(x) \in [\alpha_j, \beta_j]$. The same applies if $\beta_j - \varepsilon < \bar{u}_j(x) < \beta_j$.

(3) If $\alpha_j + \varepsilon \le \bar{u}_j(x) \le \beta_j - \varepsilon$, then on the left side, $\bar{u}_j(x) + \rho v_{j,\varepsilon}(x) \ge \alpha_j + \varepsilon - \varepsilon^2 \frac{1}{\varepsilon} = \alpha_j$, and on the right side $\bar{u}_j(x) + \rho v_{j,\varepsilon}(x) \le \beta_j - \varepsilon + \varepsilon^2 \frac{1}{\varepsilon} = \beta_j$.

Thus $\bar{\mathbf{u}} + \rho \mathbf{v}_\varepsilon$ belongs to U_{ad}. Making a second order Taylor expansion of J at $\bar{\mathbf{u}}$ and taking into account that it is a local minimum for $\rho < \varepsilon^2$ small enough, there exists $0 < \theta_\rho < \rho$ such that

$$0 \le J(\bar{\mathbf{u}} + \rho \mathbf{v}_\varepsilon) - J(\bar{\mathbf{u}}) = \rho J'(\bar{\mathbf{u}}) \mathbf{v}_\varepsilon + \frac{\rho^2}{2} J''(\bar{\mathbf{u}} + \theta_\rho \mathbf{v}_\varepsilon) \mathbf{v}_\varepsilon^2.$$

Since $\mathbf{v}_\varepsilon \in C_{\bar{\mathbf{u}}}$, (3.13) implies that $J'(\bar{\mathbf{u}}) \mathbf{v}_\varepsilon = 0$. Therefore the above inequality leads to $J''(\bar{\mathbf{u}} + \theta_\rho \mathbf{v}_\varepsilon) \mathbf{v}_\varepsilon^2 \ge 0$. Now we must take the limit as $\rho \to 0$ to get $J''(\bar{\mathbf{u}}) \mathbf{v}_\varepsilon^2 \ge 0$. Next it is enough to take the limit as $\varepsilon \to 0$. To do this, let us recall the expression of $J''(\bar{\mathbf{u}})$ provided by (3.7):

$$J''(\bar{\mathbf{u}}) \mathbf{v}_\varepsilon^2 = \int_\Omega (|\mathbf{z}_{\mathbf{v}_\varepsilon}|^2 - 2(\mathbf{z}_{\mathbf{v}_\varepsilon} \cdot \nabla) \mathbf{z}_{\mathbf{v}_\varepsilon} \cdot \boldsymbol{\Phi}_{\bar{\mathbf{u}}}) \, dx + N \int_\omega |\mathbf{v}_\varepsilon|^2 dx$$

$$\longrightarrow \int_\Omega (|\mathbf{z}_{\mathbf{v}}|^2 - 2(\mathbf{z}_{\mathbf{v}} \cdot \nabla) \mathbf{z}_{\mathbf{v}} \cdot \boldsymbol{\Phi}_{\bar{\mathbf{u}}}) \, dx + N \int_\omega |\mathbf{v}|^2 dx = J''(\bar{\mathbf{u}}) \mathbf{v}^2 \quad \text{as } \varepsilon \to 0. \qquad \square$$

The following result is an obvious consequence of the previous theorem and the expression of J'' given by (3.7).

COROLLARY 3.7. *Let $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ be a nonsingular local solution of* (P) *and let $\bar{\boldsymbol{\Phi}}$ be the corresponding adjoint state. Then*

$$\int_\Omega (|\mathbf{z}|^2 - 2(\mathbf{z} \cdot \nabla) \mathbf{z} \cdot \bar{\boldsymbol{\Phi}}) dx + N \int_\omega |\mathbf{v}|^2 dx \ge 0 \tag{3.14}$$

for every $(\mathbf{v}, \mathbf{z})$ satisfying the linearized state equation (2.6) *and $\mathbf{v} \in C_{\bar{\mathbf{u}}}$.*

To state second order sufficient conditions we will *not* suppose that $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ is a nonsingular solution of the Navier–Stokes equations (1.2). The result we are going to state is the following.

THEOREM 3.8. *Let $(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{\boldsymbol{\Phi}}) \in L^2(\omega; \mathbb{R}^m) \times \mathbf{V}_0^1(\Omega) \times \mathbf{V}_0^1(\Omega)$ satisfy* (3.2)–(3.5) *with $\bar{\lambda} = 1$. Let us suppose that*

$$\int_\Omega (\mathbf{z}^2 - 2(\mathbf{z} \cdot \nabla) \mathbf{z} \cdot \bar{\boldsymbol{\Phi}}) dx + N \int_\omega \mathbf{v}^2 dx > 0 \tag{3.15}$$

for every $(\mathbf{v},\mathbf{z}) \neq (0,0)$ *satisfying the linearized state equation* (2.6) *and* $\mathbf{v} \in C_{\bar{\mathbf{u}}}$. *Then there exist* $\varepsilon > 0$ *and* $\mu > 0$ *such that*

$$F(\bar{\mathbf{u}},\bar{\mathbf{y}}) + \frac{\mu}{2}\left(\|\mathbf{u}-\bar{\mathbf{u}}\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{y}-\bar{\mathbf{y}}\|^2_{\mathbf{L}^2(\Omega)}\right) \leq F(\mathbf{u},\mathbf{y})$$

for every $(\mathbf{u},\mathbf{y})$ *satisfying* (1.2), $\mathbf{u} \in U_{ad}$, *and* $\|\mathbf{u}-\bar{\mathbf{u}}\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{y}-\bar{\mathbf{y}}\|^2_{\mathbf{L}^2(\Omega)} \leq \varepsilon^2$.

Proof. Let us suppose the theorem is false. In that case, for all $k \in \mathbb{N}$, there exists $(\mathbf{u}_k,\mathbf{y}_k)$ satisfying (1.2), $\mathbf{u}_k \in U_{ad}$,

$$\|\mathbf{u}_k-\bar{\mathbf{u}}\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{y}_k-\bar{\mathbf{y}}\|^2_{\mathbf{L}^2(\Omega)} < \frac{1}{k^2},$$

and

$$F(\bar{\mathbf{u}},\bar{\mathbf{y}}) + \frac{1}{k}\left(\|\mathbf{u}_k-\bar{\mathbf{u}}\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{y}_k-\bar{\mathbf{y}}\|^2_{\mathbf{L}^2(\Omega)}\right) > F(\mathbf{u}_k,\mathbf{y}_k). \tag{3.16}$$

Since the sequence $\{\mathbf{u}_k\}_{k=1}^\infty$ is bounded in $L^2(\omega;\mathbb{R}^m)$, Theorem 2.2 implies that $\{\mathbf{y}_k\}_{k=1}^\infty$ is bounded in $\mathbf{H}^2(\Omega)\cap\mathbf{V}_0^1(\Omega)$. Let us set

$$\rho_k = \sqrt{\|\mathbf{u}_k-\bar{\mathbf{u}}\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{y}_k-\bar{\mathbf{y}}\|^2_{\mathbf{L}^2(\Omega)}},\ \mathbf{v}_k = \frac{\mathbf{u}_k-\bar{\mathbf{u}}}{\rho_k},\ \mathbf{z}_k = \frac{\mathbf{y}_k-\bar{\mathbf{y}}}{\rho_k}.$$

Clearly $\|\mathbf{v}_k\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{z}_k\|^2_{\mathbf{L}^2(\Omega)} = 1$, and hence there exist weakly convergent subsequences in $L^2(\omega;\mathbb{R}^m)$ and $\mathbf{L}^2(\Omega)$, still indexed by k, such that $\mathbf{v}_k \rightharpoonup \mathbf{v}$, $\mathbf{z}_k \rightharpoonup \mathbf{z}$. We are going to check that the pair $(\mathbf{v},\mathbf{z})$ satisfies the linearized equation (2.6) and $\mathbf{v} \in C_{\bar{\mathbf{u}}}$.

The pair $(\mathbf{v}_k,\mathbf{z}_k)$ satisfies the equation

$$\begin{cases} -\nu\Delta\mathbf{z}_k + (\bar{\mathbf{y}}\cdot\nabla)\mathbf{z}_k + (\mathbf{z}_k\cdot\nabla)\mathbf{y}_k + \nabla\pi_k = \mathcal{C}\mathbf{v}_k \text{ in } \Omega, \\ \operatorname{div}\mathbf{z}_k = 0 \text{ in } \Omega,\ \mathbf{z}_k = 0 \text{ on } \Gamma, \end{cases} \tag{3.17}$$

where $\pi_k = (\bar{p}-p_k)/\rho_k$, which is equivalent to the variational formulation

$$a(\mathbf{z}_k,\mathbf{z}) + b(\bar{\mathbf{y}},\mathbf{z}_k,\mathbf{z}) + b(\mathbf{z}_k,\mathbf{y}_k,\mathbf{z}) = (\mathcal{C}\mathbf{v}_k,\mathbf{z})\ \forall\,\mathbf{z} \in \mathbf{V}_0^1(\Omega). \tag{3.18}$$

Taking $\mathbf{z} = \mathbf{z}_k$ and using (2.4), we obtain

$$a(\mathbf{z}_k,\mathbf{z}_k) = (\mathcal{C}\mathbf{v}_k,\mathbf{z}_k) - \int_\Omega (\mathbf{z}_k\cdot\nabla)\mathbf{y}_k\cdot\mathbf{z}_k dx.$$

Using the equality $\|\mathbf{v}_k\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{z}_k\|^2_{\mathbf{L}^2(\Omega)} = 1$ and the imbedding $\mathbf{H}^1(\Omega)\subset\mathbf{L}^4(\Omega)$, we obtain

$$\nu\|\mathbf{z}_k\|^2_{\mathbf{H}^1(\Omega)} \leq \|\mathcal{C}\| + \|\mathbf{y}_k\|_{\mathbf{H}^1(\Omega)}\|\mathbf{z}_k\|^2_{\mathbf{L}^4(\Omega)} \leq C(1+\|\mathbf{z}_k\|^2_{\mathbf{L}^4(\Omega)}),$$

because $(\mathbf{y}_k)_k$ is bounded in $\mathbf{H}^1(\Omega)$. From the well-known interpolation inequality when $d=3$ (see Temam [20, Lemma 3.5, p. 296]),

$$\|\zeta\|_{L^4(\Omega)} \leq \sqrt{2}\|\zeta\|^{1/4}_{L^2(\Omega)}\|\zeta\|^{3/4}_{H^1(\Omega)}\quad \forall\zeta \in H_0^1(\Omega),$$

and the bound $\|\mathbf{z}_k\|_{\mathbf{L}^2(\Omega)} \leq 1$, it follows that

$$\|\mathbf{z}_k\|^2_{\mathbf{H}^1(\Omega)} \leq C\left(1+\|\mathbf{z}_k\|^{3/2}_{\mathbf{H}^1(\Omega)}\right).$$

Thus the sequence $\{\mathbf{z}_k\}$ is bounded in $\mathbf{V}_0^1(\Omega)$, and therefore $\{\mathbf{z}_k\}$ converges strongly to $\mathbf{z}$ in $\mathbf{L}^2(\Omega)$. Now we can take the limit in (3.18), and we obtain that $(\mathbf{v}, \mathbf{z})$ satisfies (2.6).

Let us now check that $\mathbf{v} \in C_{\bar{\mathbf{u}}}$. The sign condition (3.11)–(3.12) is satisfied by $v_{k,j}$, and this is conserved when we pass to the weak limit because the set of functions satisfying these sign conditions is closed and convex in $L^2(\omega; \mathbb{R}^m)$. On the other hand, using condition (3.16), for all k, we have

$$\begin{aligned}\frac{\rho_k}{k} &> \frac{F(\bar{\mathbf{u}} + \rho_k \mathbf{v}_k, \bar{\mathbf{y}} + \rho_k \mathbf{z}_k) - F(\bar{\mathbf{u}}, \bar{\mathbf{y}})}{\rho_k} \\ &= \frac{1}{2} \int_\Omega \frac{|\bar{\mathbf{y}} + \rho_k \mathbf{z}_k - \mathbf{y}_d|^2 - |\bar{\mathbf{y}} - \mathbf{y}_d|^2}{\rho_k} dx + \frac{N}{2} \int_\omega \frac{|\bar{\mathbf{u}} + \rho_k \mathbf{v}_k|^2 - |\bar{\mathbf{u}}|^2}{\rho_k} dx \\ &= \frac{1}{2} \int_\Omega (2(\bar{\mathbf{y}} - \mathbf{y}_d) + \rho_k \mathbf{z}_k) \cdot \mathbf{z}_k \, dx + \frac{N}{2} \int_\omega (2\bar{\mathbf{u}} + \rho_k \mathbf{v}_k) \cdot \mathbf{v}_k \, dx.\end{aligned}$$

Since $\|\mathbf{v}_k\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{z}_k\|^2_{\mathbf{L}^2(\Omega)} = 1$, $\rho_k < 1/k$ converges to 0, $\mathbf{z}_k \to \mathbf{z}$ in $\mathbf{L}^2(\Omega)$, and $\mathbf{v}_k \rightharpoonup \mathbf{v}$ weakly in $L^2(\omega; \mathbb{R}^m)$, we can pass to the limit when k tends to infinity, and we get

$$\int_\Omega (\bar{\mathbf{y}} - \mathbf{y}_d) \cdot \mathbf{z} \, dx + N \int_\omega \bar{\mathbf{u}} \cdot \mathbf{v} \, dx \le 0,$$

which is exactly

$$\int_\omega \bar{\mathbf{d}}(x) \cdot \mathbf{v}(x) \le 0.$$

The sign condition (3.5) implies that $\bar{d}_j(x) v_j(x) \ge 0$ for a.e. $x \in \omega$; therefore the above inequality is equivalent to

$$\sum_{j=1}^m \int_\Omega |\bar{d}_j(x) v_j(x)| \, dx \le 0.$$

Thus if $\bar{d}_j(x) \ne 0$, $v_j(x) = 0$, $1 \le j \le m$, and hence $\mathbf{v} \in C_{\bar{\mathbf{u}}}$.

Making a second order Taylor expansion of F at $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$, with condition (3.16), we obtain

$$\frac{1}{\rho_k} \left(\partial_{\mathbf{u}} F(\bar{\mathbf{u}}, \bar{\mathbf{y}}) \mathbf{v}_k + \partial_{\mathbf{y}} F(\bar{\mathbf{u}}, \bar{\mathbf{y}}) \mathbf{z}_k\right) + \frac{1}{2} \int_\Omega |\mathbf{z}_k|^2 \, dx + \frac{N}{2} \int_\omega |\mathbf{v}_k|^2 \, dx < \frac{1}{k}. \tag{3.19}$$

Notice that the pair $(\mathbf{v}_k, \mathbf{z}_k)$ satisfies (3.17), but does not satisfy the linearized equation (2.6). Thus $\frac{1}{\rho_k} \left(\partial_{\mathbf{u}} F(\bar{\mathbf{u}}, \bar{\mathbf{y}}) \mathbf{v}_k + \partial_{\mathbf{y}} F(\bar{\mathbf{u}}, \bar{\mathbf{y}}) \mathbf{z}_k\right)$ is not equal to $\int_\omega \bar{\mathbf{d}}(x) \cdot \mathbf{v}_k(x) dx$. We can write (3.17) as follows:

$$\begin{aligned}&-\nu \Delta \mathbf{z}_k + (\bar{\mathbf{y}} \cdot \nabla) \mathbf{z}_k + (\mathbf{z}_k \cdot \nabla) \bar{\mathbf{y}} + \nabla \pi_k = \mathcal{C} \mathbf{v}_k - (\mathbf{z}_k \cdot \nabla)(\mathbf{y}_k - \bar{\mathbf{y}}) \text{ in } \Omega, \\ &\operatorname{div} \mathbf{z}_k = 0 \text{ in } \Omega, \ \mathbf{z}_k = 0 \text{ on } \Gamma.\end{aligned}$$

Since

$$\partial_{\mathbf{u}} F(\bar{\mathbf{u}}, \bar{\mathbf{y}}) \mathbf{v}_k + \partial_{\mathbf{y}} F(\bar{\mathbf{u}}, \bar{\mathbf{y}}) \mathbf{z}_k = \int_\Omega (\bar{\mathbf{y}} - \mathbf{y}_d) \cdot \mathbf{z}_k dx + N \int_\omega \bar{\mathbf{u}} \cdot \mathbf{v}_k dx,$$

using the adjoint state $\bar{\mathbf{\Phi}}$ and making an integration by parts, we get that

$$\int_\Omega (\bar{\mathbf{y}} - \mathbf{y}_d) \cdot \mathbf{z}_k dx = \int_\Omega \bar{\mathbf{\Phi}} \cdot (\mathcal{C}\mathbf{v}_k - (\mathbf{z}_k \cdot \nabla)(\mathbf{y}_k - \bar{\mathbf{y}}))\, dx,$$

and therefore

$$\frac{1}{\rho_k} (\partial_{\mathbf{u}} F(\bar{\mathbf{u}}, \bar{\mathbf{y}})\mathbf{v}_k + \partial_{\mathbf{y}} F(\bar{\mathbf{u}}, \bar{\mathbf{y}})\mathbf{z}_k) = \frac{1}{\rho_k} \int_\omega \bar{\mathbf{d}}(x) \cdot \mathbf{v}_k(x) dx - \int_\Omega (\mathbf{z}_k \cdot \nabla)\mathbf{z}_k \cdot \bar{\mathbf{\Phi}} dx.$$

Since $\mathbf{v}_k$ satisfy the sign condition, we have $\bar{\mathbf{d}}(x) \cdot \mathbf{v}_k(x) \geq 0$; therefore (3.19) leads to

$$-2 \int_\Omega (\mathbf{z}_k \cdot \nabla)\mathbf{z}_k \cdot \bar{\mathbf{\Phi}} dx + \int_\Omega |\mathbf{z}_k|^2 dx + N \int_\omega |\mathbf{v}_k|^2 dx < \frac{2}{k} \quad \forall k.$$

Taking the inferior limit in this inequality we deduce

$$\int_\Omega (|\mathbf{z}|^2 - 2(\mathbf{z} \cdot \nabla)\mathbf{z} \cdot \bar{\mathbf{\Phi}}) dx + N \int_\omega |\mathbf{v}|^2 dx \leq 0.$$

Since $\mathbf{v} \in C_{\bar{\mathbf{u}}}$ and the pair $(\mathbf{v}, \mathbf{z})$ satisfies the linearized equation (2.6), this is possible only if $(\mathbf{v}, \mathbf{z}) = (0, 0)$.

The sequence $\{\mathbf{z}_k\}_{k=1}^\infty$ converges strongly in $\mathbf{L}^2(\Omega)$ and weakly in $\mathbf{V}_0^1(\Omega)$. Since $\bar{\mathbf{\Phi}} \in \mathbf{L}^\infty(\Omega)$, by passing to the limit when k tends to infinity, we obtain

$$-2 \int_\Omega (\mathbf{z}_k \cdot \nabla)\mathbf{z}_k \cdot \bar{\mathbf{\Phi}} dx + \int_\Omega |\mathbf{z}_k|^2 dx \to -2 \int_\Omega (\mathbf{z} \cdot \nabla)\mathbf{z} \cdot \bar{\mathbf{\Phi}} dx + \int_\Omega |\mathbf{z}|^2 dx = 0.$$

The last three relations imply that $\mathbf{v}_k \to 0$ strongly in $L^2(\omega; \mathbb{R}^m)$. So we have proved that $(\mathbf{v}_k, \mathbf{z}_k) \to 0$ strongly in $L^2(\omega)^m \times \mathbf{L}^2(\Omega)$, which contradicts the fact that

$$\|\mathbf{v_k}\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{z_k}\|^2_{\mathbf{L}^2(\Omega)} = 1.$$

The proof is complete. □

The sufficient condition (3.15) is the best possible. Actually the gap between (3.15) and the second order necessary condition (3.14) is the same as in finite dimension. In the case of nonsingular solutions we have the following result analogous to Theorem 3.6.

COROLLARY 3.9. *Let us assume that* $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ *is a nonsingular solution of* (1.2) *and* $(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{\mathbf{\Phi}})$ *satisfies* (3.2)–(3.5) *with* $\bar{\lambda} = 1$. *Then* (3.15) *is equivalent to* $J''(\bar{\mathbf{u}})\mathbf{v}^2 > 0$ *for every* $\mathbf{v} \in C_{\bar{\mathbf{u}}} \setminus \{0\}$.

This corollary is an immediate consequence of (3.7) and the fact that $\mathbf{z} = \mathbf{z_v}$ if $(\mathbf{v}, \mathbf{z})$ satisfies (2.6).

To make the numerical analysis of control problem (P), we will use the following equivalent condition to (3.15), which may seem stronger but is not, as we will see below. Given $\tau > 0$, let us define a bigger cone than $C_{\bar{\mathbf{u}}}$ in the following way:

$$v_j(x) = 0 \quad \text{if } |\bar{d}_j(x)| > \tau, \tag{3.20}$$

$$v_j(x) \geq 0 \quad \text{if } -\infty < \alpha_j = \bar{u}_j(x) \text{ and } |\bar{d}_j(x)| \leq \tau, \tag{3.21}$$

$$v_j(x) \leq 0 \quad \text{if } \bar{u}_j(x) = \beta_j < \infty \text{ and } |\bar{d}_j(x)| \leq \tau, \tag{3.22}$$

and

$$C^\tau_{\bar{\mathbf{u}}} = \Big\{ \mathbf{v} \in L^2(\omega; \mathbb{R}^m) \mid \mathbf{v} \text{ satisfies (3.20)–(3.22)} \Big\}.$$

THEOREM 3.10. *Let* $(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{\mathbf{\Phi}}) \in L^2(\omega;\mathbb{R}^m) \times \mathbf{V}_0^1(\Omega) \times \mathbf{V}_0^1(\Omega)$ *satisfy* (3.2)–(3.5) *with* $\bar{\lambda} = 1$. *Then the condition* (3.15) *is equivalent to the existence of* $\delta > 0$ *and* $\tau > 0$ *such that*

$$\int_\Omega (|\mathbf{z}|^2 - 2(\mathbf{z}\cdot\nabla)\mathbf{z}\cdot\bar{\mathbf{\Phi}})\,dx + N\int_\omega |\mathbf{v}|^2\,dx \geq \delta\left(\|\mathbf{v}\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{z}\|^2_{\mathbf{L}^2(\Omega)}\right) \tag{3.23}$$

for every $(\mathbf{v},\mathbf{z})$ *satisfying the linearized state equation* (2.6) *and* $\mathbf{v} \in C^\tau_{\bar{\mathbf{u}}}$.

Proof. Notice that $C_{\bar{\mathbf{u}}} = C^0_{\bar{\mathbf{u}}} \subseteq C^\tau_{\bar{\mathbf{u}}}$; therefore (3.23) implies (3.15).

Suppose that (3.15) holds and (3.23) is false. In that case, for every $k \in \mathbb{N}$ there exists a pair $(\mathbf{v}_k, \mathbf{z}_k)$ satisfying the linearized state equation (2.6), $\mathbf{v}_k \in C^{1/k}_{\bar{\mathbf{u}}}$, and

$$\int_\Omega (|\mathbf{z}_k|^2 - 2(\mathbf{z}_k\cdot\nabla)\mathbf{z}_k\cdot\bar{\mathbf{\Phi}})dx + N\int_\omega \mathbf{v}_k^2 dx < \frac{1}{k}\left(\|\mathbf{v}_k\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{z}_k\|^2_{\mathbf{L}^2(\Omega)}\right). \tag{3.24}$$

We can suppose that $\|\mathbf{v}_k\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{z}_k\|^2_{\mathbf{L}^2(\Omega)} = 1$; otherwise we can redefine $\mathbf{v}_k = \mathbf{v}_k/\rho_k$ and $\mathbf{z}_k = \mathbf{z}_k/\rho_k$, with $\rho_k = (\|\mathbf{v}_k\|^2_{L^2(\omega;\mathbb{R}^m)} + \|\mathbf{z}_k\|^2_{\mathbf{L}^2(\Omega)})^{1/2}$. Then there exist two weakly convergent subsequences in $L^2(\omega;\mathbb{R}^m)$ and $\mathbf{L}^2(\Omega)$, still indexed by k, such that $\mathbf{v}_k \rightharpoonup \mathbf{v}$ and $\mathbf{z}_k \rightharpoonup \mathbf{z}$. Repeating the argument of the proof of Theorem 3.8, we deduce that the pair $(\mathbf{v},\mathbf{z})$ satisfies the linearized equation (2.6) and $\{\mathbf{z}_k\}_{k=1}^\infty$ is bounded in $\mathbf{V}_0^1(\Omega)$. Thus $\{\mathbf{z}_k\}_{k=1}^\infty$ converges strongly in $\mathbf{L}^4(\Omega)$. Let us prove that $\mathbf{v} \in C_{\bar{\mathbf{u}}}$. The sign condition (3.11)–(3.12) is again trivial since every $\mathbf{v}_k$ satisfies it. To check condition (3.10) we are going to prove that if $|\bar{d}_j(x)| \neq 0$, then $v_j(x) = 0$. Let us fix $\varepsilon > 0$ and define $\omega_\varepsilon = \{x \in \omega : |d_j(x)| > \varepsilon\}$. Notice that $\int_{\omega_\varepsilon} v_{j,k}(x)\bar{d}_j(x)dx \to \int_{\omega_\varepsilon} v_j(x)\bar{d}_j(x)dx$ when k tends to infinity. From the definition of $C^{1/k}_{\bar{\mathbf{u}}}$ it follows that for $k > 1/\varepsilon$ all the terms of the sequence $\{\int_{\omega_\varepsilon} v_{j,k}(x)\bar{d}_j(x)dx\}_k$ are 0, and so the limit is also 0. Since $\mathbf{v}$ satisfies the sign condition (3.5), this can happen only if $v_j(x) = 0$ almost everywhere in ω_ε. Since ε is arbitrarily small, we conclude that $v_j(x) = 0$ for a.e. x such that $|\bar{d}_j(x)| \neq 0$, and so $\mathbf{v} \in C_{\bar{\mathbf{u}}}$.

Finally, taking the lower limit in (3.24) we obtain that

$$\int_\Omega (\mathbf{z}^2 - 2(\mathbf{z}\cdot\nabla)\mathbf{z}\cdot\bar{\mathbf{\Phi}})dx + N\int_\omega \mathbf{v}^2 dx \leq 0.$$

We complete the proof by arguing as at the end of the proof of Theorem 3.8. ∎

COROLLARY 3.11. *Let us assume that* $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ *is a nonsingular solution of* (1.2) *and* $(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{\mathbf{\Phi}})$ *satisfies* (3.2)–(3.5) *with* $\bar{\lambda} = 1$. *Then* (3.15) *is equivalent to the existence of* $\delta > 0$ *and* $\tau > 0$ *such that*

$$J''(\bar{\mathbf{u}})\mathbf{v}^2 \geq \delta\|\mathbf{v}\|^2_{L^2(\omega;\mathbb{R}^m)} \quad \forall\, \mathbf{v} \in C^\tau_{\bar{\mathbf{u}}}. \tag{3.25}$$

This a consequence of Theorem 3.10 and the expression of $J''(\bar{\mathbf{u}})$ stated in (3.7).

4. Numerical approximation of the control problem.

4.1. Numerical analysis of the state equation. Let $\mathbf{X}_h \subset \mathbf{H}_0^1(\Omega)$ and $M_h \subset L_0^2(\Omega)$ be two finite dimensional spaces satisfying the assumptions (H1)–(H3) stated below.

(H1) (Approximation property of $\mathbf{X}_h$). There exists an operator $r_h \in \mathcal{L}(\mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega), \mathbf{X}_h)$ such that

$$\begin{aligned}
&\text{(a) } \|\mathbf{y} - r_h\mathbf{y}\|_{\mathbf{H}_0^1(\Omega)} \le Ch\|\mathbf{y}\|_{\mathbf{H}^2(\Omega)} \ \forall\, \mathbf{y} \in \mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega),\\
&\text{(b) } \|\mathbf{y} - r_h\mathbf{y}\|_{\mathbf{L}^2(\Omega)} \le Ch^2\|\mathbf{y}\|_{\mathbf{H}^2(\Omega)} \ \forall\, \mathbf{y} \in \mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega),\\
&\text{(c) } \|\mathbf{y} - r_h\mathbf{y}\|_{\mathbf{L}^\infty(\Omega)} \le Ch^{2-d/2}\|\mathbf{y}\|_{\mathbf{H}^2(\Omega)} \ \forall\, \mathbf{y} \in \mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega),\\
&\text{(d) } \|\mathbf{y}_h\|_{\mathbf{L}^\infty(\Omega)} \le Ch^{-d/2}\|\mathbf{y}_h\|_{\mathbf{L}^2(\Omega)} \ \forall\, \mathbf{y}_h \in \mathbf{X}_h.
\end{aligned}$$

(H2) (Approximation property of M_h). There exists an operator $s_h \in \mathcal{L}(L_0^2(\Omega), M_h)$ such that

$$\|p - s_h p\|_{L_0^2(\Omega)} \le Ch\|p\|_{H^1(\Omega)} \ \forall\, p \in H^1(\Omega) \cap L_0^2(\Omega).$$

(H3) (Uniform inf-sup condition). For each $p_h \in M_h$ there exists $\mathbf{y}_h \in \mathbf{X}_h$ such that

$$(p_h, \operatorname{div} \mathbf{y}_h) = \|p_h\|^2_{L_0^2(\Omega)} \quad \text{and} \quad \|\mathbf{y}_h\|_{\mathbf{H}_0^1(\Omega)} \le C\|p_h\|_{L_0^2(\Omega)},$$

where $C > 0$ is independent of h, p_h, and $\mathbf{y}_h$.

Remark 4.1. Assumptions (H1)(b), (H1)(c), and (H1)(d) are needed to establish uniform convergence for the approximation of the state and the adjoint state (cf. Lemmas 4.10 and 4.13). In particular, if we use the finite element method when the family of triangulations is quasi-uniform, the above assumptions are satisfied for the Taylor–Hood finite element method and for the (P1-Bubble, P1) finite element method (see [12, p. 98, Lemma A.7 on p. 103, and Chapter 2]). The quasi-uniformity condition can be relaxed in some cases. For instance, Eriksson [11] gives some conditions on a locally refined family of triangulations in order to have an inverse inequality similar to (H1)(d).

Assumption (H3) is equivalent to the classical inf-sup condition. See Girault–Raviart [12, Remark II.1.4].

For $\rho > 0$, $\bar{\mathbf{y}} \in \mathbf{H}_0^1(\Omega)$, $\bar{p} \in L_0^2(\Omega)$, and $\bar{\mathbf{u}} \in L^2(\omega; \mathbb{R}^m)$, let us set

$$\begin{aligned}
B_\rho(\bar{\mathbf{y}}) &= \Big\{\mathbf{y} \in \mathbf{H}_0^1(\Omega) \mid \|\mathbf{y} - \bar{\mathbf{y}}\|_{\mathbf{H}_0^1(\Omega)} \le \rho\Big\},\\
B_\rho(\bar{p}) &= \Big\{p \in L_0^2(\Omega) \mid \|p - \bar{p}\|_{L_0^2(\Omega)} \le \rho\Big\},\\
B_\rho(\bar{\mathbf{u}}) &= \Big\{\mathbf{u} \in L^2(\omega; \mathbb{R}^m) \mid \|\mathbf{u} - \bar{\mathbf{u}}\|_{L^2(\omega;\mathbb{R}^m)} \le \rho\Big\}.
\end{aligned}$$

For all $\mathbf{u} \in L^2(\omega; \mathbb{R}^m)$, we define a discrete state equation in $\mathbf{X}_h \times M_h$, associated with (1.2), as follows:

$$\begin{aligned}
&\text{Find } (\mathbf{y}_d, p_h) \in \mathbf{X}_h \times M_h \text{ satisfying}\\
&a(\mathbf{y}_h, \mathbf{w}_h) + b(\mathbf{y}_h, \mathbf{y}_h, \mathbf{w}_h) - (p_h, \operatorname{div} \mathbf{w}_h) = (\mathbf{f} + \mathcal{C}\mathbf{u}, \mathbf{w}_h) \ \forall\, \mathbf{w}_h \in \mathbf{X}_h,\\
&(\lambda_h, \operatorname{div} \mathbf{y}_h) = 0 \ \forall\, \lambda_h \in M_h.
\end{aligned} \tag{4.1}$$

For a given $\mathbf{u} \in L^2(\omega;\mathbb{R}^m)$, this equation does not necessarily have a unique solution $\mathbf{y}_h$. Our main objective in this section is to show that there exist $\rho_1 > 0$ and $\rho_2 > 0$ independent of h, such that, for all $\mathbf{u} \in B_{\rho_2}(\bar{\mathbf{u}})$, (4.1) admits a unique solution in $B_{\rho_1}(\bar{\mathbf{y}}) \times B_{\rho_1}(\bar{p})$. Let T be the bounded linear operator from $\mathbf{H}^{-1}(\Omega)$ to $\mathbf{V}_0^1(\Omega) \times L_0^2(\Omega)$ defined by $T\mathbf{g} = (\mathbf{z}, q)$, where $(\mathbf{z}, q)$ is the solution of

$$-\nu\Delta\mathbf{z} + \nabla q = \mathbf{g} \text{ in } \Omega,\ \operatorname{div}\mathbf{z} = 0 \text{ in } \Omega,\ \mathbf{z} = 0 \text{ on } \Gamma.$$

Let $\mathcal{F}$ be the nonlinear mapping from $L^2(\omega;\mathbb{R}^m) \times \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ into $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ defined by

$$\mathcal{F}(\mathbf{u}, \mathbf{y}, p) = (\mathbf{y}, p) + T[B(\mathbf{y}) - (\mathbf{f} + \mathcal{C}\mathbf{u})].$$

Notice that $\mathcal{F}(\mathbf{u}, \mathbf{y}, p) = 0$ if and only if $A\mathbf{y} + B(\mathbf{y}) = \mathbf{f} + \mathcal{C}\mathbf{u}$ in $\mathbf{V}^{-1}(\Omega)$ and $p \in L_0^2(\Omega)$ satisfies $\nabla p = (I - P)(\mathbf{f} + \mathcal{C}\mathbf{u} + \nu\Delta\mathbf{y} - (\mathbf{y}\cdot\nabla)\mathbf{y})$. The operator $\partial_{(\mathbf{y},p)}\mathcal{F}(\mathbf{u}, \mathbf{y}, p)$ belongs to $\mathcal{L}(\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega))$ and is defined by

$$\partial_{(\mathbf{y},p)}\mathcal{F}(\mathbf{u}, \mathbf{y}, p)(\mathbf{z}, q) = (\mathbf{z}, q) + T[B'(\mathbf{y})\mathbf{z}].$$

Observe that $\partial_{(\mathbf{y},p)}\mathcal{F}(\mathbf{u}, \mathbf{y}, p)$ does not depend on $\mathbf{u} \in L^2(\omega;\mathbb{R}^m)$ and $p \in L_0^2(\Omega)$.

LEMMA 4.2. *Let $(\bar{\mathbf{u}}, \bar{\mathbf{y}}) \in L^2(\omega;\mathbb{R}^m) \times \mathbf{V}_0^1(\Omega)$ be a solution of (1.2), with associated pressure $\bar{p}$. Then $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ is a nonsingular solution if and only if $\partial_{(\mathbf{y},p)}\mathcal{F}(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{p})$ is an automorphism in $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$.*

Proof. Let us assume that $(\bar{\mathbf{u}}, \bar{\mathbf{y}}) \in L^2(\omega;\mathbb{R}^m) \times \mathbf{V}_0^1(\Omega)$ is a nonsingular solution of (1.2). Let $(\hat{\mathbf{y}}, \hat{p})$ be in $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$. We have to check that there exists a unique pair $(\mathbf{y}, p) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ such that $(\mathbf{y}, p) + T[B'(\bar{\mathbf{y}})\mathbf{y}] = (\hat{\mathbf{y}}, \hat{p})$. Let $(\mathbf{y}_0, p_0) \in \mathbf{V}_0^1(\Omega) \times L_0^2(\Omega)$ be the unique solution of the equation

$$-\nu\Delta\mathbf{y}_0 + B'(\bar{\mathbf{y}})\mathbf{y}_0 + \nabla p_0 = -B'(\bar{\mathbf{y}})\hat{\mathbf{y}} \text{ in } \Omega,\ \operatorname{div}\mathbf{y}_0 = 0 \text{ in } \Omega,\ \mathbf{y}_0 = 0 \text{ on } \Gamma.$$

Set $\mathbf{y} = \mathbf{y}_0 + \hat{\mathbf{y}}$ and $p = p_0 + \hat{p}$. The equality $(\mathbf{y}, p) + T[B'(\bar{\mathbf{y}})(\mathbf{y})] = (\hat{\mathbf{y}}, \hat{p})$, i.e., $T[-B'(\bar{\mathbf{y}})\mathbf{y}_0 - B'(\bar{\mathbf{y}})\hat{\mathbf{y}}] = (\mathbf{y}_0, p_0)$, follows from the definition of T and of $(\mathbf{y}_0, p_0)$. So we have proved the surjectivity of $\partial_{(\mathbf{y},p)}\mathcal{F}(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{p})$. For the injectivity let us assume that $(\mathbf{y}, p) + T[B'(\bar{\mathbf{y}})\mathbf{y}] = (0, 0)$. This implies that $A\mathbf{y} + B'(\bar{\mathbf{y}})\mathbf{y} = 0$ in $\mathbf{V}^{-1}(\Omega)$; then $\mathbf{y} = 0$ and therefore $p = 0$ too.

Conversely, let us assume that $\partial_{(\mathbf{y},p)}\mathcal{F}(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{p})$ is an automorphism in $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$. Let $\mathbf{g} \in \mathbf{V}^{-1}(\Omega)$. Let $(\hat{\mathbf{y}}, \hat{p}) \in \mathbf{V}_0^1(\Omega) \times L_0^2(\Omega)$ be the solution of the equation

$$-\nu\Delta\hat{\mathbf{y}} + \nabla\hat{p} = \mathbf{g}.$$

Let $(\mathbf{y}, p) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ be the solution of the equation

$$\partial_{(\mathbf{y},p)}\mathcal{F}(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{p})(\mathbf{y}, p) = (\hat{\mathbf{y}}, \hat{p}).$$

It is easy to check that $\mathbf{y} \in \mathbf{V}_0^1(\Omega)$ is the unique solution of $A\mathbf{y} + B'(\bar{\mathbf{y}})\mathbf{y} = \mathbf{g}$. ∎

Let T_h be the bounded linear operator from $\mathbf{H}^{-1}(\Omega)$ to $\mathbf{X}_h \times M_h$ defined by $T_h\mathbf{g} = (\mathbf{z}_h, q_h)$, where $(\mathbf{z}_h, q_h) \in \mathbf{X}_h \times M_h$ is the solution of

$$\begin{aligned} &a(\mathbf{z}_h, \mathbf{w}_h) - (q_h, \operatorname{div}\mathbf{w}_h) = (\mathbf{g}, \mathbf{w}_h)\ \forall\, \mathbf{w}_h \in \mathbf{X}_h, \\ &(\lambda_h, \operatorname{div}\mathbf{z}_h) = 0\ \forall\, \lambda_h \in M_h. \end{aligned}$$

Let $\mathcal{F}_h$ be the nonlinear mapping from $L^2(\omega;\mathbb{R}^m) \times \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ into $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ defined by

$$\mathcal{F}_h(\mathbf{u}, \mathbf{y}, p) = (\mathbf{y}, p) + T_h[B(\mathbf{y}) - (\mathbf{f} + \mathcal{C}\mathbf{u})].$$

Remark 4.3. Notice that if $\mathcal{F}_h(\mathbf{u},\mathbf{y},p)=0$, then $(\mathbf{y},p)$ belongs to $\mathbf{X}_h\times M_h$ and is a solution of (4.1). Conversely if $(\mathbf{y},p)\in\mathbf{X}_h\times M_h$ is a solution of (4.1), then $\mathcal{F}_h(\mathbf{u},\mathbf{y},p)=0$.

Now we want to prove that if $\bar{\mathbf{y}}$ is nonsingular and if $\|\mathbf{y}-\bar{\mathbf{y}}\|_{\mathbf{H}_0^1(\Omega)}$ is small enough, then $\partial_{(\mathbf{y},p)}\mathcal{F}_h(\mathbf{u},\mathbf{y},p)$ is an automorphism in $\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)$. For that we make the following additional and usual assumptions concerning the approximation results for the Stokes problem.

(S1) $\lim_{h\to 0}\|(T-T_h)\mathbf{g}\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}=0\ \forall\,\mathbf{g}\in\mathbf{H}^{-1}(\Omega)$.

(S2) $\|(T-T_h)\mathbf{g}\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}\le Ch\|\mathbf{g}\|_{\mathbf{L}^2(\Omega)}\ \forall\,\mathbf{g}\in\mathbf{L}^2(\Omega)$.

Before proving the desired property of $\partial_{(\mathbf{y},p)}\mathcal{F}_h(\mathbf{u},\mathbf{y},p)$, we establish several lemmas.

LEMMA 4.4. *There exists $C>0$ independent of h such that*

$$\|T_h\|_{\mathcal{L}(\mathbf{H}^{-1}(\Omega),\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}\le C.$$

Proof. We want to estimate $\sup\{\|T_h\mathbf{g}\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}\mid\|\mathbf{g}\|_{\mathbf{H}^{-1}(\Omega)}\le 1\}$. Recall that $T_h\mathbf{g}$ is the solution $(\mathbf{z}_h,q_h)$ to the discrete Stokes problem

$$\begin{aligned}&a(\mathbf{z}_h,\mathbf{w}_h)-(q_h,\operatorname{div}\mathbf{w}_h)=(g,\mathbf{w}_h)\ \forall\,\mathbf{w}_h\in\mathbf{X}_h,\\&(\lambda_h,\operatorname{div}\mathbf{z}_h)=0\ \forall\,\lambda_h\in M_h.\end{aligned}$$

Taking $\mathbf{w}_h=\mathbf{z}_h$, we obtain $\|\mathbf{z}_h\|_{\mathbf{H}_0^1(\Omega)}\le C\|\mathbf{g}\|_{\mathbf{H}^{-1}(\Omega)}$. The estimate for the pressure q_h follows from inf-sup condition (H3). Indeed if we take $\mathbf{w}_h$ such that $(q_h,\operatorname{div}\mathbf{w}_h)=\|q_h\|^2_{L_0^2(\Omega)}$ and $\|\mathbf{w}_h\|_{\mathbf{H}_0^1(\Omega)}\le C\|q_h\|_{L_0^2(\Omega)}$, it is clear that $\|q_h\|_{L_0^2(\Omega)}\le C\|g\|_{\mathbf{H}^{-1}(\Omega)}$. ∎

We will need the following standard result.

LEMMA 4.5. *Let X be a Banach space, $A\in\mathcal{L}(X)$ invertible and $B\in\mathcal{L}(X)$. If $\|A-B\|_{\mathcal{L}(X)}<1/\|A^{-1}\|_{\mathcal{L}(X)}$, then B is invertible. If $\|A-B\|_{\mathcal{L}(X)}<1/(2\|A^{-1}\|_{\mathcal{L}(X)})$, then $\|B^{-1}\|_{\mathcal{L}(X)}\le 2\|A^{-1}\|_{\mathcal{L}(X)}$.*

Proof. $A^{-1}B=I-A^{-1}(A-B)$. Since $\|A^{-1}(A-B)\|\le\|A^{-1}\|\,\|A-B\|<1$, we have that $A^{-1}B$ is invertible and so is B.

$B^{-1}A=(I-A^{-1}(A-B))^{-1}=\sum_{k=1}^{\infty}(A^{-1}(A-B))^k$. So $\|B^{-1}\|\le\|A^{-1}\|/(1-\|A^{-1}(A-B)\|)\le 2\|A^{-1}\|$. ∎

LEMMA 4.6. *Let $\bar{\mathbf{y}}\in\mathbf{V}_0^1(\Omega)$ be a nonsingular solution of (2.2). Then for every $\varepsilon>0$ there exist $h_\varepsilon>0$ and $\rho_\varepsilon>0$ such that*

$$\|T[B'(\bar{\mathbf{y}})]-T_h[B'(\mathbf{y})]\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega),\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}<\varepsilon$$

for all $0<h<h_\varepsilon$ and all $\mathbf{y}\in B_{\rho_\varepsilon}(\bar{\mathbf{y}})$.

Proof. With classical calculations we can write

$$\begin{aligned}&\|T[B'(\bar{\mathbf{y}})]-T_h[B'(\mathbf{y})]\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega),\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}\\&\le\|(T-T_h)[B'(\bar{\mathbf{y}})]\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega),\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}+\|T_h[B'(\bar{\mathbf{y}})-B'(\mathbf{y})]\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega),\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}\\&\le\sup_{\|\mathbf{z}\|_{\mathbf{H}_0^1(\Omega)}\le 1}\|(T-T_h)[B'(\bar{\mathbf{y}})\mathbf{z}]\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}\\&+\sup_{\|\mathbf{z}\|_{\mathbf{H}_0^1(\Omega)}\le 1}\|T_h[(B'(\bar{\mathbf{y}})-B'(\mathbf{y}))\mathbf{z}]\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}.\end{aligned}$$

Since $\bar{\mathbf{y}}\in\mathbf{H}^2(\Omega)$, $B'(\bar{\mathbf{y}})\mathbf{z}$ belongs to $\mathbf{L}^2(\Omega)$, and due to assumption (S2) we have

$$\begin{aligned}&\sup_{\|\mathbf{z}\|_{\mathbf{H}_0^1(\Omega)}\le 1}\|(T-T_h)[B'(\bar{\mathbf{y}})\mathbf{z}]\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}\\&\le Ch\sup_{\|\mathbf{z}\|_{\mathbf{H}_0^1(\Omega)}\le 1}\|B'(\bar{\mathbf{y}})\mathbf{z}\|_{\mathbf{L}^2(\Omega)}\le Ch\|\bar{\mathbf{y}}\|_{\mathbf{H}^2(\Omega)}.\end{aligned}$$

On the other hand, using Lemma 4.4 we have

$$\begin{aligned}
&\sup_{\|\mathbf{z}\|_{\mathbf{H}_0^1(\Omega)}\le 1} \|T_h[(B'(\bar{\mathbf{y}})-B'(\mathbf{y}))\mathbf{z}]\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}\\
&\le \|T_h\|_{\mathcal{L}(\mathbf{H}^{-1}(\Omega),\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))} \sup_{\|\mathbf{z}\|_{\mathbf{H}_0^1(\Omega)}\le 1} \|(B'(\bar{\mathbf{y}})-B'(\mathbf{y}))\mathbf{z}\|_{\mathbf{H}^{-1}(\Omega)}\\
&\le C\|\bar{\mathbf{y}}-\mathbf{y}\|_{\mathbf{H}_0^1(\Omega)}.
\end{aligned}$$

Taking h_ε and ρ_ε small enough, we obtain the desired result. ∎

THEOREM 4.7. *Let $(\bar{\mathbf{u}},\bar{\mathbf{y}})\in L^2(\omega;\mathbb{R}^m)\times\mathbf{V}_0^1(\Omega)$ be a nonsingular solution of (2.2) and $\bar p$ the associated pressure. There exist $h_0>0$ and $\rho_0>0$ such that for all $0<h<h_0$ and all $\mathbf{y}\in B_{\rho_0}(\bar{\mathbf{y}})$, $\partial_{(\mathbf{y},p)}\mathcal{F}_h(\mathbf{u},\mathbf{y},p)$ is an automorphism in $\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)$, and*

$$\|\partial_{(\mathbf{y},p)}\mathcal{F}_h(\mathbf{u},\mathbf{y},p)^{-1}\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}\le 2\|\partial_{(\mathbf{y},p)}\mathcal{F}(\bar{\mathbf{u}},\bar{\mathbf{y}},\bar p)^{-1}\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}.$$

Proof. The proof is a straightforward consequence of the previous lemmas. Take

$$\varepsilon=\frac{1}{2\|\partial_{(\mathbf{y},p)}\mathcal{F}(\bar{\mathbf{u}},\bar{\mathbf{y}},\bar p)^{-1}\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}},$$

and set $(h_0,\rho_0)=(h_\varepsilon,\rho_\varepsilon)$, where $(h_\varepsilon,\rho_\varepsilon)$ is the pair corresponding to ε and defined in Lemma 4.6. For every $0<h<h_0$ and all $\mathbf{y}\in B_{\rho_0}(\bar{\mathbf{y}})$, we have

$$\begin{aligned}
&\|\partial_{(\mathbf{y},p)}\mathcal{F}(\bar{\mathbf{u}},\bar{\mathbf{y}},\bar p)-\partial_{(\mathbf{y},p)}\mathcal{F}_h(\mathbf{u},\mathbf{y},p)\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}=\\
&\|T[B'(\bar{\mathbf{y}})]-T_h[B'(\mathbf{y})]\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega),\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}<\varepsilon,
\end{aligned}$$

and the result follows from Lemma 4.5. ∎

THEOREM 4.8. *Let $(\bar{\mathbf{u}},\bar{\mathbf{y}})$ be as in Theorem 4.7; then there exist $\rho_1>0$, $\rho_2>0$, and $h_1>0$ such that for all $0<h<h_1$ and $\mathbf{u}\in B_{\rho_2}(\bar{\mathbf{u}})$, the equation $\mathcal{F}_h(\mathbf{u},\mathbf{y}_h,p_h)=0$ admits a unique solution in $B_{\rho_1}(\bar{\mathbf{y}})\times B_{\rho_1}(\bar p)$.*

Proof. Let ρ_0 and h_0 be the positive constants given by Theorem 4.7. For $\rho\le\rho_0$, $h\le h_0$, and $\mathbf{u}\in B_{\rho^2}(\bar{\mathbf{u}})$, we define the mapping $\Psi_{\mathbf{u}}$ from $B_\rho(\bar{\mathbf{y}})\times B_\rho(\bar p)$ into $\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)$ by

$$\Psi_{\mathbf{u}}(\mathbf{y},p)=(\mathbf{y},p)-\left[\partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}},\bar{\mathbf{y}},\bar p)\right]^{-1}\mathcal{F}_h(\mathbf{u},\mathbf{y},p).$$

It is clear that any fixed point of $\Psi_{\mathbf{u}}$ is a solution of $\mathcal{F}_h(\mathbf{u},\mathbf{y},p)=0$. Let us show that $\Psi_{\mathbf{u}}$ is a strict contraction if ρ is small enough.

(i) First, we show that $\Psi_{\mathbf{u}}$ is a mapping from $B_\rho(\bar{\mathbf{y}})\times B_\rho(\bar p)$ into itself. With the identity $\mathcal{F}(\bar{\mathbf{u}},\bar{\mathbf{y}},\bar p)=0$, and a Taylor formula we obtain

$$\begin{aligned}
&\|\Psi_{\mathbf{u}}(\mathbf{y},p)-(\bar{\mathbf{y}},\bar p)\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}\\
&=\|\left[\partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}},\bar{\mathbf{y}},\bar p)\right]^{-1}\left\{\partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}},\bar{\mathbf{y}},\bar p)(\mathbf{y}-\bar{\mathbf{y}},p-\bar p)\right.\\
&\qquad\left.+[-\mathcal{F}_h(\mathbf{u},\mathbf{y},p)+\mathcal{F}_h(\mathbf{u},\bar{\mathbf{y}},\bar p)]+[-\mathcal{F}_h(\mathbf{u},\bar{\mathbf{y}},\bar p)+\mathcal{F}(\bar{\mathbf{u}},\bar{\mathbf{y}},\bar p)]\right\}\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}\\
&\le C\|\partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}},\bar{\mathbf{y}},\bar p)(\mathbf{y}-\bar{\mathbf{y}},p-\bar p)\\
&\qquad-\int_0^1\partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}},\mathbf{y}_\theta,p_\theta)(\mathbf{y}-\bar{\mathbf{y}},p-\bar p)d\theta\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}\\
&\qquad+C\|\mathcal{F}_h(\mathbf{u},\bar{\mathbf{y}},\bar p)-\mathcal{F}(\bar{\mathbf{u}},\bar{\mathbf{y}},\bar p)\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}\\
&\le C\int_0^1\|\partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}},\mathbf{y}_\theta,p_\theta)-\partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}},\bar{\mathbf{y}},\bar p)\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}d\theta\\
&\qquad\times\|(\mathbf{y}-\bar{\mathbf{y}},p-\bar p)\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}\\
&\qquad+C\|(T_h-T)[B(\bar{\mathbf{y}})-\mathbf{f}]\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}+\ C\|(T-T_h)[\mathcal{C}\bar{\mathbf{u}}]\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}\\
&\qquad+C\|T_h[\mathcal{C}(\bar{\mathbf{u}}-\mathbf{u})]\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)},
\end{aligned}$$

where $(\mathbf{y}_\theta, p_\theta) = (\bar{\mathbf{y}} + \theta(\mathbf{y} - \bar{\mathbf{y}}), \bar{p} + \theta(p - \bar{p}))$.

Let us estimate each of the terms. Using the definition of $\mathcal{F}_h$ and Lemma 4.4 we get

$$\begin{aligned}
&\|\partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}}, \bar{\mathbf{y}} + \theta(\mathbf{y} - \bar{\mathbf{y}}), \bar{p} + \theta(p - \bar{p})) - \partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{p})\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))} \\
&\quad = \|T_h[B'(\bar{\mathbf{y}} + \theta(\mathbf{y} - \bar{\mathbf{y}})) - B'(\bar{\mathbf{y}})]\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))} \\
&\quad \le C\|B'(\mathbf{y} - \bar{\mathbf{y}})\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega),\mathbf{H}^{-1}(\Omega))} \le C\|\mathbf{y} - \bar{\mathbf{y}}\|_{\mathbf{H}_0^1(\Omega)}. \qquad (4.2)
\end{aligned}$$

With assumption (S2) we have

$$\|(T_h - T)[B(\bar{\mathbf{y}}) - \mathbf{f}]\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)} \le Ch(\|\bar{\mathbf{y}}\|_{\mathbf{H}^2(\Omega)} + \|\mathbf{f}\|_{\mathbf{L}^2(\Omega)}),$$

and

$$\|(T - T_h)[\mathcal{C}\bar{\mathbf{u}}]\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)} \le Ch\|\bar{\mathbf{u}}\|_{L^2(\omega;\mathbb{R}^m)}.$$

Finally, from Lemma 4.4 it follows that

$$\|T_h[\mathcal{C}(\bar{\mathbf{u}} - \mathbf{u})]\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)} \le C\|\bar{\mathbf{u}} - \mathbf{u}\|_{L^2(\omega;\mathbb{R}^m)}.$$

Collecting these estimates all together, we have proved that there exists a constant $\hat{C} > 0$ independent of h and ρ such that

$$\|\Psi_{\mathbf{u}}(\mathbf{y}, p) - (\bar{\mathbf{y}}, \bar{p})\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)} \le \hat{C}(h + \rho^2).$$

We choose $\hat{\rho}_1 \le \min\{\rho_0, 1/(2\hat{C})\}$, $\hat{\rho}_2 = \hat{\rho}_1^2$, and $\hat{h}_1 = \min\{h_0, \hat{\rho}_1/(2\hat{C})\}$. It is clear that for all $0 < h < \hat{h}_1$ and all $\mathbf{u} \in B_{\hat{\rho}_2}(\bar{\mathbf{u}})$, $\Psi_{\mathbf{u}}$ is a mapping from $B_{\hat{\rho}_1}(\bar{\mathbf{y}}) \times B_{\hat{\rho}_1}(\bar{p})$ into itself.

(ii) Now we look for conditions to have a strict contraction. Take $(\mathbf{y}_1, p_1)$, $(\mathbf{y}_2, p_2) \in B_{\hat{\rho}_1}(\bar{\mathbf{y}}) \times B_{\hat{\rho}_1}(\bar{p})$, $0 < h < \hat{h}_1$, and $\mathbf{u} \in B_{\hat{\rho}_2}(\bar{\mathbf{u}})$. Classical calculations lead to

$$\begin{aligned}
&\| \Psi_{\mathbf{u}}(\mathbf{y}_1, p_1) - \Psi_{\mathbf{u}}(\mathbf{y}_2, p_2)\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)} \\
&= \Big\| (\mathbf{y}_1 - \mathbf{y}_2, p_1 - p_2) \\
&- \left[\partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{p})\right]^{-1} \Big\{ \mathcal{F}_h(\mathbf{u}, \mathbf{y}_1, p_1) - \mathcal{F}_h(\mathbf{u}, \mathbf{y}_2, p_2) \Big\} \Big\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)} \\
&= \Big\| \left[\partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{p})\right]^{-1} \Big\{ \partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{p})(\mathbf{y}_1 - \mathbf{y}_2, p_1 - p_2) \\
&- \int_0^1 \partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}}, \mathbf{y}_1 + \theta(\mathbf{y}_2 - \mathbf{y}_1), p_1 + \theta(p_2 - p_1))(\mathbf{y}_1 - \mathbf{y}_2, p_1 - p_2) d\theta \Big\} \Big\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)}.
\end{aligned}$$

The norm $\| \left[\partial_{(\mathbf{y},p)}\mathcal{F}_h(\bar{\mathbf{u}}, \bar{\mathbf{y}}, \bar{p})\right]^{-1} \|_{\mathcal{L}(\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}$ can be estimated by a constant C independent of h; see Theorem 4.7. To estimate the expression in brackets we can repeat the argument of inequalities (4.2), since $\mathbf{y} = \mathbf{y}_1 + \theta(\mathbf{y}_2 - \mathbf{y}_1) \in B_{\hat{\rho}_1}(\bar{\mathbf{y}})$. There then exists $\tilde{C} > 0$ independent of $\hat{\rho}_1$ and h such that

$$\|\Psi_{\mathbf{u}}(\mathbf{y}_1, p_1) - \Psi_{\mathbf{u}}(\mathbf{y}_2, p_2)\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)} \le \tilde{C}\hat{\rho}_1^2.$$

Choosing $\rho_1 = \min\left\{\rho_0, 1/(2\hat{C}), 1/\sqrt{2\tilde{C}}\right\}$, $\rho_2 = \rho_1^2$, and $h_1 = \min\left\{h_0, \rho_1/(2\hat{C})\right\}$, we have established that, for all $0 < h < h_1$ and all $\mathbf{u} \in B_{\rho_2}(\bar{\mathbf{u}})$, $\Psi_{\mathbf{u}}$ is a strict contraction in $B_{\rho_1}(\bar{\mathbf{y}}) \times B_{\rho_1}(\bar{p})$. ∎

Remark 4.9. We have proved that, for all $0 < h < h_1$ and all $\mathbf{u} \in B_{\rho_2}(\bar{\mathbf{u}})$, the equation $\mathcal{F}_h(\mathbf{u}, \mathbf{y}_h, p_h) = 0$ admits a unique solution $(\mathbf{y}_h(\mathbf{u}), p_h(\mathbf{u}))$ in $(B_{\rho_1}(\bar{\mathbf{y}}) \times B_{\rho_1}(\bar{p})) \cap (\mathbf{X}_h \times M_h)$, and that $\partial_{(\mathbf{y},p)}\mathcal{F}_h(\mathbf{u}, \mathbf{y}_h(\mathbf{u}), p_h(\mathbf{u}))$ is an automorphism in $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$. Therefore the mapping G_h from $B_{\rho_2}(\bar{\mathbf{u}})$ into $(B_{\rho_1}(\bar{\mathbf{y}}) \times B_{\rho_1}(\bar{p})) \cap (\mathbf{X}_h \times M_h)$ defined by $G_h(\mathbf{u}) = (\mathbf{y}_h(\mathbf{u}), p_h(\mathbf{u}))$, obeys $\mathcal{F}_h(\mathbf{u}, G_h(\mathbf{u})) = 0$, and the implicit function theorem implies that it is of class C^∞ in the interior of the ball $B_{\rho_2}(\bar{\mathbf{u}})$. Notice that G_h is not an approximation of G because $G(\mathbf{u}) = \mathbf{y}_\mathbf{u}$ is a velocity field, while $G_h(\mathbf{u})$ stands for a velocity field and a pressure.

4.2. Discretization of the control problem. For simplicity throughout the following we assume that ω is a polygonal domain. But we could consider a more general situation if we take into account the error we introduce by approximating ω by a polygonal domain.

For $h > 0$, let $\mathcal{T}_h$ be a triangulation of ω. Although the discretization of the control can be done independently of the discretization of the state equation, in practice, when we use the finite element method to approximate the state and adjoint state equation, the same family of triangulations is used. Some assumptions must be made on the family of triangulations in order to have the inverse estimate of assumption (H1)(d). We will suppose that the family is quasi-uniform (see, e.g., [9, p. 135]): In this case $h = \max_{T \in \mathcal{T}_h} \rho(T)$, where $\rho(T)$ is the diameter of the set T. We denote by $\sigma(T)$ the diameter of the largest ball contained in T. We assume there exist two positive constants ρ and σ such that

$$\frac{\rho(T)}{\sigma(T)} \le \sigma, \qquad \frac{h}{\rho(T)} \le \rho$$

hold for all $T \in \mathcal{T}_h$ and all $0 < h$.

In the following we would like to treat in the same way the cases when the control set is discretized and when it is not. We shall see that we obtain better estimates when the control set is not discretized. For that we set

$$U_h = \Big\{ \mathbf{u} \in L^2(\omega; \mathbb{R}^m) \mid u_i|_T \in P_0(T) \ \forall\, T \in \mathcal{T}_h \Big\},$$
$$U_{ad}^h = \Big\{ \mathbf{u} \in U_h \mid \alpha_i \le u_i \le \beta_i \ \forall\ 1 \le i \le d \Big\}.$$

In the discrete control problem stated below, the case when the control set is not discretized corresponds to the choice $U_{ad,h} = U_{ad}$, while the case when the control set is discretized corresponds to $U_{ad,h} = U_{ad}^h$.

We can now define the discrete control problem associated with (P) in the following way:

$$(\mathrm{P}_h) \qquad \inf\Big\{ F(\mathbf{u}, \mathbf{y}) \mid (\mathbf{u}, \mathbf{y}, p) \in U_{ad,h} \times \mathbf{X}_h \times M_h \text{ and } (\mathbf{u}, \mathbf{y}, p) \text{ satisfies (4.1)} \Big\}.$$

Let us recall that $(\mathbf{u}, \mathbf{y}, p)$ satisfies (4.1) if and only if

$$\mathcal{F}_h(\mathbf{u}, \mathbf{y}, p) = 0. \tag{4.3}$$

Our aim is to study the existence of local minima of problems (P_h) which approximate the local minima of (P). This can be proved for nonsingular local solutions of (P). Let us start by proving some error estimates for the state equation. Given a nonsingular solution $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ of (1.2), let $h_1 > 0$ and $\rho_2 > 0$ be given by Theorem 4.8.

By using the function G_h from $B_{\rho_2}(\bar{\mathbf{u}})$ into $(B_{\rho_1}(\bar{\mathbf{y}}) \times B_{\rho_1}(\bar{p})) \cap (\mathbf{X}_h \times M_h)$ introduced at the end of the previous section in Remark 4.9, we set $(\mathbf{y}_{\mathbf{u}}^h, p_{\mathbf{u}}^h) = G_h(\mathbf{u}) = (\mathbf{y}_h(\mathbf{u}), p_h(\mathbf{u}))$. Now we have the following result.

LEMMA 4.10. *Let $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ be as in Theorem 4.7. There exists a constant $C > 0$ such that, for all $\mathbf{u}$, $\hat{\mathbf{u}} \in \bar{B}_{\rho_2}(\bar{\mathbf{u}})$, and $0 < h < h_1$, the following estimates hold:*

$$\|\mathbf{y}_{\mathbf{u}} - \mathbf{y}_{\mathbf{u}}^h\|_{\mathbf{L}^2(\Omega)} \le Ch^2 \|\mathbf{y}_{\mathbf{u}}\|_{\mathbf{H}^2(\Omega)}, \tag{4.4}$$

$$\|\mathbf{y}_{\mathbf{u}} - \mathbf{y}_{\hat{\mathbf{u}}}^h\|_{\mathbf{H}_0^1(\Omega)} + \|p_{\mathbf{u}} - p_{\hat{\mathbf{u}}}^h\|_{L_0^2(\Omega)} \le C(h + \|\mathbf{u} - \hat{\mathbf{u}}\|_{L^2(\omega;\mathbb{R}^m)}). \tag{4.5}$$

Moreover, if $\mathbf{u}_h \in B_{\rho_2}(\bar{\mathbf{u}})$ and $\mathbf{u}_h \rightharpoonup \mathbf{u}$ weakly in $L^2(\omega;\mathbb{R}^m)$, then $\mathbf{y}_{\mathbf{u}_h}^h \to \mathbf{y}_{\mathbf{u}}$ in $C(\bar{\Omega};\mathbb{R}^d)$.

Proof. (i) The estimate (4.4) directly follows from usual estimates for the approximation of the Navier–Stokes equations by a finite element method. See, for instance, Girault–Raviart [12, Theorem IV.4.2].

(ii) To prove (4.5), let us write

$$\|\mathbf{y}_{\mathbf{u}} - \mathbf{y}_{\hat{\mathbf{u}}}^h\|_{\mathbf{H}_0^1(\Omega)} \le \|\mathbf{y}_{\mathbf{u}} - \mathbf{y}_{\mathbf{u}}^h\|_{\mathbf{H}_0^1(\Omega)} + \|\mathbf{y}_{\mathbf{u}}^h - \mathbf{y}_{\hat{\mathbf{u}}}^h\|_{\mathbf{H}_0^1(\Omega)},$$
$$\|p_{\mathbf{u}} - p_{\hat{\mathbf{u}}}^h\|_{L_0^2(\Omega)} \le \|p_{\mathbf{u}} - p_{\mathbf{u}}^h\|_{L_0^2(\Omega)} + \|p_{\mathbf{u}}^h - p_{\hat{\mathbf{u}}}^h\|_{L_0^2(\Omega)}.$$

Usual finite element estimates [12, estimate (4.7)] give us

$$\|\mathbf{y}_{\mathbf{u}} - \mathbf{y}_{\mathbf{u}}^h\|_{\mathbf{H}_0^1(\Omega)} + \|p_{\mathbf{u}} - p_{\mathbf{u}}^h\|_{L_0^2(\Omega)} \le Ch.$$

If $\mathbf{u}$ belongs to the interior of $B_{\rho_2}(\bar{\mathbf{u}})$, from the definition of G_h it follows that

$$G_h'(\mathbf{u})\mathbf{v} = -[\partial_{(\mathbf{y},p)} \mathcal{F}_h(\mathbf{u}, \mathbf{y}_{\mathbf{u}}^h, p_{\mathbf{u}}^h)]^{-1} T_h[\mathcal{C}\mathbf{v}].$$

Hence, with Lemma 4.4 and Theorem 4.7 we obtain

$$\begin{aligned}
&\|G_h(\mathbf{u}) - G_h(\hat{\mathbf{u}})\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)} \\
&= \left\| \int_0^1 [\partial_{(\mathbf{y},p)} \mathcal{F}_h(\mathbf{u}_\theta, \mathbf{y}_{\mathbf{u}_\theta}^h, p_{\mathbf{u}_\theta}^h)]^{-1} T_h[\mathcal{C}(\mathbf{u} - \hat{\mathbf{u}})] \right\|_{\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)} \\
&\le C\|\mathbf{u} - \hat{\mathbf{u}}\|_{L^2(\omega;\mathbb{R}^m)},
\end{aligned}$$

where $\mathbf{u}_\theta = \hat{\mathbf{u}} + \theta(\mathbf{u} - \hat{\mathbf{u}})$. Collecting the previous estimates, the proof of (4.5) is complete.

(iii) Let $(\mathbf{u}_h)_h$ be a sequence in $B_{\rho_2}(\bar{\mathbf{u}}) \cap U_{ad}$, weakly converging to $\mathbf{u}$ in $L^2(\omega;\mathbb{R}^m)$. Due to Theorem 2.2, $\mathbf{y}_{\mathbf{u}}$ belongs to $\mathbf{W}^{2,\bar{r}}(\Omega)$ and $\{\mathbf{y}_{\mathbf{u}_h}\}_h$ is bounded in $\mathbf{W}^{2,\bar{r}}(\Omega)$. Thus it converges to $\mathbf{y}_{\mathbf{u}}$ in $\mathbf{L}^p(\Omega)$ for all $2 \le p < \infty$, and the sequence $\{\mathbf{y}_{\mathbf{u}_h} \otimes \mathbf{y}_{\mathbf{u}_h}\}_h$ converges to $\mathbf{y}_{\mathbf{u}} \otimes \mathbf{y}_{\mathbf{u}}$ in $(\mathbf{L}^p(\Omega))^d$ for all $2 \le p < \infty$. The function $\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}_{\mathbf{u}}$ satisfies the equation

$$A(\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}_{\mathbf{u}}) = \operatorname{div}\big((\mathbf{y}_{\mathbf{u}} \otimes \mathbf{y}_{\mathbf{u}_h}) - (\mathbf{y}_{\mathbf{u}_h} \otimes \mathbf{y}_{\mathbf{u}})\big) + \mathcal{C}(\mathbf{u}_h - \mathbf{u}) \quad \text{in } \mathbf{V}^{-1}(\Omega).$$

Let p satisfy $d < p < 6$. From classical estimates for the Stokes equations it follows that

$$\begin{aligned}
&\|\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}_{\mathbf{u}}\|_{\mathbf{W}^{1,p}(\Omega)} \le C\|\operatorname{div}\big((\mathbf{y}_{\mathbf{u}} \otimes \mathbf{y}_{\mathbf{u}_h}) - (\mathbf{y}_{\mathbf{u}_h} \otimes \mathbf{y}_{\mathbf{u}})\big) + \mathcal{C}(\mathbf{u}_h - \mathbf{u})\|_{\mathbf{W}^{-1,p}(\Omega)} \\
&\le C\big(\|(\mathbf{y}_{\mathbf{u}} \otimes \mathbf{y}_{\mathbf{u}_h}) - (\mathbf{y}_{\mathbf{u}_h} \otimes \mathbf{y}_{\mathbf{u}})\|_{\mathbf{L}^p(\Omega)} + \|\mathcal{C}(\mathbf{u}_h - \mathbf{u})\|_{\mathbf{W}^{-1,p}(\Omega)}\big).
\end{aligned}$$

Since $\mathbf{W}^{1,p}(\Omega) \hookrightarrow \mathbf{L}^\infty(\Omega)$, and $\mathbf{L}^2(\Omega)$ is compactly embedded in $\mathbf{W}^{-1,p}(\Omega)$ (because $p < 6$), it is clear that $\{\mathbf{y}_{\mathbf{u}_h}\}_h$ tends to $\mathbf{y}_{\mathbf{u}}$ in $\mathbf{L}^\infty(\Omega)$.

We have

$$\|\mathbf{y}^h_{\mathbf{u}_h} - \mathbf{y}_{\mathbf{u}}\|_{\mathbf{L}^\infty(\Omega)} \le \|\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}_{\mathbf{u}}\|_{\mathbf{L}^\infty(\Omega)} + \|\mathbf{y}^h_{\mathbf{u}_h} - \mathbf{y}_{\mathbf{u}_h}\|_{\mathbf{L}^\infty(\Omega)}$$
$$\le \|\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}_{\mathbf{u}}\|_{\mathbf{L}^\infty(\Omega)} + \|\mathbf{y}^h_{\mathbf{u}_h} - r_h\mathbf{y}_{\mathbf{u}_h}\|_{\mathbf{L}^\infty(\Omega)} + \|r_h\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}_{\mathbf{u}_h}\|_{\mathbf{L}^\infty(\Omega)}.$$

From (H1)(c) and (H1)(d) we deduce that

$$\|\mathbf{y}_{\mathbf{u}_h} - r_h\mathbf{y}_{\mathbf{u}_h}\|_{\mathbf{L}^\infty(\Omega)} \le Ch^{2-d/2}\|\mathbf{y}_{\mathbf{u}_h}\|_{\mathbf{H}^2(\Omega)},$$

and

$$\|r_h\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}^h_{\mathbf{u}_h}\|_{\mathbf{L}^\infty(\Omega)} \le Ch^{-d/2}\|r_h\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}^h_{\mathbf{u}_h}\|_{\mathbf{L}^2(\Omega)}$$
$$\le Ch^{-d/2}\|r_h\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}_{\mathbf{u}_h}\|_{\mathbf{L}^2(\Omega)} + Ch^{-d/2}\|\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}^h_{\mathbf{u}_h}\|_{\mathbf{L}^2(\Omega)}.$$

With (H1)(b) and (4.4) we have

$$\|r_h\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}_{\mathbf{u}_h}\|_{\mathbf{L}^2(\Omega)} \le Ch^2\|\mathbf{y}_{\mathbf{u}_h}\|_{\mathbf{H}^2(\Omega)},$$
$$\|\mathbf{y}_{\mathbf{u}_h} - \mathbf{y}^h_{\mathbf{u}_h}\|_{\mathbf{L}^2(\Omega)} \le Ch^2\|\mathbf{y}_{\mathbf{u}_h}\|_{\mathbf{H}^2(\Omega)}.$$

Collecting together these estimates and the previous convergence result we have proved that $\{\mathbf{y}^h_{\mathbf{u}_h}\}_h$ converges to $\mathbf{y}_{\mathbf{u}}$ in $\mathbf{L}^\infty(\Omega)$. □

THEOREM 4.11. *Let us assume that* (P) *has a nonsingular local minimum* $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$. *Then there exists* $h_2 > 0$ *such that, for all* $0 < h < h_2$, (P_h) *has at least one solution. If, furthermore,* $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ *is a strict local minimum of* (P), *then* (P_h) *has a local minimum* $(\bar{\mathbf{u}}_h, \bar{\mathbf{y}}_h)$ *in a neighborhood of* $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ *for all* $0 < h < h_2$ *and the following identities hold:*

$$\lim_{h\to 0} J_h(\bar{\mathbf{u}}_h) = J(\bar{\mathbf{u}}),\ \lim_{h\to 0}\|\bar{\mathbf{u}} - \bar{\mathbf{u}}_h\|_{\mathbf{L}^2(\omega)} = 0,\ \textit{and}\ \lim_{h\to 0}\|\bar{\mathbf{y}} - \bar{\mathbf{y}}_h\|_{\mathbf{H}^1_0(\Omega)} = 0,$$

where $J_h(\bar{\mathbf{u}}_h) = F(\bar{\mathbf{u}}_h, \bar{\mathbf{y}}_h)$.

Proof. Let us start by proving that the set of feasible pairs $(\mathbf{u}, \mathbf{y})$ for problem (P_h) is nonempty for h small enough. We prove it only in the case when $U_{ad,h} = U^h_{ad}$. The case when $U_{ad,h} = U_{ad}$ is obvious.

Since $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ is a nonsingular local minimum, with the aid of Theorem 4.8 we derive the existence of $\rho \le \rho_2$ such that

$$J(\bar{\mathbf{u}}) \le J(\mathbf{u}) \quad \forall \mathbf{u} \in U_{ad} \cap B_\rho(\bar{\mathbf{u}}). \tag{4.6}$$

Let us define $\Pi_h\bar{\mathbf{u}} \in U_h$ by

$$\Pi_h\bar{\mathbf{u}}|_T = \frac{1}{|T|}\int_T \bar{\mathbf{u}}(x)dx. \tag{4.7}$$

It is clear that $\Pi_h\bar{\mathbf{u}} \in U_{ad,h}$. Let us prove that it belongs to $B_\rho(\bar{\mathbf{u}})$ if h is small enough. Since $\bar{\mathbf{u}}$ is Lipschitz continuous (see Theorem 3.5), we can write

$$\int_\omega (\Pi_h\bar{u}_i(s) - \bar{u}_i(s))^2 ds = \sum_{T\subset\omega}\int_T \left(\frac{1}{|T|}\int_T \bar{u}_i(x)dx - \bar{u}_i(s)\right)^2 ds$$
$$= \sum_{T\subset\omega}\int_T (\bar{u}_i(x_T) - \bar{u}_i(s))^2 ds \le |\omega|\|\bar{\mathbf{u}}\|^2_{W^{1,\infty}(\omega;\mathbb{R}^m)}h^2.$$

Therefore if

$$h_2 = \min\left\{h_1, \frac{\rho}{\|\bar{\mathbf{u}}\|_{W^{1,\infty}(\omega;\mathbb{R}^m)}|\omega|^{1/2}}\right\},$$

then $\Pi_h\bar{\mathbf{u}}$ belongs to $U_{ad,h}\cap B_\rho(\bar{\mathbf{u}})$ for all $h\le h_2$. Now setting $\mathbf{u}_h=\Pi_h\bar{\mathbf{u}}$ and $(\mathbf{y}^h_{\mathbf{u}_h},p^h_{\mathbf{u}_h})=G_h(\mathbf{u}_h)$, we have that $(\mathbf{u}_h,\mathbf{y}^h_{\mathbf{u}_h},p^h_{\mathbf{u}_h})$ satisfies (4.3) and $(\mathbf{u}_h,\mathbf{y}^h_{\mathbf{u}_h})$ is a feasible pair for (P_h) for any $h\le h_2$.

Since the set of feasible points of (P_h) is nonempty and closed, and F_h is continuous, convex on $U_{ad,h}\times\mathbf{X}_h$, and coercive with respect to $\mathbf{u}\in U_{ad,h}$, then (P_h) has at least one solution.

Now let us assume that $(\bar{\mathbf{u}},\bar{\mathbf{y}})$ is a strict local solution of (P) in $(U_{ad}\cap B_\rho(\bar{\mathbf{u}}))\times B_\rho(\bar{\mathbf{y}})$. We consider the problems

$$(\mathrm{Q}_h)\begin{cases}\min J_h(\mathbf{u}),\\ \mathbf{u}\in U_{ad,h}\cap B_\rho(\bar{\mathbf{u}}),\end{cases}$$

where $J_h(\mathbf{u})=F(\mathbf{u},\mathbf{y}^h_{\mathbf{u}})$ with $(\mathbf{y}^h_{\mathbf{u}},p^h_{\mathbf{u}})=G_h(\mathbf{u})$, G_h being defined in Remark 4.9. Above we have proved that $U_{ad,h}\cap B_\rho(\bar{\mathbf{u}})$ is nonempty for $h\le h_2$. Observe that $U_{ad,h}\cap B_\rho(\bar{\mathbf{u}})$ is convex, bounded, and closed in $L^2(\omega;\mathbb{R}^m)$, the mapping $\mathbf{u}\mapsto\int_\omega|\mathbf{u}|^2$ is lower semicontinuous for the weak topology of $L^2(\omega;\mathbb{R}^m)$, and from Remark 4.9 it follows that the mapping $\mathbf{u}\mapsto\int_\Omega|\mathbf{y}^h_{\mathbf{u}}-\mathbf{y}_d|^2$ is continuous for the weak topology of $L^2(\omega;\mathbb{R}^m)$. Therefore (Q_h) has at least one solution $\bar{\mathbf{u}}_h$. From any subsequence of $\{\bar{\mathbf{u}}_h\}_h$, we can extract another subsequence, still indexed by h to simplify the notation, converging weakly in $L^2(\omega;\mathbb{R}^m)$ to some $\tilde{\mathbf{u}}\in B_\rho(\bar{\mathbf{u}})$. Let us check that $\tilde{\mathbf{u}}=\bar{\mathbf{u}}$. Let us take again $\mathbf{u}_h=\Pi_h\bar{\mathbf{u}}\in U_{ad,h}\cap B_\rho(\bar{\mathbf{u}})$ for all $h<h_2$. By passing to the limit when h tends to zero, with the convergence result stated in Lemma 4.10, we can write

$$J(\tilde{\mathbf{u}})\le\liminf_{h\to0}J_h(\bar{\mathbf{u}}_h)\le\limsup_{h\to0}J_h(\bar{\mathbf{u}}_h)\le\limsup_{h\to0}J_h(\Pi_h\bar{\mathbf{u}})=J(\bar{\mathbf{u}}).$$

Since $\tilde{\mathbf{u}}\in B_\rho(\bar{\mathbf{u}})$ and the inequality in (4.6) is strict for $\mathbf{u}\ne\bar{\mathbf{u}}$, the above inequality implies that $\tilde{\mathbf{u}}=\bar{\mathbf{u}}$. Thus we have

$$\lim_{h\to0}J_h(\bar{\mathbf{u}}_h)=J(\bar{\mathbf{u}}),$$

and still with Lemma 4.10, we deduce that

$$\lim_{h\to0}\int_\omega|\bar{\mathbf{u}}_h|^2=\int_\omega|\bar{\mathbf{u}}|^2.$$

Therefore the subsequence $\{\bar{\mathbf{u}}_h\}_h$ converges to $\bar{\mathbf{u}}$ in $L^2(\omega;\mathbb{R}^m)$. Since $\bar{\mathbf{u}}$ is the only cluster point for the weak topology of $L^2(\omega;\mathbb{R}^m)$ of the original sequence $\{\bar{\mathbf{u}}_h\}_h$, it is clear that the convergence properties stated in the theorem hold for the whole sequence $\{\bar{\mathbf{u}}_h\}_h$. The convergence of the corresponding states is a consequence of Lemma 4.10. Finally, the strong convergence $\bar{\mathbf{u}}_h\to\bar{\mathbf{u}}$ in $L^2(\omega;\mathbb{R}^m)$ implies that $\bar{\mathbf{u}}_h$ belongs to the interior of the ball $B^h_\rho(\bar{\mathbf{u}})$, which implies that $(\bar{\mathbf{u}}_h,\bar{\mathbf{y}}_h)$ is a local minimum of (P_h). ∎

4.3. Discrete adjoint equation. We define the discrete adjoint state $(\boldsymbol{\Phi}^h_{\mathbf{u}},\pi^h_{\mathbf{u}})\in\mathbf{X}_h\times M_h$ associated with a control $\mathbf{u}\in B_{\rho_2}(\bar{\mathbf{u}})$ as the solution to the problem

$$\begin{aligned}&a(\boldsymbol{\Phi}^h_{\mathbf{u}},\mathbf{w}_h)+b(\mathbf{y}^h_{\mathbf{u}},\mathbf{w}_h,\boldsymbol{\Phi}^h_{\mathbf{u}})+b(\mathbf{w}_h,\mathbf{y}^h_{\mathbf{u}},\boldsymbol{\Phi}^h_{\mathbf{u}})-(\pi^h_{\mathbf{u}},\operatorname{div}\mathbf{w}_h)\\ &=(\mathbf{y}^h_{\mathbf{u}}-\mathbf{y}_d,\mathbf{w}_h)\quad\forall\,\mathbf{w}_h\in\mathbf{X}_h,\\ &(\lambda_h,\operatorname{div}\boldsymbol{\Phi}^h_{\mathbf{u}})=0\quad\forall\,\lambda_h\in M_h.\end{aligned}\tag{4.8}$$

LEMMA 4.12. *Let* $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ *be as in Theorem* 4.7. *There exist* $0 < h_3 \le h_2$ *and* $0 < \rho_3 \le \rho_2$ *such that, for all* $\mathbf{u} \in B_{\rho_3}(\bar{\mathbf{u}})$ *and all* $0 < h \le h_3$, *the system* (4.8) *admits a unique solution* $(\boldsymbol{\Phi}_{\mathbf{u}}^h, \pi_{\mathbf{u}}^h) \in \mathbf{X}_h \times M_h$.

Proof. (i) For $\mathbf{y} \in \mathbf{H}_0^1(\Omega)$, consider the mapping $\mathcal{G}_{\mathbf{y}}$ from $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ into itself defined by

$$\mathcal{G}_{\mathbf{y}}(\boldsymbol{\Phi}, \pi) = (\boldsymbol{\Phi}, \pi) + T[B'(\mathbf{y})^* \boldsymbol{\Phi}].$$

As in Lemma 4.2, we can easily show that $\mathcal{G}_{\mathbf{y}}$ is an automorphism in $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ if and only if $\mathbf{y}$ is a nonsingular solution of (2.2). Thus $\mathcal{G}_{\bar{\mathbf{y}}}$ is an automorphism in $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$. We also introduce the mapping $\mathcal{G}_{\mathbf{y},h}$ from $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ into itself defined by

$$\mathcal{G}_{\mathbf{y},h}(\boldsymbol{\Phi}, \pi) = (\boldsymbol{\Phi}, \pi) + T_h[B'(\mathbf{y})^* \boldsymbol{\Phi}].$$

Arguing as in the proof of Theorem 4.7, we can assume that h_0 is chosen so that, for all $0 < h < h_0$ and all $\mathbf{y} \in B_{\rho_0}(\bar{\mathbf{y}})$, $\mathcal{G}_{\mathbf{y},h}$ is an automorphism in $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$. In particular, according to estimate (4.5), there exist $0 < h_3 \le h_2$ and $0 < \rho_3 \le \rho_2$ such that, for all $0 < h \le h_3$ and all $\mathbf{u} \in B_{\rho_3}(\bar{\mathbf{u}})$, $\mathcal{G}_{\mathbf{y}_{\mathbf{u}}^h,h}$ is an automorphism in $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ and

$$\|\mathcal{G}_{\mathbf{y}_{\mathbf{u}}^h,h}^{-1}\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))} \le 2\|\mathcal{G}_{\bar{\mathbf{y}}}^{-1}\|_{\mathcal{L}(\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega))}.$$

Without loss of generality we can also assume that $\mathcal{G}_{\mathbf{y}(\mathbf{u})}$ is an automorphism in $\mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ for all $\mathbf{u} \in B_{\rho_3}(\bar{\mathbf{u}})$.

(ii) Now we are going to show that $(\boldsymbol{\Phi}_{\mathbf{u}}^h, \pi_{\mathbf{u}}^h) \in \mathbf{X}_h \times M_h$ is a solution of (4.8) if and only if

$$\mathcal{G}_{\mathbf{y}_{\mathbf{u}}^h,h}(\boldsymbol{\Phi}_{\mathbf{u}}^h, \pi_{\mathbf{u}}^h) = (\mathbf{z}_h, q_h), \tag{4.9}$$

where $(\mathbf{z}_h, q_h) \in \mathbf{X}_h \times M_h$ is the solution of the discrete Stokes problem,

$$\begin{aligned} &a(\mathbf{z}_h, \mathbf{w}_h) - (q_h, \operatorname{div} \mathbf{w}_h) = (\mathbf{y}_h(\mathbf{u}) - \mathbf{y}_d, \mathbf{w}_h)\ \forall\, \mathbf{w}_h \in \mathbf{X}_h, \\ &(\lambda_h, \operatorname{div} \mathbf{z}_h) = 0\ \forall\, \lambda_h \in M_h. \end{aligned} \tag{4.10}$$

To prove this result, we notice that (4.9) is satisfied if and only if

$$(\boldsymbol{\Phi}_{\mathbf{u}}^h - \mathbf{z}_h, \pi_{\mathbf{u}}^h - q_h) = -T_h[B'(\mathbf{y}_{\mathbf{u}}^h)^* \boldsymbol{\Phi}_{\mathbf{u}}^h],$$

which is equivalent to

$$\begin{aligned} &a(\boldsymbol{\Phi}_{\mathbf{u}}^h - \mathbf{z}_h, \mathbf{w}_h) - (\pi_{\mathbf{u}}^h - q_h, \operatorname{div} \mathbf{w}_h) = (B'(\mathbf{y}_{\mathbf{u}}^h)^* \boldsymbol{\Phi}_{\mathbf{u}}^h, \mathbf{w}_h)\ \forall\, \mathbf{w}_h \in \mathbf{X}_h, \\ &(\lambda_h, \operatorname{div} \boldsymbol{\Phi}_{\mathbf{u}}^h - \operatorname{div} \mathbf{z}_h) = 0\ \forall\, \lambda_h \in M_h. \end{aligned} \tag{4.11}$$

Now using equation (4.10), we see that (4.11) is equivalent to (4.8). This completes the proof. ∎

We are going to prove error estimates for the discrete adjoint state. Set

$$\mathbf{V}_h = \Big\{ \boldsymbol{\Phi}_h \in X_h \mid (\lambda_h, \operatorname{div} \boldsymbol{\Phi}_h) = 0\ \forall\, \lambda_h \in M_h \Big\}.$$

LEMMA 4.13. *Let* $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ *be as in Theorem* 4.7. *There exists a constant* $C > 0$ *such that, for all* $\mathbf{u}$, $\hat{\mathbf{u}} \in B_{\rho_3}(\bar{\mathbf{u}})$ *and all* $0 < h < h_3$, *the solution* $(\boldsymbol{\Phi}_{\mathbf{u}}, \pi_{\mathbf{u}})$ *to* (3.8)

and the solutions $(\mathbf{\Phi}^h_{\mathbf{u}}, \pi^h_{\mathbf{u}})$ *and* $(\mathbf{\Phi}^h_{\hat{\mathbf{u}}}, \pi^h_{\hat{\mathbf{u}}})$ *to* (4.8) *obey the following estimates:*

$$\|\mathbf{\Phi}_{\mathbf{u}} - \mathbf{\Phi}^h_{\mathbf{u}}\|_{\mathbf{L}^2(\Omega)} \le Ch^2, \tag{4.12}$$

$$\|\mathbf{\Phi}_{\mathbf{u}} - \mathbf{\Phi}^h_{\mathbf{u}}\|_{\mathbf{H}^1_0(\Omega)} + \|\pi_{\mathbf{u}} - \pi^h_{\mathbf{u}}\|_{L^2_0(\Omega)} \le Ch, \tag{4.13}$$

$$\|\mathbf{\Phi}_{\mathbf{u}} - \mathbf{\Phi}^h_{\hat{\mathbf{u}}}\|_{\mathbf{H}^1_0(\Omega)} + \|\pi_{\mathbf{u}} - \pi^h_{\hat{\mathbf{u}}}\|_{L^2_0(\Omega)} \le C(h + \|\mathbf{u} - \hat{\mathbf{u}}\|_{L^2(\omega;\mathbb{R}^m)}). \tag{4.14}$$

Moreover, if $\mathbf{u}_h \in B_{\rho_3}(\bar{\mathbf{u}})$ *and* $\mathbf{u}_h \rightharpoonup \mathbf{u}$ *weakly in* $L^2(\omega;\mathbb{R}^m)$*, then* $\mathbf{\Phi}^h_{\mathbf{u}} \to \mathbf{\Phi}_{\mathbf{u}}$ *strongly in* $C(\bar{\Omega};\mathbb{R}^d)$.

Proof. (i) We first show (4.13). From the proof of Lemma 4.12 it follows that $\mathcal{G}_{\mathbf{y_u}}$ is an automorphism in $\mathbf{H}^1_0(\Omega) \times L^2_0(\Omega)$, and that, for all $\mathbf{u} \in B_{\rho_3}(\bar{\mathbf{u}})$ and all $0 < h < h_3$, $\mathcal{G}_{\mathbf{y}^h_{\mathbf{u}}}$ is an automorphism in $\mathbf{H}^1_0(\Omega) \times L^2_0(\Omega)$ and

$$\|\mathcal{G}^{-1}_{\mathbf{y}^h_{\mathbf{u}},h}\|_{\mathcal{L}(\mathbf{H}^1_0(\Omega)\times L^2_0(\Omega))} \le 2\|\mathcal{G}^{-1}_{\bar{\mathbf{y}}}\|_{\mathcal{L}(\mathbf{H}^1_0(\Omega)\times L^2_0(\Omega))}.$$

Let us recall that $(\mathbf{\Phi}_{\mathbf{u}}, \pi_{\mathbf{u}})$ is the solution of the equation

$$\mathcal{G}_{\mathbf{y_u}}(\mathbf{\Phi}_{\mathbf{u}}, \pi_{\mathbf{u}}) = T(\mathbf{y_u} - \mathbf{y}_d),$$

and that $(\mathbf{\Phi}^h_{\mathbf{u}}, \pi^h_{\mathbf{u}})$ is the solution of

$$\mathcal{G}_{\mathbf{y}^h_{\mathbf{u}},h}(\mathbf{\Phi}^h_{\mathbf{u}}, \pi^h_{\mathbf{u}}) = T_h(\mathbf{y}^h_{\mathbf{u}} - \mathbf{y}_d).$$

Thus we have

$$\begin{aligned}
&\mathcal{G}_{\mathbf{y}^h_{\mathbf{u}},h}(\mathbf{\Phi}_{\mathbf{u}} - \mathbf{\Phi}^h_{\mathbf{u}}, \pi_{\mathbf{u}} - \pi^h_{\mathbf{u}}) \\
&= (\mathbf{\Phi}_{\mathbf{u}}, \pi_{\mathbf{u}}) + T_h(B'(\mathbf{y}^h_{\mathbf{u}})^*\mathbf{\Phi}_{\mathbf{u}}) - \mathcal{G}_{\mathbf{y}^h_{\mathbf{u}},h}(\mathbf{\Phi}^h_{\mathbf{u}}, \pi^h_{\mathbf{u}}) \\
&= T_h[B'(\mathbf{y}^h_{\mathbf{u}})^*\mathbf{\Phi}_{\mathbf{u}}] - T[B'(\mathbf{y_u})^*\mathbf{\Phi}_{\mathbf{u}}] + T(\mathbf{y_u} - \mathbf{y}_d) - T_h(\mathbf{y}^h_{\mathbf{u}} - \mathbf{y}_d) \\
&= (T - T_h)[\mathbf{y}^h_{\mathbf{u}} - B'(\mathbf{y}^h_{\mathbf{u}})^*\mathbf{\Phi}_{\mathbf{u}} - \mathbf{y}_d] \\
&\qquad + T[B'(\mathbf{y}^h_{\mathbf{u}})^*\mathbf{\Phi}_{\mathbf{u}} - B'(\mathbf{y_u})^*\mathbf{\Phi}_{\mathbf{u}}] + T[\mathbf{y_u} - \mathbf{y}^h_{\mathbf{u}}],
\end{aligned}$$

which yields

$$\begin{aligned}
&(\mathbf{\Phi}_{\mathbf{u}} - \mathbf{\Phi}^h_{\mathbf{u}}, \pi_{\mathbf{u}} - \pi^h_{\mathbf{u}}) \\
&= \Big(\mathcal{G}_{\mathbf{y}^h_{\mathbf{u}},h}\Big)^{-1}\Big((T - T_h)[\mathbf{y}^h_{\mathbf{u}} - B'(\mathbf{y}^h_{\mathbf{u}})^*\mathbf{\Phi}_{\mathbf{u}} - \mathbf{y}_d] \\
&\qquad + T[B'(\mathbf{y}^h_{\mathbf{u}})^*\mathbf{\Phi}_{\mathbf{u}} - B'(\mathbf{y_u})^*\mathbf{\Phi}_{\mathbf{u}}] + T[\mathbf{y_u} - \mathbf{y}^h_{\mathbf{u}}]\Big).
\end{aligned}$$

With estimate (4.5) and assumption (S2), we obtain

$$\|\mathbf{\Phi}_{\mathbf{u}} - \mathbf{\Phi}^h_{\mathbf{u}}\|_{\mathbf{H}^1_0(\Omega)} + \|\pi_{\mathbf{u}} - \pi^h_{\mathbf{u}}\|_{L^2_0(\Omega)} \le Ch.$$

(ii) To prove (4.12) we proceed as in [12, Chapter 2, Theorems 1.2 and 1.9]. The solution $(\mathbf{\Phi}_{\mathbf{u}}, \pi_{\mathbf{u}})$ to (3.8) and the solution $(\mathbf{\Phi}^h_{\mathbf{u}}, \pi^h_{\mathbf{u}})$ to (4.8) satisfy

$$\begin{aligned}
&a(\mathbf{\Phi}_{\mathbf{u}} - \mathbf{\Phi}^h_{\mathbf{u}}, \mathbf{w}_h) + b(\mathbf{y_u}, \mathbf{w}_h, \mathbf{\Phi}_{\mathbf{u}} - \mathbf{\Phi}^h_{\mathbf{u}}) + b(\mathbf{w}_h, \mathbf{y_u}, \mathbf{\Phi}_{\mathbf{u}} - \mathbf{\Phi}^h_{\mathbf{u}}) \\
&- (\pi_{\mathbf{u}} - \pi^h_{\mathbf{u}}, \operatorname{div} \mathbf{w}_h) = (\mathbf{y_u} - \mathbf{y}^h_{\mathbf{u}}, \mathbf{w}_h) + b(\mathbf{y_u} - \mathbf{y}^h_{\mathbf{u}}, \mathbf{w}_h, \mathbf{\Phi}^h_{\mathbf{u}}) \\
&+ b(\mathbf{w}_h, \mathbf{y_u} - \mathbf{y}^h_{\mathbf{u}}, \mathbf{\Phi}^h_{\mathbf{u}}) \ \forall\ \mathbf{w}_h \in \mathbf{X}_h, \\
&(\lambda_h, \operatorname{div} \mathbf{\Phi}_{\mathbf{u}} - \operatorname{div} \mathbf{\Phi}^h_{\mathbf{u}}) = 0 \ \forall\, \lambda_h \in M_h.
\end{aligned} \tag{4.15}$$

For all $\mathbf{g} \in \mathbf{L}^2(\Omega)$, let us consider the solution $(\mathbf{z_g}, q_\mathbf{g}) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ to

$$\begin{aligned} &a(\mathbf{z_g}, \mathbf{w}) + b(\mathbf{y_u}, \mathbf{z_g}, \mathbf{w}) + b(\mathbf{z_g}, \mathbf{y_u}, \mathbf{w}) - (q_\mathbf{g}, \operatorname{div} \mathbf{w}) = (\mathbf{g}, \mathbf{w}) \ \forall \mathbf{w} \in \mathbf{H}_0^1(\Omega), \\ &(\lambda, \operatorname{div} \mathbf{z_g}) = 0 \ \forall \lambda \in L_0^2(\Omega), \end{aligned} \tag{4.16}$$

and the solution $(\mathbf{z}_\mathbf{g}^h, q_\mathbf{g}^h) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ to

$$\begin{aligned} &a(\mathbf{z}_\mathbf{g}^h, \mathbf{w}_h) + b(\mathbf{y_u}, \mathbf{z}_\mathbf{g}^h, \mathbf{w}_h) + b(\mathbf{z}_\mathbf{g}^h, \mathbf{y_u}, \mathbf{w}_h) - (q_\mathbf{g}^h, \operatorname{div} \mathbf{w}_h) = (\mathbf{g}, \mathbf{w}_h) \ \forall \mathbf{w} \in \mathbf{X}_h, \\ &(\lambda, \operatorname{div} \mathbf{z}_\mathbf{g}^h) = 0 \ \forall \lambda \in M_h. \end{aligned}$$

Choosing $\mathbf{w}_h = \mathbf{z}_\mathbf{g}^h$ in (4.15) and $\mathbf{w} = \mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h$ in (4.16) and combining the two identities, we obtain

$$\begin{aligned} &(\mathbf{g}, \mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h) = a(\mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h, \mathbf{z_g} - \mathbf{z}_\mathbf{g}^h) + b(\mathbf{y_u}, \mathbf{z_g} - \mathbf{z}_\mathbf{g}^h, \mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h) \\ &+ b(\mathbf{z_g} - \mathbf{z}_\mathbf{g}^h, \mathbf{y_u}, \mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h) + (\pi_\mathbf{u} - \pi_\mathbf{u}^h, \operatorname{div} \mathbf{z}_\mathbf{g}^h) - (q_\mathbf{g}, \operatorname{div} \mathbf{\Phi_u} - \operatorname{div} \mathbf{\Phi}_\mathbf{u}^h) \\ &+ (\mathbf{y_u} - \mathbf{y}_\mathbf{u}^h, \mathbf{z}_\mathbf{g}^h) + b(\mathbf{y_u} - \mathbf{y}_\mathbf{u}^h, \mathbf{z}_\mathbf{g}^h, \mathbf{\Phi}_\mathbf{u}^h) + b(\mathbf{z}_\mathbf{g}^h, \mathbf{y_u} - \mathbf{y}_\mathbf{u}^h, \mathbf{\Phi}_\mathbf{u}^h) \\ &= a(\mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h, \mathbf{z_g} - \mathbf{z}_\mathbf{g}^h) - b(\mathbf{y_u}, \mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h, \mathbf{z_g} - \mathbf{z}_\mathbf{g}^h) - b(\mathbf{z_g} - \mathbf{z}_\mathbf{g}^h, \mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h, \mathbf{y_u}) \\ &+ (\mathbf{y_u} - \mathbf{y}_\mathbf{u}^h, \mathbf{z}_\mathbf{g}^h) - b(\mathbf{y_u} - \mathbf{y}_\mathbf{u}^h, \mathbf{\Phi}_\mathbf{u}^h, \mathbf{z}_\mathbf{g}^h) - b(\mathbf{z}_\mathbf{g}^h, \mathbf{\Phi}_\mathbf{u}^h, \mathbf{y_u} - \mathbf{y}_\mathbf{u}^h) \\ &+ (\pi_\mathbf{u} - \pi_\mathbf{u}^h, \operatorname{div} \mathbf{z}_\mathbf{g}^h - \operatorname{div} \mathbf{z_g}) - (q_\mathbf{g} - q_\mathbf{g}^h, \operatorname{div} \mathbf{\Phi_u} - \operatorname{div} \mathbf{\Phi}_\mathbf{u}^h). \end{aligned}$$

Thus we have

$$\begin{aligned} &\|\mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h\|_{\mathbf{L}^2(\Omega)} = \sup_{\|\mathbf{g}\|_{\mathbf{L}^2(\Omega)}=1} (\mathbf{g}, \mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h) \\ &\le C \sup_{\|\mathbf{g}\|_{\mathbf{L}^2(\Omega)}=1} \Big\{ \|\mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h\|_{\mathbf{H}_0^1(\Omega)} \|\mathbf{z_g} - \mathbf{z}_\mathbf{g}^h\|_{\mathbf{H}_0^1(\Omega)} \\ &\quad + \|\mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h\|_{\mathbf{H}_0^1(\Omega)} \|\mathbf{z_g} - \mathbf{z}_\mathbf{g}^h\|_{\mathbf{L}^2(\Omega)} \|\mathbf{y_u}\|_{\mathbf{L}^\infty(\Omega)} \\ &\quad + \|\mathbf{y_u} - \mathbf{y}_\mathbf{u}^h\|_{\mathbf{L}^2(\Omega)} \|\mathbf{z}_\mathbf{g}^h\|_{\mathbf{L}^2(\Omega)} + \|\mathbf{y_u} - \mathbf{y}_\mathbf{u}^h\|_{\mathbf{L}^2(\Omega)} \|\mathbf{z}_\mathbf{g}^h\|_{\mathbf{L}^\infty(\Omega)} \|\mathbf{\Phi}_\mathbf{u}^h\|_{\mathbf{H}_0^1(\Omega)} \\ &\quad + \|\pi_\mathbf{u} - \pi_\mathbf{u}^h\|_{L_0^2(\Omega)} \|\mathbf{z_g} - \mathbf{z}_\mathbf{g}^h\|_{\mathbf{H}_0^1(\Omega)} + \|q_\mathbf{g} - q_\mathbf{g}^h\|_{L_0^2(\Omega)} \|\mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h\|_{\mathbf{H}_0^1(\Omega)} \Big\}. \end{aligned} \tag{4.17}$$

To complete estimate (4.12), we are going to use (4.13) and a similar error estimate for $(\mathbf{z_g}, q_\mathbf{g})$:

$$\|\mathbf{z_g} - \mathbf{z}_\mathbf{g}^h\|_{\mathbf{H}_0^1(\Omega)} + \|q_\mathbf{g} - q_\mathbf{g}^h\|_{L_0^2(\Omega)} \le Ch(\|\mathbf{z_g}\|_{\mathbf{H}^2(\Omega)} + \|q_\mathbf{g}\|_{H^1(\Omega)}). \tag{4.18}$$

With (4.17), (4.13), (4.18), and (4.4), we obtain

$$\|\mathbf{\Phi_u} - \mathbf{\Phi}_\mathbf{u}^h\|_{\mathbf{L}^2(\Omega)} \le Ch^2.$$

The proof of (4.12) is complete. Estimate (4.14) and the last statement in the lemma can now be proved in the same way as we did it for the state. ∎

Let $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ be a nonsingular strict local minimum of (P) and $\{(\bar{\mathbf{u}}_h, \bar{\mathbf{y}}_h)\}_{h \le h_3}$ be a sequence of local minima of problems (P_h) converging to $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ in $\mathbf{L}^2(\omega; \mathbb{R}^m) \times \mathbf{H}_0^1(\Omega)$, with $\bar{\mathbf{u}}_h \in B_{\rho_3}(\bar{\mathbf{u}})$, where h_3 and ρ_3 are given by Lemma 4.12. Then every element $\bar{\mathbf{u}}_h$ from a sequence $\{\bar{\mathbf{u}}_h\}_{h \le h_3}$ is a local solution of the problem

$$(\hat{\mathrm{P}}_h) \begin{cases} \min J_h(\mathbf{u}) = F(\mathbf{u}, \mathbf{y}_\mathbf{u}^h), \\ \mathbf{u} \in U_{ad,h}, \end{cases}$$

where $(\mathbf{y}_{\mathbf{u}}^h, p_{\mathbf{u}}^h) = G_h(\mathbf{u})$, G_h being defined in Remark 4.9.

LEMMA 4.14. *Let $\bar{\mathbf{u}}_h$ be a solution to problem $(\hat{\mathrm{P}}_h)$, and let $(\bar{\mathbf{y}}_h, \bar{p}_h) \in \mathbf{X}_h \times M_h$ be the corresponding state and pressure. Then $\bar{\mathbf{u}}_h$ satisfies*

$$\int_\omega (\mathcal{C}^* \bar{\mathbf{\Phi}}_h + N\bar{\mathbf{u}}_h) \cdot (\mathbf{u}_h - \bar{\mathbf{u}}_h)\, dx \geq 0 \ \forall\, \mathbf{u}_h \in U_{ad,h},$$

where $(\bar{\mathbf{\Phi}}_h, \bar{\pi}_h) = (\mathbf{\Phi}^h_{\bar{\mathbf{u}}_h}, \pi^h_{\bar{\mathbf{u}}_h}) \in \mathbf{X}_h \times M_h$ is the discrete adjoint state associated with $\bar{\mathbf{u}}_h$, that is, the solution to the system (4.8) *where $\mathbf{u}$ is replaced by $\bar{\mathbf{u}}_h$.*

Proof. The lemma is a consequence of the following identity:

$$J_h'(\bar{\mathbf{u}}_h)(\mathbf{u}_h - \bar{\mathbf{u}}_h) = \int_\omega (\mathcal{C}^* \bar{\mathbf{\Phi}}_h + N\bar{\mathbf{u}}_h) \cdot (\mathbf{u}_h - \bar{\mathbf{u}}_h)\, dx. \qquad \square$$

Now we can establish uniform convergence for the controls.

LEMMA 4.15. *Let $\bar{\mathbf{u}}_h$ be as in Lemma* 4.14*; then* $\lim_{h\to 0} \|\bar{\mathbf{u}}_h - \bar{\mathbf{u}}\|_{L^\infty(\omega;\mathbb{R}^m)} = 0$.

Proof. Let us start with the case where $U_{ad,h} = U_{ad}^h$. Since the components of the elements of U_h are constant on every triangle, for all $T \in \mathcal{T}_h$ and $1 \leq i \leq m$, we have

$$\bar{u}_{i,h}|_T = \mathrm{Proj}_{[\alpha_i,\beta_i]} \left(-\frac{1}{N|T|} \int_T (\mathcal{C}^* \bar{\mathbf{\Phi}}_h)_i(x) dx \right).$$

For all $x \in T$, using (3.9), the integral mean value theorem, and the Lipschitz continuity of $\bar{\mathbf{\Phi}}$, we can write

$$\begin{aligned}
|\bar{u}_{i,h}(x) - \bar{u}_i(x)| &\leq \left| \frac{1}{N|T|} \int_T (\mathcal{C}^* \bar{\mathbf{\Phi}}_h)_i(s) ds - \frac{1}{N} (\mathcal{C}^* \bar{\mathbf{\Phi}})_i(x) \right| \\
&= \frac{1}{N} |(\mathcal{C}^* \bar{\mathbf{\Phi}}_h)_i(x_T) - (\mathcal{C}^* \bar{\mathbf{\Phi}})_i(x)| \\
&\leq \frac{1}{N} |(\mathcal{C}^* \bar{\mathbf{\Phi}}_h)_i(x_T) - (\mathcal{C}^* \bar{\mathbf{\Phi}})_i(x_T)| + \frac{1}{N} |(\mathcal{C}^* \bar{\mathbf{\Phi}})_i(x_T) - (\mathcal{C}^* \bar{\mathbf{\Phi}})_i(x)| \\
&\leq C \|\bar{\mathbf{\Phi}}_h - \bar{\mathbf{\Phi}}\|_{\mathbf{L}^\infty(\Omega)} + C|x_T - x| \leq C(\|\bar{\mathbf{\Phi}}_h - \bar{\mathbf{\Phi}}\|_{\mathbf{L}^\infty(\Omega)} + h)
\end{aligned}$$

for some $x_T \in T$. The uniform convergence of the adjoint states allows us to complete the proof in the case when $U_{ad,h} = U_{ad}^h$.

In the case when $U_{ad,h} = U_{ad}$ we have

$$\bar{u}_{i,h}(x) = \mathrm{Proj}_{[\alpha_i,\beta_i]} \left(-\frac{1}{N|T|} (\mathcal{C}^* \bar{\mathbf{\Phi}}_h)_i(x) \right).$$

The convergence of $\bar{\mathbf{u}}_h$ follows from Lemma 4.13. $\square$

4.4. Error estimates. Let $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ be a nonsingular local solution of (P) satisfying the sufficient second order optimality conditions (3.15) or, equivalently, (3.25). As a consequence of these conditions, we know that $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$ is a strict local minimum of (P). Let $\{(\bar{\mathbf{u}}_h, \bar{\mathbf{y}}_h)\}_h$ be a sequence of local solutions of problems (P_h) converging to $(\bar{\mathbf{u}}, \bar{\mathbf{y}})$; see Theorem 4.11 and Lemma 4.15. We assume that $h \leq h_3$ and $\bar{\mathbf{u}}_h \in B_{\rho_3}(\bar{\mathbf{u}})$, so that $\bar{\mathbf{u}}_h$ is a local minimum of $(\hat{\mathrm{P}}_h)$. The goal of this section is to estimate the order of convergence of this sequence.

LEMMA 4.16. *Let $\delta > 0$ be the constant defined in Corollary* 3.11*. There exists $0 < h_4 \leq h_3$ such that*

$$\frac{\delta}{2} \|\bar{\mathbf{u}} - \bar{\mathbf{u}}_h\|^2_{L^2(\omega;\mathbb{R}^m)} \leq (J'(\bar{\mathbf{u}}_h) - J'(\bar{\mathbf{u}}))(\bar{\mathbf{u}}_h - \bar{\mathbf{u}}) \ \forall\, 0 < h < h_4.$$

Proof. First, let us check that for $h > 0$ small enough, $\bar{\mathbf{u}}_h - \bar{\mathbf{u}}$ belongs to $C^\tau_{\bar{\mathbf{u}}}$. The sign condition (3.21)–(3.22) is trivial since $\bar{\mathbf{u}}_h \in U_{ad}$. We have to check condition (3.20). Let us set

$$\bar{\mathbf{d}}_h(x) = (\mathcal{C}^* \bar{\mathbf{\Phi}}_h)(x) + N\bar{\mathbf{u}}_h(x).$$

Take h_4 small enough to have

$$\|\bar{\mathbf{d}} - \bar{\mathbf{d}}_h\|_{L^\infty(\omega)} < \frac{\tau}{4}, \quad \text{and} \quad \|\bar{\mathbf{d}}(x_1) - \bar{\mathbf{d}}(x_2)\|_{\mathbb{R}^m} < \frac{\tau}{4} \quad \text{if} \quad \|x_1 - x_2\|_{\mathbb{R}^d} < h,$$

for all $0 < h \le h_4$. First consider the case where $U_{ad,h} = U_{ad}$. In that case, if $\bar{d}_i(\xi) > \tau$ (respectively, $\bar{d}_i(\xi) < -\tau$), we have $\bar{d}_{i,h}(\xi) > 3\tau/4$ (respectively, $\bar{d}_{i,h}(\xi) < -3\tau/4$), and $\bar{u}_i(\xi) = \alpha_i$ and $\bar{u}_{i,h}(\xi) = \alpha_i > -\infty$ (respectively, $\bar{u}_i(\xi) = \beta_i$ and $\bar{u}_{i,h}(\xi) = \beta_i < \infty$). Thus $\mathbf{u}_{i,h}(\xi) = \mathbf{u}_i(\xi)$ if $|\bar{d}_i(\xi)| > \tau$, and condition (3.20) is satisfied.

Now consider the case where $U_{ad,h} = U^h_{ad}$. For all $T \in \mathcal{T}_h$ and all $1 \le i \le m$ let us set

$$I_{i,T} = \int_T \bar{d}_{i,h}(x)dx.$$

Take $\xi \in \omega$ such that $\bar{d}_i(\xi) > \tau$. In this case $\bar{u}_i(\xi) = \alpha_i > -\infty$. Choose x in the same triangle T as ξ. Then

$$\bar{d}_{i,h}(x) = \bar{d}_{i,h}(x) - \bar{d}_i(x) + \bar{d}_i(x) - \bar{d}_i(\xi) + \bar{d}_i(\xi) > -\frac{\tau}{4} - \frac{\tau}{4} + \tau = \frac{\tau}{2}.$$

Therefore $I_{i,T} > 0$ and $\bar{u}_{i,h}|_T = \alpha_i$. In particular $\bar{u}_{i,h}(\xi) = \alpha_i$ and $\bar{u}_{i,h}(\xi) - \bar{u}_i(\xi) = 0$. Similarly if $\bar{d}_i(\xi) < -\tau$, we have $\bar{u}_{i,h}(\xi) = \beta_i < \infty$ and $\bar{u}_{i,h}(\xi) - \bar{u}_i(\xi) = 0$, and condition (3.20) is still satisfied in that case.

Thus second order sufficient conditions stated in Corollary 3.11 can be applied, and we have

$$J''(\bar{\mathbf{u}})(\bar{\mathbf{u}}_h - \bar{\mathbf{u}})^2 \ge \delta \|\bar{\mathbf{u}}_h - \bar{\mathbf{u}}\|^2_{L^2(\omega;\mathbb{R}^m)}.$$

On the other hand, with the mean value theorem, we obtain

$$(J'(\bar{\mathbf{u}}_h) - J'(\bar{\mathbf{u}}))(\bar{\mathbf{u}}_h - \bar{\mathbf{u}}) = J''(\bar{\mathbf{u}} + \theta_h(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h))(\bar{\mathbf{u}}_h - \bar{\mathbf{u}})^2$$

for some $0 < \theta_h < 1$. Due to the uniform convergence properties stated for the control and the adjoint state and the explicit form of the second derivative of J, it is clear that we can choose h_4 small enough to have

$$J''(\bar{\mathbf{u}} + \theta_h(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h))(\bar{\mathbf{u}}_h - \bar{\mathbf{u}})^2 \ge \frac{\delta}{2} \|\bar{\mathbf{u}}_h - \bar{\mathbf{u}}\|^2_{L^2(\omega;\mathbb{R}^m)}$$

for all $0 < h \le h_4$. The proof is complete. □

LEMMA 4.17. *Assume that $U_{ad,h} = U^h_{ad}$. There exists $0 < h_5 \le h_4$ such that for every $0 < h \le h_5$ there exist $\mathbf{u}^*_h \in U_h$ and a constant $C > 0$ independent of h such that*

(1) $\mathbf{u}^*_h \in U_{ad,h}$,
(2) $J'(\bar{\mathbf{u}})\bar{\mathbf{u}} = J'(\bar{\mathbf{u}})\mathbf{u}^*_h$,
(3) $\|\bar{\mathbf{u}} - \mathbf{u}^*_h\|_{L^\infty(\omega;\mathbb{R}^m)} \le Ch$.

Proof. For every triangle $T \in \mathcal{T}_h$ and $1 \le i \le m$, define

$$I_{i,T} = \int_T \bar{d}_i(x)dx$$

and

$$u^*_{i,h}|_T = \begin{cases} \dfrac{1}{I_{i,T}} \displaystyle\int_T d_i(x)\bar{u}_i(x)dx & \text{if } I_{i,T} \neq 0, \\ \dfrac{1}{|T|} \displaystyle\int_T \bar{u}_i(x)dx & \text{if } I_{i,T} = 0. \end{cases}$$

Due to the Lipschitz continuity of $\bar{\mathbf{u}}$, there exists $0 < h_5 \le h_4$ such that, for $0 < h \le h_5$, each component $\bar{u}_i$ cannot achieve both values α and β in the same triangle. Hence, for each $T \in \mathcal{T}_h$, either $\bar{d}_i(x)$ is nonnegative for all $x \in T$ or $\bar{d}_i(x)$ is nonpositive for all $x \in T$. Therefore, $I_{i,T} = 0$ if and only if $\bar{d}_i(x) = 0$ for all $x \in T$. Moreover, if $I_{i,T} \neq 0$, then $\bar{d}_i(x)/I_{i,T} \ge 0$ for all $x \in T$. So applying the integral mean value theorem if $I_{i,T} = 0$ or the generalized mean value theorem if $I_{i,T} \neq 0$, we have $u^*_{i,h}|_T = \bar{u}_i(x_T)$ for some $x_T \in T$. As a first consequence, $\mathbf{u}^*_h \in U_{ad,h}$. Moreover, due to the Lipschitz continuity of $\bar{\mathbf{u}}$, we have that for $x \in \omega$, if we fix the triangle T such that $x \in T$,

$$|\bar{u}_i(x) - u^*_{i,h}(x)| = |\bar{u}_i(x) - \bar{u}_i(x^i_T)| \le C\|x - x^i_T\|_{\mathbb{R}^d} \le Ch,$$

and we have proved statement 3.

Since $I_{i,T} = 0$ if and only if $\bar{d}_i(x) = 0$ for all $x \in T$, we can claim that

$$I_{i,T}u^*_{i,h}|_T = \int_T \bar{d}_i(x)\bar{u}_i(x)dx$$

for all $T \in \mathcal{T}_h$ and all $1 \le i \le m$. A straightforward calculation yields statement 2:

$$J'(\bar{\mathbf{u}})\mathbf{u}^*_h = \int_\omega \bar{\mathbf{d}}(x) \cdot \bar{\mathbf{u}}^*_h(x)dx = \sum_{i=1}^m \sum_{T\in\mathcal{T}_h} \int_T \bar{d}_i(x)u^*_{i,h}(x)dx$$
$$= \sum_{i=1}^m \sum_{T\in\mathcal{T}_h} I_{i,T}u^*_{i,h}|_T = \sum_{i=1}^m \sum_{T\in\mathcal{T}_h} \int_T \bar{d}_i(x)\bar{u}_i(x)dx = J'(\bar{\mathbf{u}})\bar{\mathbf{u}}. \qquad \square$$

THEOREM 4.18. *There exists a constant $C > 0$ such that, for all $0 < h \le h_5$, we have*

$$\|\bar{\mathbf{u}} - \bar{\mathbf{u}}_h\|_{L^2(\omega;\mathbb{R}^m)} \le Ch^2 \qquad \text{if } \ U_{ad,h} = U_{ad},$$

while

$$\|\bar{\mathbf{u}} - \bar{\mathbf{u}}_h\|_{L^2(\omega;\mathbb{R}^m)} \le Ch \qquad \text{if } \ U_{ad,h} = U^h_{ad}.$$

Proof. (i) Let us start with the case where $U_{ad,h} = U^h_{ad}$. For $0 < h \le h_5$, we have

$$\begin{aligned} &\frac{\delta}{2}\|\bar{\mathbf{u}} - \bar{\mathbf{u}}_h\|^2_{L^2(\omega;\mathbb{R}^m)} \le (J'(\bar{\mathbf{u}}) - J'(\bar{\mathbf{u}}_h))(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h) \\ &= (J'(\bar{\mathbf{u}}) - J'_h(\bar{\mathbf{u}}_h))(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h) + (J'_h(\bar{\mathbf{u}}_h) - J'(\bar{\mathbf{u}}_h))(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h). \end{aligned} \tag{4.19}$$

With (4.12) in Lemma 4.13, we can estimate the last term as follows:

$$
\begin{aligned}
&(J_h'(\bar{\mathbf{u}}_h) - J'(\bar{\mathbf{u}}_h))(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h) \\
&= \int_\omega (\mathcal{C}^*(\bar{\boldsymbol{\Phi}}_h - \boldsymbol{\Phi}_{\bar{\mathbf{u}}_h}) + N(\bar{\mathbf{u}}_h - \bar{\mathbf{u}}_h)) \cdot (\bar{\mathbf{u}} - \bar{\mathbf{u}}_h) dx \\
&\le C\|\bar{\boldsymbol{\Phi}}_h - \boldsymbol{\Phi}_{\bar{\mathbf{u}}_h}\|_{\mathbf{L}^2(\Omega)} \|\bar{\mathbf{u}} - \bar{\mathbf{u}}_h\|_{L^2(\omega;\mathbb{R}^m)} \\
&\le Ch^2 \|\bar{\mathbf{u}} - \bar{\mathbf{u}}_h\|_{L^2(\omega;\mathbb{R}^m)}.
\end{aligned} \tag{4.20}
$$

Let us check what happens with the first term. From first order optimality conditions for problems (P) and (P$_h$) we have

$$
\begin{aligned}
J'(\bar{\mathbf{u}})(\bar{\mathbf{u}}_h - \bar{\mathbf{u}}) &\ge 0, \\
J_h'(\bar{\mathbf{u}}_h)(\mathbf{u}_h^* - \bar{\mathbf{u}}_h) = J_h'(\bar{\mathbf{u}}_h)(\mathbf{u}_h^* - \bar{\mathbf{u}}) &+ J_h'(\bar{\mathbf{u}}_h)(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h) \ge 0.
\end{aligned}
$$

Making the sum of these two expressions and using Lemma 4.17(2)–(3), we have

$$
\begin{aligned}
&J'(\bar{\mathbf{u}})(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h) - J_h'(\bar{\mathbf{u}}_h)(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h) \le J_h'(\bar{\mathbf{u}}_h)(\mathbf{u}_h^* - \bar{\mathbf{u}}) \\
&= J_h'(\bar{\mathbf{u}}_h)(\mathbf{u}_h^* - \bar{\mathbf{u}}) - J'(\bar{\mathbf{u}})(\mathbf{u}_h^* - \bar{\mathbf{u}}) \\
&= \int_\omega (\mathcal{C}(\bar{\boldsymbol{\Phi}}_h - \bar{\boldsymbol{\Phi}}) + N(\bar{\mathbf{u}}_h - \bar{\mathbf{u}})) \cdot (\mathbf{u}_h^* - \bar{\mathbf{u}}) dx \\
&\le C(\|\bar{\boldsymbol{\Phi}}_h - \bar{\boldsymbol{\Phi}}\|_{\mathbf{L}^2(\Omega)} + \|\bar{\mathbf{u}}_h - \bar{\mathbf{u}}\|_{L^2(\omega;\mathbb{R}^m)}) \|\mathbf{u}_h^* - \bar{\mathbf{u}}\|_{L^2(\omega;\mathbb{R}^m)} \\
&\le Ch(\|\bar{\boldsymbol{\Phi}}_h - \boldsymbol{\Phi}_{\bar{\mathbf{u}}_h}\|_{\mathbf{L}^2(\Omega)} + \|\boldsymbol{\Phi}_{\bar{\mathbf{u}}_h} - \bar{\boldsymbol{\Phi}}\|_{\mathbf{L}^2(\Omega)} + \|\bar{\mathbf{u}}_h - \bar{\mathbf{u}}\|_{L^2(\omega;\mathbb{R}^m)}) \\
&\le Ch(h^2 + \|\bar{\mathbf{u}}_h - \bar{\mathbf{u}}\|_{L^2(\omega;\mathbb{R}^m)}).
\end{aligned} \tag{4.21}
$$

From (4.19), (4.20), and (4.21), we deduce that therefore there exists a constant $C > 0$, independent of h, such that

$$
\frac{\delta}{2} \|\bar{\mathbf{u}} - \bar{\mathbf{u}}_h\|^2_{L^2(\omega;\mathbb{R}^m)} \le Ch^3 + Ch\|\bar{\mathbf{u}}_h - \bar{\mathbf{u}}\|_{L^2(\omega;\mathbb{R}^m)}.
$$

We conclude with Young's inequality.

(ii) Now let us consider the case where $U_{ad,h} = U_{ad}$. We rewrite the previous steps by introducing the simplifications corresponding to this case. For $0 < h \le h_5$, we have

$$
\begin{aligned}
\frac{\delta}{2} \|\bar{\mathbf{u}} - \bar{\mathbf{u}}_h\|^2_{L^2(\omega;\mathbb{R}^m)} &\le (J'(\bar{\mathbf{u}}) - J'(\bar{\mathbf{u}}_h))(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h) \\
&= (J'(\bar{\mathbf{u}}) - J_h'(\bar{\mathbf{u}}_h))(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h) + (J_h'(\bar{\mathbf{u}}_h) - J'(\bar{\mathbf{u}}_h))(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h).
\end{aligned}
$$

Since $U_{ad,h} = U_{ad}$, from the first order optimality conditions satisfied by $\bar{\mathbf{u}}$ and $\bar{\mathbf{u}}_h$ we have

$$
(J'(\bar{\mathbf{u}}) - J_h'(\bar{\mathbf{u}}_h))(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h) \le 0.
$$

We have already seen that

$$
(J_h'(\bar{\mathbf{u}}_h) - J'(\bar{\mathbf{u}}_h))(\bar{\mathbf{u}} - \bar{\mathbf{u}}_h) \le Ch^2 \|\bar{\mathbf{u}} - \bar{\mathbf{u}}_h\|_{L^2(\omega;\mathbb{R}^m)}.
$$

Therefore there exists a constant $C > 0$ independent of h such that

$$
\frac{\delta}{2} \|\bar{\mathbf{u}} - \bar{\mathbf{u}}_h\|^2_{L^2(\omega;\mathbb{R}^m)} \le Ch^2 \|\bar{\mathbf{u}}_h - \bar{\mathbf{u}}\|_{L^2(\omega;\mathbb{R}^m)}.
$$

The proof is complete. □

From the previous theorem and Lemmas 4.10 and 4.13 we deduce

$$\|\bar{\mathbf{y}} - \bar{\mathbf{y}}_h\|_{\mathbf{H}_0^1(\Omega)} + \|\bar{p} - \bar{p}_h\|_{L_0^2(\Omega)} \leq Ch,$$
$$\|\bar{\mathbf{\Phi}} - \bar{\mathbf{\Phi}}_h\|_{\mathbf{H}_0^1(\Omega)} + \|\bar{\pi} - \bar{\pi}_h\|_{L_0^2(\Omega)} \leq Ch.$$

REFERENCES

[1] F. Abergel and E. Casas, *Some optimal control problems of multistate equations appearing in fluid mechanics*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 223–247.

[2] N. Arada, E. Casas, and F. Tröltzsch, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Comput. Optim. Appl., 23 (2002), pp. 201–229.

[3] E. Casas, *Optimality conditions for some control problems of turbulent flows*, in Flow Control, IMA Vol. Math. Appl. 68, M. Gunzburger, ed., Springer-Verlag, New York, 1995, pp. 127–147.

[4] E. Casas, *Error estimates for the numerical approximation of semilinear elliptic control problems with finitely many state constraints*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 345–374.

[5] E. Casas, *Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems*, Adv. Comput. Math., 26 (2007), pp. 137–153.

[6] E. Casas, M. Mateos, and F. Tröltzsch, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comput. Optim. Appl., 31 (2005), pp. 193–219.

[7] E. Casas and J.-P. Raymond, *Error estimates for the numerical approximation of Dirichlet boundary control for semilinear elliptic equations*, SIAM J. Control Optim., 45 (2006), pp. 1586–1611.

[8] L. Cattabriga, *Su un problema al contorno relativo al sistema di equazioni di Stokes*, Rend. Sem. Mat. Univ. Padova, 31 (1961), pp. 308–340.

[9] P. G. Ciarlet and J.-L. Lions, eds., *Handbook of Numerical Analysis*, Vol. II, Handb. Numer. Anal. II, Finite Element Methods. Part I, North–Holland, Amsterdam, 1991.

[10] K. Deckelnick and M. Hinze, *Semidiscretization and error estimates for distributed control of the instationary Navier–Stokes equations*, Numer. Math., 97 (2004), pp. 297–320.

[11] K. Eriksson *Improved accuracy by adapted mesh-refinements in the finite element method*, Math. Comp. 44 (1985), pp. 321–343.

[12] P. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[13] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with distributed and Neumann controls*, Math. Comp., 57 (1991), pp. 123–151.

[14] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with Dirichlet controls*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 711–748.

[15] M. D. Gunzburger, L. Hou, and T. P. Svobodny, *Boundary velocity control of incompressible flow with an application to viscous drag reduction*, SIAM J. Control Optim., 30 (1992), pp. 167–181.

[16] M. Hinze, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–63.

[17] H.-C. Lee and S. Kim, *Finite element approximation and computations of optimal Dirichlet boundary control problems for the Boussinesq equations*, J. Korean Math. Soc., 41 (2004), pp. 681–715.

[18] J. C. De los Reyes, *A primal-dual active set method for bilaterally control constrained optimal control of the Navier-Stokes equations*, Numer. Funct. Anal. Optim., 25 (2004), pp. 657–683.

[19] T. Roubiček and F. Tröltzsch, *Lipschitz stability of optimal controls for the steady-state Navier–Stokes equations*, Control Cybernet., 32 (2003), pp. 683–705.

[20] R. Temam, *Navier-Stokes Equations*, North–Holland, Amsterdam, 1979.

[21] F. Tröltzsch and D. Wachsmuth, *Second-order sufficient optimality conditions for the optimal control of Navier-Stokes equations*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 93–119.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 983–998

A SAMPLING METHOD AND APPROXIMATION RESULTS FOR IMPULSIVE SYSTEMS*

PETER R. WOLENSKI† AND STANISLAV ŽABIĆ†

Abstract. This paper studies an impulsive dynamical system that is in the form of a measure-driven differential inclusion. The employed solution concept depends upon a graph completion of the measure. The first main result shows that a subsequence of discrete-time trajectories *graph-converges* to a solution, and, under Lipschitz hypotheses, that every solution can be obtained in this manner. The second main result pursues a similar development with classical systems that approximate the given one.

Key words. impulsive systems, sampling, approximate solutions, optimal control, calculus of variations

AMS subject classification. 34A37

DOI. 10.1137/040620734

1. Introduction.

This paper studies a *measure-driven* dynamical system of the form

$$\begin{cases} dx \in F\big(x(t)\big)\,dt + G\big(x(t)\big)\,\mu(dt), \\ x(0-) = x_0, \end{cases} \tag{1.1}$$

where $F(\cdot)$ and $G(\cdot)$ are multifunctions (set-valued maps) whose values, respectively, are subsets of $\mathbb{R}^n$ and $\mathcal{M}_{n\times m}$ (= the $n\times m$ matrices), and μ is a vector-valued measure with values in a closed convex cone $K \subseteq \mathbb{R}^m$. The system (1.1) is also referred to as an *impulsive* system since the measure may have atoms (i.e., *impulses*) which in effect may force the state trajectory $x(\cdot)$ to be discontinuous. Analogously to classical ODE theory, one can also consider the following integral inclusion:

$$\begin{cases} x(t) \in x_0 + \displaystyle\int_0^t F\big(x(\tau)\big)\,d\tau + \int_{[0,t)} G\big(x(\tau)\big)\,\mu(d\tau), \\ x(0-) = x_0. \end{cases} \tag{1.2}$$

One expects and certainly desires that the solution sets of (1.1) and (1.2) coincide; however, in both cases the precise notion of solution needs further explanation (see section 2 below). The mathematical formalisms (1.1) and (1.2) combine the following: (i) the system studied in [20, 21, 23] where the measure is positive and scalar-valued (that is, $m = 1$), and (ii) the system in [7] where the the set-valued maps are singletons (that is, (1.1) is an impulsive differential *equation*). The goal of the present paper is to prove closure and approximation results. More specifically, we develop an analogue of the Euler time-discretization method and prove results on absolute continuous approximations of μ. There are subtle modeling issues between the vector and scalar cases that we will discuss in more detail below. We first give a brief review of the relevant literature.

*Received by the editors December 13, 2004; accepted for publication (in revised form) January 23, 2007; published electronically June 28, 2007.
http://www.siam.org/journals/sicon/46-3/62073.html

†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918 (wolenski@math.lsu.edu, zabic@math.lsu.edu).

The early study of impulsive systems dates back to Rishel's paper [17], where the idea of handling impulses through time reparametrization was introduced. Warga [24] soon after extended this technique to a more general context. Another key insight was observed by Dal Maso and Rampazzo [7] (see also [3, 4, 10]), where the subtlety in defining the multiplication of a point-mass measure with a state-dependent term was observed as not well-defined unless a graph completion was also provided. Loosely speaking, a graph completion is a relation in graph space that extends the graph of the distribution function $u(\cdot)$ of μ to a connected subset by prescribing an arc to connect the right- and left-hand limits of $u(\cdot)$ at the points of discontinuity. This additional information is crucial since different graph completions can give rise to different solutions. The notion of a solution to a measure-driven differential *equation* (that is, $F(\cdot)$ and $G(\cdot)$ in (1.1) are singleton-valued) was defined in [7] as the solution to an auxiliary system that is reparameterized in time and depends on a given graph completion. The measure in [7] is vector-valued and is interpreted as the derivative of a control function of bounded variation. A natural extension to differential *inclusions* is considered in [21], where the main goals are to extend the robust solution concept to set-valued dynamics and to prove a closure property of the set of solutions. The measure here, however, is real-valued and positive, which essentially only uses the canonical completion. This will be discussed further in the next paragraph.

The idea of a reparameterized solution is further developed in a control setting by Bressan [2], Bressan and Rampazzo [3, 4], Motta and Rampazzo [14], Rampazzo and Sartori [16], and others. For the applications in mind, it is natural in these papers to use vector-valued measures since they appear as the derivative of a vector-valued control. We adopt that viewpoint as well. The major technical difference between a vector-valued versus a scalar-valued measure is the prominent feature of a graph completion in the former. In the [21] formulation, the measure is scalar-valued and so the graph completion is a straight-line scalar completion of the distribution of μ, and the behavior of the trajectory during a jump is driven by a differential inclusion involving only G during a time interval with length equal to the magnitude of the measure's atom. In [2, 4, 25] and in the present paper, μ is vector-valued and the choice of a graph completion is incorporated into the definition of solution. This means the behavior during a jump depends on the particular graph completion. The relationship between the two solution concepts is related but can be different if the underlying vector fields appearing as the columns of $G(\cdot)$ do not commute; see [3, 4]. These papers write $G(x)d\mu$ as $\sum_{i=1}^{m} g_i(x,u)\dot{u}^i$; again there are some subtle modeling issues as to whether the formalisms are equivalent to (1.1), which will be the topic of future work. If μ is vector valued, then the scalar-valued measure $\frac{d\mu}{d|\mu|}d|\mu|$ as developed in [21] can capture the same dynamics only when a straight-line completion is used. One can also incorporate the cone constraint into $G(\cdot)$ and obtain a solution set consisting of all solutions that exist for *some* graph completion; this is discussed further in the conclusion of the paper. Thus the solution set defined in [21] does not distinguish between solutions that arise from different graph completions, and thus are essentially only widely applicable in systems where the vector fields $\{g_i(\cdot)\}$ commute.

There is a vast and rapidly growing literature mostly in the engineering community of so-called hybrid systems. A consensus is perhaps forming as to what constitutes a hybrid system, but there is no universally accepted formulation. The key feature of all hybrid systems is that discrete and continuous variables interact over time; we refer the reader to the literature [12, 19, 9] and their many references. Although (1.1) has discrete features that may appear through the presence of atoms in the measure μ, the relationship between (1.1) and other formulations of hybrid systems found in

the engineering literature partially overlaps but is not exact.

We also mention related and independent work by Murray [15] who studied a proper extension of integral functionals from absolutely continuous arcs to ones of bounded variation. This approach can handle the above dynamics by encoding them through the technique of infinite penalization. The extension required the same type of graph completion of a vector-valued measure and the arcs are of bounded variation of the type we employ here.

We have proven in [25] that a related concept of solution, called a *direct* solution, can be formulated directly from the differential form (1.1) of the inclusion by matching the components of the decomposition of each measure into its absolutely continuous, continuous singular, and discrete parts (cf. [8, p. 102]). We show in [25] that it is equivalent to the other solution concepts. The advantage, in our view, of the direct solution concept is that it provides insight into formulating invariance concepts, developing a Hamilton–Jacobi theory, and proving stability results. These topics will be developed in future work.

The two major issues addressed in the present paper involve (1) time discretization of (1.1), and (2) absolutely continuous approximation to (1.1). We show in Theorem 3.1 that an analogue to the Euler one-step method can be developed to produce a sequence of sampled solutions that *graph-converge* to a solution of (1.1). The second part of this theorem shows that under additional Lipschitz assumptions, *every* solution of (1.1) is the limit of such a sequence. Time discretizations play a major role in classical optimal control theory in several ways. For example, sampling techniques are employed in proving invariance results (see [6]), deriving refined necessary conditions (see [13]), etc. For the same reasons that time-discretion has been a fruitful development in classical control systems, it seems desirable that such techniques are available for impulsive systems.

The second issue is the approximation of the measure μ by measures that are absolutely continuous with respect to Lebesgue measure. This is also natural and desirable, for it validates the interpretation that impulsive systems are the "completions" of classical ones. Theorem 4.1 has two parts analogous to the two parts of Theorem 3.1. The first part shows that if a sequence of absolutely continuous measures *graph-converge* to a given graph completion of μ, then any sequence $\{x(\cdot)\}$ of associated solutions contains a graph-convergent subsequence. The second part of the theorem shows that under additional Lipschitz assumptions, every solution of (1.1) is the limit of such a sequence. Further motivation and comparison of this result with the closure theorem in [20] is given in the conclusion.

Impulsive systems are introduced precisely in section 2, and the main result of [25] is reviewed to provide motivational background. Section 3 develops our sampling method, and section 4 relates solutions of (1.1) with solutions of approximate systems that have absolutely continuous measures. In section 5, we conclude the paper by comparing our limiting theorems with the main robustness result in [20]. The distinction will be more clear after we have precisely developed our results.

Throughout the paper, the following data with accompanying assumptions are given:

(H1) A closed convex cone $K \subseteq \mathbb{R}^m$.

(H2) A multifunction $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ with closed graph and convex values, and satisfying

$$f \in F(x) \implies |f| \leq c(1 + |x|) \quad \forall x \in \mathbb{R}^n$$

(where $c > 0$ is a given constant).

(H3) A multifunction $G : \mathbb{R}^n \rightrightarrows \mathcal{M}_{n\times m}$ (where $\mathcal{M}_{n\times m}$ denotes the $n \times m$ dimensional matrices with real entries) with closed graph and closed convex values, and satisfying

$$g \in G(x) \implies \|g\| \leq c(1 + |x|) \quad \forall x \in \mathbb{R}^n.$$

The set of vector-valued Borel measures defined on the interval $[0, T] \subset \mathbb{R}$ with values in K is denoted by $\mathcal{B}_K([0, T])$.

2. Impulsive systems and their trajectories. Suppose $\mu \in \mathcal{B}_K([0, T])$ is given. The impulsive system considered in this paper is described by the differential inclusion (1.1). Basic theory of differential inclusions without impulses is covered in [1, 22, 5, 6]. The trajectory $x(\cdot)$ is a function of bounded variation; however, as described in the introduction, further information is required to frame an unambiguous solution concept. Recall that the (right continuous) distribution function $u(\cdot) : [0, T] \to \mathbb{R}^m$ of μ is given by $u(t) = \mu([0, t])$. Following [4, 2], a graph completion of $u(\cdot)$ consists of a Lipschitz continuous map $(\phi_0, \phi) : [0, S] \to [0, T] \times \mathbb{R}^m$ so that $\phi_0(\cdot)$ is nondecreasing and mapping onto $[0, T]$, and for every $t \in [0, T]$, there exists an $s \in [0, S]$ with $\big(\phi_0(s), \phi(s)\big) = (t, u(t))$. The role of the graph completion is to pin down the behavior of the trajectory $x(\cdot)$ during the "jumps" of $u(\cdot)$ so that multiplication by $G(x)$ during this fast time movement is unambiguous. The function ϕ_0 is a reparameterized time variable, and in this paper we avoid additional technical issues by always choosing it as the "filled-in" inverse of

$$\eta(t) := t + |\mu|([0, t]); \tag{2.1}$$

that is,

$$\phi_0(s) = t \quad \Leftrightarrow \quad \eta(t-) \leq s \leq \eta(t+), \tag{2.2}$$

where $\eta(t-)$ and $\eta(t+)$, respectively, denote the left-hand limit $\lim_{t\nearrow t_0} \eta(t)$ and right-hand limit $\lim_{t\searrow t_0} \eta(t)$. These left- and right-hand limits are equal if and only if t is not an atom of μ. If 0 is an atom of μ, then $\eta(0-) = 0$ by convention. We let $\mathcal{I}$ be the at most countable index set of atoms $\mathcal{T} := \{t_i\}_{i\in\mathcal{I}}$, and $I_i := [s_i^-, s_i^+] := \phi_0^{-1}(t_i)$ the "fast" time consumed during the jump. Since ϕ_0 will always be specified as in (2.1) and (2.2), only the spatial component $\phi(\cdot)$ will be referred to as the graph completion, and it is uniquely determined at each $s \notin \cup_{i\in\mathcal{I}} I_i$, since then $t = \phi_0(s)$ is uniquely determined and $\phi(s) = u(t)$. We also require *cone adherence* of $\phi(\cdot)$, which means that

$$\dot{\phi}(s) \in K \tag{2.3}$$

hold for almost all $s \in [0, S]$.

Suppose we are now given $\mu \in \mathcal{B}_K([0, T])$. Consider a three-tuple

$$X_\mu := \big(x(\cdot), \phi(\cdot), \{y_i(\cdot)\}_{i\in\mathcal{I}}\big) \tag{2.4}$$

with the following constituents: $x(\cdot) : [0, T] \to \mathbb{R}^n$ is of bounded variation with its points of discontinuity contained in the set $\mathcal{T}$ of μ's atoms, $\phi(\cdot) : [0, S] \to \mathbb{R}^m$ is a graph completion of μ's distribution function $u(\cdot)$, and $\{y_i(\cdot)\}_{i\in\mathcal{I}}$ is a collection of Lipschitz functions, each defined on the nondegenerate interval $I_i := [s_i^-, s_i^+] := \phi_0^{-1}(t_i)$ and satisfying $y_i(s_i^\pm) = x(t_i\pm)$.

The following is a slight modification of a definition given in [4, 2].

DEFINITION 2.1. *Consider a three-tuple X_μ as in* (2.4)*, and let*

$$y(s) = \begin{cases} x(t) & \text{if } s \notin \cup_{i\in\mathcal{I}} I_i, \quad t = \phi_0(s), \\ y_i(s) & \text{if } s \in I_i. \end{cases} \tag{2.5}$$

Then X_μ is a reparameterized solution *of* (1.1) *provided $y(\cdot)$ is Lipschitz on $[0, S]$ and satisfies*

$$\begin{cases} \dot{y}(s) \in F\big(y(s)\big)\dot{\phi}_0(s) + G\big(y(s)\big)\dot{\phi}(s) & \text{a.e. } s \in [0, S], \\ y(0) = x_0. \end{cases} \tag{2.6}$$

One may observe that $y(\cdot)$ defined by (2.5) is a graph completion of the vector-valued function $x(\cdot)$.

We next introduce a solution concept with the same data structure as in (2.4), but which requires properties stated directly in the original time frame. Recall that an arc $x(\cdot)$ of bounded variation induces a measure dx that can be decomposed into absolutely continuous, continuous singular, and discrete (that is, purely atomic) parts, and so can be written as

$$dx = \dot{x}(t)\,dt + dx_\sigma + dx_D,$$

where dx_σ is a singular continuous measure and $dx_D := \sum_{i\in\mathcal{I}} \delta^x_{t_i}$ is the discrete part with $\delta^x_{t_i}$ denoting the point mass jump of the vector $x(t_i+) - x(t_i-)$. If 0 is an atom, then the initial point of the jump is denoted by $x(0-)$. Likewise, the measure $\mu \in \mathcal{B}_K([0,T])$ decomposes into $\mu = \dot{u}(t)\,dt + \mu_\sigma + \mu_D$, where $\mu_D = \sum_{i\in\mathcal{I}} \delta^u_{t_i}$.

DEFINITION 2.2. *The three-tuple X_μ in* (2.4) *is a* solution *of* (1.1) *provided*

(i) *for almost all $t \in [0,T]$,*

$$\begin{cases} \dot{x}(t) \in F\big(x(t)\big) + G\big(x(t)\big)\dot{u}(t), \\ x(0-) = x_0; \end{cases}$$

(ii) *there exists a bounded μ_σ-measurable selection $\gamma(t) \in G\big(x(t)\big)$ with*

$$dx_\sigma = \gamma(t)\,\mu_\sigma \quad \textit{(as measures on } [0,T]\textit{)};$$

(iii) *the set of atoms of dx is contained in $\mathcal{T} = \{t_i\}_{i\in\mathcal{I}}$, and for each $i \in \mathcal{I}$, $y_i(s_i^-) = x(t_i-)$, $y_i(s_i^+) = x(t_i+)$, and*

$$\dot{y}_i(s) \in G\big(y_i(s)\big)\dot{\phi}(s) \quad \text{a.e. } s \in I_i.$$

The fundamental role played by the graph completion in this definition surfaces in the differential inclusions stated in (iii), and in effect circumscribes the fast velocities that are available during that jump in t time. A simple concrete example is given in [2] where different graph completions give different reachable sets.

The following theorem is proven in [25]. Although this result is not required in the proofs that follow, we state it for completeness.

THEOREM 2.1. *Suppose $\mu \in \mathcal{B}_K([0,T])$ and X_μ is as in* (2.4)*. Then X_μ is a reparameterized solution of* (1.1) *if and only if X_μ is a solution of* (1.1)*.*

3. A sampling method. An Euler-type discretization procedure is introduced in this section that produces approximate discrete solutions (called *sampled trajectories*) when the measure μ and a graph completion are given. The limit of a subsequence of approximations will be shown to graph-converge in the Hausdorff metric to some solution X_μ of (1.1). In future work, we shall describe a sampling method that produces the measure and graph completion as well.

With X_μ as in (2.4), its graph is defined as the set

$$\operatorname{gr} X_\mu := \{(t, x(t)) : t \in [0,T]\} \cup \{(t_i, y_i(s)) : s \in I_i,\ i \in \mathcal{I}\}.$$

The idea is to discretize the ordinary trajectory $y(\cdot)$ that is defined in (2.5), where the "compactness of trajectories" is known to hold, and to project it down into t-space.

Let N be a positive integer, and let $h := \frac{S}{N}$ be the step-size parameter. Let $s_0 = 0 = t_0$, and for each $j = 1, \dots, N$, let $s_j = jh$, $t_j = \phi_0(s_j)$, and $\lambda_j = t_j - t_{j-1}$. Sampled points $\{x_j\}_{j=1}^N$ are defined and "velocity" data are selected as follows (the parameter N is suppressed in this notation):

$$\begin{array}{lll}
x_0 = x_0 & f_0 \in F(x_0) & g_0 \in G(x_0) \\
x_1 = x_0 + \lambda_1 f_0 + g_0\big(\phi(s_1) - \phi(s_0)\big) & & \\
 & f_1 \in F(x_1) & g_1 \in G(x_1) \\
\vdots & \vdots & \vdots \\
x_{j+1} = x_j + \lambda_j f_j + g_j\big(\phi(s_j) - \phi(s_{j-1})\big) & & \\
 & f_{j+1} \in F(x_{j+1}) & g_{j+1} \in G(x_{j+1}) \\
\vdots & \vdots & \vdots \\
x_N = x_{N-1} + \lambda_N f_{N-1} + g_{N-1}\big(\phi(s_N) - \phi(s_{N-1})\big) & &
\end{array}$$

We denote by Ω^N the graph of a sampled trajectory:

$$\Omega^N := \big\{(t_j, x_j) : j = 0, \dots, N\big\}. \tag{3.1}$$

Recall that the Hausdorff distance $\operatorname{dist}_{\mathcal{H}}(A_1, A_2)$ between two compact subsets A_1, A_2 of $\mathbb{R}^n$ is defined by

$$\operatorname{dist}_{\mathcal{H}}(A_1,\, A_2) = \min\big\{\delta \geq 0 : A_1 \subseteq A_2 + \delta\overline{\mathbb{B}} \text{ and } A_2 \subseteq A_1 + \delta\overline{\mathbb{B}}\big\},$$

and that any multifunction $M : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ with compact values is locally Lipschitz if for every bounded set $C \subset \mathbb{R}^n$ there exists a constant c so that

$$\operatorname{dist}_{\mathcal{H}}\big(M(x),\, M(y)\big) \leq c\|x - y\| \quad \forall\, x, y \in C.$$

The main result of this section follows.

THEOREM 3.1. *Suppose $\mu \in \mathcal{B}_K([0,T])$ and a graph completion $\phi(\cdot)$ are given.*

(a) *For every sequence $\{\Omega^N\}_N$ of graphs of sampled trajectories, there is a solution X_μ of (1.1) and a subsequence $\{\Omega^{N_k}\}_k$ of $\{\Omega^N\}_N$ such that*

$$\operatorname{dist}_{\mathcal{H}}\big(\Omega^{N_k},\, \operatorname{gr} X_\mu\big) \to 0 \quad \textit{as} \quad k \to \infty.$$

(b) *Assume F and G are locally Lipschitz. For every solution X_μ of* (1.1), *there exists a sequence $\{\Omega^N\}_N$ of graphs of sampled trajectories so that*

$$\operatorname{dist}_{\mathcal{H}}\big(\Omega^N, \operatorname{gr} X_\mu\big) \to 0 \quad \textit{as} \quad N \to \infty.$$

Proof. Suppose the sequences $\{f_j\}$, $\{g_j\}$, $\{x_j\}$ are constructed by the sampling method described above. We first show there exists a constant c_1 independent of N so that

$$\max_j \big\{\|x_j\|, \|f_j\|, \|g_j\|\big\} \le c_1 \tag{3.2}$$

for all j and $N \in \mathbb{N}$. Indeed, with r as in (2.3) (which is the Lipschitz constant of $\phi(\cdot)$) and c as in (H2) and (H3), we have

$$\begin{aligned} |x_{j+1}| &\le |x_j| + h|f_j| + \|g_j\| rh \\ &\le |x_j| + \big[c(1+|x_j|) + c(1+|x_j|r)\big]h \\ &= h\alpha + \big[1 + h\alpha\big]|x_j|, \end{aligned}$$

where $\alpha := c(1+r)$. It follows from the discrete Gronwall inequality that

$$|x_j| \le e^{\alpha S}\big(1 + |x_0|\big) - 1,$$

and that then (3.2) holds by (H2) and (H3) with $c_1 := c[e^{\alpha S}(1+|x_0|)]$.

Define the multifunction $M : [0,S] \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ by

$$M(s,y) = F(y)\dot{\phi}_0(s) + G(y)\dot{\phi}(s), \tag{3.3}$$

which is $\mathcal{L} \times \mathcal{B}$ measurable, has nonempty compact convex values, and has linear growth. Moreover, $M(s,\cdot)$ has closed graph for almost all $s \in [0,S]$. For each $N \in \mathbb{N}$, let $\tilde{\Omega}^N$ be the sampled trajectory in s-time:

$$\tilde{\Omega}^N := \big\{(s_j, x_j) : j = 0, \dots, N\big\}. \tag{3.4}$$

Also consider its related polygonal arc $y^N(\cdot)$ defined on $[0,S]$ given by

$$y^N(s) := x_j + \frac{s - s_j}{h}(x_{j+1} - x_j) \quad \text{for } s \in [s_j, s_{j+1}]. \tag{3.5}$$

Note for later use that

$$\operatorname{dist}_{\mathcal{H}}\big(\tilde{\Omega}^N, \operatorname{gr} y^N(\cdot)\big) \le \max\big\{h, c_1(1+r)h\big\}. \tag{3.6}$$

We claim there exist the sequences of

- positive numbers δ_N and r_N so that $\delta_N \to 0$ and $r_N \to 0$, and
- measurable sets $A_N \subseteq [0,S]$ so that $m(A_N) \to 0$,

where the limits are as $N \to \infty$, and that they satisfy

$$\inf\big\{\|\dot{y}^N(s) - v\| : v \in M\big(s, y^N(s) + \delta_N \overline{\mathbb{B}}\big)\big\} \;\le\; r_N \quad \text{a.e. } s \in A_N. \tag{3.7}$$

To see this, let $\delta_N = \frac{S}{N} c_1(1+r)$, where c_1 is as in (3.2). Note for each $j = 1, 2, \dots, N-1$ and $s \in [s_{j-1}, s_j]$ that

$$\begin{aligned} \big|y^N(s) - x_j\big| &\le \big|x_{j+1} - x_j\big| \\ &= \big|\lambda_{j+1} f_j + g_j\big(\phi(s_{j+1}) - \phi(s_j)\big)\big| \\ &\le h\big[|f_j| + \|g_j\| r\big] \\ &\le \delta_N. \end{aligned}$$

Next, for $s \in [0, S-h]$, define

$$\Phi_0^N(s) := \frac{1}{h}\int_s^{s+h} \dot\phi_0(s')ds' \quad \text{and} \quad \Phi^N(s) := \frac{1}{h}\int_s^{s+h} \dot\phi(s')ds'$$

and recall that $\Phi_0^N(s) \to \dot\phi_0(s)$ and $\Phi_0^N(s) \to \dot\phi(s)$ for almost all $s \in [0,S]$ as $N \to \infty$. By Egoroff's theorem, there exist measurable sets $A_N \subseteq [0,S]$ with $m(A_N) \to 0$ (and for notational simplicity, we may assume $[S-h, S] \subseteq A_N$) and satisfying

$$r_N := c_1 \max_{s\in[0,S]\setminus A_N} \left\{ \left|\Phi_0^N(s) - \dot\phi_0(s)\right|, \left|\Phi^N(s) - \dot\phi(s)\right| \right\} \to 0$$

as $N \to \infty$. Now let

$$v^N(s) := f_j\dot\phi_0(s) + g_j\dot\phi(s) \quad \text{for } s \in [s_j, s_{j+1}],$$

and note that $v^N(s) \in M(s, x_j)$ for almost all $s \in [s_j, s_{j+1}]$. Recall that $\dot y^N(s) = \Phi_0^N(s_j)f_j + g_j\Phi^N(s_j)$, and thus

$$\max_{s\in[0,S]\setminus A_N} \left|\dot y^N(s) - v^N(s)\right| \le \max_{\substack{j=1,\dots,N \\ s\in[s_j,s_{j+1}]\setminus A_N}} \left| \left(\Phi_0^N(s) - \dot\phi_0(s)\right)f_j + g_j\left(\Phi^N(s) - \dot\phi(s)\right)\right| \le r_N.$$

We have shown that (3.7) holds.

From the compactness of trajectories theorem [6, Theorem 4.1.11], there exists a trajectory $y(\cdot)$ of M and a subsequence (labeled $\{y^{N_k}(\cdot)\}_k$) of $\{y^N(\cdot)\}_N$ so that $y^{N_k}(\cdot) \to y(\cdot)$ uniformly on $[0,S]$. One sees easily that this means

$$\operatorname{dist}_{\mathcal H}\left(\operatorname{gr} y^{N_k}(\cdot), \operatorname{gr} y(\cdot)\right) \to 0 \tag{3.8}$$

as $k \to \infty$. We define the components of a solution X_μ to (1.1) as follows. Let $x(\cdot) : [0,T] \to \mathbb{R}^n$ be given by $x(t) = y\big(\eta(t)\big)$, and define the functions $y_i(\cdot)$ (for each $i \in \mathcal{I}$) as the restriction of $y(\cdot)$ to I_i.

Now recall Ω^N as in (3.1) and $\tilde\Omega^N$ as in (3.4), and observe the second coordinates are the same for each $j = 1, \dots, N$. Similarly, the second coordinates of $\operatorname{gr} X_\mu$ and $\operatorname{gr} y(\cdot) := \left\{\big(s, y(s)\big) : s \in [0,S]\right\}$ are the same for each $t \notin \mathcal{T}$, $t = \phi_0(s)$; and when $t \in \mathcal{T}$, the set of projections onto the second coordinate are the same. Thus the difference between the Hausdorff distances of Ω^N and $\operatorname{gr} X_\mu$ on the one hand and $\tilde\Omega^N$ and $\operatorname{gr} y(\cdot)$ on the other is affected by only the first coordinate. It follows that

$$\operatorname{dist}_{\mathcal H}\big(\Omega^N, \operatorname{gr} X_\mu\big) \le \operatorname{dist}_{\mathcal H}\big(\tilde\Omega^N, \operatorname{gr} y(\cdot)\big), \tag{3.9}$$

where the right-hand side is at most h larger than the left-hand side. By the triangle inequality, one has

$$\operatorname{dist}_{\mathcal H}\big(\tilde\Omega^N, \operatorname{gr} y(\cdot)\big) \le \operatorname{dist}_{\mathcal H}\big(\tilde\Omega^N, \operatorname{gr} y^N(\cdot)\big) + \operatorname{dist}_{\mathcal H}\big(\operatorname{gr} y^N(\cdot), \operatorname{gr} y(\cdot)\big). \tag{3.10}$$

Finally, passing to the subsequence $\{N_k\}$ and starting from (3.9), it follows from (3.10), (3.6), and (3.8) that

$$\operatorname{dist}_{\mathcal H}\big(\Omega^{N_k}, \operatorname{gr} X_\mu\big) \to 0,$$

which finishes the proof of part (a).

To prove part (b), assume now that F and G are locally Lipschitz, and X_μ is as in (2.4) and is a solution of (1.1). Let $y(\cdot)$ be defined as in (2.5), and so there exist measurable selections $f(\cdot)$ and $g(\cdot)$ of $F\big(y(\cdot)\big)$ and $G\big(y(\cdot)\big)$, respectively, so that

$$\dot{y}(s) = f(s)\dot{\phi}_0(s) + g(s)\dot{\phi}(s) \quad \text{a.e. } s \in [0, S].$$

In a manner similar to proving the discrete bound (3.2), one can show there exists a constant c_2 so that $|y(s)| \le c_2$. Observe that for $0 \le \bar{s} < \hat{s} \le S$, one has

$$\big|y(\hat{s}) - y(\bar{s})\big| \le \int_{\bar{s}}^{\hat{s}} |\dot{y}(s)|\, ds \le (1 + c_2)(1 + r)(\hat{s} - \bar{s}) =: c_3(\hat{s} - \bar{s}). \tag{3.11}$$

Let $L > 0$ be the Lipschitz constant for F and G on $c_2\overline{\mathbb{B}}$, and denote by $\mathrm{proj}_{F(y)}(f)$ the projection of f into $F(y)$ (which is unique since $F(y)$ is convex). If $|y_j| \le c_2$ $(j = 1, 2)$ and $f \in F(y_1)$, then $|f - \mathrm{proj}_{F(y_2)}(f)| \le L|y_1 - y_2|$. Similar considerations hold with F replaced by G.

We use the notation of the sampling method and will show there exists a sequence $\{\Omega^N\}$ that graph-converges to $\mathrm{gr}\, X_\mu$.

Let $f_0 = \frac{1}{h}\int_0^{s_1} \mathrm{proj}_{F(x_0)}\big(f(s)\big)\, ds$, let $g_0 = \frac{1}{h}\int_0^{s_1} \mathrm{proj}_{G(x_0)}\big(g(s)\big)\, ds$, and let x_1 be defined as in the sampling method. We observe

$$\begin{aligned} x_1 - y(s_1) \;=\; & \frac{\phi_0(s_1) - \phi_0(0)}{h} \int_0^{s_1} \big[\mathrm{proj}_{F(x_0)}\big(f(s)\big) - f(s)\big]\, ds \\ & + \int_0^{s_1} \big[\mathrm{proj}_{G(x_0)}\big(g(s)\big) - g(s)\big] \left(\frac{\phi(s_1) - \phi(0)}{h}\right) ds \\ & + \int_0^{s_1} \left(\frac{\phi_0(s_1) - \phi_0(0)}{h} - \dot{\phi}_0(s)\right) f(s)\, ds \\ & + \int_0^{s_1} g(s) \left(\frac{\phi(s_1) - \phi(0)}{h} - \dot{\phi}(s)\right) ds \\ =: \; & I + II + III + IV. \end{aligned}$$

Recall that $\phi_0(\cdot)$ is Lipschitz of rank 1, and so by the Lipschitz property of F, we have

$$|I| \le L \int_0^{s_1} \big|y(s) - x_0\big|\, ds \le Lc_3 \int_0^{s_1} s\, ds = \frac{Lc_3}{2} h^2,$$

where the second inequality follows from (3.11). In the same way, one can show $|II| \le \frac{Lc_3 r}{2} h^2$ since $\phi(\cdot)$ is Lipschitz of rank r. To estimate III and IV, we reuse earlier notation to redefine $\Phi^N(\cdot)$ on $[0, S]$ by setting

$$\Phi^N(s) := \max\left\{ \left| \frac{\phi_0(s_{j+1}) - \phi_0(s_j)}{h} - \dot{\phi}_0(s) \right|, \left| \frac{\phi(s_{j+1}) - \phi(s_j)}{h} - \dot{\phi}(s) \right| \right\}$$

whenever $s \in [s_j, s_{j+1}]$. Then it follows that both $|III|$ and $|IV|$ are bounded above by $c(1 + c_2)\int_0^{s_1} \Phi^N(s)\, ds$. Putting all this together, we have

$$|x_1 - y(s_1)| \le \frac{Lc_3(1 + r)}{2} h^2 + 2c(1 + c_2) \int_0^{s_1} \Phi^N(s)\, ds.$$

Inductively, one proceeds by setting $f_j = \frac{1}{h}\int_{s_j}^{s_{j+1}} f(s)\,ds$ and $g_j = \frac{1}{h}\int_{s_j}^{s_{j+1}} g(s)\,ds$, and letting x_{j+1} be as in the sampling method construction. The same argument used above can operate at each iteration, and inductively one has the following estimate:

$$|x_j - y(s_j)| \le \frac{Lc_3(1+r)}{2} jh^2 + 2c(1+c_2)\int_0^{s_j} \Phi^N(s)\,ds.$$

Since $\Phi^N(s)$ is bounded above and converges to 0 almost everywhere, it follows that $\tilde{\Omega}^N := \{(s_j, x_j) : j = 1, \dots, N\}$ satisfies $\operatorname{dist}_{\mathcal{H}}\big(\tilde{\Omega}^N, \operatorname{gr} y(\cdot)\big) \to 0$ as $N \to \infty$. The bound in (3.9) is still valid here, and the conclusion of (b) readily follows. □

4. Approximate controls. The original and perhaps most natural approach to defining solutions to the impulsive inclusion (1.1) is to consider limits of a sequence of solutions $x^N(\cdot)$ of an *approximate* control problem of the form

$$\dot{x}^N(t) \in F\big(x(t)\big)\dot{\phi}_0(t) + G\big(x(t)\big)\dot{u}^N(t), \tag{4.1}$$

where $d\mu^N = \dot{u}^N(\cdot)dt$ are absolutely continuous measures that approximate μ in some sense. See, for example, the discussion in [2]. We introduce in this section a concept of "graph convergence" of measures that is appropriate to carry out such an analysis. Graph convergence as defined below is perhaps considerably stronger than would be desirable, but we mention that even when the solutions of (4.1) are unique (which happens, for example, in the singleton case $F(x) = \{f(x)\}$ and $G(x) = \{g(x)\}$ with $f(\cdot)$ and $g(\cdot)$ smooth functions), the limit arc may not be unique if the measures converge in some weaker sense.

Suppose we are given the following: a measure $\mu \in \mathcal{B}_K([0,T])$, an associated graph completion $\phi(\cdot) : [0,S] \to \mathbb{R}^n$ that is Lipschitz of rank r, and a sequence $\{\mu^N\}$ of absolutely continuous measures belonging to $\mathcal{B}_K([0,T])$ whose associated distribution functions $u^N(t) := \mu^N\big([0,t]\big)$ are Lipschitz.

DEFINITION 4.1. *The sequence $\{\mu^N\}_N$ of absolutely continuous measures* graph-converges *to (μ, ϕ) provided*

(i) *there exist numbers $S^N > 0$ such that $S^N \to S$;*

(ii) *for each N, there exists a strictly increasing function $\phi_0^N(\cdot) : [0, S^N] \to [0,T]$ that is onto and Lipschitz of rank at most one, and such that*

$$\int_0^{\min\{S,S^N\}} \big|\dot{\phi}_0^N(s) - \dot{\phi}_0(s)\big|\,ds \to 0 \quad as \quad N \to \infty;$$

(iii) *for each N, the sequence of functions defined by $\phi^N(s) := \big(u^N \circ \phi_0^N\big)(s)$ is Lipschitz with $\limsup_{N\to\infty} \|\dot{\phi}^N(\cdot)\|_\infty \le r$ and satisfies*

$$\int_0^{\min\{S,S^N\}} \big|\dot{\phi}^N(s) - \dot{\phi}(s)\big|\,ds \to 0 \quad as \quad N \to \infty.$$

The main result in this section follows.

THEOREM 4.1. *Suppose the measure $\mu \in \mathcal{B}_K([0,T])$ and an associated graph completion $\phi(\cdot) : [0,S] \to \mathbb{R}^n$ are given.*

(a) *Suppose $\{\mu^N\}$ is a sequence of absolutely continuous measures that graph-converges to $\big(\mu, \phi(\cdot)\big)$, and $\{x^N(\cdot)\}$ is a sequence of absolutely continuous arcs satisfying*

$$\dot{x}^N(t) \in F\big(x^N(t)\big) + G\big(x^N(t)\big)\dot{u}^N(t). \tag{4.2}$$

Then there exists a solution X_μ *of* (1.1) *and a subsequence* $\{x^{N_k}(\cdot)\}$ *of* $\{x^N(\cdot)\}$ *such that*

$$\operatorname{dist}_{\mathcal{H}}\big(\operatorname{gr} x^{N_k}(\cdot),\, \operatorname{gr} X_\mu\big) \to 0 \quad \textit{as} \quad k \to \infty.$$

(b) *Conversely, suppose* F *and* G *are locally Lipschitz multifunctions and* $X_\mu := \big(x(\cdot), \phi(\cdot), \{y_i(\cdot)\}_{i\in\mathcal{I}}\big)$ *is a solution of* (1.1). *Then there is a sequence* $\{\mu^N\}$ *of absolutely continuous measures that graph converge to* $\big(\mu, \phi(\cdot)\big)$, *and a sequence* $x^N(\cdot)$ *of solutions to* (4.2) *so that*

$$\operatorname{dist}_{\mathcal{H}}\big(\operatorname{gr} x^{N}(\cdot),\, \operatorname{gr} X_\mu\big) \to 0 \quad \textit{as} \quad N \to \infty.$$

Proof. Suppose we are given the measures $d\mu^N = \dot{u}^N(t)dt$, the functions $\phi_0^N(\cdot)$ and $\phi^N(\cdot)$ satisfying Definition 4.1, and solutions $x^N(\cdot)$ of (4.2). Set $\bar{S}^N := \min\{S, S^N\}$. Let $y^N(s) = (x^N \circ \phi_0^N)(s)$, which for almost all $s \in [0, \bar{S}^N]$ satisfies

$$\begin{aligned}
\dot{y}^N(s) &= \dot{x}^N\big(\phi_0^N(s)\big)\dot{\phi}_0^N(s) \\
&\in F\big(y^N(s)\big)\dot{\phi}_0^N(s) + G\big(y^N(s)\big)\dot{u}^N\big(\phi_0^N(s)\big)\dot{\phi}_0^N(s) \\
&= F\big(y^N(s)\big)\dot{\phi}_0^N(s) + G\big(y^N(s)\big)\dot{\phi}^N(s),
\end{aligned}$$

where the last equality follows since $\dot{\phi}^N(s) = \dot{u}^N\big(\phi_0^N(s)\big)\dot{\phi}_0^N(s)$ almost everywhere. It follows that there exist measurable selections $f^N(s) \in F\big(y^N(s)\big)$ and $g^N(s) \in G\big(y^N(s)\big)$ so that

$$\dot{y}^N(s) = f^N(s)\dot{\phi}_0^N(s) + g^N(s)\dot{\phi}^N(s).$$

Recall that Definition 4.1 imposes a priori bounds on the Lipschitz rank of $\phi_0^N(\cdot)$ and $\phi^N(\cdot)$, and that $F(\cdot)$ and $G(\cdot)$ satisfy linear growth assumptions. A standard argument involving Gronwall's inequality implies there exists a constant c_4 independent of N that is an upper bound of both $\|f^N(\cdot)\|_\infty$ and $\|g^N(\cdot)\|_\infty$.

Let $M : [0, S] \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be defined as in (3.3), define $\dot{z}^N(\cdot) : [0, \bar{S}^N] \to \mathbb{R}^n$ by

$$\dot{z}^N(s) := f^N(s)\dot{\phi}_0(s) + g^N(s)\dot{\phi}(s),$$

and define $z^N(\cdot) : [0, \bar{S}^N] \to \mathbb{R}^n$ by $z^N(s) := x_0 + \int_0^s \dot{z}^N(s')\, ds'$. It is clear from the definitions that

$$\dot{z}^N(s) \in M\big(s, y^N(s)\big) \quad \text{a.e. } s \in [0, \bar{S}^N]. \tag{4.3}$$

Furthermore, it is readily seen that

$$\sup_{s\in[0,\bar{S}^N]} \big|z^N(s) - y^N(s)\big| \le c_4\big\{\|\dot{\phi}_0^N - \dot{\phi}_0\|_1 + \|\dot{\phi}^N - \dot{\phi}\|_1\big\},$$

which implies via the assumption of the graph convergence of the measures that $y^N - z^N$ approaches zero uniformly. In view of (4.3) and the compactness of trajectories theorem [6, Theorem 4.1.11], there exists $y(\cdot) : [0, S] \to \mathbb{R}^n$ that is a trajectory of M and to which a subsequence of $\{z^N(\cdot)\}$, and hence also of $\{y^N(\cdot)\}$, converges uniformly. That is, there exists a subsequence N_k for which

$$\operatorname{dist}_{\mathcal{H}}\big(\operatorname{gr} y^{N_k}(\cdot),\, \operatorname{gr} y(\cdot)\big) \to 0 \quad \text{as } k \to \infty. \tag{4.4}$$

We now define X_μ as before; see the paragraph containing (3.8) in the previous section. Similar reasoning as employed there shows also that $\operatorname{dist}_{\mathcal{H}}\big(\operatorname{gr} x^{N_k}(\cdot), \operatorname{gr} X_\mu\big)$ is bounded above by

$$\operatorname{dist}_{\mathcal{H}}\big(\operatorname{gr} y^{N_k}(\cdot),\, \operatorname{gr} y(\cdot)\big) + \sup_{s\in[0,\bar{S}^{N_k}]} |\phi_0^{N_k}(s) - \phi_0(s)|,$$

which goes to zero as $k \to \infty$ by (4.4) and the assumption contained in Definition 4.1(ii). This finishes the proof of part (a).

We turn to part (b). Suppose F and G are now locally Lipschitz and X_μ is a solution to (1.1). For $N = 1, \dots,$ we proceed to construct the absolutely continuous measures μ^N and solutions $x^N(\cdot)$ of (4.2) that will converge in graph to X_μ. Fix $N > 0$ and set $h = \frac{S}{N}$, and for $j = 1, \dots, N$, set $s_j = jh$ and $t_j = \phi_0(s_j)$. We will first introduce a new partition $\{\bar{t}_j\}$ of $[0, T]$ consisting of N distinct points that resembles the partition $\{t_j\}$ but has repeated nodes "pulled apart" and indexed accordingly. To this end, let $\mathcal{J}_0^N$ be those indices j for which $t_{j-1} < t_j < t_{j+1}$ (to treat the endpoints, by convention, we take $t_{-1} < t_0$ and $t_{N+1} > t_N$; thus $t_0 \in \mathcal{J}_0^N$ if $t_0 < t_1$ and $t_N \in \mathcal{J}_0^N$ if $t_{N-1} < t_N$). We set $\bar{t}_j = t_j$ whenever $j \in \mathcal{J}_0^N$. Let $\mathcal{J}^N$ be those indices j for which $t_{j-1} < t_j = t_{j+1}$ (by convention, then, $t_0 \in \mathcal{J}^N$ if $t_0 = t_1$ and t_N cannot belong to $\mathcal{J}^N$). For these latter j, let $k_j \geq 1$ be such that $t_j = t_{j+1} = \cdots = t_{j+k_j} < t_{j+k_j+1}$, and

$$\lambda_j := \frac{1}{2} \min\big\{h^2, t_j - t_{j-1}, t_{j+k_j+1} - t_j\big\}$$

(if $0 \in \mathcal{J}^N$, then $\lambda_0 := \min\{h^2, \frac{t_{j+k_j+1}-t_j}{2}\}$). If $j \notin \mathcal{J}_0^N$, then $j = \bar{j} + k$, where there exists precisely one pair $(\bar{j}, k)$ with $\bar{j} \in \mathcal{J}^N$ and $0 \leq k \leq k_{\bar{j}}$. In this case $\bar{t}_j$ is defined by

$$\bar{t}_j := \begin{cases} t_j + \left[\frac{2k}{k_j} - 1\right] \lambda_j & \text{if } j \neq 0, \\ \frac{k}{k_0}\lambda_0 & \text{if } j = 0. \end{cases}$$

Thus a new partition $\{\bar{t}_j\}$ of $[0, T]$ has been constructed consisting of N distinct points, and which satisfy

$$|\bar{t}_j - t_j| \leq h^2 \quad \forall j. \tag{4.5}$$

Next, we define $\phi_0^N(\cdot) : [0, S] \to [0, T]$ by

$$\phi_0^N(s) = \bar{t}_j + \frac{s - s_j}{h}(\bar{t}_{j+1} - \bar{t}_j) \quad \text{whenever } s \in [s_j, s_{j+1}],$$

which is onto and Lipschitz of rank at most 1. We claim that $\dot{\phi}_0^N(\cdot)$ converges to $\dot{\phi}_0(\cdot)$ in $L^1[0, S]$. Indeed, let $\tilde{\phi}_0^N(\cdot) : [0, S] \to [0, T]$ be given by

$$\tilde{\phi}_0^N(s) = t_j + \frac{s - s_j}{h}(t_{j+1} - t_j) \quad \text{whenever } s \in [s_j, s_{j+1}].$$

The difference between the linear interpolations $\phi_0^N(\cdot)$ and $\tilde{\phi}_0^N(\cdot)$ is that $\phi_0^N(\cdot)$ maps s_j to $\bar{t}_j$, whereas $\tilde{\phi}_0^N(\cdot)$ maps s_j to t_j. For $s \in [s_j, s_{j+1}]$, we have

$$\big|\dot{\phi}_0^N(s) - \dot{\tilde{\phi}}_0^N(s)\big| = \frac{1}{h}|\bar{t}_{j+1} - \bar{t}_j - t_{j+1} + t_j| \leq 2h, \tag{4.6}$$

where the inequality is justified by (4.5). The Lebesgue differentiation theorem says that $\dot{\bar{\phi}}_0^N(s) \to \dot{\phi}_0(s)$ as $N \to \infty$ for almost all $s \in [0, S]$, and since these functions are bounded above by 1, the dominated convergence theorem implies that $\dot{\bar{\phi}}_0^N(\cdot) \to \dot{\phi}_0(\cdot)$ in $L^1[0, S]$. It follows from this and (4.6) that $\dot{\phi}_0^N(\cdot) \to \dot{\phi}_0(\cdot)$ in $L^1[0, S]$, as claimed.

Now define $u^N(\cdot) : [0, T] \to \mathbb{R}^n$ as the piecewise linear interpolation satisfying $u^N(\bar{t}_j) = \phi(s_j)$; that is,

$$u^N(t) = \phi(s_j) + \frac{t - \bar{t}_j}{\bar{t}_{j+1} - \bar{t}_j}\big(\phi(s_{j+1}) - \phi(s_j)\big) \quad \text{whenever } t \in [\bar{t}_j, \bar{t}_{j+1}].$$

Let $\phi^N(\cdot) := (u^N \circ \phi_0^N)(\cdot)$, and note $\phi^N(s_j) = \phi(s_j)$ for all j, and for $s \in [s_j, s_{j+1}]$ that

$$\dot{\phi}^N(s) = \dot{u}^N\big(\phi_0^N(s)\big)\dot{\phi}_0^N(s) = \frac{\phi(s_{j+1}) - \phi(s_j)}{\bar{t}_{j+1} - \bar{t}_j}\,\frac{\bar{t}_{j+1} - \bar{t}_j}{h} = \frac{\phi(s_{j+1}) - \phi(s_j)}{h}.$$

Since $\phi(\cdot)$ is Lipschitz of rank r, it follows that each of $\phi^N(\cdot)$ are also of rank at most r. Completely analogous to the proof above showing $\dot{\phi}_0^N(\cdot) \to \dot{\phi}_0(\cdot)$ in $L^1[0, S]$ as $N \to \infty$, one has that $\dot{\phi}^N(\cdot) \to \dot{\phi}(\cdot)$ in $L^1[0, S]$ as $N \to \infty$. Therefore, with μ^N the absolutely continuous measure satisfying $d\mu^N = \dot{u}^N(t)dt$, we have shown that μ^N graph-converges to $(\mu, \phi(\cdot))$ as $N \to \infty$ (where $S^N = S$ for all N in Definition 4.1).

We now turn to approximating a given solution X_μ by a solution of (4.2). By Theorem 3.1(b), there exists a sequence of sampled trajectories whose graphs converge to $\operatorname{gr} X_\mu$. Denote these graphs by

$$\Omega^N := \{(t_j, x_j) \mid j = 1, \dots, N\},$$

where $x_{j+1} = x_j + (t_{j+1} - t_j)f_j + (g_j)\big(\phi(s_{j+1}) - \phi(s_j)\big)$, $f_j \in F(x_j)$, and $g_j \in G(x_j)$, and they satisfy

$$\operatorname{dist}_{\mathcal{H}}\big(\Omega^N, \operatorname{gr} X_\mu\big) \to 0 \quad \text{as} \quad N \to \infty. \tag{4.7}$$

For simplicity of notation, the dependence of x_j, f_j, and g_j on N has been suppressed. A new sampled set of points $\{\bar{x}_j\}$ is defined by replacing the partition $\{t_j\}$ by $\{\bar{t}_j\}$ and "tracking" the given sampled data. This is done as follows. Let $\bar{f}_0 = f_0$ and $\bar{g}_0 = g_0$ and define

$$\bar{x}_1 = \bar{x}_0 + (\bar{t}_1 - \bar{t}_0)\bar{f}_0 + (\bar{g}_0)\big(\phi(s_1) - \phi(s_0)\big).$$

Having chosen the data at stage $j - i$, inductively let $\bar{f}_j \in F(\bar{x}_j)$ and $\bar{g}_j \in G(\bar{x}_j)$ be the projections of f_j and g_j onto $F(\bar{x}_j)$ and $G(\bar{x}_j)$, respectively. That is, $\bar{f}_j \in F(\bar{x}_j)$ and satisfies

$$|\bar{f}_j - f_j| = \inf_{f \in F(\bar{x}_j)} |f - f_j|,$$

and similarly for $\bar{g}_j$. Define the next node by

$$\bar{x}_{j+1} = \bar{x}_j + (\bar{t}_{j+1} - \bar{t}_j)\bar{f}_j + (\bar{g}_j)\big(\phi(s_{j+1}) - \phi(s_j)\big).$$

The linear growth assumptions on F and G guarantee that all of the sampled data remain in a bounded set, and let c_1 be as in (3.2) but such that it also bounds the newly sampled data. With L a Lipschitz constant for both F and G on $c_1\overline{\mathbb{B}}$, one has

$$|\bar{f}_j - f_j| \le L|\bar{x}_j - x_j| \quad \text{and} \quad \|\bar{g}_j - g_j\| \le L|\bar{x}_j - x_j|. \tag{4.8}$$

The estimate between the nodes x_j and $\bar{x}_j$ is calculated by

$$\begin{aligned}\left|\bar{x}_{j+1}-x_{j+1}\right| &\le |\bar{x}_j-x_j|+|t_{j+1}-t_j-\bar{t}_{j+1}+\bar{t}_j|\,|f_j|\\ &\quad+|\bar{t}_{j+1}-\bar{t}_j|\,|\bar{f}_j-f_j|+\|\bar{g}_j-g_j\|\,|\phi(s_{j+1})-\phi(s_j)|\\ &\le |\bar{x}_j-x_j|+2h^2c_1+hL|\bar{x}_j-x_j|+hLr|\bar{x}_j-x_j|\\ &= 2h^2c_1+(1+hL+hLr)|\bar{x}_j-x_j|,\end{aligned}$$

where the second inequality was deduced using (4.5), (4.8), and the fact that $\phi(\cdot)$ is Lipschitz of rank r. Gronwall's inequality implies

$$|\bar{x}_j-x_j|\le 2hc_1\frac{e^{LS(1+r)}-1}{L(1+r)}$$

for each $j=0,1,\dots,N$ and in particular implies that

$$\operatorname{dist}_{\mathcal{H}}(\Omega^N,\bar{\Omega}^N)\to 0 \quad \text{as } N\to\infty, \tag{4.9}$$

where $\bar{\Omega}^N$ is the newly sampled graph:

$$\bar{\Omega}^N:=\{(\bar{t}_j,\bar{x}_j)\mid j=1,\dots,N\}.$$

Next, let $\bar{x}^N(\cdot)$ be the piecewise linear arc interpolating the points in $\bar{\Omega}^N(\cdot)$, which specifically means

$$\bar{x}^N(t)=\bar{x}_j+(t-\bar{t}_j)\bar{f}_j+(t-\bar{t}_j)\bar{g}_j\frac{\phi(s_{j+1})-\phi(s_j)}{\bar{t}_{j+1}-\bar{t}_j} \quad \text{and}$$

$$\dot{\bar{x}}^N(t)=\bar{f}_j+\bar{g}_j\dot{u}^N(t)\in F(\bar{x}_j)+G(\bar{x}_j)\dot{u}^N(t) \quad \text{whenever } t\in(\bar{t}_j,\bar{t}_{j+1}). \tag{4.10}$$

Let $\Gamma^N(\cdot):[0,T]\times\mathbb{R}^n\rightrightarrows\mathbb{R}^n$ be given by $\Gamma^N(t,x):=F(x)+G(x)\dot{u}^N(t)$, which is the multifunction appearing in (4.2). It has convex compact values, is measurably Lipschitz (see [5]), and has linear growth in x. We will find a trajectory $x^N(\cdot)$ of Γ^N that is close to $\bar{x}^N(\cdot)$. Following the notation in [5], we have

$$\begin{aligned}\rho_\Gamma\big(\bar{x}^N(\cdot)\big) &:= \int_0^T \operatorname{dist}\Big(\dot{\bar{x}}^N(t),\Gamma^N\big(t,\bar{x}^N(t)\big)\Big)\,dt\\ &= \sum_{j=0}^{N-1}\int_{t_j}^{t_{j+1}} \operatorname{dist}\Big(\dot{\bar{x}}^N(t),\Gamma^N\big(t,\bar{x}^N(t)\big)\Big)\,dt\\ &\le \sum_{j=0}^{N-1}\int_{t_j}^{t_{j+1}} \operatorname{dist}_{\mathcal{H}}\Big(\Gamma^N\big(t,\bar{x}_j\big),\Gamma^N\big(t,\bar{x}^N(t)\big)\Big)\,dt\\ &\le L\sum_{j=0}^{N-1}\int_{t_j}^{t_{j+1}}\big(1+|\dot{u}^N(t)|\big)\big|\bar{x}^N(t)-\bar{x}_j\big|\,dt,\end{aligned} \tag{4.11}$$

where (4.10) was used in the first inequality, and the Lipschitz property of F and G in the second. For $t\in[\bar{t}_j,\bar{t}_{j+1}]$, one has

$$\begin{aligned}\big|\dot{\bar{x}}^N(t)-\bar{x}_j\big| &\le \frac{t-\bar{t}_j}{\bar{t}_{j+1}-\bar{t}_j}\big|\bar{x}_{j+1}-\bar{x}_j\big|\\ &\le \frac{t-\bar{t}_j}{\bar{t}_{j+1}-\bar{t}_j}\big[(\bar{t}_{j+1}-\bar{t}_j)|\bar{f}_j|+\|\bar{g}_j\|\,|\phi(s_{j+1})-\phi(s_j)|\big]\\ &\le c_1\left[1+\frac{rh}{\bar{t}_{j+1}-\bar{t}_j}\right](t-\bar{t}_j)\end{aligned}$$

and

$$|\dot{u}^N(t)| = \left| \frac{\phi(s_{j+1}) - \phi(s_j)}{\bar{t}_{j+1} - \bar{t}_j} \right| \leq \frac{rh}{\bar{t}_{j+1} - \bar{t}_j}.$$

We thus have

$$\begin{aligned}\int_{t_j}^{t_{j+1}} \big(1 + |\dot{u}^N(t)|\big) \big|\bar{x}^N(t) - \bar{x}_j\big|\, dt &\leq \left[1 + \frac{rh}{\bar{t}_{j+1} - \bar{t}_j}\right] c_1 \left[1 + \frac{rh}{\bar{t}_{j+1} - \bar{t}_j}\right] \int_{t_j}^{t_{j+1}} (t - \bar{t}_j)\, dt \\ &= c_1 \left[1 + \frac{rh}{\bar{t}_{j+1} - \bar{t}_j}\right]^2 \frac{(\bar{t}_{j+1} - \bar{t}_j)^2}{2} \\ &\leq c_6 h^2\end{aligned}$$

for some constant c_6. Combined with (4.11), this estimate yields that

$$\rho_\Gamma\big(\bar{x}^N(\cdot)\big) \leq LSc_6 h,$$

and so by Filippov's theorem (see [5, Theorem 3.1.6, p. 115]), for each N there exists a trajectory $x^N(\cdot)$ of Γ^N such that $x^N(0) = x_0$ and for which

$$\operatorname{dist}_{\mathcal{H}}\big(\operatorname{gr} x^N(\cdot),\, \operatorname{gr} \bar{x}^N(\cdot)\big) \to 0 \tag{4.12}$$

as $N \to \infty$. Finally, we have by the triangular inequality

$$\begin{aligned}\operatorname{dist}_{\mathcal{H}}\big(\operatorname{gr} x^N(\cdot),\, \operatorname{gr} X_\mu\big) \leq{}& \operatorname{dist}_{\mathcal{H}}\big(\operatorname{gr} x^N(\cdot),\, \operatorname{gr} \bar{x}^N(\cdot)\big) + \operatorname{dist}_{\mathcal{H}}\big(\operatorname{gr} \bar{x}^N(\cdot),\, \bar{\Omega}^N\big) \\ &+ \operatorname{dist}_{\mathcal{H}}\big(\bar{\Omega}^N,\, \Omega^N\big) + \operatorname{dist}_{\mathcal{H}}\big(\Omega^N,\, \operatorname{gr} X_\mu\big),\end{aligned}$$

which approaches 0 as $N \to \infty$ by (4.12), (4.9), and (4.7). This finishes the proof. ∎

5. Conclusion. We developed a sampling method for impulsive systems that is analogous to the classical Euler one-step method of ODEs. These techniques will be employed in forthcoming work on invariance and Hamilton–Jacobi theory.

The second major result of the paper is concerned with approximation of the given measure by systems without impulses. We showed such approximation was possible if the measures graph-converged, which is a considerably stronger property than weak-$\star$ convergence of the measures. It is interesting to compare this result with the closure theorem in [21], which is the analogue of the compactness of trajectories theorem. It is shown in Theorem 5.1 of [21] that if a sequence of (positive, scalar-valued) measures $\{\mu^N\}_N$ converges weak-$\star$ to μ, then a subsequence of a given sequence of associated solutions to the inclusion involving μ^N converges to a solution of (1.1). The difference between this result and our Theorem 4.1 lies in the fact that the limiting arc in [21] would in general be a solution associated to *some* graph completion, whereas our goal in Theorem 4.1 was to tie the convergence to a particular graph completion.

REFERENCES

[1] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer, Berlin, 1984.

[2] A. Bressan, *Impulsive control systems*, in Nonsmooth Analysis and Geometric Methods in Deterministic Optimal Control, B. Mordukhovich and H. J. Sussmann, eds., Springer, New York, 1996, pp. 1–22.

[3] A. Bressan and F. Rampazzo, *Impulsive systems with commutative vector fields*, J. Optim. Theory Appl., 71 (1991), pp. 67–83.

[4] A. Bressan and F. Rampazzo, *Impulsive control systems without commutativity assumptions*, J. Optim. Theory Appl., 81 (1994), pp. 435–457.
[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canadian Mathematical Society, Wiley-Interscience, Toronto, 1983.
[6] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer, New York, 1998.
[7] G. Dal Maso and F. Rampazzo, *On systems of ordinary differential equations with measures as controls*, Differential Integral Equations, 4 (1991), pp. 739–765.
[8] G. B. Folland, *Real Analysis, Modern Techniques and Their Applications*, John Wiley, New York, 1984.
[9] R. Goebel and A. R. Teel, *Solutions to hybrid inclusions via set and graphical convergence with stability theory applications*, Automatica, 24 (2006), pp. 573–587.
[10] O. Hajek, *Review of "Differential Systems Involving Impulses"* (by S. G. Pandit and S. G. Deo), Bull. Amer. Math. Soc. (N.S.), 12 (1985), pp. 272–279.
[11] E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer, New York, 1965.
[12] A. S. Matveev and A. V. Savkin, *Qualitative Theory of Hybrid Dynamical Systems*, Birkhäuser, Boston, 2000.
[13] B. S. Mordukhovich, *Discrete approximations and refined Euler–Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.
[14] M. Motta and F. Rampazzo, *Dynamic programming for nonlinear systems driven by ordinary and impulsive controls*, SIAM J. Control Optim., 34 (1996), pp. 199–225.
[15] J. M. Murray, *Existence theorems for optimal control and calculus of variations problems where the states can jump*, SIAM J. Control Optim., 24 (1986), pp. 412–438.
[16] F. Rampazzo and C. Sartori, *The minimum time function with unbounded controls*, J. Math. Systems Estim. Control, 8 (1998), pp. 1–34.
[17] R. W. Rishel, *An extended Pontryagin principle for control systems whose control laws contain measures*, J. Soc. Indust. Appl. Math. Ser. A Control, 3 (1965), pp. 191–205.
[18] R. T. Rockafellar, *Dual problems of Lagrange for arcs of bounded variation*, in Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976, pp. 155–192.
[19] A. V. Savkin and R. J. Evans, *Hybrid Dynamical Systems, Controller and Sensor Switching Problems*, Birkhäuser, Boston, 2002.
[20] G. N. Silva and R. B. Vinter, *Measure driven differential inclusions*, J. Math. Anal. Appl., 202 (1996), pp. 767–746.
[21] G. N. Silva and R. B. Vinter, *Necessary conditions for optimal impulsive control problems*, SIAM J. Control Optim., 35 (1997), pp. 1829–1846.
[22] G. V. Smirnov, *Introduction to the Theory of Differential Inclusions*, AMS, Providence, RI, 2002.
[23] R. B. Vinter and F. M. F. L. Pereira, *A maximum principle for optimal processes with discontinuous trajectories*, SIAM J. Control Optim., 26 (1988), pp. 205–229.
[24] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[25] P. R. Wolenski and S. Žabić, *A differential solution concept for impulsive systems*, Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal., 13B (2006), pp. 199–210.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 999–1021

CONVEX PROGRAMS FOR TEMPORAL VERIFICATION OF NONLINEAR DYNAMICAL SYSTEMS*

STEPHEN PRAJNA† AND ANDERS RANTZER‡

Abstract. A methodology for safety verification of continuous and hybrid systems using barrier certificates has been proposed recently. Conditions that must be satisfied by a barrier certificate can be formulated as a convex program, and the feasibility of the program implies system safety in the sense that there is no trajectory starting from a given set of initial states that reaches a given unsafe region. The dual of this problem, i.e., the reachability problem, concerns proving the existence of a trajectory starting from the initial set that reaches another given set. Using insights from the linear programming duality appearing in the discrete shortest path problem, we show in this paper that reachability of continuous systems can also be verified through convex programming. Several convex programs for verifying safety and reachability, as well as other temporal properties such as eventuality, avoidance, and their combinations, are formulated. Some examples are provided to illustrate the application of the proposed methods. Finally, we exploit the convexity of our methods to derive a converse theorem for safety verification using barrier certificates.

Key words. temporal verification, safety verification, reachability analysis, barrier certificate, density function, convex programming, duality

AMS subject classifications. 93C10, 68Q60, 90C90

DOI. 10.1137/050645178

1. Introduction.

Consider a continuous-time dynamical system of the form

$$\dot{x}(t) = f(x(t)),$$

where $x(t)$ is the state of the system, taking its value in the set $\mathcal{X} \subseteq \mathbb{R}^n$. Also given are the set of possible initial states $\mathcal{X}_0 \subseteq \mathcal{X}$, the set of "bad" states $\mathcal{X}_u \subseteq \mathcal{X}$, and the set of "good" states $\mathcal{X}_r \subseteq \mathcal{X}$. In this paper, we will be concerned with methods for verifying or proving temporal properties of the system such as the following:

- *safety*: all trajectories of the system starting from $\mathcal{X}_0$ will never reach $\mathcal{X}_u$;
- *avoidance*: at least one trajectory of the system starting from $\mathcal{X}_0$ will never reach $\mathcal{X}_u$;
- *eventuality*: all trajectories of the system starting from $\mathcal{X}_0$ will reach $\mathcal{X}_r$ in finite time;
- *reachability*: at least one trajectory of the system starting from $\mathcal{X}_0$ will reach $\mathcal{X}_r$ in finite time.

They will be defined more precisely later in the paper. In addition, we will look at more complex temporal properties, which are the combinations of the above, and will also consider systems with uncertain time-varying disturbance inputs.

When the system under consideration is a discrete transition system, such as a finite automaton, the problem described above has been studied extensively in the

*Received by the editors November 15, 2005; accepted for publication (in revised form) September 26, 2006; published electronically June 29, 2007. Preliminary versions of this paper appeared in *Hybrid Systems: Computation and Control* 2005 and *Proceedings of the IFAC World Congress* 2005.

http://www.siam.org/journals/sicon/46-3/64517.html

†Control and Dynamical Systems, California Institute of Technology, Pasadena, CA 91125 (prajna@cds.caltech.edu). Current address: Credit Suisse, One Cabot Square, London E14 4QJ, United Kingdom.

‡Automatic Control LTH, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden (rantzer@control.lth.se).

computer science literature, and has applications, e.g., in the verification of correctness of computer protocols, algorithms, and software. See [8, 10, 12, 18]. The methods that have been proposed fall into two mainstream approaches: *model checking* [8] and *deductive verification* (or *theorem proving*) [18]. Model checking performs an exhaustive exploration of all possible system behaviors in a fully automated way, but is applicable only to finite state systems. Deductive verification, on the other hand, verifies system properties through formal deduction based on a set of inference rules. It is applicable to infinite state systems, but has a drawback in the sense that guidance from users is often needed in the process.

Uncountable state space and continuous dynamics are introduced when we consider applications in control, since they usually involve physical plants whose dynamics is governed by differential equations. Here the need for temporal verification arises as the complexity of the system increases, especially in safety-critical applications such as air traffic management [29], automated highway systems [11], and life support systems [9]. For such systems, exact verification cannot be performed through simulation, due to the infinite number of possibilities taken by the continuous state and also the uncertainties of the system.

The success of model checking techniques in verification of discrete, finite state transition systems has prompted the development of analogous approaches for continuous and hybrid systems. These approaches (see, e.g., [1, 2, 3, 5, 7, 15, 16, 30, 31]) require explicit computation of the states reachable from the initial set, which, for example, is performed by propagating the set of states. Unfortunately, although they allow us to compute an exact or nearly exact approximation of reachable sets, it is very difficult to perform such a computation due to the uncountability of the state space, especially when the system is nonlinear and uncertain. Note also that most of the existing literature focuses on the verification of safety property, although some of their techniques can be used to verify other temporal properties stated at the beginning of this paper.

In a different vein, a deductive method for safety verification that does not require explicit computation of the reachable sets, but instead is based on functions of states termed barrier certificates, has been recently proposed in a work by the first author [21]. The idea here is to study properties of the system without the need to compute the flow explicitly. Our conditions for safety can be stated as follows. Suppose that the vector field $f(x)$ is continuous and that there exists a continuously differentiable function $B : \mathbb{R}^n \to \mathbb{R}$ such that the inequalities

$$B(x) \leq 0 \quad \forall x \in \mathcal{X}_0, \tag{1.1}$$

$$B(x) > 0 \quad \forall x \in \mathcal{X}_u, \tag{1.2}$$

$$\frac{\partial B}{\partial x}(x) f(x) \leq 0 \quad \forall x \in \mathcal{X} \tag{1.3}$$

are satisfied. Then the safety property is verified, namely, there is no trajectory $x(t)$ of the system $\dot{x} = f(x)$ such that $x(0) \in \mathcal{X}_0$, $x(T) \in \mathcal{X}_u$ for some $T \geq 0$, and $x(t) \in \mathcal{X}$ for all $t \in [0, T]$. The function $B(x)$ here is called a *barrier certificate*. When $f(x)$ is polynomial and the sets $\mathcal{X}$, $\mathcal{X}_0$, $\mathcal{X}_u$ are semialgebraic, a polynomial barrier certificate $B(x)$ can be efficiently searched using sum of squares programming [22]—a convex optimization framework based on sum of squares decompositions of multivariate polynomials [20] and semidefinite programming [6]. Because of this, our method appears to be more scalable than many other methods. The method has also

been extended to handle hybrid, uncertain, and stochastic systems [21], and successful application to a NASA life support system, which is a nonlinear hybrid system with 6 discrete modes and 10 continuous state variables, has been reported [9]. To the best of our knowledge, all other verification methods that can handle nonlinear hybrid systems are practically limited to about 5 continuous state variables.

The above method is analogous to the Lyapunov method for stability analysis [14]. Contrary to stability analysis, however, no notion of equilibrium, stability, or convergence is required in temporal verification. For example, the system does not need to have an equilibrium, and also for the eventuality and reachability properties the system is not required to stay in $\mathcal{X}_r$ once the set is reached. Our method is also related to the viability theory [4], the smallest invariant set [13], and the invariant generation [26, 27, 28] approaches to safety verification. However, one of the distinctive features of our approach is that we use convex programming to verify properties of interest, which gives benefit in terms of *computation* and in terms of its inherent *duality structure*.

In the present paper, we use insights from the linear programming duality appearing in the discrete shortest path problem [19] and the concept of density function [23, 24] to formulate a convex program for proving reachability. In fact, not only safety and reachability, but also other temporal properties such as eventuality, avoidance, and their combinations can be verified through convex programming. Several convex programs for this will be formulated. Similar to before, when the description of the system is polynomial and the sets are semialgebraic, polynomial solutions to these programs can be searched using sum of squares programming. In addition to this, we will exploit strong duality to prove a converse theorem for safety verification using barrier certificates.

The outline of the paper is as follows. In section 2, we give an intuitive illustration of some main ideas by addressing the verification of a simple discrete transition system. The convex programs for verification of continuous-time systems are presented and proven in section 3. In section 4, some examples will be presented to illustrate the applications of the proposed method. A converse theorem for barrier certificates will be stated and proven in section 5, and we offer some conclusions in section 6.

2. Discrete example. Let us consider the verification of a simple discrete transition system, shown in Figure 2.1. The system has four states, labeled 1 through 4, and three transitions between states, represented by the directed edges in the graph. We assume that node 1 is the initial state and node 4 is the bad/unsafe state.

For verifying the safety property, conditions analogous to (1.1)–(1.3) that must be satisfied by a barrier certificate can be formulated. One way to find a barrier certificate which proves safety is by solving the linear program (LP)[1]

$$\max\ s^T B$$

$$\text{subject to } A^T B \leq 0,$$

where $B \triangleq \mathrm{col}(B_1, B_2, B_3, B_4) \in \mathbb{R}^4$ is the decision variable of the LP (i.e., the barrier

[1]Here we assume that there are only one initial state and one unsafe state. A generalization of this can be formulated by considering a bigger graph obtained by augmenting an extra "source" node and edges that connect it to all initial states, as well as an extra "sink" node and edges that connect all unsafe states to it.

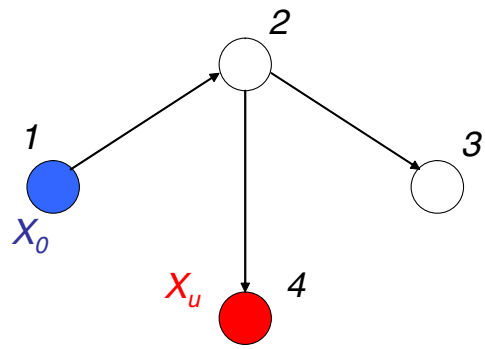


FIG. 2.1. *Verification of a simple discrete transition system. The nodes represent the states of the system, while the directed edges represent transitions between states. Node* 1 *is the initial state and node* 4 *is the unsafe state.*

certificate); A is the incidence matrix of the graph, in this case given by

$$A = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}^T;$$

and s is a 4×1 column vector whose ith entry is equal to 1 if the ith node is the unsafe state, and equal to -1 if the ith node is the initial state. This formulation is similar to the continuous case. Analogous to (1.3), we ask that $B_j \leq B_i$ if there is a directed edge from node i to node j. The objective function of the LP is just the difference between the values of B at the unsafe state and at the initial state. If there is a feasible solution to the above LP such that the objective function is strictly positive, then the safety property can be inferred; i.e., we prove that there is no path going from node 1 to node 4.

The dual of the above LP is as follows:

$$\begin{aligned} &\min\ 0 \\ &\text{subject to } A\rho = s, \\ &\qquad \rho \geq 0, \end{aligned}$$

where $\rho \triangleq \text{col}(\rho_{12}, \rho_{23}, \rho_{24}) \in \mathbb{R}^3$ is the dual decision variables, whose entries correspond to the edges in the graph. The dual decision variable ρ_{ij} can be interpreted as the transportation density from node i to node j. The equality constraints basically state that conservation of flows holds at each node, namely, that the total flow into a node is equal to the total flow out. In addition, the first and third equality constraints indicate that there exist a unit source at node 1, i.e., the initial state, and a unit sink at node 4, i.e., the unsafe state. This duality interpretation has been studied extensively in the past; see, e.g., [19] and the references therein.

The existence of a feasible solution to the dual LP implies the existence of a path from the initial state to the unsafe state. This can be shown using the facts that the flows are conserved and that there are a unit source and a unit sink at the initial state and unsafe state, respectively. Hence, solving the dual LP can be used for verifying reachability. As a matter of fact, we obtain a linear programming formulation of the shortest path problem if we also add the objective function $\sum \rho_{ij}$ to the dual LP. In

this case, the nonzero entries corresponding to any optimal vertex solution to the LP will indicate a shortest path from the initial node to the unsafe node [19].

This duality argument can also be used to prove that the existence of a barrier certificate is both sufficient and necessary for safety. For this, suppose that there exists no barrier certificate for the system, which is equivalent to the maximum objective value of the primal LP being equal to zero. This objective value is attained by, e.g., $B_i = 0$ for all i. The linear programming duality [6] implies that there exists a feasible solution to the dual LP, from which we can further conclude the existence of a path from the initial state to the unsafe state, as explained in the previous paragraph. In the continuous case, a strong duality argument will also be used to prove a converse theorem for barrier certificates later in this paper.

For the above example, the optimal objective value of the primal LP is equal to zero, and hence the safety property does not hold. The unique feasible solution to the dual LP is given by $\rho_{12} = 1$, $\rho_{23} = 0$, $\rho_{24} = 1$, which shows the path from node 1 to node 4. If the direction of the edge from node 2 to node 4 were reversed, for example, the optimal objective value of the corresponding primal LP would be ∞, and there would be no feasible solution to the dual LP.

Other properties of this discrete transition system such as eventuality and avoidance can also be verified by solving some appropriate LPs. We will not state them here, but instead we will now proceed to discuss the corresponding convex programs for continuous systems.

3. Continuous systems. We denote the space of m-times continuously differentiable functions mapping $X \subseteq \mathbb{R}^n$ to $\mathbb{R}^p$ by $C^m(X, \mathbb{R}^p)$. When $p = 1$, we will simply write $C^m(X)$, and for continuous functions ($m = 0$), we will omit the superscript. The solution $x(t)$ of $\dot{x} = f(x)$ starting from $x(0) = x_0$, if unique, is denoted by $\phi_t(x_0)$. For a set Z, we define $\phi_t(Z) \triangleq \{\phi_t(x) : x \in Z\}$.

The divergence of a vector field $f \in C^1(X, \mathbb{R}^n)$ is denoted by $\nabla \cdot f(x)$. Finally, let $\mathrm{cl}(X)$ denote the closure of a set X, and let ∂X denote the boundary of X.

The following version of Liouville's theorem (from [23]) will be used in the proofs of the main theorems.

LEMMA 3.1. *Let $f \in C^1(D, \mathbb{R}^n)$, where $D \subseteq \mathbb{R}^n$ is open, and let $\rho \in C^1(D, \mathbb{R})$ be integrable, i.e., $\int_D \rho(x)dx$ is finite. Consider the system $\dot{x} = f(x)$. For a measurable set Z, assume that $\phi_\tau(Z)$ is a subset of D for all τ between 0 and T. Then*

$$\int_{\phi_T(Z)} \rho(x)dx - \int_Z \rho(z)dz = \int_0^T \int_{\phi_\tau(Z)} \left[\nabla \cdot (f\rho)\right](x)dxd\tau. \tag{3.1}$$

3.1. Safety and reachability verification. At this point, we are ready to state and prove the first pair of convex programs that verify safety and reachability for continuous systems. The first convex program was proposed in [21] but will be repeated here for completeness.

THEOREM 3.2. *Consider the system $\dot{x} = f(x)$ with $f \in C(\mathbb{R}^n, \mathbb{R}^n)$. Let the sets $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{X}_0 \subseteq \mathcal{X}$, and $\mathcal{X}_u \subseteq \mathcal{X}$ be given. Suppose that there exists a function $B \in C^1(\mathbb{R}^n)$ satisfying*

$$B(x) \leq 0 \quad \forall x \in \mathcal{X}_0, \tag{3.2}$$

$$B(x) > 0 \quad \forall x \in \mathcal{X}_u, \tag{3.3}$$

$$\frac{\partial B}{\partial x}(x) f(x) \leq 0 \quad \forall x \in \mathcal{X}. \tag{3.4}$$

Then the safety property holds; i.e., there exists no trajectory $x(t)$ of the system such that $x(0) \in \mathcal{X}_0$, $x(T) \in \mathcal{X}_u$ for some $T \geq 0$, and $x(t) \in \mathcal{X}$ for all $t \in [0, T]$.

Proof. Our proof is by contradiction. Assume that there exists a barrier certificate $B(x)$ satisfying conditions (3.2)–(3.4), while at the same time the system is not safe; i.e., there exist a time instance $T \geq 0$ and an initial condition $x_0 \in \mathcal{X}_0$ such that a trajectory $x(t)$ of the system starting at $x(0) = x_0$ satisfies $x(t) \in \mathcal{X}$ for all $t \in [0, T]$ and $x(T) \in \mathcal{X}_u$. Condition (3.4) implies that the derivative of $B(x(t))$ with respect to time is nonpositive on the time interval $[0, T]$. A direct consequence of this (which, for example, can be shown using the mean value theorem) is that $B(x(T))$ must be less than or equal to $B(x(0))$, which is contradictory to (3.2)–(3.3). Thus the initial hypothesis is not correct: the safety property must hold. ∎

We will next present a convex program for verifying reachability. It can be viewed as a continuous-time version of the dual LP in section 2. The decision variable $\rho(x)$ in this convex program is termed *density function* and has an interpretation as the stationary density of a substrate that is generated and consumed in various parts of the state space, and that is transported according to the vector field of the system. It has been previously used in an almost global stability criterion in [23].

THEOREM 3.3. *Consider the system $\dot{x} = f(x)$ with $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Let the sets $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X}_0 \subseteq \mathcal{X}$, and $\mathcal{X}_r \subseteq \mathcal{X}$ be given. Assume that the sets are bounded and that $\mathcal{X}_0$ has a nonempty interior. If there exists a function $\rho \in C^1(\mathbb{R}^n)$ satisfying*

$$\int_{\mathcal{X}_0} \rho(x)dx \geq 0, \tag{3.5}$$

$$\rho(x) < 0 \quad \forall x \in \mathrm{cl}(\partial\mathcal{X} \setminus \partial\mathcal{X}_r), \tag{3.6}$$

$$\nabla \cdot (\rho f)(x) > 0 \quad \forall x \in \mathrm{cl}(\mathcal{X} \setminus \mathcal{X}_r), \tag{3.7}$$

then the reachability property holds; i.e., there exists a trajectory $x(t)$ of the system such that $x(0) \in \mathcal{X}_0$, $x(T) \in \mathcal{X}_r$ for some $T \geq 0$, and $x(t) \in \mathcal{X}$ for all $t \in [0, T]$.

Proof. Let $X \subseteq \mathcal{X}_0$ be an open set on which $\rho(x) \geq 0$. We will first prove that there must be an initial condition $x_0 \in X$ whose flow $\phi_t(x_0)$ leaves $\mathcal{X} \setminus \mathcal{X}_r$ in finite time. In fact, the set of all initial conditions in X whose flows do not leave $\mathcal{X} \setminus \mathcal{X}_r$ in finite time is a set of measure zero. To show this, let Y be an open neighborhood of $\mathcal{X} \setminus \mathcal{X}_r$ such that $\nabla \cdot (\rho f)(x) > 0$ on $\mathrm{cl}(Y)$. Now define

$$Z = \bigcap_{i=1,2,\ldots} \{x_0 \in X : \phi_t(x_0) \in Y \quad \forall t \in [0, i]\}.$$

The set Z is an intersection of countable open sets and hence is measurable. It contains all initial conditions in X for which the trajectories stay in Y for all $t \geq 0$. That Z is a set of measure zero can be shown using Lemma 3.1 as follows. Since $\phi_t(Z) \subset Y$, Y is bounded, and $\rho(x)$ is continuous, the left-hand side of

$$\int_{\phi_t(Z)} \rho(x)dx - \int_Z \rho(x)dx = \int_0^t \int_{\phi_\tau(Z)} \left[\nabla \cdot (f\rho)\right](x)dxd\tau$$

is bounded for all $t \geq 0$. Therefore, for the above equation to hold, we must have $\int_{\phi_\tau(Z)} \left[\nabla \cdot (f\rho)\right](x)dx \to 0$ as $\tau \to \infty$, or, equivalently, the measure of $\phi_\tau(Z)$ converges to zero as $\tau \to \infty$. Suppose now that Z has nonzero measure. We have a contradiction since $\lim_{t\to\infty} \int_{\phi_t(Z)} \rho(x)dx = 0$, whereas $\lim_{t\to\infty} \int_0^t \int_{\phi_\tau(Z)} \left[\nabla \cdot (f\rho)\right](x)dxd\tau +$

$\int_Z \rho(x)dx$ is strictly positive, as implied by (3.5) and (3.7). Using this argument, we conclude that Z has measure zero. Since $\mathcal{X} \setminus \mathcal{X}_r \subset Y$, it follows immediately that the set of all initial conditions in X whose flows stay in $\mathcal{X} \setminus \mathcal{X}_r$ for all time is a set of measure zero.

Now take any $x_0 \in X$ whose flow leaves $\mathcal{X} \setminus \mathcal{X}_r$ in finite time; we will show that such a flow must enter $\mathcal{X}_r$ before leaving $\mathcal{X}$. Suppose to the contrary that the flow $\phi_t(x_0)$ leaves $\mathcal{X}$ without entering $\mathcal{X}_r$ first. Let $T > 0$ be the "first" time instant $\phi_t(x_0)$ leaves $\mathcal{X}$. By this we mean that either $\phi_t(x_0) \in \mathcal{X} \setminus \mathcal{X}_r$ for all $t \in [0, T)$ and $\phi_T(x_0) \notin \mathcal{X}$, or $\phi_t(x_0) \in \mathcal{X} \setminus \mathcal{X}_r$ for all $t \in [0, T]$ and $\phi_{T+\epsilon}(x_0) \notin \mathcal{X}$ for any $\epsilon > 0$. From conditions (3.6)–(3.7), it follows that for a sufficiently small neighborhood U of x_0 we have

$$\rho(x) \geq 0 \quad \forall x \in U,$$

$$\rho(x) < 0 \quad \forall x \in \phi_T(U),$$

$$\nabla \cdot (\rho f)(x) > 0 \quad \forall x \in \phi_t(U), t \in [0, T].$$

Apply Lemma 3.1 again to obtain a contradiction. According to the above, the left-hand side of

$$\int_{\phi_T(U)} \rho(x)dx - \int_U \rho(x)dx = \int_0^T \int_{\phi_\tau(U)} [\nabla \cdot (f\rho)](x)dxd\tau$$

is negative, while the right-hand side is positive. Thus there is a contradiction, and we conclude that for $x(0) = x_0$ there must exist $T \geq 0$ such that $x(T) \in \mathcal{X}_r$ and $x(t) \in \mathcal{X}$ for all $t \in [0, T]$. □

Remark 3.4. Modulo the following modification on the assertion of Theorem 3.3, the conclusion will still hold even when the sets are not bounded. In particular, we need to add the condition that $\rho(x)$ is integrable on $\mathcal{X}$ (i.e., $\int_{\mathcal{X}} \rho(x)dx$ is finite) and replace (3.7) by

$$\nabla \cdot (\rho f)(x) \geq \epsilon \quad \forall x \in (\mathcal{X} \setminus \mathcal{X}_r)$$

for a positive number ϵ.

Notice that all the conditions presented in the above theorems (as well as in the theorems that will be presented later) form convex programming problems, as the sets of $B(x)$'s satisfying (3.2)–(3.4) or $\rho(x)$'s satisfying (3.5)–(3.7) are convex. This just follows from the definition of convexity. For example, if $B_1(x)$ and $B_2(x)$ satisfy (3.2)–(3.4), then for any $\alpha \in [0, 1]$, the function $\alpha B_1(x) + (1-\alpha)B_2(x)$ also satisfies the conditions. The convexity of the programs opens the possibility of computing $B(x)$ and $\rho(x)$ using convex optimization. For systems whose vector fields are polynomial and whose set descriptions are semialgebraic (i.e., described by polynomial equalities and inequalities), a computational method called sum of squares programming can be utilized if we use a polynomial parameterization for $B(x)$ or $\rho(x)$. The method is based on the sum of squares decomposition of multivariate polynomials [20] and semidefinite programming [6]. Software tools [22] are helpful for this purpose. See [21] for details.

When we set $\mathcal{X}_u = \mathcal{X}_r$, the convex programs in Theorems 3.2 and 3.3 form a pair of *weak alternatives*: at most one of them can be feasible. Nevertheless, strictly speaking it should be noted that these convex programs are not pairs of *Lagrange dual* problems [6] in the sense of convex optimization. We deliberately do not use Lagrange dual

problems to avoid computational problems when we postulate $B(x)$ or $\rho(x)$ as polynomials. For example, the Lagrange dual problem of the safety test in Theorem 3.2 will require $\nabla \cdot (\rho f)(x)$ to be zero on $\mathcal{X} \setminus (\mathcal{X}_0 \cup \mathcal{X}_u)$ (cf. section 5). Although useful for theoretical purposes, this will hinder the computation of $\rho(x)$ through polynomial parameterization and sum of squares programming. In this regard, some interesting future directions would be to see if a pair of Lagrange dual problems can be formulated so that both problems can be solved using sum of squares programming, or, more importantly, to see if the dual infeasibility certificate of one convex program can be interpreted directly as a feasible solution to the dual convex program.

3.2. Eventuality and avoidance verification. We will now present two other convex programs for verifying the eventuality and avoidance properties. Analogous to what we have in the previous subsection, when $\mathcal{X}_u = \mathcal{X}_r$ these programs form a pair of weak alternatives.

THEOREM 3.5. *Consider the system $\dot{x} = f(x)$ with $f \in C(\mathbb{R}^n, \mathbb{R}^n)$. Let $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X}_0 \subseteq \mathcal{X}$, and $\mathcal{X}_r \subseteq \mathcal{X}$ be bounded sets. If there exists a function $B \in C^1(\mathbb{R}^n)$ satisfying*

$$B(x) \leq 0 \quad \forall x \in \mathcal{X}_0, \tag{3.8}$$

$$B(x) > 0 \quad \forall x \in \mathrm{cl}(\partial\mathcal{X} \setminus \partial\mathcal{X}_r), \tag{3.9}$$

$$\frac{\partial B}{\partial x}(x) f(x) < 0 \quad \forall x \in \mathrm{cl}(\mathcal{X} \setminus \mathcal{X}_r), \tag{3.10}$$

then the eventuality property holds; i.e., for all initial conditions $x_0 \in \mathcal{X}_0$, the trajectory $x(t)$ of the system starting at $x(0) = x_0$ satisfies $x(T) \in \mathcal{X}_r$ and $x(t) \in \mathcal{X}$ for all $t \in [0, T]$ for some $T \geq 0$.

Proof. Consider any point $x_0 \in \mathcal{X}_0$, for which $B(x_0) \leq 0$, and let $x(t)$ be a trajectory of the system starting at $x(0) = x_0$. The trajectory $x(t)$ must leave $\mathcal{X} \setminus \mathcal{X}_r$ in finite time, since the derivative inequality (3.10) holds and $B(x)$ is bounded below on $\mathcal{X}$. Now, suppose that $x(t)$ leaves $\mathcal{X}$ without entering $\mathcal{X}_r$ first, and consider the "first" time instant $t = T$ at which it happens. Similar to the proof of Theorem 3.3, by this we mean that either $x(t) \in \mathcal{X} \setminus \mathcal{X}_r$ for all $t \in [0, T)$ and $x(T) \notin \mathcal{X}$, or $x(t) \in \mathcal{X} \setminus \mathcal{X}_r$ for all $t \in [0, T]$ and $x(T + \epsilon) \notin \mathcal{X}$ for any $\epsilon > 0$. From (3.10) and $B(x_0) \leq 0$, it follows that $B(x(T))$ is nonpositive, which is contradictory to (3.9). Thus we conclude that for any trajectory $x(t)$ starting at $x(0) = x_0$ there must exist $T \geq 0$ such that $x(T) \in \mathcal{X}_r$ and $x(t) \in \mathcal{X}$ for all $t \in [0, T]$. Since x_0 is an arbitrary point in $\mathcal{X}_0$, the conclusion of the theorem follows. ∎

Remark 3.6. Similarly to before, with some modifications to the assertion of the theorem, the conclusion of Theorem 3.5 will still hold even when the sets are not bounded. In particular, we need to add the condition that $B(x)$ is bounded below on $\mathcal{X}$ and replace (3.10) by

$$\frac{\partial B}{\partial x}(x) f(x) \leq -\epsilon \quad \forall x \in (\mathcal{X} \setminus \mathcal{X}_r)$$

for a positive number ϵ.

THEOREM 3.7. *Consider the system $\dot{x} = f(x)$ with $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Let $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{X}_0 \subseteq \mathcal{X}$, and $\mathcal{X}_u \subseteq \mathcal{X}$ be some given sets, with $\mathcal{X}_0$ having a nonempty interior. If*

there exist an open set $\tilde{\mathcal{X}}$ *and a function* $\rho \in C^1(\mathbb{R}^n)$ *such that* $\mathcal{X} \subseteq \tilde{\mathcal{X}}$ *and*

$$\int_{\mathcal{X}_0} \rho(x)dx \geq 0, \tag{3.11}$$

$$\rho(x) < 0 \quad \forall x \in \mathcal{X}_u, \tag{3.12}$$

$$\nabla \cdot (\rho f)(x) \geq 0 \quad \forall x \in \tilde{\mathcal{X}}, \tag{3.13}$$

then the avoidance property holds; i.e., for some initial condition $x_0 \in \mathcal{X}_0$*, there exists no* $T \geq 0$ *such that trajectory* $x(t)$ *of the system starting at* $x(0) = x_0$ *satisfies* $x(T) \in \mathcal{X}_u$ *and* $x(t) \in \mathcal{X}$ *for all* $t \in [0, T]$.

Proof. From (3.11), it follows that there exists an open set $X \subseteq \mathcal{X}_0$ on which $\rho(x) \geq 0$. Take any x_0 in X—we will show that the trajectory starting from this point will never reach $\mathcal{X}_u$. Suppose to the contrary that there exists a $T \geq 0$ such that $\phi_T(x_0) \in \mathcal{X}_u$ and $\phi_t(x_0) \in \mathcal{X}$ for $t \in [0, T]$. Then it follows from (3.12)–(3.13) that for a sufficiently small neighborhood Z of x_0 we have

$$\rho(x) \geq 0 \quad \forall x \in Z,$$

$$\rho(x) < 0 \quad \forall x \in \phi_T(Z),$$

$$\nabla \cdot (\rho f)(x) \geq 0 \quad \forall x \in \phi_t(Z), t \in [0, T].$$

Now apply Lemma 3.1 to obtain a contradiction. Use a bounded but sufficiently large $D \subset \mathbb{R}^n$ such that $\phi_t(Z) \subset D$ for all $t \in [0, T]$; then $\rho(x)$ is integrable on D. According to the above, the left-hand side of

$$\int_{\phi_T(Z)} \rho(x)dx - \int_Z \rho(x)dx = \int_0^T \int_{\phi_\tau(Z)} \left[\nabla \cdot (f\rho)\right](x)dxd\tau$$

is negative and the right-hand side is nonnegative. Hence there is a contradiction and the proof is complete. ∎

In applications where the system has stable equilibrium points, it is often convenient to exclude a neighborhood of the equilibria from the region where the divergence inequality (3.13) must be satisfied, since the inequality is otherwise impossible to satisfy without a singularity in $\rho(x)$. This does not make the conclusion of the theorem weaker, as long as the excluded set does not intersect $\mathcal{X}_u$ and is entirely surrounded by a region of positive $\rho(x)$.

Similarly, the Lie derivative inequality (3.10) is impossible to satisfy when the system has equilibrium points in $\mathcal{X} \setminus \mathcal{X}_r$. In this case, a neighborhood of the equilibria should also be excluded from the region where the inequality is to be satisfied. The conclusion of the theorem is still valid as long as the excluded set is entirely surrounded by a region of positive $B(x)$.

3.3. Some extensions. Whereas the convex programs for safety and reachability as well as eventuality and avoidance are related since they form pairs of weak alternatives, the safety property is also related to avoidance, and eventuality to reachability, via replacing the universal quantifier with an existential quantifier. As a consequence, reachability and avoidance verification can also be performed using the

barrier certificate $B(x)$. The conditions are similar to those in Theorems 3.5 and 3.2, respectively, but with conditions (3.8) and (3.2) replaced by

$$\int_{\mathcal{X}_0} B(x)dx \leq 0,$$

where we also ask that $\mathcal{X}_0$ has a nonempty interior. The proof is similar to the proofs of Theorems 3.5 and 3.2, noting that $B(x)$ will then be less than or equal to zero in some open set contained in $\mathcal{X}_0$.

A modification of the convex program involving $\rho(x)$ in Theorem 3.7 can also be used to verify the safety property. For this, we need to replace (3.11) by

$$\rho(x) \geq 0 \quad \forall x \in \tilde{\mathcal{X}}_0,$$

where $\tilde{\mathcal{X}}_0$ is an open set containing $\mathcal{X}_0$. Note that in this case $\mathcal{X}_0$ no longer needs to have a nonempty interior.

On the other hand, an analogous modification of Theorem 3.3 can only be used to verify the eventuality property in the *weak* sense: that *almost all* trajectories of the system starting from $\mathcal{X}_0$ will reach $\mathcal{X}_r$ in finite time. This is stated in the corollary below.

COROLLARY 3.8. *Consider the system $\dot{x} = f(x)$ with $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Let the sets $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X}_0 \subseteq \mathcal{X}$, and $\mathcal{X}_r \subseteq \mathcal{X}$ be given. Assume that the sets are bounded, and let $\tilde{\mathcal{X}}_0$ be an open set containing $\mathcal{X}_0$. If there exists a function $\rho \in C^1(\mathbb{R}^n)$ satisfying*

$$\rho(x) \geq 0 \quad \forall x \in \tilde{\mathcal{X}}_0, \tag{3.14}$$

$$\rho(x) < 0 \quad \forall x \in \mathrm{cl}(\partial\mathcal{X} \setminus \partial\mathcal{X}_r), \tag{3.15}$$

$$\nabla \cdot (\rho f)(x) > 0 \quad \forall x \in \mathrm{cl}(\mathcal{X} \setminus \mathcal{X}_r), \tag{3.16}$$

then the weak eventuality property holds; i.e., for almost all[2] *initial conditions $x_0 \in \mathcal{X}_0$, there exists $T \geq 0$ such that the trajectory $x(t)$ of the system starting at $x(0) = x_0$ satisfies $x(T) \in \mathcal{X}_r$ and $x(t) \in \mathcal{X}$ for all $t \in [0, T]$.*

Proof. Using an argument similar to the proof of Theorem 3.3, it can be shown that for almost all initial conditions $x_0 \in \tilde{\mathcal{X}}_0$, there exists $T \geq 0$ such that the trajectory $x(t)$ of the system starting at $x(0) = x_0$ satisfies $x(T) \in \mathcal{X}_r$ and $x(t) \in \mathcal{X}$ for all $t \in [0, T]$. Since $\mathcal{X}_0 \subseteq \tilde{\mathcal{X}}_0$, the corollary follows. □

Example 3.9. To show that the weak eventuality property mentioned above cannot in general be strengthened to eventuality, consider the system $\dot{x} = x$, with $\mathcal{X} = (-5, 5) \subset \mathbb{R}$, $\mathcal{X}_0 = (-1, 1)$, $\mathcal{X}_r = (-5, -4) \cup (4, 5)$. The function $\rho(x) = 1$ satisfies all the conditions that guarantee weak eventuality; hence almost all trajectories starting from $\mathcal{X}_0$ will reach $\mathcal{X}_r$ in finite time. The only exception in this case is the trajectory $x(t) = 0$.

Let us now consider the verification of a system with disturbance input $\dot{x} = f(x, d)$, where $f \in C(\mathbb{R}^{n+m}, \mathbb{R}^n)$; the disturbance signal $d(t)$ is assumed to be piecewise continuous, bounded on any finite time interval, and take its value in a set $\mathcal{D} \subseteq \mathbb{R}^m$. Then solving the convex program in Theorem 3.2 with the Lie derivative inequality (3.4) replaced by

$$\frac{\partial B}{\partial x}(x) f(x, d) \leq 0 \quad \forall (x, d) \in \mathcal{X} \times \mathcal{D}$$

[2]This is in the sense that the set of initial conditions which do not satisfy the property is a set of measure zero.

will prove safety under all possible disturbances $d(t)$. Also, solving the convex program in Theorem 3.5 with the Lie derivative inequality (3.10) replaced by

$$\frac{\partial B}{\partial x}(x)f(x,d) \leq -\epsilon \quad \forall (x,d) \in (\mathcal{X} \setminus \mathcal{X}_r) \times \mathcal{D}$$

for some positive ϵ will prove eventuality under all possible disturbances $d(t)$. Similar adaptations can be applied to the convex programs that verify reachability and avoidance using $B(x)$.

At present, it is unclear how a similar worst-case analysis for systems with time-varying disturbance can be formulated using $\rho(x)$. However, as pointed out in [23], the density function $\rho(x)$ seems to have a better convexity property that is more beneficial for controller design. For a system $\dot{x} = f(x) + g(x)u(x)$, where $u(x)$ is the control input (assumed to be in a state feedback form), the inequalities (3.5)–(3.6) and

$$\nabla \cdot [\rho(f + ug)](x) > 0 \quad \forall x \in \mathrm{cl}(\mathcal{X} \setminus \mathcal{X}_r)$$

(and similarly for (3.11)–(3.13)) are certainly convex conditions on the pair $(\rho, \rho u)$. It is therefore natural to introduce $\psi = \rho u$ as a search variable and use convex optimization to find a feasible pair (ρ, ψ), then recover the control law as $u(x) = \psi(x)/\rho(x)$. Some results along this direction are available in [25].

While one may argue that the reachability and avoidance properties can be shown by running a numerical simulation of $\dot{x} = f(x)$ starting from a properly chosen $x_0 \in \mathcal{X}_0$, the merit of the convex programming tests presented before is twofold. First, a solution to the convex programs for reachability or avoidance will automatically indicate a set from which all points (or almost all points) can be chosen as the initial state. Second, the use of these convex programs allows us to also consider the verification of systems with disturbance (which obviously cannot be performed using simulation), or even the controller design problem, as we have seen above.

3.4. Other temporal properties. It is clear that the convex programs in the previous subsections can also be extended to prove combined temporal properties such as reachability–safety:

> there exists a trajectory $x(t)$ such that $x(0) \in \mathcal{X}_0$, $x(T) \in \mathcal{X}_r$ for some $T \geq 0$, and $x(t) \notin \mathcal{X}_u$, $x(t) \in \mathcal{X}$ for all $t \in [0, T]$;

and eventuality–safety[3] (or weak eventuality–safety):

> for all (or almost all) initial states $x_0 \in \mathcal{X}_0$, the trajectory $x(t)$ starting at $x(0) = x_0$ will satisfy $x(T) \in \mathcal{X}_r$ for some $T \geq 0$ and $x(t) \notin \mathcal{X}_u$, $x(t) \in \mathcal{X}$ for all $t \in [0, T]$.

Note that for the above temporal specifications, the system can reach $\mathcal{X}_u$ *after* it reaches $\mathcal{X}_r$ first.

For instance, convex programs for verifying the eventuality–safety and weak eventuality–safety properties are stated in the following corollaries.

COROLLARY 3.10. *Consider the system $\dot{x} = f(x)$ with $f \in C(\mathbb{R}^n, \mathbb{R}^n)$ and let $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X}_0 \subseteq \mathcal{X}$, $\mathcal{X}_u \subseteq \mathcal{X}$, $\mathcal{X}_r \subseteq \mathcal{X}$ be bounded. Suppose that there exists a function*

[3] In linear temporal logic (LTL), for example, this property corresponds to the "until" operator.

$B \in C^1(\mathbb{R}^n)$ satisfying

$$B(x) \le 0 \quad \forall x \in \mathcal{X}_0, \tag{3.17}$$

$$B(x) > 0 \quad \forall x \in \mathrm{cl}(\partial\mathcal{X} \setminus \partial\mathcal{X}_r) \cup \mathcal{X}_u, \tag{3.18}$$

$$\frac{\partial B}{\partial x}(x) f(x) < 0 \quad \forall x \in \mathrm{cl}(\mathcal{X} \setminus \mathcal{X}_r). \tag{3.19}$$

Then the eventuality–safety property holds; i.e., for all initial states $x_0 \in \mathcal{X}_0$, the trajectory $x(t)$ starting at $x(0) = x_0$ will satisfy $x(T) \in \mathcal{X}_r$ for some $T \ge 0$ and $x(t) \notin \mathcal{X}_u$, $x(t) \in \mathcal{X}$ for all $t \in [0, T]$.

COROLLARY 3.11. *Consider the system $\dot{x} = f(x)$ with $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ and let $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X}_0 \subseteq \mathcal{X}$, $\mathcal{X}_u \subseteq \mathcal{X}$, $\mathcal{X}_r \subseteq \mathcal{X}$ be bounded. If there exist an open set $\tilde{\mathcal{X}}_0$ containing $\mathcal{X}_0$ and a function $\rho \in C^1(\mathbb{R}^n)$ satisfying*

$$\rho(x) \ge 0 \quad \forall x \in \tilde{\mathcal{X}}_0, \tag{3.20}$$

$$\rho(x) < 0 \quad \forall x \in \mathrm{cl}(\partial\mathcal{X} \setminus \partial\mathcal{X}_r) \cup \mathcal{X}_u, \tag{3.21}$$

$$\nabla \cdot (\rho f)(x) > 0 \quad \forall x \in \mathrm{cl}(\mathcal{X} \setminus \mathcal{X}_r), \tag{3.22}$$

then the weak eventuality–safety property holds; i.e., for almost all initial states $x_0 \in \mathcal{X}_0$, the trajectory $x(t)$ starting at $x(0) = x_0$ will satisfy $x(T) \in \mathcal{X}_r$ for some $T \ge 0$ and $x(t) \notin \mathcal{X}_u$, $x(t) \in \mathcal{X}$ for all $t \in [0, T]$. In this case, the safety property holds also for trajectories that do not reach $\mathcal{X}_r$ in finite time.

4. Examples. We will now consider some examples to illustrate the application of the proposed methods. The MATLAB m-files for solving these examples can be found at http://www.cds.caltech.edu/∼prajna/files/PraR06.

4.1. Successive safety and reachability refinements. Consider the two-dimensional system

$$\dot{x}_1 = x_2,$$

$$\dot{x}_2 = -x_1 + \frac{1}{3}x_1^3 - x_2,$$

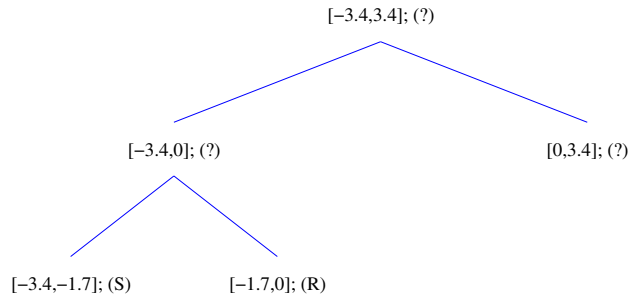
and let the set of states be $\mathcal{X} = [-3.5, 3.5] \times [-3.5, 3.5] \subset \mathbb{R}^2$. Furthermore, define

$$\mathcal{X}_0 = [-3.4, 3.4] \times [3.35, 3.45], \qquad \mathcal{X}_2 = [-3.5, 3.5] \times \{-3.5\},$$

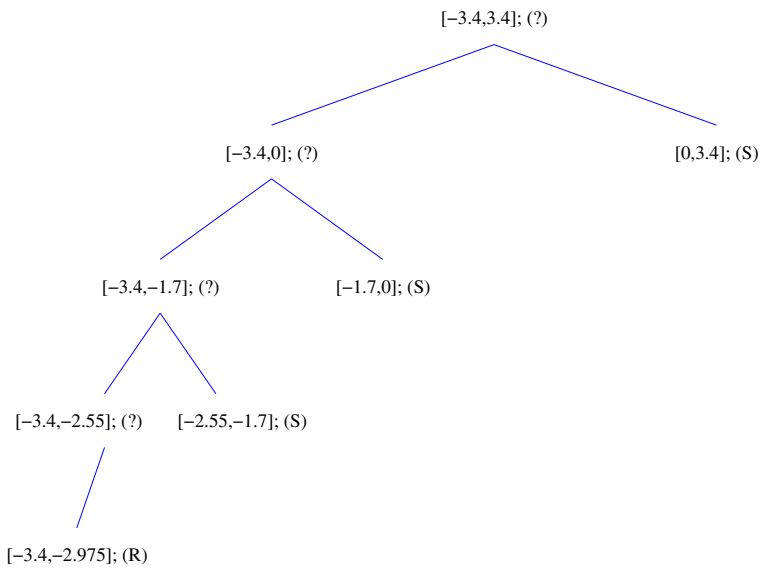
$$\mathcal{X}_1 = \{3.5\} \times [-3.5, 3.5], \qquad \mathcal{X}_3 = \{-3.5\} \times [-3.5, 3.5].$$

In this example, we will investigate the reachability of $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{X}_3$ from $\mathcal{X}_0$. Such facet-to-facet analysis is encountered when constructing a discrete abstraction of continuous or hybrid systems, or when analyzing a counterexample found during the verification of such an abstraction [1].

The convex programs in Theorems 3.2 and 3.3 will be used for our analysis. Since the vector field is polynomial and the sets are semialgebraic, we use polynomial parameterization for $B(x)$ and $\rho(x)$, and then utilize sum of squares programming to compute them. A degree bound equal to 8 is imposed on $B(x)$ and $\rho(x)$. Because of this, we might not be able to find a single $B(x)$ or $\rho(x)$ that proves safety or



(a) $\mathcal{X}_0 \to \mathcal{X}_1$



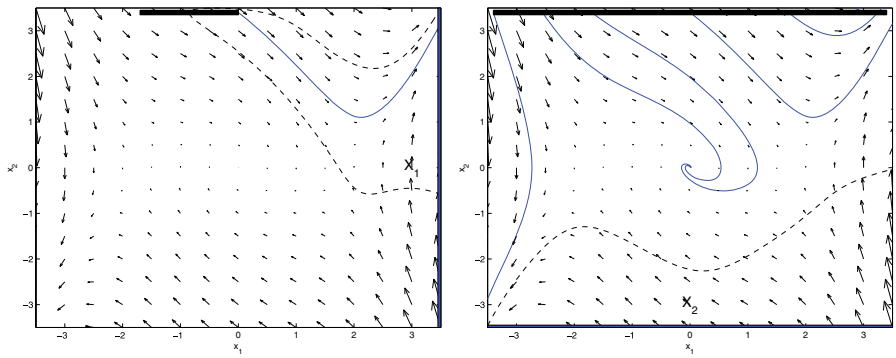
(b) $\mathcal{X}_0 \to \mathcal{X}_3$

FIG. 4.1. *Proving the reachability of $\mathcal{X}_1$ and $\mathcal{X}_3$ from $\mathcal{X}_0$ in the example of section* 4.1. *At each node we indicate the range of x_1 in $\mathcal{X}_0$ for which safety and reachability are tested. If neither is verified (denoted by ?), then the x_1-interval is divided into two and the tests are applied to the smaller sets. The annotation S (respectively, R) indicates that $B(x)$ (respectively, $\rho(x)$) is found. Breadth-first search starting from the leftmost branch is used. The verification of $\mathcal{X}_0 \nrightarrow \mathcal{X}_2$ terminates at the top node, since a barrier certificate $B(x)$ can be found directly.*

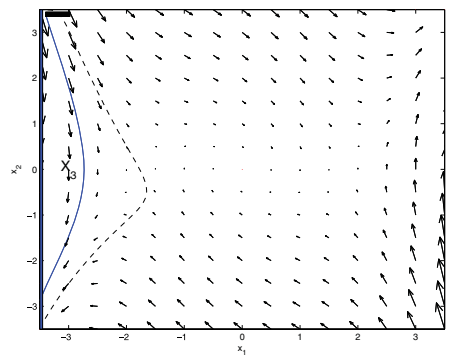
reachability for the whole $\mathcal{X}_0$. If neither $B(x)$ nor $\rho(x)$ can be found, we divide the interval of x_1 in $\mathcal{X}_0$ into two parts and apply the tests again to the smaller sets. A set is pruned if $B(x)$ is found, and this process is repeated until a $\rho(x)$ is found or the whole $\mathcal{X}_0$ is proven safe.

The result is as follows.

- We prove that the set $\mathcal{X}_1$ is reachable from $\mathcal{X}_0$. The verification progress is shown in Figure 4.1(a).



(a) $\mathcal{X}_0 \to \mathcal{X}_1$ (b) $\mathcal{X}_0 \nrightarrow \mathcal{X}_2$



(c) $\mathcal{X}_0 \to \mathcal{X}_3$

FIG. 4.2. *Possible transitions from $\mathcal{X}_0$ to $\mathcal{X}_1$, $\mathcal{X}_2$, and $\mathcal{X}_3$ in the example of section* 4.1. *In* (a) *and* (c), *dashed curves are the zero level sets of $\rho(x)$'s that certify reachability. In* (b), *the dashed curve is the zero level set of $B(x)$ that certifies safety; trajectories starting from $\mathcal{X}_0$ cannot cross this level set to reach $\mathcal{X}_2$. Thick solid lines at the top of the figures are the initial sets for which the certificates are computed. Some trajectories of the system are depicted by solid curves.*

- It can be proven directly that $\mathcal{X}_2$ is not reachable from $\mathcal{X}_0$.
- It is proven that the set $\mathcal{X}_3$ is reachable from $\mathcal{X}_0$. See Figure 4.1(b).

For proofs of the corresponding reachability and safety, see Figure 4.2.

Obviously, the above bisection algorithm is just a simple, straightforward approach to refine and prune the initial set, and other algorithms that are more efficient can be proposed in the future.

4.2. Van der Pol oscillator. Consider the van der Pol oscillator with disturbance input:

$$\dot{x}_1 = x_2,$$

$$\dot{x}_2 = x_2(1 - x_1^2) - x_1 + d,$$

where d is the disturbance input, taking its value in $\mathcal{D} = [-0.25, 0.25] \subset \mathbb{R}$. Let $\mathcal{X} = \{x \in \mathbb{R}^2 : 0.5 \le \|x\|_2 \le 5\}$. In addition, let

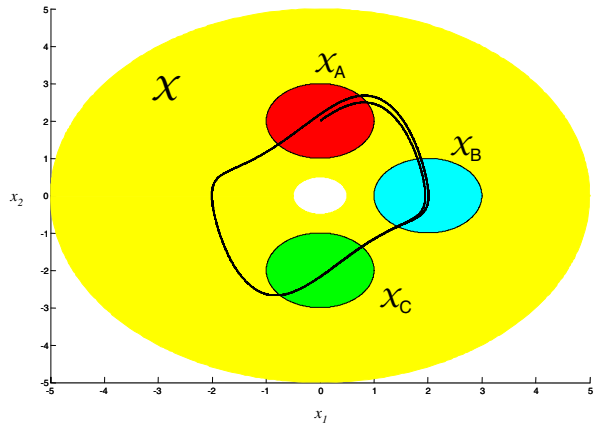


FIG. 4.3. *Verifying temporal properties of the van der Pol oscillator with disturbance. It is to be verified that under all possible disturbance input, if the system starts in $\mathcal{X}_A$, then both $\mathcal{X}_B$ and $\mathcal{X}_C$ are reached in finite time, but $\mathcal{X}_C$ will not be reached before the system reaches $\mathcal{X}_B$. The nominal trajectory of the system (i.e., for $d = 0$) starting at $x = (0, 2)$ is depicted by the solid curve.*

$$\mathcal{X}_A = \{x \in \mathbb{R}^2 : (x_1)^2 + (x_2 - 2)^2 \leq 1\},$$

$$\mathcal{X}_B = \{x \in \mathbb{R}^2 : (x_1 - 2)^2 + (x_2)^2 \leq 1\},$$

$$\mathcal{X}_C = \{x \in \mathbb{R}^2 : (x_1)^2 + (x_2 + 2)^2 \leq 1\}.$$

These sets are depicted in Figure 4.3, where a nominal trajectory of the system starting at $x = (0, 2)$ is also shown. Our objective in this example is to verify that under all possible piecewise continuous and bounded disturbances $d(t)$, if the system starts in $\mathcal{X}_A$, then both $\mathcal{X}_B$ and $\mathcal{X}_C$ are reached in finite time, but $\mathcal{X}_C$ will not be reached before the system reaches $\mathcal{X}_B$.

To verify this temporal specification, we will search for two barrier certificates $B_1(x)$ and $B_2(x)$ satisfying the following conditions:

$$\begin{cases} B_1(x) \leq 0 & \forall x \in \mathcal{X}_A, \\ B_1(x) > 0 & \forall x \in \partial\mathcal{X} \cup \mathcal{X}_C, \\ \frac{\partial B_1}{\partial x} f(x, d) \leq -\epsilon & \forall (x, d) \in (\mathcal{X} \setminus \mathcal{X}_B) \times \mathcal{D}, \end{cases}$$

$$\begin{cases} B_2(x) \leq 0 & \forall x \in \mathcal{X}_A, \\ B_2(x) > 0 & \forall x \in \partial\mathcal{X}, \\ \frac{\partial B_2}{\partial x} f(x, d) \leq -\epsilon & \forall x \in (\mathcal{X} \setminus \mathcal{X}_C) \times \mathcal{D} \end{cases}$$

for some positive ϵ. Using sum of squares programming, polynomials $B_1(x)$ and $B_2(x)$ of degree 10 can be found, and thus the temporal specification is verified.

5. A converse theorem. In this section, we will prove a converse theorem for safety verification using barrier certificates by exploiting the convexity of the problem formulation. The main result of the section can be stated as follows.

THEOREM 5.1. *Consider the system $\dot{x} = f(x)$ with $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Let $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X}_0 \subseteq \mathcal{X}$, $\mathcal{X}_u \subseteq \mathcal{X}$ be compact sets, and suppose that there exists a function $\tilde{B} \in C^1(\mathbb{R}^n)$ such that $\frac{\partial \tilde{B}}{\partial x}(x)f(x) < 0$ for all $x \in \mathcal{X}$. Then there exists a function $B \in C^1(\mathbb{R}^n)$ that satisfies*

$$B(x) \le 0 \quad \forall x \in \mathcal{X}_0, \tag{5.1}$$

$$B(x) > 0 \quad \forall x \in \mathcal{X}_u, \tag{5.2}$$

$$\frac{\partial B}{\partial x}(x)f(x) \le 0 \quad \forall x \in \mathcal{X} \tag{5.3}$$

if and only if the safety property holds, i.e., if there exists no trajectory $x(t)$ of the system such that $x(0) \in \mathcal{X}_0$, $x(T) \in \mathcal{X}_u$ for some $T \ge 0$, and $x(t) \in \mathcal{X}$ for all $t \in [0, T]$.

Notice that in the theorem we have used a seemingly strong assumption that there exists a function $\tilde{B} \in C^1(\mathbb{R}^n)$ such that $\frac{\partial \tilde{B}}{\partial x}(x)f(x) < 0$ for all $x \in \mathcal{X}$. Later in the section we will show that in many cases of interest the existence of such $\tilde{B}(x)$ is actually guaranteed.

Our proof of the converse statement in Theorem 5.1 consists of two parts, given in Lemmas 5.2 and 5.4 below. In the first lemma, we use the Hahn–Banach theorem to show that the nonexistence of a $B(x)$ satisfying the conditions in Theorem 5.1 implies the existence of measures ψ_0, ψ_u, ρ satisfying some appropriate conditions. Then in Lemma 5.4 we show that the existence of such ψ_0, ψ_u, ρ actually implies that there exists an unsafe trajectory of the system.

LEMMA 5.2. *Let $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, and let $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X}_0 \subseteq \mathcal{X}$, $\mathcal{X}_u \subseteq \mathcal{X}$ be compact sets. Suppose there exists a function $\tilde{B} \in C^1(\mathbb{R}^n)$ such that $\frac{\partial \tilde{B}}{\partial x}(x)f(x) < 0$ for all $x \in \mathcal{X}$. Then there exists no $B \in C^1(\mathbb{R}^n)$ satisfying (5.1)–(5.3) only if there are measures of bounded variation ψ_0, ψ_u, ρ (each defined on $\mathbb{R}^n$) such that ψ_0, ψ_u, ρ are nonnegative on $\mathbb{R}^n$ and equal to zero outside $\mathcal{X}_0$, $\mathcal{X}_u$, and $\mathcal{X}$, respectively; and*

$$\int_{\mathcal{X}_0} d\psi_0 = 1,$$

$$\int_{\mathcal{X}_u} d\psi_u = 1,$$

$$\nabla \cdot (\rho f) = \psi_0 - \psi_u,$$

where $\nabla \cdot (\rho f)$ is interpreted as a distributional derivative.

Proof. Let us consider the convex optimization problem

$$\begin{aligned} &\sup\ B_u - B_0 \\ &\text{subject to } B(x) - B_0 \le 0 \quad \forall x \in \mathcal{X}_0, \\ &\qquad B(x) - B_u \ge 0 \quad \forall x \in \mathcal{X}_u, \\ &\qquad \frac{\partial B}{\partial x}(x)f(x) \le 0 \quad \forall x \in \mathcal{X}, \end{aligned}$$

with the supremum denoted by γ, and taken over all $B_0 \in \mathbb{R}$, $B_u \in \mathbb{R}$, and $B \in C^1(\mathbb{R}^n)$. Since $B_0 = 0$, $B_u = 0$, and $B(x) = 0$ satisfy the constraint, γ must be greater

than or equal to zero. In addition, since the objective function and the constraints are all linear, the value of γ is either zero or ∞. There exists no $B \in C^1(\mathbb{R}^n)$ satisfying (5.1)–(5.3) if and only if the value of γ is equal to zero.

Now suppose that $\gamma = 0$. Let $\mathcal{K} = \mathbb{R} \times (C(\mathcal{X}))^3$, $\mathcal{B} = \mathbb{R}^2 \times C_0^1(\mathbb{R}^n)$, and define $\mathcal{K}_1$, $\mathcal{K}_2$ as follows:

$$\mathcal{K}_1 = \Bigg\{ (z, h_0, h_u, h) \in \mathcal{K} : h_0 = B_0 - B,\ h_u = B - B_u,\ h = -\frac{\partial B}{\partial x} f \text{ on } \mathcal{X};$$
$$z = B_u - B_0;\ \text{ and } (B_0, B_u, B) \in \mathcal{B} \Bigg\},$$

$$\mathcal{K}_2 = \{(z, h_0, h_u, h) \in \mathcal{K} : z \geq 0,\ h_0 \geq 0 \text{ on } \mathcal{X}_0,\ h_u \geq 0 \text{ on } \mathcal{X}_u,\ h \geq 0 \text{ on } \mathcal{X}\}.$$

Then both $\mathcal{K}_1$ and $\mathcal{K}_2$ are convex sets, and $\mathcal{K}_2$ has a nonempty interior in $\mathcal{K}$. Furthermore, since $\gamma = 0$, it follows that the first component in $\mathcal{K}_1$ is less than or equal to zero when the second, third, and fourth components are greater than or equal to zero, and therefore $\mathcal{K}_1 \cap \operatorname{int}(\mathcal{K}_2) = \emptyset$. Now, by the Hahn–Banach theorem [17], there exists a nonzero $k^* = (a, \tilde{\psi}_0, \tilde{\psi}_u, \tilde{\rho}) \in \mathcal{K}^* = \mathbb{R} \times (C(\mathcal{X})^*)^3$ such that

$$\sup_{k_1 \in \mathcal{K}_1} \langle k^*, k_1 \rangle \leq \inf_{k_2 \in \mathcal{K}_2} \langle k^*, k_2 \rangle, \tag{5.4}$$

where $C(\mathcal{X})^*$ in this case is the set of measures on $\mathcal{X}$ with bounded variation. The right-hand side of the inequality can be expanded as follows:

$$\inf_{k_2 \in \mathcal{K}_2} \langle k^*, k_2 \rangle = \inf_{(z, h_0, h_u, h) \in \mathcal{K}_2} az + \langle \tilde{\psi}_0, h_0 \rangle + \langle \tilde{\psi}_u, h_u \rangle + \langle \tilde{\rho}, h \rangle$$
$$= \begin{cases} 0 & \text{if } a \geq 0;\ \tilde{\psi}_0, \tilde{\psi}_u, \tilde{\rho} \geq 0; \text{ and} \\ & \quad \tilde{\psi}_0, \tilde{\psi}_u \text{ are zero outside } \mathcal{X}_0, \mathcal{X}_u, \text{ respectively;} \\ -\infty & \text{otherwise.} \end{cases}$$

Now denote the extension of $\tilde{\psi}_0$, $\tilde{\psi}_u$, $\tilde{\rho}$ to the whole $\mathbb{R}^n$ by ψ_0, ψ_u, ρ, which are obtained by letting them be equal to zero outside of $\mathcal{X}$. Then, for the left-hand side of (5.4), we have the following equality:

$$\sup_{k_1 \in \mathcal{K}_1} \langle k^*, k_1 \rangle = \sup_{(B_0, B_u, B) \in \mathcal{B}} a(B_u - B_0) + \langle \psi_0, B_0 - B \rangle$$
$$+ \langle \psi_u, B - B_u \rangle + \left\langle \rho, -\frac{\partial B}{\partial x} f \right\rangle$$
$$= \sup_{(B_0, B_u, B) \in \mathcal{B}} \left(-a + \int d\psi_0 \right) B_0 + \left(a - \int d\psi_u \right) B_u$$
$$+ \langle -\psi_0 + \psi_u + \nabla \cdot (\rho f), B \rangle$$
$$= \begin{cases} 0 & \text{if } \int_{\mathbb{R}^n} d\psi_0 = a, \int_{\mathbb{R}^n} d\psi_u = a, \text{ and} \\ & \quad -\psi_0 + \psi_u + \nabla \cdot (\rho f) = 0; \\ \infty & \text{otherwise,} \end{cases}$$

where $\nabla \cdot (\rho f)$ is interpreted as a distributional derivative. Thus, for the supremum to be less than or equal to the infimum, we must have a nonzero $(a, \psi_0, \psi_u, \rho)$, where ψ_0, ψ_u, ρ are measures of bounded variation on $\mathbb{R}^n$, such that $a \geq 0$; ψ_0, ψ_u, ρ are nonnegative; ψ_0, ψ_u, ρ are equal to zero outside $\mathcal{X}_0$, $\mathcal{X}_u$, and $\mathcal{X}$, respectively; and

$$\int_{\mathbb{R}^n} d\psi_0 = a,$$

$$\int_{\mathbb{R}^n} d\psi_u = a,$$

$$\nabla \cdot (\rho f) = \psi_0 - \psi_u.$$

We will next show that because of the assumption that there exists a $\tilde{B} \in C^1(\mathbb{R}^n)$ such that $\frac{\partial \tilde{B}}{\partial x}(x) f(x) < 0$ for all $x \in \mathcal{X}$, we must have $a > 0$. For this, let $\mathcal{L} = (C(\mathcal{X}))^3$, and define

$$\mathcal{L}_1 = \Bigg\{ (h_0, h_u, h) \in \mathcal{L} : h_0 = B_0 - B,\ h_u = B - B_u,\ h = -\frac{\partial B}{\partial x} f \text{ on } \mathcal{X};$$

$$\text{and } (B_0, B_u, B) \in \mathcal{B} \Bigg\},$$

$$\mathcal{L}_2 = \{ (h_0, h_u, h) \in \mathcal{L} : h_0 \geq 0 \text{ on } \mathcal{X}_0,\ h_u \geq 0 \text{ on } \mathcal{X}_u,\ h \geq 0 \text{ on } \mathcal{X} \}.$$

Note in particular that due to the above assumption and the compactness of $\mathcal{X}_0$, $\mathcal{X}_u$, $\mathcal{X}$, we have $\mathcal{L}_1 \cap \text{int}(\mathcal{L}_2) \neq \emptyset$. Now consider $k^* = (a, \tilde{\psi}_0, \tilde{\psi}_u, \tilde{\rho})$ that we have before. Suppose that $a = 0$ and substitute this to (5.4). Then we have a nonzero $(\tilde{\psi}_0, \tilde{\psi}_u, \tilde{\rho}) \in (C(\mathcal{X})^*)^3$, such that

$$\sup_{\ell_1 \in \mathcal{L}_1} \langle (\tilde{\psi}_0, \tilde{\psi}_u, \tilde{\rho}), \ell_1 \rangle \leq \inf_{\ell_2 \in \mathcal{L}_2} \langle (\tilde{\psi}_0, \tilde{\psi}_u, \tilde{\rho}), \ell_2 \rangle.$$

This implies that $\mathcal{L}_1 \cap \text{int}(\mathcal{L}_2) = \emptyset$, which is contradictory to the above. Thus a must be strictly positive. Without loss of generality, assume that k^* is scaled such that $a = 1$. This completes the proof of our lemma. ∎

Next, we will show that the existence of ψ_0, ψ_u, ρ in the conclusion of Lemma 5.2 implies that there exists an unsafe trajectory of the system. Since in this case we have a density function ρ which is in fact a measure, we need a version of the Liouville theorem which applies to measures.

LEMMA 5.3. *Let $f \in C^1(D, \mathbb{R}^n)$, where $D \subseteq \mathbb{R}^n$ is open. For a measurable set Z, assume that $\phi_t(Z)$ is a subset of D for all t between 0 and T. If ρ is a measure of bounded variation on D such that ρ has a compact support and the distributional derivative $\nabla \cdot (\rho f)$ is also a measure of bounded variation with compact support, then*

$$\int_{\phi_T(Z)} d\rho - \int_Z d\rho = \int_0^T \int_{\phi_t(Z)} d(\nabla \cdot (\rho f)) dt.$$

Proof. Choose $\rho_1, \rho_2, \ldots \in C_0^\infty(D)$ such that $\rho_k \to \rho$ in the (weak) topology of distributions. Then also $\nabla \cdot (\rho_k f) \to \nabla \cdot (\rho f)$ in the sense of distributions. In

particular

$$\lim_{k\to\infty}\int_X d|\rho_k-\rho|=0,$$

$$\lim_{k\to\infty}\int_X d|\nabla\cdot(\rho_k f)-\nabla\cdot(\rho f)|=0$$

for every $X\subset D$. The lemma (cf. Lemma 3.1) was proven for the case of smooth ρ in [23], i.e.,

$$\int_{\phi_T(Z)}\rho_k(x)dx-\int_Z\rho_k(x)dx=\int_0^T\int_{\phi_t(Z)}[\nabla\cdot(\rho_k f)(x)]dxdt.$$

So the desired equality is obtained in the limit as $k\to\infty$. □

LEMMA 5.4. *Consider the system $\dot{x}=f(x)$ with $f\in C^1(\mathbb{R}^n,\mathbb{R}^n)$, and let $\mathcal{X}\subset\mathbb{R}^n$, $\mathcal{X}_0\subseteq\mathcal{X}$, $\mathcal{X}_u\subseteq\mathcal{X}$ be compact sets. Suppose there exist measures of bounded variations ψ_0, ψ_u, ρ such that ψ_0,ψ_u,ρ are nonnegative on $\mathbb{R}^n$ and equal to zero outside $\mathcal{X}_0$, $\mathcal{X}_u$, and $\mathcal{X}$, respectively; and $\int_{\mathcal{X}_0}d\psi_0=1$, $\int_{\mathcal{X}_u}d\psi_u=1$, $\nabla\cdot(\rho f)=\psi_0-\psi_u$. Then there exists a $T\geq 0$ and a trajectory $x(t)$ of the system such that $x(0)\in\mathcal{X}_0$, $x(T)\in\mathcal{X}_u$, and $x(t)\in\mathcal{X}$ for all $t\in[0,T]$.*

Proof. Let $X_1,X_2,\ldots\subseteq\mathbb{R}^n$ be a sequence of open sets such that $\mathcal{X}_0\subseteq X_i$ for all i and $\lim_{i\to\infty}X_i=\mathcal{X}_0$. In addition, define the measurable sets

$$Z_i=\bigcup_{x_0\in X_i}\{x\in\mathbb{R}^n : x=\phi_t(x_0)\text{ for some }t\geq 0\}\text{ for }i=1,2,\ldots.$$

By the assertions of the lemma, both ρ and $\nabla\cdot(\rho f)$ are measures with bounded variation and compact support, so we can use Lemma 5.3 and $\nabla\cdot(\rho f)=\psi_0-\psi_u$ to obtain the relation

$$\int_{\phi_t(Z_i)}d\rho-\int_{Z_i}d\rho=\int_0^t\int_{\phi_\tau(Z_i)}d(\psi_0-\psi_u)d\tau$$

for all $t\geq 0$. Since $\rho\geq 0$ and $\phi_t(Z_i)\subseteq Z_i$ for all $t\geq 0$, the left-hand side of the above expression is less than or equal to zero. It follows from $\int_{\mathcal{X}_0}d\psi_0=1$ and $\psi_0\geq 0$ that $\mathcal{X}_u\cap Z_i\neq\emptyset$ for all $i=1,2,\ldots$, for otherwise the right-hand side of the expression can be made strictly greater than zero by taking some $t>0$, and we obtain a contradiction. Since the sets $\mathcal{X}_0$ and $\mathcal{X}_u$ are closed, we conclude that $\phi_T(x_0)\in\mathcal{X}_u$ for some $T\geq 0$ and $x_0\in\mathcal{X}_0$. For our purposes, let T be the first time instance such that $\phi_T(x_0)\in\mathcal{X}_u$.

The case in which $T=0$ is trivial since $\mathcal{X}_0\subseteq\mathcal{X}$. Consider now the case in which $T>0$. We will show that $\phi_t(x_0)\in\mathcal{X}$ for all $t\in[0,T]$ by a contradiction. Suppose to the contrary that there exists $\tilde{T}\in(0,T)$ such that $\phi_{\tilde{T}}(x_0)\notin\mathcal{X}$. Then, for a sufficiently small open neighborhood U of x_0, we have

$$\phi_{\tilde{T}}(U)\subset\mathbb{R}^n\setminus(\mathcal{X}),$$

$$\phi_t(U)\cap\mathcal{X}_u=\emptyset\quad\forall t\in[0,\tilde{T}].$$

Using Lemma 5.3 again we obtain

$$\int_{\phi_{\tilde{T}}(U)}d\rho-\int_U d\rho=\int_0^{\tilde{T}}\int_{\phi_\tau(U)}d(\psi_0-\psi_u)d\tau.$$

Since $\rho = 0$ on $\mathbb{R}^n \setminus (\mathcal{X})$, the first term on the left is equal to zero, and therefore the left-hand side is nonpositive, which leads to a contradiction since the right-hand side is strictly greater than zero. This lets us conclude that $\phi_t(x_0) \in \mathcal{X}$ for all $t \in [0, T]$, thus finishing the proof of the lemma. □

We are now ready to present the proof of the main theorem.

Proof of Theorem 5.1.

($\Rightarrow$): This has been proven in Theorem 3.2.

($\Leftarrow$): This follows from Lemmas 5.2 and 5.4. □

5.1. Some remarks. The result stated in Theorem 5.1 uses the assumption that the following Slater-like condition [6] is fulfilled: there exists a function $\tilde{B} \in C^1(\mathbb{R}^n)$ such that $\frac{\partial \tilde{B}}{\partial x}(x) f(x) < 0$ for all $x \in \mathcal{X}$. While in the discrete case strong duality holds (and hence so does the necessity of barrier certificates) without such an assumption, its proof depends on a special property of polyhedral convex sets, which does not carry over to the continuous case. Eliminating this condition in the continuous case will presumably require a different proof technique than the one presented in this paper. Nevertheless, there are cases in which the condition is automatically fulfilled—for instance, when the trajectories of the system starting from any $x_0 \in \mathcal{X}$ leave a neighborhood of $\mathcal{X}$ at least once, as shown in the following proposition.

PROPOSITION 5.5. *Consider the system $\dot{x} = f(x)$ with $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ and let $\mathcal{X} \subset \mathbb{R}^n$ be a compact set. Suppose there exist an open neighborhood $\tilde{\mathcal{X}}$ of $\mathcal{X}$ and a time instant $T > 0$ such that for all initial conditions $x_0 \in \mathcal{X}$, we have the flow $\phi_t(x_0)$ outside of $\mathrm{cl}(\tilde{\mathcal{X}})$ for some $t \in [0, T]$. Then there exists a function $\tilde{B} \in C^1(\mathbb{R}^n)$ such that $\frac{\partial \tilde{B}}{\partial x}(x) f(x) < 0$ for all $x \in \mathcal{X}$.*

Proof. Let $\mathcal{Y}$ be an open neighborhood of $\mathcal{X}$ such that its closure is contained in $\tilde{\mathcal{X}}$. In addition, let $\xi \in C^1(\mathbb{R}^n)$ be a nonnegative function such that $\xi(x) = 1$ for all $x \in \mathcal{Y}$ and $\xi(x) = 0$ for all $x \notin \tilde{\mathcal{X}}$; also let $\psi \in C^1(\mathbb{R}^n)$ be a function such that $\psi(x) > 0$ for all $x \in \mathcal{X}$ and $\psi(x) = 0$ for all $x \notin \mathcal{Y}$. Now consider the differential equation $\dot{x} = \xi(x) f(x)$. Denote the flow of $\dot{x} = \xi(x) f(x)$ starting at x_0 by $\tilde{\phi}_t(x_0)$. Modulo a time reparameterization, the flows $\tilde{\phi}_t(x_0)$ and $\phi_t(x_0)$ are identical up to some finite time. Next define

$$\tilde{B}(x_0) = \int_0^\infty \psi(\tilde{\phi}_t(x_0)) dt.$$

For all x_0 in a neighborhood of $\mathcal{X}$, the flow $\tilde{\phi}_t(x_0)$ is outside of $\mathcal{Y}$ for large t and thus by its construction $\psi(\tilde{\phi}_t(x_0))$ is equal to zero for large t and for all such x_0. It follows that $\tilde{B}(x)$ is well defined on a neighborhood of $\mathcal{X}$. The function $\tilde{B}(x)$ is continuously differentiable on $\mathcal{X}$ since both $\psi(x)$ and $\tilde{\phi}_t(x)$ are also continuously differentiable. Taking the total derivative of $\tilde{B}(x)$ with respect to time, we obtain

$$\frac{\partial \tilde{B}}{\partial x}(x) \xi(x) f(x) = -\psi(x),$$

which is strictly less than zero, on $\mathcal{X}$. Finally, recall that on $\mathcal{X}$ we have $\xi(x) = 1$. This completes the proof of the proposition. □

While the above Slater-like condition excludes the possibility of applying Theorem 5.1 when there is, e.g., an equilibrium point in $\mathcal{X}$, analysis can still be performed by excluding a neighborhood of the equilibrium point from $\mathcal{X}$ in the condition (3.4). If the excluded region is either backward or forward invariant, and does not intersect $\mathcal{X}_0$ and $\mathcal{X}_u$, then the safety criterion (5.1)–(5.3) will still apply in terms of the original sets.

Finally, note also that when *all* the connected components of $\mathbb{R}^n \setminus \mathcal{X}$ are either forward or backward invariant, an even stronger safety criterion can be obtained, as in the following proposition.

PROPOSITION 5.6. *Let the system $\dot{x} = f(x)$ with $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ and the compact sets $\mathcal{X}_0 \subset \mathbb{R}^n$, $\mathcal{X}_u \subset \mathbb{R}^n$ be given, with $0 \notin \mathcal{X}_0 \cup \mathcal{X}_u$. Suppose that the origin is a globally asymptotically stable equilibrium of the system with a global strict Lyapunov function $V(x)$.*[4] *Let $\epsilon_1 = \min_{x \in \mathcal{X}_0 \cup \mathcal{X}_u} V(x)$ and $\epsilon_2 = \max_{x \in \mathcal{X}_0 \cup \mathcal{X}_u} V(x)$. Then there exists a function $B \in C^1(\mathbb{R}^n)$ satisfying*

$$B(x) \leq 0 \quad \forall x \in \mathcal{X}_0, \tag{5.5}$$

$$B(x) > 0 \quad \forall x \in \mathcal{X}_u, \tag{5.6}$$

$$\frac{\partial B}{\partial x}(x) f(x) \leq 0 \quad \forall x \in \{x \in \mathbb{R}^n : \epsilon_1 \leq V(x) \leq \epsilon_2\} \tag{5.7}$$

if and only if there exists no trajectory $x(t)$ of the system such that

$$x(0) \in \mathcal{X}_0, \tag{5.8}$$

$$x(T) \in \mathcal{X}_u \text{ for some } T \geq 0. \tag{5.9}$$

Proof. Define $\mathcal{X} = \{x \in \mathbb{R}^n : \epsilon_1 \leq V(x) \leq \epsilon_2\}$. In this case, the existence of a function $\tilde{B} \in C^1(\mathbb{R}^n)$ such that $\frac{\partial \tilde{B}}{\partial x}(x) f(x) < 0$ for all $x \in \mathcal{X}$ is guaranteed by Proposition 5.5, and even the Lyapunov function $V(x)$ can be used as $\tilde{B}(x)$. By Theorem 5.1, there exists a function $B \in C^1(\mathbb{R}^n)$ satisfying (5.5)–(5.7) if and only if there exists no trajectory $x(t)$ of the system such that $x(0) \in \mathcal{X}_0$, $x(T) \in \mathcal{X}_u$ for some $T \geq 0$, and $x(t) \in \mathcal{X}$ for all $t \in [0, T]$.

Since the connected components of $\mathbb{R}^n \setminus \mathcal{X}$ are either forward or backward invariant, however, there can be no trajectory $x(t)$ of the system and time instants T_1, T_2, T_3 such that $T_1 < T_2 < T_3$, $x(T_1) \in \mathcal{X}$, $x(T_2) \in \mathbb{R}^n \setminus \mathcal{X}$, and $x(T_3) \in \mathcal{X}$. This combined with the fact that $\mathcal{X}_0, \mathcal{X}_u \subseteq \mathcal{X}$ implies that the set of trajectories satisfying $x(0) \in \mathcal{X}_0$, $x(T) \in \mathcal{X}_u$ for some $T \geq 0$ and $x(t) \in \mathcal{X}$ for all $t \in [0, T]$ is the same as the set of trajectories satisfying (5.8)–(5.9), and therefore the statement of the proposition follows. ∎

6. Conclusions. In the previous sections, we have used insights from the linear programming duality appearing in the shortest path problem and the concept of density function to formulate a convex program for reachability, which together with a convex program for safety verification using barrier certificates proposed in an earlier work form a pair of weak alternatives for safety and reachability verification. We have additionally shown that other temporal properties such as eventuality and avoidance can also be verified via convex programming and have presented convex programs to do so. This opens the possibility of performing the verification using convex optimization. In particular, sum of squares programming can be used for this purpose when the vector field of the system is polynomial and the sets are semialgebraic.

We have further commented on the use of this methodology for worst-case verification or controller synthesis. It was pointed out that the convex programs can be combined to verify properties such as reachability–safety and eventuality–safety. Some

[4] That is, $V \in C^1(\mathbb{R}^n)$ is radially unbounded, $V(x) > 0$ for all $x \neq 0$, and $\frac{\partial V}{\partial x}(x) f(x) < 0$ for all $x \neq 0$.

examples have been presented for illustration. At the end of the paper, a converse theorem in safety verification using barrier certificates was proven.

Even though the present tests are aimed for continuous systems, they are useful for constructing discrete abstractions of hybrid systems. In addition, we expect that all of them can also be extended to handle hybrid systems directly, using an approach similar to the one presented in [21].

REFERENCES

[1] R. Alur, T. Dang, and F. Ivancic, *Progress on reachability analysis of hybrid systems using predicate abstraction*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 2623, Springer-Verlag, Heidelberg, 2003, pp. 4–19.

[2] H. Anai and V. Weispfenning, *Reach set computations using real quantifier elimination*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 2034, Springer-Verlag, Berlin, 2001, pp. 63–76.

[3] E. Asarin, T. Dang, and O. Maler, *The d/dt tool for verification of hybrid systems*, in Computer Aided Verification, Lecture Notes in Comput. Sci. 2404, Springer-Verlag, Berlin, 2002, pp. 365–370.

[4] J.-P. Aubin, *Viability Theory*, Birkhäuser, Boston, MA, 1991.

[5] A. Bemporad, F. D. Torrisi, and M. Morari, *Optimization-based verification and stability characterization of piecewise affine and hybrid systems*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 1790, Springer-Verlag, Berlin, 2000, pp. 45–58.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[7] A. Chutinan and B. H. Krogh, *Computational techniques for hybrid system verification*, IEEE Trans. Automat. Control, 48 (2003), pp. 64–75.

[8] E. M. Clarke, Jr., O. Grumberg, and D. A. Peled, *Model Checking*, MIT Press, Cambridge, MA, 2000.

[9] S. Glavaski, A. Papachristodoulou, and K. Ariyur, *Safety verification of controlled advanced life support system using barrier certificates*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 3414, Springer-Verlag, Heidelberg, 2005, pp. 306–321.

[10] G. J. Holzmann, *Design and Validation of Computer Protocols*, Prentice–Hall, Englewood Cliffs, NJ, 1991.

[11] R. Horowitz and P. Varaiya, *Control design of an automated highway system*, Proc. IEEE, 88 (2000), pp. 913–925.

[12] M. Huth and M. Ryan, *Logic in Computer Science: Modelling and Reasoning about Systems*, Cambridge University Press, Cambridge, UK, 2000.

[13] M. Jirstrand, *Invariant sets for a class of hybrid systems*, in Proceedings of the 37th IEEE Conference on Decision and Control, 1998, pp. 3699–3704.

[14] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice–Hall, Upper Saddle River, NJ, 1996.

[15] A. Kurzhanski and P. Varaiya, *Ellipsoidal techniques for reachability analysis*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 1790, Springer-Verlag, Heidelberg, 2000, pp. 203–213.

[16] G. Lafferriere, G. J. Pappas, and S. Yovine, *Symbolic reachability computations for families of linear vector fields*, J. Symbolic Comput., 32 (2001), pp. 231–253.

[17] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, New York, 1969.

[18] Z. Manna and A. Pnueli, *Temporal Verification of Reactive Systems: Safety*, Springer-Verlag, New York, 1995.

[19] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Dover, Mineola, NY, 1998.

[20] P. A. Parrilo, *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 2000.

[21] S. Prajna, A. Jadbabaie, and G. J. Pappas, *A framework for worst-case and stochastic safety verification using barrier certificates*, IEEE Trans. Automat. Control, to appear (2007).

[22] S. Prajna, A. Papachristodoulou, and P. A. Parrilo, *Introducing SOSTOOLS: A general purpose sum of squares programming solver*, in Proceedings of the 41st IEEE Conference on Decision and Control, 2002, pp. 741–746; available at http://www.cds.caltech.edu/sostools

and http://www.mit.edu/∼parrilo/sostools.

[23] A. Rantzer, *A dual to Lyapunov's stability theorem*, Systems Control Lett., 42 (2001), pp. 161–168.

[24] A. Rantzer and S. Hedlund, *Duality between cost and density in optimal control*, in Proceedings of the 42nd IEEE Conference on Decision and Control, 2003, pp. 1218–1221.

[25] A. Rantzer and S. Prajna, *On analysis and synthesis of safe control laws*, in Proceedings of the 42nd Allerton Conference on Communication, Control, and Computing, 2004.

[26] S. Sankaranarayanan, H. Sipma, and Z. Manna, *Constructing invariants for hybrid systems*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 2993, Springer-Verlag, Berlin, 2004, pp. 539–554.

[27] A. Tiwari, *Approximate reachability for linear systems*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 2623, Springer-Verlag, Berlin, 2003, pp. 514–525.

[28] A. Tiwari and G. Khanna, *Nonlinear systems: Approximating reach sets*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 2993, Springer-Verlag, Berlin, 2004, pp. 600–614.

[29] C. Tomlin, I. Mitchell, and R. Ghosh, *Safety verification of conflict resolution maneuvers*, IEEE Trans. Intelligent Transportation Systems, 2 (2001), pp. 110–120.

[30] C. J. Tomlin, I. Mitchell, A. M. Bayen, and M. Oishi, *Computational techniques for the verification of hybrid systems*, Proc. IEEE, 91 (2003), pp. 986–1001.

[31] H. Yazarel and G. Pappas, *Geometric programming relaxations for linear system reachability*, in Proceedings of the American Control Conference, 2004, pp. 553–559.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 1022–1051

LOCAL EXACT BOUNDARY CONTROLLABILITY FOR NONLINEAR WAVE EQUATIONS*

YI ZHOU† AND ZHEN LEI‡

Abstract. This paper deals with the local exact boundary controllability for dynamics governed by nonlinear wave equations, subject to Dirichlet, Neumann, or any other kind of boundary controls which result in well-posedness of the corresponding initial-boundary value problem. A constructive method is developed. The local exact boundary controllability for semilinear wave equations is constructed in the case of both three (odd) and two (even) space dimensions, and the boundary control is time optimal when the space dimension is three (odd). Especially, the local exact boundary controllability is established for quasi-linear wave equations in several space dimensions by using the constructive method.

Key words. local exact boundary controllability, semilinear wave equations, quasi-linear wave equations, time optimal

AMS subject classifications. 93B05, 35L05

DOI. 10.1137/060650222

1. Introduction.

The problems of controllability are clearly of significant practical interest. There is an extremely large number of publications on these topics. Some classical references are Lions [15] and Russell [17].

The aim of this paper is to study the local exact boundary controllability for semilinear and quasi-linear wave equations in several space dimensions. The problem can be described as follows: Let Ω_0 be a bounded open subset of $\mathbb{R}^n$ ($n = 1, 2, 3$ in the applications) with a smooth boundary Γ. The equations under consideration take the form

$$\Box u = F(t, x, u, u'), \quad 0 < t < T,\ x \in \Omega_0, \tag{1.1}$$

in the semilinear case, and the form

$$\Box u = G(t, x, u, u', u''), \quad 0 < t < T,\ x \in \Omega_0, \tag{1.2}$$

in the quasi-linear case, where $\Box = \partial_t^2 - \Delta$ is the wave operator, the Laplace Δ is taken with respect to the spatial variables $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, and F and G are smooth functions of their arguments, vanishing together with their first-order derivatives with respect to (u, u') or (u, u', u'') at $(u, u') = 0$ or $(u, u', u'') = 0$; namely,

*Received by the editors January 18, 2006; accepted for publication (in revised form) January 23, 2007; published electronically June 29, 2007.
http://www.siam.org/journals/sicon/46-3/65022.html

†School of Mathematical Sciences, Fudan University, Shanghai 200433, People's Republic of China, and Key Laboratory of Mathematics for Nonlinear Sciences (Fudan University), Ministry of Education, Shanghai 200433, People's Republic of China (yizhou@fudan.ac.cn). This author was partially supported by the National Science Foundation of China under grant 10225102, by a 973 project of the National Science Foundation of China, and by the Doctoral Program Foundation of the Ministry of Education of China.

‡School of Mathematical Sciences, Fudan University, Shanghai 200433, People's Republic of China, and School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, People's Republic of China (leizhn@yahoo.com). This author was partially supported by LGH under grant 1140003 and by the Foundation of Excellent Doctoral Dissertation of China.

the following holds in a neighborhood of $(u, u') = 0$ or $(u, u', u'') = 0$:

$$\begin{cases} F(t,x,u,u') = O\big(|u|^2 + |u'|^2\big), \\ G(t,x,u,u',u'') = O\big(|u|^2 + |u'|^2 + |u''|^2\big), \end{cases} \tag{1.3}$$

and $u' = (u_t, \nabla u)$, $u'' = (u_{tt}, \nabla u_t, \nabla^2 u)$ are the first-order and second-order space-time derivatives of $u(t,x)$. Without loss of generality, we shall assume that

$$G(t,x,u,u',u'') = g_{i\alpha}(t,x,u,u_t,\nabla u)\partial^2_{i\alpha} u + F(t,x,u,u'), \tag{1.4}$$

where $g_{i\alpha}$ are smooth functions with

$$g_{i\alpha}(t,x,0,0,0) = 0 \tag{1.5}$$

for all $1 \le i \le n$ and $0 \le \alpha \le n$. We point out here that in what follows, Latin indices $i, j, k, \dots$ range over $1, 2, \dots, n$, Greek indices $\alpha, \beta, \gamma, \dots$ over $0, 1, 2, \dots, n$, and summations over repeated indices are always well understood.

Consider the initial data

$$u(0,x) = f_0(x), \quad u_t(0,x) = f_1(x), \quad x \in \Omega_0, \tag{1.6}$$

and the final data

$$u(T,x) = g_0(x), \quad u_t(T,x) = g_1(x), \quad x \in \Omega_0, \tag{1.7}$$

with $f_0, g_0 \in H^{s+1}(\Omega_0)$ and $f_1, g_1 \in H^s(\Omega_0)$, $H^s(\Omega_0)$ being the standard Sobolev space of order s. We impose any one of the following boundary conditions for $0 \le t \le T$ and $x \in \partial\Omega_0$:

$$\begin{cases} u = h(t,x) & \text{of Dirichlet type,} \\ \frac{\partial u}{\partial n} = h(t,x) & \text{of Neumann type,} \\ \frac{\partial u}{\partial n} + bu = h(t,x) & \text{of the third type,} \\ \frac{\partial u}{\partial n} + \overline{b} u_t = h(t,x) & \text{of the dissipative type,} \end{cases} \tag{1.8}$$

where b and $\overline{b}$ are given positive constants.

Then the problem of exact boundary controllability for (1.1) or (1.2) is stated as follows: given $T > 0$, is it possible, for any initial data (f_0, f_1) and final data (g_0, g_1), to find an appropriate boundary control $h(t,x)$, such that the solution $u(t,x)$ of the semilinear system (1.1) with (1.6) and (1.8) (the quasi-linear system (1.2) with (1.6) and (1.8), respectively) satisfies the final state (1.7)?

This problem has received considerable attention in the literature, with numerous contributions over the past several decades, by using the so-called Hilbert uniqueness method introduced by Lions [14]. When $F = G \equiv 0$, the problem is by now well understood; for example, see [15, 17] and their references. In the case of linear wave equations with variable coefficients, there are also the works by Bardos, Lebeau, and Rauch [1], Cavalcanti [2], Tataru [18], and Yao [21]. For the semilinear case, there are plenty of results when $F = F(u)$ [8, 24, 25]. These works succeed in the framework of weak solutions such that the Hilbert uniqueness method can work. In the framework of classical solutions, Li [10] and Li et al. [12, 13] recently studied local exact boundary controllability for one-dimensional quasi-linear wave equations and hyperbolic systems with both one-sided and two-sided boundary controls. Yu [23] also

studied local exact boundary controllability for higher-order one-dimensional quasilinear hyperbolic equations. However, even in the weak sense, to our knowledge few results are known for the semilinear wave equations of the form (1.1) and the quasilinear wave equations of the form (1.2) in the multidimensional case. After completion of this work, we discovered that similar results were announced in Yao [22], a recently completed manuscript dealing with exact boundary controllability for quasi-linear wave equations under some geometrical conditions on the domain. His proof relies on the observability inequalities.

Our results are obtained in the framework of the classical solutions of wave equations. A constructive method is developed. We point out that the methods presented here are completely different from the recent work by Yao [22]. In the case of semilinear wave equations in three (odd) space dimensions, the local exact boundary controllability is established and the boundary control is time optimal. The proof relies on the so-called Huygens principle for the linear wave equation, which had been used by Russell [17] to treat the exact boundary controllability in the linear case. In the case of semilinear wave equations in two (even) space dimensions, we proved the fact that the energy is linearly dissipative when one adds a boundary condition of dissipative type by using Morawetz energy estimates. Then the desired local exact boundary controllability is established. We point out here that a similar idea and results have appeared in [17] in the linear case. The quasi-linear case is similar to the even space–dimensional semilinear case, but it is much more complicated and difficult. The proof relies on an exponentially dissipative energy estimate by using a shift technique, which reveals the underlying conservation law governing the system. In what follows, for convenience, the proofs will be illustrated in the cases of two and three space dimensions. However, the results presented here are also valid for general multidimensional cases.

Without loss of generality, we assume that

$$\Omega_0 \subset\subset \mathbb{B}_\rho, \tag{1.9}$$

where $\mathbb{B}_\rho$ is a ball of radius $\rho > 0$ centered at the origin. Define

$$R(T) = \Omega_0 \times (0, T).$$

The main result on the local exact boundary controllability for the three-dimensional semilinear wave equations can be stated as follows.

THEOREM 1.1. *Assume that $n = 3$ and F is a smooth function with respect to its arguments. Assume, furthermore, that* (1.3) *and* (1.9) *hold. Let $T > 2\rho$. Then, for any given initial data (f_0, f_1) and final data (g_0, g_1) with $f_0, g_0 \in H^{s+1}(\Omega_0)$, $f_1, g_1 \in H^s(\Omega_0)$, $s \geq 2$, there exist boundary controls h such that the mixed initial-boundary value problem* (1.1) *and* (1.6), (1.8) *admits a unique $\cap_{j=0}^{s+1} C^j\big([0,T]; H^{s+1-j}(\Omega_0)\big)$ solution $u = u(t,x)$ on the domain $R(T)$ satisfying the final condition* (1.7), *provided that*

$$\|(f_0, g_0)\|_{H^{s+1}(\Omega_0)} + \|(f_1, g_1)\|_{H^s(\Omega_0)} \leq \varepsilon_1, \tag{1.10}$$

where ε_1 is a sufficiently small positive constant.

REMARK 1.2. *The control time $T > 2\rho$ can be replaced by $T > \mathrm{diam}(\Omega_0)$ in Theorem* 1.1. *See also Remark* 2.2.

In the case of semilinear wave equations in two space dimensions, the boundary control is obtained for an appropriately large time.

THEOREM 1.3. *Assume that $n = 2$ and F is a smooth function with respect to its arguments. Assume, furthermore, that* (1.3) *and* (1.9) *hold. Then, for any given initial data (f_0, f_1) and final data (g_0, g_1) with $f_0, g_0 \in H^{s+1}(\Omega_0)$, $f_1, g_1 \in H^s(\Omega_0)$, $s \geq 2$, there exist boundary controls h such that the mixed initial-boundary value problem* (1.1) *and* (1.6), (1.8) *admit a unique $\cap_{j=0}^{s+1} C^j\big([0,T]; H^{s+1-j}(\Omega_0)\big)$ solution $u = u(t,x)$ on the domain $R(T)$ satisfying the final condition* (1.7), *provided that T is a big enough constant and*

$$\|(f_0, g_0)\|_{H^{s+1}(\Omega_0)} + \|(f_1, g_1)\|_{H^s(\Omega_0)} \leq \varepsilon_2, \tag{1.11}$$

where ε_2 is a sufficiently small positive constant.

REMARK 1.4. *There are results in the literature (see, e.g.,* [15, 7, 16]*) where the control time $T > \mathrm{diam}(\Omega_0)$ for a class of linear wave equations. Unfortunately, we cannot get this kind of sharp result for nonlinear wave equations.*

REMARK 1.5. *For semilinear wave equations in n space dimensions, the results in Theorems* 1.1 *and* 1.3 *are also valid if integer $s > \frac{n}{2}$. This amount of regularity is required for the classical local existence theorem of semilinear wave equations.*

Moreover, the local exact boundary controllability is also established for the quasi-linear wave equation in several space dimensions.

THEOREM 1.6. *Assume that $n = 2$ or 3 and G is a smooth function with respect to its arguments. Assume, furthermore, that* (1.9) *and* (1.4)–(1.5) *hold. Then, for any given initial data (f_0, f_1) and final data (g_0, g_1) with $f_0, g_0 \in H^{s+1}(\Omega_0)$, $f_1, g_1 \in H^s(\Omega_0)$, $s \geq 3$, there exist boundary controls h such that the mixed initial-boundary value problem* (1.2) *and* (1.6), (1.8) *admits a unique $\cap_{j=0}^{s+1} C^j\big([0,T]; H^{s+1-j}(\Omega_0)\big)$ solution $u = u(t,x)$ on the domain $R(T)$ satisfying the final condition* (1.7), *provided that T is a big enough constant and*

$$\|(f_0, g_0)\|_{H^{s+1}(\Omega_0)} + \|(f_1, g_1)\|_{H^s(\Omega_0)} \leq \varepsilon_3, \tag{1.12}$$

where ε_3 is a sufficiently small positive constant.

REMARK 1.7. *For quasi-linear wave equations in n space dimensions, the results in Theorem* 1.6 *are also valid if integer $s > \frac{n}{2} + 1$. This amount of regularity is required for the classical local existence theorem of quasi-linear wave equations.*

We point out here that the boundary control is not unique. By way of our construction, it does not matter what kind of boundary condition in (1.8) we use as long as the initial-boundary value problem is well-posed. However, the type of boundary condition may be relevant for other different approaches (especially when the control is assumed to be applied only on a part of the boundary, in which case the type of boundary control may lead to differences; for an example, see Lasiecka, Triggiani, and Zhang [9]). In what follows, we will concentrate on boundary controls of Dirichlet type.

The paper is organized as follows: Section 2 is devoted to establishing exact boundary controllability for linear wave equations. In the case of three space dimensions, one can utilize the so-called Huygens principle and then get time-optimal control. In the case of two space dimensions, one can construct the desired controllability by using Morawetz energy estimates, but one does not know how large the control time is. The results are then extended to the semilinear cases by the contraction mapping principle in section 3. The proofs of Theorems 1.1 and 1.3 can be found there. In section 4, the exponentially dissipative energy estimates of solutions for quasi-linear wave equations are established, and then the local exact boundary controllability is obtained for quasi-linear wave equations in both two and three space dimensions using a direct constructive method.

2. Exact boundary controllability for linear wave equations. In this section we investigate the controllability problem for the linear wave equation

$$\Box u = 0, \quad 0 \le t \le T,\ x \in \Omega_0, \tag{2.1}$$

with initial data (1.6) and final data (1.7), and boundary condition of Dirichlet type

$$u|_{\partial\Omega_0} = h, \quad 0 \le t \le T,\ x \in \partial\Omega_0. \tag{2.2}$$

We point out here that exact boundary controllability for linear wave equations has been studied by many authors and is now well known (see, e.g., [1, 16, 17, 15]). We study exact boundary controllability for linear wave equations in order to establish a basis for semilinear and quasi-linear wave equations.

Without loss of generality, we can assume that $\rho = 1$ and

$$\Omega_0 \subset\subset \mathbb{B}_1.$$

In fact, if $\rho \neq 1$, we can define $\widehat{u}(t,x) = u(\rho t, \rho x)$, $\widehat{f}_0(x) = f_0(\rho x)$, $\widehat{f}_1(x) = f_1(\rho x)$, $\widehat{g}_0(x) = g_0(\rho x)$, $\widehat{g}_1(x) = g_1(\rho x)$, and then consider the controllability problems for the $\widehat{u}$ system with initial data $\widehat{f}_0, \widehat{f}_1$ and final data $\widehat{g}_0, \widehat{g}_1$. Then we can always extend the functions f_0, g_0, f_1, g_1 to $\widetilde{f}_0, \widetilde{g}_0, \widetilde{f}_1, \widetilde{g}_1$ such that

$$\operatorname{supp}(\widetilde{f}_0, \widetilde{g}_0, \widetilde{f}_1, \widetilde{g}_1) \subset\subset \mathbb{B}_1 \tag{2.3}$$

and

$$\begin{cases} \|\widetilde{f}_0\|_{H^{s+1}(\mathbb{B}_1)} + \|\widetilde{g}_0\|_{H^{s+1}(\mathbb{B}_1)} \le C_s\big(\|f_0\|_{H^{s+1}(\Omega_0)} + \|g_0\|_{H^{s+1}(\Omega_0)}\big), \\ \|\widetilde{f}_1\|_{H^{s}(\mathbb{B}_1)} + \|\widetilde{g}_1\|_{H^{s}(\mathbb{B}_1)} \le C_s\big(\|f_1\|_{H^{s}(\Omega_0)} + \|g_1\|_{H^{s}(\Omega_0)}\big) \end{cases} \tag{2.4}$$

for some constants $C_s > 0$ and $s \ge 0$ (for example, see [4]). In what follows, we will use the same extension several times and always denote the extension operator by $\widetilde{\ }: f \longrightarrow \widetilde{f}$.

In three space dimensions, with the aid of the so-called Huygens principle, it is rather easy to construct the exact boundary control for the linear wave equation, and the control time is optimal in the sense of $T > \operatorname{diam}(\Omega_0)$. But in two space dimensions, the Huygens principle is invalid, the exact boundary controllability is not trivial, and we do not know how large the control time T is based on our methods. We begin with the easier case of three space dimensions.

2.1. Exact boundary controllability for linear wave equations in three space dimensions. First of all, we consider the Cauchy problem

$$\begin{cases} \Box v = 0, \quad 0 \le t \le T, \quad x \in \mathbb{R}^3; \\ v(0,x) = \widetilde{f}_0, \quad v_t(0,x) = \widetilde{f}_1, \quad x \in \mathbb{R}^3, \end{cases} \tag{2.5}$$

and the inverted Cauchy problem from some positive time T to 0,

$$\begin{cases} \Box w = 0, \quad 0 \le t \le T, \quad x \in \mathbb{R}^3; \\ w(T,x) = \widetilde{g}_0, \quad w_t(T,x) = \widetilde{g}_1, \quad x \in \mathbb{R}^3. \end{cases} \tag{2.6}$$

By the Huygens principle, we have that

$$\begin{cases} v(T,x) = v_t(T,x) = 0, \\ w(0,x) = w_t(0,x) = 0, \end{cases} \tag{2.7}$$

provided $T > 2$ and $x \in \mathbb{B}_1$.

Thus, if one defines

$$\mathcal{L}\big(\widetilde{f}_0, \widetilde{g}_0, \widetilde{f}_1, \widetilde{g}_1\big) = v + w, \quad 0 \le t \le T,\ x \in \Omega_0, \tag{2.8}$$

then it is obvious that $u = \mathcal{L}\big(\widetilde{f}_0, \widetilde{g}_0, \widetilde{f}_1, \widetilde{g}_1\big)$ solves (2.1) in $R(T)$ and verifies the initial and final data (1.6) and (1.7) for $x \in \Omega_0$. Consequently,

$$h = u|_{\partial\Omega_0} \tag{2.9}$$

is the boundary control we want. We point out here that a similar method was used by Russell [16, 17] in the seventies.

Summing up, we have the following lemma.

LEMMA 2.1. *Consider the linear wave equation* (2.1) *in three space dimensions. Then, for any given initial data* (f_0, f_1) *and final data* (g_0, g_1) *with* $f_0, g_0 \in H^{s+1}(\Omega_0)$, $f_1, g_1 \in H^s(\Omega_0)$, $s \ge 1$, *there exist boundary controls* h *such that the mixed initial-boundary value problem* (2.1), (2.2) *and* (1.6) *admits a unique* $\cap_{j=0}^{s+1} C^j\big([0,T]; H^{s+1-j}(\Omega_0)\big)$ *solution* $u = u(t,x)$ *on the domain* $R(T)$ *which arrives at the final condition* (1.7), *provided* $T > 2$. *Moreover, we have*

$$\sup_{0\le t\le T} \sum_{j=0}^{s+1} \|\partial_t^j u\|^2_{H^{s+1-j}(\Omega_0)} \le CT\Big(\|(\widetilde{f}_0, \widetilde{g}_0)\|^2_{H^{s+1}(\mathbb{B}_1)} + \|(\widetilde{f}_1, \widetilde{g}_1)\|^2_{H^s(\mathbb{B}_1)}\Big), \tag{2.10}$$

where C *is a positive constant independent of* T, u, *and the initial and final data.*

Proof. It remains to prove (2.10). By energy estimates, it is rather easy to see that

$$\sum_{j+k=0}^{s} \frac{d}{dt}\Big(\|\nabla^k \partial_t^{j+1} v\|^2_{L^2(\mathbb{R}^3)} + \|\nabla\nabla^k \partial_t^j v\|^2_{L^2(\mathbb{R}^3)}\Big) = 0,$$

which results in

$$\sum_{j=1}^{s+1} \|\partial_t^j v\|^2_{H^{s+1-j}(\mathbb{R}^3)} + \|\nabla v\|^2_{H^s(\mathbb{R}^3)} \le C\big(\|\widetilde{f}_0\|^2_{H^{s+1}(\mathbb{B}_1)} + \|\widetilde{f}_1\|^2_{H^s(\mathbb{B}_1)}\big).$$

On the other hand, by using the identity

$$v(t,x) = \widetilde{f}_0 + \int_0^t v_s(s,x)ds,$$

we have

$$\|v\|^2_{L^2(\mathbb{R}^3)} \le 2\|\widetilde{f}_0\|^2_{L^2(\mathbb{R}^3)} + CT \sup_{0\le t\le T} \|v_t\|^2_{L^2(\mathbb{R}^3)} \le CT\Big(\|\widetilde{f}_0\|^2_{H^{s+1}(\mathbb{B}_1)} + \|\widetilde{f}_1\|^2_{H^s(\mathbb{B}_1)}\Big).$$

Consequently, we have

$$\sum_{j=0}^{s+1} \|\partial_t^j v\|^2_{H^{s+1-j}(\mathbb{R}^3)} \le CT\Big(\|(\widetilde{f}_0, \widetilde{g}_0)\|^2_{H^{s+1}(\mathbb{B}_1)} + \|(\widetilde{f}_1, \widetilde{g}_1)\|^2_{H^s(\mathbb{B}_1)}\Big).$$

Obviously, the same result holds for w. Thus, (2.10) is proved. ∎

REMARK 2.2. *It is obvious that the result in Lemma* 2.1 *is also valid by replacing* $\mathbb{B}_1$ *by any open, connected set* $\Omega_1 \supset \Omega_0$*; then* $T > 2$ *can be replaced by* $T > \mathrm{diam}(\Omega_0)$ *by choosing* Ω_1 *such that the distance of* $\partial\Omega_1$ *and* $\partial\Omega_0$ *is small enough. Thus, the boundary controls for three-dimensional semilinear wave equations in section* 3 *are in fact time-optimal in the sense that* $T > \mathrm{diam}(\Omega_0)$*. In section* 3*, we will present the results in two- and three-dimensional cases in the uniform way since we do not pursue the time-optimal controls in the two-dimensional case.*

2.2. Exact boundary controllability for linear wave equations in two space dimensions. Now we shall construct the boundary controls for the linear wave equation (2.1) in two space dimensions. There are many well-known results on this subject, and time-optimal controllability especially has also been obtained for a class of hyperbolic equations (see, e.g., [7]). The result we present below is not time-optimal. However, our method has the flexibility to work in the quasi-linear case.

Compared to three space dimensions, the difficulty in two space dimensions lies in the invalidity of the Huygens principle. The proof relies first on the linearly dissipative energy estimates of solutions for a mixed initial-boundary value linear wave equation with a dissipative boundary condition, and then on a constructive method.

First of all, we focus our attention on the following initial-boundary value problem with a dissipative boundary condition:

$$\begin{cases} \Box\psi = 0, \quad 0 \le t \le T, \quad x \in \mathbf{B}_1, \\ \psi_t + \psi_r = 0, \quad 0 \le t \le T, \quad x \in \mathbf{S}^1, \\ \psi(0,x) = \widetilde{f}_0, \quad \psi_t(0,x) = \widetilde{f}_1, \quad x \in \mathbf{B}_1, \end{cases} \tag{2.11}$$

where ψ_r is the outer normal derivative of ψ on $\mathbf{S}^1$.

For system (2.11), we desire to establish the following lemma. See also [1], where similar results are obtained even for linear wave equations with variant coefficients. However, the methods we use below can be generalized to nonlinear cases.

LEMMA 2.3. *There exists a constant* $T_0 > 0$*, such that every solution* $\psi(t,x)$ *of system* (2.11) *satisfies*

$$\|\nabla\psi(T,\cdot)\|^2_{L^2(\mathbb{B}_1)} + \|\psi_t(T,\cdot)\|^2_{L^2(\mathbb{B}_1)} \le \lambda_0\Big(\|\widetilde{f}_0\|^2_{H^1(\mathbb{B}_1)} + \|\widetilde{f}_1\|^2_{L^2(\mathbb{B}_1)}\Big) \tag{2.12}$$

for $T > T_0$ *and some* $\lambda_0 \in (0,1)$*. Thus, we have*

$$\|\nabla\psi(T,\cdot)\|^2_{L^2(\Omega_0)} + \|\psi_t(T,\cdot)\|^2_{L^2(\Omega_0)} \le \lambda_0\Big(\|\widetilde{f}_0\|^2_{H^1(\mathbb{B}_1)} + \|\widetilde{f}_1\|^2_{L^2(\mathbb{B}_1)}\Big) \tag{2.13}$$

for $T > T_0$ *and some* $\lambda_0 \in (0,1)$*.*

Proof. Let $\psi(t,x)$ be the solution of system (2.11). First of all, we claim that

$$\int_{\mathbb{B}_1} \psi_t(t,x)dx + \int_{\mathbb{S}^1} \psi(t,y)d\sigma_y = \int_{\mathbb{B}_1} \widetilde{f}_1 dx. \tag{2.14}$$

To carry out the details, let us integrate the linear wave equation with respect to x on the unit ball $\mathbb{B}_1$ to yield

$$\frac{d}{dt}\int_{\mathbb{B}_1} \psi_t dx - \int_{\mathbb{S}^1} \psi_r d\sigma_y = 0,$$

where we used the so-called Green's formula. By using the boundary condition in (2.11), we get

$$\frac{d}{dt}\Big(\int_{\mathbb{B}_1}\psi_t dx+\int_{\mathbb{S}^1}\psi d\sigma_y\Big)=0. \tag{2.15}$$

Noting (2.3), we have

$$\int_{\mathbb{S}^1}\widetilde{f}_0 d\sigma_y=0.$$

Then we have proved the claim (2.14).

Let us continue the proof of Lemma 2.3 under (2.14). By taking the L^2 inner product of the linear wave equation in system (2.11) with ψ_t, one obtains the following energy estimate:

$$\frac{1}{2}\frac{d}{dt}\Big(\|\nabla\psi(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}\Big)-\int_{\mathbb{S}^1}\psi_t\psi_r d\sigma_y=0.$$

With the aid of the boundary condition in (2.11), we have

$$\frac{1}{2}\frac{d}{dt}\Big(\|\nabla\psi(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}\Big)+\|\psi_t\|^2_{L^2(\mathbb{S}^1)}=0, \tag{2.16}$$

which implies that

$$\begin{aligned}&\|\nabla\psi(T,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(T,\cdot)\|^2_{L^2(\mathbb{B}_1)}\\&\le\|\nabla\psi(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}\\&\le\|\nabla\psi(0,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(0,\cdot)\|^2_{L^2(\mathbb{B}_1)}\end{aligned} \tag{2.17}$$

for $0\le t\le T$.

Next we do the following energy estimate of Morawetz type. By taking the L^2 inner product of the linear wave equation in system (2.11) with $x\cdot\nabla\psi$, one gets

$$\begin{aligned}&\frac{d}{dt}\int_{\mathbb{B}_1}(x\cdot\nabla\psi)\psi_t dx-\int_{\mathbb{B}_1}x\cdot\nabla\frac{|\psi_t|^2}{2}dx\\&=\int_{\mathbb{B}_1}\nabla_k\big(\nabla_k\psi x\cdot\nabla\psi\big)-|\nabla\psi|^2-x\cdot\nabla\frac{|\nabla\psi|^2}{2}dx.\end{aligned} \tag{2.18}$$

Thus, by using Green's formula and integration by parts, one has

$$\begin{aligned}&\frac{d}{dt}\int_{\mathbb{B}_1}(x\cdot\nabla\psi)\psi_t dx+\frac{1}{2}\Big(\|\nabla\psi(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}\Big)\\&\le-\frac{1}{2}\int_{\mathbb{B}_1}\big(|\psi_t|^2-|\nabla\psi|^2\big)dx+\|\psi_t\|^2_{L^2(\mathbb{S}^1)}-\frac{1}{2}\|\Omega\psi\|^2_{L^2(\mathbb{S}^1)},\end{aligned} \tag{2.19}$$

where we used the boundary condition in system (2.11) and the following decomposition of spatial derivatives into radial and angular components:

$$\nabla=\frac{x}{r}\partial_r-\frac{x\wedge\Omega}{r^2}. \tag{2.20}$$

On the other hand, a straightforward calculation shows that

$$\begin{aligned}(2.21)\qquad &\int_{\mathbb{B}_1}\big(|\psi_t|^2-|\nabla\psi|^2\big)dx\\ &=\frac{d}{dt}\int_{\mathbb{B}^1}\psi\psi_t dx-\int_{\mathbb{B}_1}\psi\Delta\psi+|\nabla\psi|^2dx\\ &=\frac{d}{dt}\int_{\mathbb{B}_1}\psi\psi_t dx-\frac{1}{2}\int_{\mathbb{S}^1}\partial_r\psi^2 d\sigma_y\\ &=\frac{d}{dt}\Big(\int_{\mathbb{B}_1}\psi\psi_t dx+\frac{1}{2}\int_{\mathbb{S}^1}\psi^2 d\sigma_y\Big).\end{aligned}$$

Combining (2.19) and (2.21), and with the aid of (2.16), we arrive at

$$\begin{aligned}(2.22)\qquad &\frac{1}{2}\Big(\|\nabla\psi(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}\Big)+\frac{1}{2}\|\Omega\psi\|^2_{L^2(\mathbb{S}^1)}\\ &\le\|\psi_t\|^2_{L^2(\mathbb{S}^1)}-\frac{d}{dt}\Big[\int_{\mathbb{B}_1}\Big(\frac{1}{2}\psi+x\cdot\nabla\psi\Big)\psi_t dx+\frac{1}{4}\|\psi\|^2_{L^2(\mathbb{S}^1)}\Big]\\ &=-\frac{d}{dt}\Big[\int_{\mathbb{B}_1}\Big(\frac{1}{2}\psi+x\cdot\nabla\psi\Big)\psi_t dx+\frac{1}{4}\|\psi\|^2_{L^2(\mathbb{S}^1)}\\ &\qquad+\frac{1}{2}\Big(\|\nabla\psi(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}\Big)\Big].\end{aligned}$$

To estimate the right side of (2.22), we calculate

$$\begin{aligned}(2.23)\qquad &\Big|\int_{\mathbb{B}_1}\Big(\frac{1}{2}\psi+x\cdot\nabla\psi\Big)\psi_t dx\Big|\\ &\le\frac{1}{2}\int_{\mathbb{B}_1}\psi_t^2dx+\frac{1}{2}\int_{\mathbb{B}_1}\Big[\frac{1}{4}\psi^2+(x\cdot\nabla\psi)^2+\frac{1}{2}x\cdot\nabla\psi^2\Big]dx\\ &\le\frac{1}{2}\Big(\|\nabla\psi(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}\Big)\\ &\quad+\frac{1}{4}\|\psi\|^2_{L^2(\mathbb{S}^1)}-\frac{3}{8}\|\psi\|^2_{L^2(\mathbb{B}_1)}.\end{aligned}$$

Noting (2.17), and with the aid of (2.23), we integrate (2.22) with respect to t from 0 to T, which yields

$$\begin{aligned}(2.24)\qquad &\frac{T}{2}\Big(\|\nabla\psi(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}\Big)\\ &\le\Big|\int_{\mathbb{B}_1}\Big(\frac{n-1}{2}\psi+x\cdot\nabla\psi\Big)\psi_t dx\Big|_{t=0}+\frac{1}{4}\|\psi(0,\cdot)\|^2_{L^2(\mathbb{S}^1)}\\ &\quad+\Big|\int_{\mathbb{B}_1}\Big(\frac{n-1}{2}\psi+x\cdot\nabla\psi\Big)\psi_t dx\Big|_{t=T}-\frac{1}{4}\|\psi(T,\cdot)\|^2_{L^2(\mathbb{S}^1)}\\ &\quad+\Big(\|\nabla\psi(0,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(0,\cdot)\|^2_{L^2(\mathbb{B}_1)}\Big)\\ &\le 2\Big(\|\nabla\psi(0,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\|\psi_t(0,\cdot)\|^2_{L^2(\mathbb{B}_1)}\Big)+\frac{1}{2}\|\psi(0,\cdot)\|^2_{L^2(\mathbb{S}^1)}.\end{aligned}$$

In view of the so-called Poincaré inequality, we have

$$(2.25)\qquad \|\psi(0,\cdot)\|_{L^2(\mathbb{B}_1)}\le C\Big[\|\nabla\psi(0,\cdot)\|_{L^2(\mathbb{B}_1)}+\Big|\int_{\mathbb{S}^1}\psi(0,\cdot)dx\Big|\Big].$$

Thus, by (2.14), (2.25), and the trace theorem, it follows that

$$\begin{cases} \|\psi(0,\cdot)\|_{L^2(\mathbb{B}_1)} \le C\big(\|\nabla \widetilde{f}_0\|_{L^2(\mathbb{B}_1)} + \|\widetilde{f}_1\|_{L^2(\mathbb{B}_1)}\big), \\ \|\psi(0,\cdot)\|_{L^2(\mathbb{S}^1)} \le C\big(\|\nabla \widetilde{f}_0\|_{L^2(\mathbb{B}_1)} + \|\widetilde{f}_1\|_{L^2(\mathbb{B}_1)}\big). \end{cases} \tag{2.26}$$

Finally, combining (2.24) and (2.26), we arrive at

$$T\Big(\|\nabla\psi(T,\cdot)\|^2_{L^2(\mathbb{B}_1)} + \|\psi_t(T,\cdot)\|^2_{L^2(\mathbb{B}_1)}\Big) \le C\Big(\|\widetilde{f}_0\|^2_{H^1(\mathbb{B}_1)} + \|\widetilde{f}_1\|^2_{L^2(\mathbb{B}_1)}\Big).$$

We find that (2.12) is satisfied, provided $T_0 = 2C$ and $\lambda_0 = \frac{1}{2}$. Then the proof of Lemma 2.3 is complete. □

REMARK 2.4. *It is obvious that the conclusion in Lemma* 2.3 *is also valid for arbitrary time $T > 0$ if λ_0 is replaced by $\frac{C_0}{T}$ for some constant $C_0 > 0$.*

We also need the following improved regularity.

THEOREM 2.5. *Consider the initial-boundary value problem* (2.11). *Suppose that $f_0 \in H^{s+1}(\Omega_0)$, $f_1 \in H^s(\Omega_0)$, $s \ge 1$. Then there exists a positive constant T_1, such that if $T \ge T_1$, then*

$$\|\nabla\psi(T,\cdot)\|^2_{H^s(\Omega_0)} + \sum_{j=1}^{s+1} \|\partial_t^j \psi(T,\cdot)\|^2_{H^{s+1-j}(\Omega_0)} \le \lambda_1\big(\|\widetilde{f}_0\|^2_{H^{s+1}(\mathbb{B}_1)} + \|\widetilde{f}_1\|^2_{H^s(\mathbb{B}_1)}\big) \tag{2.27}$$

holds for some $\lambda_1 \in (0,1)$.

Proof. This follows from Lemma 2.3 and the interior elliptic estimates. In fact, applying ∂_t^l, $1 \le l \le s$, to system (2.11) and then using Lemma 2.3, we deduce that

$$\begin{aligned} &\|\partial_t^{s+1}\psi(T,\cdot)\|^2_{L^2(\mathbb{B}_1)} + \sum_{j=1}^{s} \|\partial_t^j\psi(T,\cdot)\|^2_{H^1(\mathbb{B}_1)} + \|\nabla\psi(T,\cdot)\|^2_{L^2(\mathbb{B}_1)} \\ &\le C\lambda_0\big(\|\widetilde{f}_0\|^2_{H^{s+1}(\mathbb{B}_1)} + \|\widetilde{f}_1\|^2_{H^s(\mathbb{B}_1)}\big). \end{aligned}$$

On the other hand, by regarding the wave equation as an elliptic equation

$$\Delta\psi = \psi_{tt}$$

and using the interior elliptic estimates, we have

$$\begin{cases} \|\partial_t^{s-1}\psi(T,\cdot)\|^2_{H^2(\mathbb{B}_{1-\frac{1}{s}d})} \le C\|\partial_t^{s+1}\psi(T,\cdot)\|^2_{L^2(\mathbb{B}_1)} \\ \qquad + C\|\partial_t^{s-1}\psi(T,\cdot)\|^2_{L^2(\mathbb{B}_1)}, \\ \|\partial_t^{s-2}\psi(T,\cdot)\|^2_{H^3(\mathbb{B}_{1-\frac{2}{s}d})} \le C\|\partial_t^{s}\psi(T,\cdot)\|^2_{H^1(\mathbb{B}_{1-\frac{1}{s}d})} \\ \qquad + C\|\partial_t^{s-2}\psi(T,\cdot)\|^2_{L^2(\mathbb{B}_{1-\frac{1}{s}d})}, \\ \|\partial_t^{s-3}\psi(T,\cdot)\|^2_{H^4(\mathbb{B}_{1-\frac{3}{s}d})} \le C\|\partial_t^{s-1}\psi(T,\cdot)\|^2_{H^2(\mathbb{B}_{1-\frac{2}{s}d})} \\ \qquad + C\|\partial_t^{s-3}\psi(T,\cdot)\|^2_{L^2(\mathbb{B}_{1-\frac{2}{s}d})}, \\ \quad \cdots\cdots, \\ \|\nabla\psi(T,\cdot)\|^2_{H^s(\Omega_0)} \le C\|\partial_t^2\psi(T,\cdot)\|^2_{H^{s-1}(\mathbb{B}_{1-\frac{s-1}{s}d})} \\ \qquad + C\|\nabla\psi(T,\cdot)\|^2_{L^2(\mathbb{B}_{1-\frac{s-1}{s}d})}, \end{cases} \tag{2.28}$$

where $d = \mathrm{dist}(\mathbb{S}^1, \partial\Omega_0)$, $\mathrm{dist}(\cdot,\cdot)$ being the standard distance function. Then (2.27) follows by selecting a small enough λ_0. □

REMARK 2.6. *It is obvious that the conclusion in Theorem* 2.5 *is also valid for arbitrary time $T > 0$ if λ_1 is replaced by $\frac{C_1}{T}$ for some constant $C_1 > 0$.*

At last, we construct the exact boundary control for the linear wave equation (2.1) in two space dimensions.

THEOREM 2.7. *Consider the linear wave equation* (2.1) *in two space dimensions. Then, for any given initial data (f_0, f_1) and final data (g_0, g_1) with $f_0, g_0 \in H^{s+1}(\Omega_0)$, $f_1, g_1 \in H^s(\Omega_0)$, $s \geq 1$, there exist a positive constant T_0 and a boundary control function h such that the mixed initial-boundary value problem* (2.1), (2.2) *and* (1.6) *admits a unique $\cap_{j=0}^{s+1} C^j\big([0,T]; H^{s+1-j}(\Omega_0)\big)$ solution $u = u(t,x)$ on the domain $R(T)$ which arrives at the final condition* (1.7), *provided $T > T_0$. Moreover, we have*

$$\sup_{0\leq t\leq T}\sum_{j=0}^{s+1}\|\partial_t^j u\|^2_{H^{s+1-j}(\Omega_0)} \leq C\big(\|(\widetilde{f}_0,\widetilde{g}_0)\|^2_{H^{s+1}(\mathbb{B}_1)} + \|(\widetilde{f}_1,\widetilde{g}_1)\|^2_{H^s(\mathbb{B}_1)}\big). \tag{2.29}$$

Proof. For $m \geq 1$, define

$$u^{(m)} = \sum_{k=1}^{m}(-1)^{k-1}\big(v^{(k)} + w^{(k)}\big), \tag{2.30}$$

where $v^{(1)}$ and $w^{(1)}$ are defined as the solutions of the following linear initial-boundary value problems:

$$\begin{cases} \Box v^{(1)} = 0, \quad 0 \leq t \leq T, \quad x \in \mathbf{B}_1, \\ v_t^{(1)} + v_r^{(1)} = 0, \quad 0 \leq t \leq T, \quad x \in \mathbf{S}^1, \\ v^{(1)}(0,x) = \widetilde{f}_0, \quad v_t^{(1)}(0,x) = \widetilde{f}_1, \quad x \in \mathbf{B}_1, \end{cases} \tag{2.31}$$

and

$$\begin{cases} \Box w^{(1)} = 0, \quad 0 \leq t \leq T, \quad x \in \mathbf{B}_1, \\ -w_t^{(1)} + w_r^{(1)} = 0, \quad 0 \leq t \leq T, \quad x \in \mathbf{S}^1, \\ w^{(1)}(T,x) = \widetilde{g}_0, \quad w_t^{(1)}(T,x) = \widetilde{g}_1, \quad x \in \mathbf{B}_1. \end{cases} \tag{2.32}$$

For $k \geq 2$, $v^{(k)}$ and $w^{(k)}$ are inductively defined as the solutions of the following linear initial-boundary value problems:

$$\begin{cases} \Box v^{(k)} = 0, \quad 0 \leq t \leq T, \quad x \in \mathbf{B}_1, \\ v_t^{(k)} + v_r^{(k)} = 0, \quad 0 \leq t \leq T, \quad x \in \mathbf{S}^1, \\ v^{(k)}(0,x) = [\chi w^{(k-1)}]\tilde{}\,(0,x), \quad v_t^{(k)}(0,x) = [\chi w_t^{(k-1)}]\tilde{}\,(0,x), \quad x \in \mathbf{B}_1, \end{cases} \tag{2.33}$$

and

$$\begin{cases} \Box w^{(k)} = 0, \quad 0 \leq t \leq T, \quad x \in \mathbf{B}_1, \\ -w_t^{(k)} + w_r^{(k)} = 0, \quad 0 \leq t \leq T, \quad x \in \mathbf{S}^1, \\ w^{(k)}(T,x) = [\chi v^{(k-1)}]\tilde{}\,(T,x), \quad w_t^{(k)}(T,x) = [\chi v_t^{(k-1)}]\tilde{}\,(T,x), \quad x \in \mathbf{B}_1, \end{cases} \tag{2.34}$$

where χ is the characteristic function defined as

$$\begin{cases} \chi \equiv 1, & x \in \Omega_0, \\ \chi \equiv 0, & x \in \mathbb{B}_1 \setminus \Omega_0, \end{cases}$$

and $[\cdot]^{\sim}$ represents the extension operator defined at the beginning of section 2.

First of all, we observe that

$$\Box u^{(m)} = 0, \quad 0 < t < T, \quad x \in \Omega_0, \tag{2.35}$$

and

$$\begin{cases} u^{(m)}(0,x) = f_0 + (-1)^{m-1} w^{(m)}(0,x), \\ u_t^{(m)}(0,x) = f_1 + (-1)^{m-1} w_t^{(m)}(0,x), \\ u^{(m)}(T,x) = g_0 + (-1)^{m-1} v^{(m)}(T,x), \\ u_t^{(m)}(T,x) = g_1 + (-1)^{m-1} v_t^{(m)}(T,x) \end{cases} \tag{2.36}$$

for $m \geq 1$.

To show that $u^{(m)}$ defined in (2.30) is convergent, together with

$$\begin{cases} \big(u_t^{(m)}(0,x), u_t^{(m)}(T,x)\big) \longrightarrow (f_1, g_1) \text{ in } H^s(\Omega_0), \\ \big(u^{(m)}(0,x), u^{(m)}(T,x)\big) \longrightarrow (f_0, g_0) \text{ in } H^{s+1}(\Omega_0), \end{cases} \tag{2.37}$$

as $m \longrightarrow \infty$, we need the following lemma.

LEMMA 2.8. *Let $v^{(k)}$ and $w^{(k)}$ be defined as the solutions of systems* (2.31)–(2.34); *$s \geq 1$ is an integer. Then the estimates*

$$\begin{aligned} &\sum_{j=0}^{s+1} \Big(\|\partial_t^j v^{(k)}(T,\cdot)\|_{H^{s+1-j}(\Omega_0)} + \|\partial_t^j w^{(k)}(0,\cdot)\|_{H^{s+1-j}(\Omega_0)} \Big) \\ &\leq \frac{1}{8} \sum_{j=0}^{s+1} \Big(\|\partial_t^j v^{(k-1)}(T,x)\|_{H^{s+1-j}(\Omega_0)} + \|\partial_t^j w^{(k-1)}(0,x)\|_{H^{s+1-j}(\Omega_0)} \\ &\qquad + \|\partial_t^j v^{(k-2)}(T,x)\|_{H^{s+1-j}(\Omega_0)} + \|\partial_t^j w^{(k-2)}(0,x)\|_{H^{s+1-j}(\Omega_0)} \Big) \end{aligned}$$

hold, provided that λ_1 is small enough and T is big enough.

Proof. Integrate (2.31) with respect to x on the unit ball $\mathbb{B}_1$ to yield

$$\frac{d}{dt}\Big[\int_{\mathbb{B}_1} \partial_t v^{(k)} dx + \int_{\mathbb{S}^1} v^{(k)} d\sigma_y \Big] = 0.$$

Then, by integrating the above equality with respect to t from 0 to T, we have

$$\begin{aligned} &\|\nabla v^{(k)}(T,\cdot)\|_{L^2(\mathbb{B}_1)} + \Big| \int_{\mathbb{S}^1} v^{(k)}(T,x) d\sigma_y \Big| \\ &\leq \|\nabla v^{(k)}(T,\cdot)\|_{L^2(\mathbb{B}_1)} + \Big| \int_{\mathbb{B}_1} \partial_t v^{(k)}(T,x) dx \Big| \\ &\quad + \Big| \int_{\mathbb{S}^1} v^{(k)}(0,x) d\sigma_y + \int_{\mathbb{B}_1} \partial_t v^{(k)}(0,x) dx \Big| \\ &\leq \|\nabla v^{(k)}(T,\cdot)\|_{L^2(\mathbb{B}_1)} + C\|\partial_t v^{(k)}(T,x)\|_{L^2(\mathbb{B}_1)} \\ &\quad + \Big| \int_{\mathbb{S}^1} [\chi w^{(k-1)}]^{\sim}(0,x) d\sigma_y + \int_{\mathbb{B}_1} [\chi \partial_t w^{(k-1)}]^{\sim}(0,x) dx \Big| \\ &= \|\nabla v^{(k)}(T,\cdot)\|_{L^2(\mathbb{B}_1)} + C\|\partial_t v^{(k)}(T,x)\|_{L^2(\mathbb{B}_1)} \end{aligned}$$

$$
\begin{aligned}
&+\Big|\int_{\mathbb{B}_1}[\chi\partial_t w^{(k-1)}]\tilde{}\,(0,x)dx\Big|\\
&\le C\Big(\|\nabla v^{(k)}(T,\cdot)\|_{L^2(\mathbb{B}_1)}+\|\partial_t v^{(k)}(T,x)\|_{L^2(\mathbb{B}_1)}+\ \|[\chi\partial_t w^{(k-1)}]\tilde{}\,(0,x)\|_{L^2(\mathbb{B}_1)}\Big)\\
&\le C\Big(\|\nabla v^{(k)}(T,\cdot)\|_{L^2(\mathbb{B}_1)}+\|\partial_t v^{(k)}(T,x)\|_{L^2(\mathbb{B}_1)}+\ \|\partial_t w^{(k-1)}(0,x)\|_{L^2(\mathbb{B}_1)}\Big).
\end{aligned}
$$

By (2.12), Remark 2.4, and (initial or final data in) (2.31)–(2.34), we arrive at

$$
\begin{aligned}
&\|\nabla v^{(k)}(T,\cdot)\|_{L^2(\mathbb{B}_1)}+\Big|\int_{\mathbb{S}^1}v^{(k)}(T,x)d\sigma_y\Big|\\
&\le\frac{C}{T}\Big(\|v^{(k)}(0,x)\|_{H^1(\mathbb{B}_1)}+\|\partial_t v^{(k)}(0,x)\|_{L^2(\mathbb{B}_1)}\\
&\qquad+\|w^{(k-1)}(T,x)\|_{H^1(\mathbb{B}_1)}+\|\partial_t w^{(k-1)}(T,x)\|_{L^2(\mathbb{B}_1)}\Big)\\
&=\frac{C}{T}\Big(\|[\chi w^{(k-1)}]\tilde{}\,(0,x)\|_{H^1(\mathbb{B}_1)}+\|[\chi\partial_t w^{(k-1)}]\tilde{}\,(0,x)\|_{L^2(\mathbb{B}_1)}\\
&\qquad+\|[\chi v^{(k-2)}]\tilde{}\,(T,x)\|_{H^1(\mathbb{B}_1)}+\|[\chi\partial_t v^{(k-2)}]\tilde{}\,(T,x)\|_{L^2(\mathbb{B}_1)}\Big)\\
&\le\frac{C}{T}\Big(\|w^{(k-1)}(0,x)\|_{H^1(\Omega_0)}+\|\partial_t w^{(k-1)}(0,x)\|_{L^2(\Omega_0)}\\
&\qquad+\|v^{(k-2)}(T,x)\|_{H^1(\Omega_0)}+\|\partial_t v^{(k-2)}(T,x)\|_{L^2(\Omega_0)}\Big).
\end{aligned}
\tag{2.38}
$$

The same results are also valid for $\|\nabla w^{(k)}(0,\cdot)\|_{L^2(\mathbb{B}_1)}+|\int_{\mathbb{S}^1}w^{(k)}(0,x)d\sigma_y|$. Thus, by Poincaré's inequality, we have

$$
\begin{aligned}
&\|v^{(k)}(T,\cdot)\|_{L^2(\mathbb{B}_1)}+\|w^{(k)}(0,\cdot)\|_{L^2(\mathbb{B}_1)}\\
&\le C\Big(\|\nabla v^{(k)}(T,\cdot)\|_{L^2(\mathbb{B}_1)}+\Big|\int_{\mathbb{S}^1}v^{(k)}(T,x)d\sigma_y\Big|\Big)\\
&\quad+C\Big(\|\nabla w^{(k)}(0,\cdot)\|_{L^2(\mathbb{B}_1)}+\Big|\int_{\mathbb{S}^1}w^{(k)}(0,x)d\sigma_y\Big|\Big)\\
&\le\frac{C}{T}\Big(\|w^{(k-1)}(0,x)\|_{H^1(\Omega_0)}+\|\partial_t w^{(k-1)}(0,x)\|_{L^2(\Omega_0)}\\
&\qquad+\|v^{(k-2)}(T,x)\|_{H^1(\Omega_0)}+\|\partial_t v^{(k-2)}(T,x)\|_{L^2(\Omega_0)}\\
&\qquad+\|v^{(k-1)}(T,x)\|_{H^1(\Omega_0)}+\|\partial_t v^{(k-1)}(T,x)\|_{L^2(\Omega_0)}\\
&\qquad+\|w^{(k-2)}(0,x)\|_{H^1(\Omega_0)}+\|\partial_t w^{(k-2)}(0,x)\|_{L^2(\Omega_0)}\Big).
\end{aligned}
\tag{2.39}
$$

The combination of (2.39) and (2.27) gives

$$
\begin{aligned}
&\sum_{j=0}^{s+1}\Big(\|\partial_t^j v^{(k)}(T,\cdot)\|_{H^{s+1-j}(\Omega_0)}+\|\partial_t^j w^{(k)}(0,\cdot)\|_{H^{s+1-j}(\Omega_0)}\Big)\\
&\le\frac{1}{8}\sum_{j=0}^{s+1}\Big(\|\partial_t^j w^{(k-1)}(0,x)\|_{H^{s+1-j}(\Omega_0)}+\|\partial_t^j v^{(k-1)}(T,x)\|_{H^{s+1-j}(\Omega_0)}\\
&\qquad+\|\partial_t^j v^{(k-2)}(T,x)\|_{H^{s+1-j}(\Omega_0)}+\|\partial_t^j w^{(k-2)}(0,x)\|_{H^{s+1-j}(\Omega_0)}\Big),
\end{aligned}
$$

provided that λ_1 is small enough and T is big enough. □

Let us continue the proof of Theorem 2.7. By Lemma 2.8, we have

$$\begin{aligned}(2.40)\qquad &\sum_{j=0}^{s+1}\Big(\|\partial_t^j v^{(k)}(T,\cdot)\|^2_{H^{s+1-j}(\Omega_0)}+\|\partial_t^j w^{(k)}(0,\cdot)\|^2_{H^{s+1-j}(\Omega_0)}\Big)\\ &\le \frac{C}{2^k}\Big(\|(f_0,g_0)\|^2_{H^{s+1}(\Omega_0)}+\|(f_1,g_1)\|^2_{H^s(\Omega_0)}\Big)\end{aligned}$$

for any $k\ge 1$. Consequently, it follows that

$$(2.41)\qquad \begin{cases}\|\partial_t w^{(m)}(0,x)\|_{H^s(\Omega_0)}+\|\partial_t v^{(m)}(T,x)\|_{H^s(\Omega_0)}\longrightarrow 0,\\ \|w^{(m)}(0,x)\|_{H^{s+1}(\Omega_0)}+\|v^{(m)}(T,x)\|_{H^{s+1}(\Omega_0)}\longrightarrow 0\end{cases}$$

as $m\to+\infty$, which proves (2.37).

On the other hand, (2.41) is also valid with T being replaced by t for $0\le t\le T$ by an argument similar to that above. Thus, the series $\{u^{(m)}\}$ is convergent to a limit function $u(t,x)$. And, moreover, by (2.35), (2.40), and (2.37), the limit function $u(t,x)$ satisfies the linear wave equation (2.1), estimates (2.29), and verifies the initial data (1.6) and final data (1.7). Thus, if we let

$$(2.42)\qquad h(t,x)=u(t,x)|_{\partial\Omega_0},$$

then $h(t,x)$ is the desired boundary control. □

3. Local exact boundary control for semilinear wave equations. In this section we intend to construct the boundary controls for the semilinear wave equations (1.1). Applying the contraction mapping theorem, we can extend the results for the linear wave equations in section 2 to the semilinear case and prove Theorems 1.1 and 1.3.

Let $s\ge 2$. Define

$$\begin{aligned}\Sigma_\theta=\Big\{&v:[0,T]\times\Omega_0\to\mathbb{R}\big|v(0,x)=f_0,v_t(0,x)=f_1,\\ &(3.1)\qquad v(T,x)=g_0,v_t(T,x)=g_1,\mathcal{D}_\Sigma(v)\le\theta\Big\},\end{aligned}$$

where

$$(3.2)\qquad \mathcal{D}_\Sigma(v)=\sup_{0\le t\le T}\Big(\sum_{j=0}^{s+1}\|\partial_t^j v(t,\cdot)\|^2_{H^{s+1-j}(\Omega_0)}\Big)^{\frac12},$$

and $T>2$ in three space dimensions and $T>T_0$, with T_0 being determined in Theorem 2.7 in two space dimensions. We want to find a map

$$(3.3)\qquad \Pi_\Sigma: v\to u=\Pi_\Sigma v,$$

such that Π_Σ is a strict contraction from Σ_θ to itself, provided θ is small enough.

Let the operator Π_Σ be defined as follows. For $n=2$ or 3, given a function $v\in\Sigma_\theta$, let φ solve the Cauchy problem

$$(3.4)\qquad \begin{cases}\Box\varphi=F(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v}),\quad 0\le t\le T,\quad x\in\mathbb{B}_1,\\ \varphi(t,x)=0,\quad 0\le t\le T,\quad x\in\mathbb{S}^{n-1},\\ \varphi(0,x)=0,\quad \varphi_t(0,x)=0,\quad x\in\mathbb{B}_1.\end{cases}$$

Then, by using $\varphi(T,x)$ and $\varphi_t(T,x)$, we define ϕ via the following exact boundary controllability problem:

$$\begin{cases} \Box\phi = 0, \quad 0 \le t \le T, \quad x \in \Omega_0, \\ \phi(0,x) = f_0, \quad \phi_t(0,x) = f_1, \quad x \in \Omega_0, \\ \phi(T,x) = g_0 - \varphi(T,x), \quad \phi_t(T,x) = g_1 - \varphi_t(T,x), \quad x \in \Omega_0. \end{cases} \tag{3.5}$$

The theory set forth in section 2 ensures that ϕ is well-defined. At last, for $(t,x) \in [0,T] \times \Omega_0$, we define

$$\Pi_\Sigma v(t,x) = \varphi(t,x) + \phi(t,x). \tag{3.6}$$

We claim the following.

LEMMA 3.1. *Π_Σ is a strict contraction from Σ_θ to Σ_θ, provided that $\theta > 0$ is sufficiently small.*

To complete the proof of the above lemma, we need the following lemma.

LEMMA 3.2. *Suppose that $H = H(t,x,w)$ is a sufficiently smooth function of its arguments with $H(t,x,0) = 0$ and $\Omega_0 \subseteq \mathbb{R}^n$ is a bounded domain with smooth boundary. Then, for any given integer $s \ge 0$ and integer $m > 0$, if a smooth vector function $w = w(t,x) \in \mathbb{R}^m$ satisfies*

$$w(t,\cdot) \in H^s(\Omega_0) \quad \text{and} \quad \|w(t,\cdot)\|_{L^\infty(\Omega_0)} \le 1,$$

then the composite function $H(t,x,w) \in H^s(\Omega_0)$ and satisfies

$$\|H\big(t,\cdot,w(\cdot)\big)\|_{H^s(\Omega_0)} \le C(T)\|w(t,\cdot)\|_{H^s(\Omega_0)}$$

for $0 \le t \le T$.

The proof of Lemma 3.2 relies on the following proposition (see [20, Chapter 1, Proposition 3.9] or [11, Chapter 1, Theorem 4.3]).

PROPOSITION 3.3. *Suppose that $H_0 = H_0(w)$ is a sufficiently smooth function with $H_0(0) = 0$. Then, for any given integer $s \ge 0$ and integer $m > 0$, if a smooth vector function $w = w(x) \in \mathbb{R}^m$ satisfies*

$$w(\cdot) \in H^s(\mathbb{R}^n) \quad \text{and} \quad \|w\|_{L^\infty(\mathbb{R}^n)} \le 1,$$

then the composite function $H_0(w) \in H^s(\mathbb{R}^n)$ and satisfies

$$\|H_0\big(w(\cdot)\big)\|_{H^s(\mathbb{R}^n)} \le C\|w(\cdot)\|_{H^s(\mathbb{R}^n)},$$

where the constant $C > 0$ depends only on H_0.

Now we apply Proposition 3.3 to prove Lemma 3.2. First of all, we generalize Proposition 3.3 to the case when $H_0 = H_0\big(t,x,w(x)\big)$ with $H_0(t,x,0) = 0$ and $H_0(t,x,w) = 0$ for $|x| > 2\mathrm{diam}(\Omega_0)$. In fact, it is obvious that

$$\partial_t^\mu \nabla_x^{(l_0)} H_0(t,x,0) = 0,$$

where, in the context of the proof of Lemma 3.2, $(l_0),(l_1),\dots,(k_0),(k_1),\dots$ are used to represent multi-indices. Then, by the chain rule, the following holds:

$$\begin{aligned} &\|\nabla_x^{(l_0)}\big[H_0\big(t,\cdot,w(\cdot)\big)\big]\|_{L^2(\mathbb{R}^n)} \\ &\le C \sum_{(l_1)+(l_2)+\cdots+(l_m)\le(l_0)} \Big\{\|\nabla_x^{(l_1)} w_1(\cdot)\nabla_x^{(l_2)} w_2(\cdot)\cdots\nabla_x^{(l_m)} w_m(\cdot)\|_{L^2(\mathbb{R}^n)} \\ &\quad \times \sum_{|(k_0)|\le|(l_0)|,|(k_0)|+|(k_1)|\le|(l_0)|} \|\nabla_w^{(k_0)}[\nabla_x^{(k_1)} H_0]\big(t,\cdot,w(\cdot)\big)\|_{L^\infty(\mathbb{R}^n)}\Big\}. \end{aligned}$$

Since $H_0(t,x,w)$ is smooth enough, $\|w\|_{L^\infty(\mathbb{R}^n)} \le 1$, and $H(t,x,w) = 0$ for $|x| > 2\mathrm{diam}(\Omega_0)$, we have

$$\begin{aligned}&\|\nabla_x^{(l_0)}\big[H_0\big(t,\cdot,w(\cdot)\big)\big]\|_{L^2(\mathbb{R}^n)}\\ &\le C(T)\sum_{(l_1)+(l_2)+\cdots+(l_m)\le(l_0)}\|\nabla_x^{(l_1)}w_1(\cdot)\nabla_x^{(l_2)}w_2(\cdot)\cdots\nabla_x^{(l_m)}w_m(\cdot)\|_{L^2(\mathbb{R}^n)}.\end{aligned}$$

Then the desired generalization follows from the following lemma (see [20, Chapter 1, Lemma 3.10]).

LEMMA 3.4. *Let* $s = \sum_{i=1}^m |(l_i)|$, $|(l_i)| > 0$. *Then*

$$\begin{aligned}&\|\nabla_x^{(l_1)}w_1(\cdot)\nabla_x^{(l_2)}w_2(\cdot)\cdots\nabla_x^{(l_m)}w_m(\cdot)\|_{L^2(\mathbb{R}^n)}\\ &\le C\sum_{1\le k\le m}\frac{\|w_1(\cdot)\|_{L^\infty(\mathbb{R}^n)}\|w_2(\cdot)\|_{L^\infty(\mathbb{R}^n)}\cdots\|w_m(\cdot)\|_{L^\infty(\mathbb{R}^n)}}{\|w_k(\cdot)\|_{L^\infty(\mathbb{R}^n)}}\|w(\cdot)\|_{H^s(\mathbb{R}^n)}\end{aligned}$$

holds if the right-hand side is bounded.

Next, for any function $w \in \{w | w(\cdot) \in H^s(\Omega_0),\ \|w\|_{L^\infty(\Omega_0)} \le 1\}$, we can use the extension operator to get $\widetilde{w}$ (see (2.3) and (2.4)) such that

$$\|\widetilde{w}(\cdot)\|_{H^s(\mathbb{R}^n)} \le C_s\|w(\cdot)\|_{H^s(\Omega_0)},$$

where C_s is independent of w for $w \in \{w | w(\cdot) \in H^s(\Omega_0),\ \|w\|_{L^\infty(\Omega_0)} \le 1\}$. Thus, we have

$$\begin{aligned}&\|H_0\big(t,\cdot,w(\cdot)\big)\|_{H^s(\Omega_0)} \le \|H_0\big(t,\cdot,\widetilde{w}(\cdot)\big)\|_{H^s(\mathbb{R}^n)}\\ &\le C(T)\|\widetilde{w}(\cdot)\|_{H^s(\mathbb{R}^n)} \le C(T)C_s\|w(\cdot)\|_{H^s(\Omega_0)}.\end{aligned}$$

Note that the constants in the above inequalities are uniform for functions with bounded $H^s(\Omega_0)$ and $L^\infty(\Omega_0)$ norms. Consequently, they are also valid for $w = w(t,x)$, which satisfies the restrictions in Proposition 3.3. We complete the proof of Lemma 3.2.

Similarly, we also have the following lemma.

LEMMA 3.5. *Suppose that* $H = H(t,x,w)$ *is a sufficiently smooth function of its arguments with* $H(t,x,0) = 0$ *and* $\nabla_w H(t,x,0) = 0$, *and* $\Omega_0 \subseteq \mathbb{R}^n$ *is a bounded domain with smooth boundary. Then, for any given integer* $s \ge 0$ *and integer* $m > 0$, *if a smooth vector function* $w = w(t,x) \in \mathbb{R}^m$ *satisfies*

$$w(t,\cdot) \in H^s(\Omega_0) \quad \textit{and} \quad \|w(t,\cdot)\|_{L^\infty(\Omega_0)} \le 1,$$

then the composite function $H(t,x,w) \in H^s(\Omega_0)$ *and satisfies*

$$\|H\big(t,\cdot,w(\cdot)\big)\|_{H^s(\Omega_0)} \le C(T)\|w(t,\cdot)\|_{L^\infty(\Omega_0)}\|w(t,\cdot)\|_{H^s(\Omega_0)}$$

for $0 \le t \le T$.

LEMMA 3.6. *Suppose that* $H = H(t,x,w)$ *is a sufficiently smooth function of its arguments with* $H(t,x,0) = 0$ *and* $\nabla_w H(t,x,0) = 0$, *and* $\Omega_0 \subseteq \mathbb{R}^n$ *is a bounded domain with smooth boundary. Then, for any given integer* $s \ge 0$ *and integer* $m > 0$ *if two smooth vector functions* $w^1 = w^1(t,x) \in \mathbb{R}^m$, $w^2 = w^2(t,x) \in \mathbb{R}^m$ *satisfy*

$$(w^1, w^2)(t,\cdot) \in H^s(\Omega_0) \quad \textit{and} \quad \|(w^1, w^2)(t,\cdot)\|_{L^\infty(\Omega_0)} \le 1,$$

then we have

$$\begin{aligned}&\|H\big(t,\cdot,w^1(\cdot)\big)-H\big(t,\cdot,w^2(\cdot)\big)\|_{H^s(\Omega_0)}\\&\le C(T)\big(\|w^1(t,\cdot)\|_{L^\infty(\Omega_0)}+\|w^2(t,\cdot)\|_{L^\infty(\Omega_0)}\big)\|w^1(t,\cdot)-w^2(t,\cdot)\|_{H^s(\Omega_0)}\end{aligned}$$

for $0\le t\le T$.

Lemmas 3.5 and 3.6 can be proved by arguments similar to those of Lemma 3.2. We refer the reader to [11, Chapter 1, Corollary 4.3 and Theorem 4.6] for details.

Now let us begin to prove Lemma 3.1.

Proof. First of all, it is rather easy to see that

$$\begin{cases}u(0,x)=f_0, \quad u_t(0,x)=f_1,\\ u(T,x)=g_0, \quad u_t(T,x)=g_1\end{cases}\tag{3.7}$$

for $x\in\Omega_0$.

To confirm that $u=\Pi_\Sigma v\in\Sigma_\theta$, we need to go to estimate $\mathcal{D}_\Sigma(u)$. Applying ∂_t^j to (3.4), and then taking the L^2 inner product of the resulting equation with $\partial_t^{j+1}\varphi$ for $j=0,1,\dots,s$, respectively, and then adding them up, we have

$$\begin{aligned}&\frac12\frac{d}{dt}\Big(\|\partial_t^{s+1}\varphi(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\sum_{j=1}^s\|\partial_t^j\varphi(t,\cdot)\|^2_{H^1(\mathbb{B}_1)}+\|\nabla\varphi(t,\cdot)\|^2_{L^2(\mathbb{B}_1}\Big)\\&\le C\sum_{j=0}^s\|\partial_t^{j+1}\varphi(t,\cdot)\|_{L^2(\mathbb{B}_1)}\sum_{j=0}^s\|\frac{d^j}{dt^j}F(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\|_{L^2(\mathbb{B}_1)}.\end{aligned}\tag{3.8}$$

Write $w=(\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})$. By (1.3),

$$\begin{cases}\partial_t^jF(t,x,0)=0, \quad \partial_t^j\nabla_wF(t,x,0)=0,\\ \sum_{j=2}^s\sum_{2\le|(k)|\le j}\|\partial_t^{j-|(k)|}\nabla_w^{(k)}F(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\|_{L^\infty(\mathbb{B}_1)}\le C(T)\end{cases}$$

holds, since F is smooth, where integer $j\ge0$ and (k) is a multi-index similar to that in the proof of Lemma 3.2. Thus, by (3.1), and looking at $\nabla_wF=0$, ∂_t^jF, and $\partial_t^j\nabla_wF$ as new functions appearing in Lemma 3.2 or Lemma 3.5, we have

$$\begin{aligned}&\sum_{j=0}^s\Big\|\frac{d^j}{dt^j}F(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\Big\|_{H^{s-j}(\mathbb{B}_1)}\le\|F(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\|_{H^s(\mathbb{B}_1)}\\&\quad+\Big(\|F_t(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\|_{H^{s-1}(\mathbb{B}_1)}+\|\nabla_wF(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})w_t\|_{H^{s-1}(\mathbb{B}_1)}\Big)\\&\quad+\sum_{j=2}^s\Big(\|\partial_t^jF(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\|_{H^{s-j}(\mathbb{B}_1)}+\|\partial_t^{j-1}\nabla_wF(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})w_t\|_{H^{s-j}(\mathbb{B}_1)}\\&\quad+\sum_{2\le|(k)|\le j}\|\partial_t^{j-|(k)|}\nabla_w^{(k)}F(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\partial_t^{(k)}w\|_{H^{s-j}(\mathbb{B}_1)}\Big)\\&\le C(T)\theta^2.\end{aligned}\tag{3.9}$$

Consequently,

$$\|\partial_t^{s+1}\varphi(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}+\sum_{j=1}^s\|\partial_t^j\varphi(t,\cdot)\|^2_{H^1(\mathbb{B}_1)}+\|\nabla\varphi(t,\cdot)\|^2_{L^2(\mathbb{B}_1)}\le C(T)\theta^4.\tag{3.10}$$

Invoking (3.9) and the elliptic estimates in (2.28), and by arguments similar to those in Theorem 2.5, we have

$$\sup_{0\le t\le T}\Big(\sum_{j=0}^{s+1}\|\partial_t^j\varphi(t,\cdot)\|^2_{H^{s+1-j}(\Omega_0)}\Big)^{\frac{1}{2}}\le\frac{1}{4}\theta, \tag{3.11}$$

provided that θ is sufficiently small.

It is obvious that the estimate (3.11) also holds for ϕ, provided ε_1 and ε_2 are small enough. Thus, we conclude that $\Pi_\Sigma v\in\Sigma_\theta$.

It remains to demonstrate that Π_Σ is a strict contraction. Given $v_i\in\Sigma_\theta$, $i=1,2$, let φ_i solve

$$\begin{cases}\Box\varphi_i=F(t,x,\widetilde{v}_i,\partial_t\widetilde{v}_i,\nabla\widetilde{v}_i), & 0<t<T,\quad x\in\mathbb{B}_1,\\ \varphi_i(t,x)=0, & 0\le t\le T,\quad x\in\mathbb{S}^{n-1},\\ \varphi_i(0,x)=0,\quad \partial_t\varphi_i(0,x)=0, & x\in\mathbb{B}_1.\end{cases} \tag{3.12}$$

Then, by using $\varphi_i(T,x)$ and $\partial_t\varphi(T,x)$, we define ϕ_i via the following exact boundary controllability problem:

$$\begin{cases}\Box\phi_i=0, & 0<t<T,\quad x\in\Omega_0,\\ \phi_i(0,x)=f_0,\quad \partial_t\phi_i(0,x)=f_1, & x\in\Omega_0,\\ \phi_i(T,x)=g_0-\varphi_i(T,x),\quad \partial_t\phi_i(T,x)=g_1-\partial_t\varphi_i(T,x), & x\in\Omega_0.\end{cases} \tag{3.13}$$

Of course ϕ_i may not be unique in general; we point out here that the ones we used here are constructed by Lemma 2.1 and Theorem 2.7. Let

$$u_i=\varphi_i+\phi_i.$$

We must show that

$$\mathcal{D}_\Sigma(u_1-u_2)\le\lambda\mathcal{D}_\Sigma(v_1-v_2) \tag{3.14}$$

for some positive constant $0<\lambda<1$.

To confirm (3.14), it is enough to show

$$\begin{cases}\mathcal{D}_\Sigma(\varphi_1-\varphi_2)\le\frac{\lambda}{2}\mathcal{D}_\Sigma(v_1-v_2),\\ \mathcal{D}_\Sigma(\phi_1-\phi_2)\le\frac{\lambda}{2}\mathcal{D}_\Sigma(v_1-v_2)\end{cases} \tag{3.15}$$

for some positive constant $0<\lambda<1$.

On one hand, note that

$$\begin{cases}\Box(\phi_1-\phi_2)=0, & 0<t<T,\quad x\in\Omega_0,\\ (\phi_1-\phi_2)(0,x)=0,\quad \partial_t(\phi_1-\phi_2)(0,x)=0, & x\in\Omega_0,\\ (\phi_1-\phi_2)(T,x)=(\varphi_i-\varphi_1)(T,x), & x\in\Omega_0,\\ \partial_t(\phi_1-\phi_2)(T,x)=\partial_t(\varphi_2-\varphi_1)(T,x), & x\in\Omega_0.\end{cases}$$

By Lemma 2.1 and Theorem 2.7, we have

$$\begin{aligned}\mathcal{D}_\Sigma(\phi_1-\phi_2)&\le C\Big(\|\partial_t(\varphi_2-\varphi_1)(T,\cdot)\|^2_{H^s(\Omega_0)}+\|(\varphi_2-\varphi_1)(T,\cdot)\|^2_{H^{s+1}(\Omega_0)}\Big)\\ &\le C\mathcal{D}_\Sigma(\varphi_1-\varphi_2).\end{aligned}$$

Thus, to confirm (3.15), it is enough to show that

$$\mathcal{D}_\Sigma(\varphi_1-\varphi_2)\le\frac{\lambda}{M}\mathcal{D}_\Sigma(v_1-v_2) \tag{3.16}$$

for an appropriately large constant M.

On the other hand, it is easy to see that

$$\begin{cases}\Box(\varphi_1-\varphi_2)=F(t,x,\widetilde{v}_1,\partial_t\widetilde{v}_1,\nabla\widetilde{v}_1)-F(t,x,\widetilde{v}_2,\partial_t\widetilde{v}_2,\nabla\widetilde{v}_2), & 0<t<T,\quad x\in\mathbb{B}_1,\\ (\varphi_1-\varphi_2)(t,x)=0, & 0\le t\le T,\quad x\in\mathbb{S}^{n-1},\\ (\varphi_1-\varphi_2)(0,x)=0,\quad \partial_t(\varphi_1-\varphi_2)(0,x)=0, & x\in\mathbb{B}_1.\end{cases}$$

By standard energy estimates, it is easy to verify (3.16) since (by Lemma 3.6 and arguments similar to those in (3.9))

$$\begin{aligned}&\sum_{j=0}^{s}\Big\|\frac{d^j}{dt^j}\Big(F(t,x,\widetilde{v_1},\partial_t\widetilde{v_1},\nabla\widetilde{v_1})-F(t,x,\widetilde{v_2},\partial_t\widetilde{v_2},\nabla\widetilde{v_2})\Big)\Big\|_{H^{s-j}(\mathbb{B}_1)}\\ &\le C(T)\theta\mathcal{D}_\Sigma(v_1-v_2).\end{aligned} \tag{3.17}$$

Hence, we complete the proof of Lemma 3.4. □

By the standard contraction mapping theorem, there exists a point $u\in\Sigma_\theta$, such that $u=\Pi_\Sigma u$. For $(t,x)\in[0,T]\times\partial\Omega_0$, let

$$h(t,x)=u(t,x);$$

then h is the desired control and the proofs of Theorems 1.1 and 1.3 are complete.

4. Local exact boundary controllability for quasi-linear wave equations. In this section, we study local exact boundary controllability for the quasi-linear wave equations (1.2). The argument is similar to the two-dimensional semilinear case in section 3, but much more complicated. The difficulty lies in the quasi-linearity, which leads to the invalidity of the equality (2.14). To overcome the difficulty, we introduce a shift variable below. This shift variable can only be used in the basic Morawetz energy estimates, not in the standard energy estimates or any higher-order energy estimates, because it will lead to the loss of derivatives. The proof relies on the careful estimates of both the original and shift variables.

For simplicity, we assume that $n=2$ or 3, and $F(t,x,u,u')\equiv0$ in (1.4). The following presentation can be easily generalized to the case that $F(t,x,u,u')$ is not identically zero by arguments similar to those in section 3.

Let $s\ge3$ be an integer. Define

$$\begin{aligned}\Lambda_\theta=\Big\{&v:[0,T]\times\Omega_0\to\mathbb{R}\big|v(0,x)=f_0,v_t(0,x)=f_1,\\ &v(T,x)=g_0,v_t(T,x)=g_1,\mathcal{D}_\Lambda(v)\le\theta\Big\},\end{aligned} \tag{4.1}$$

where

$$\mathcal{D}_\Lambda(v)=\sup_{0\le t\le T}\sum_{j=0}^{s+1}\Big(\|\partial_t^jv(t,\cdot)\|^2_{H^{s+1-j}(\Omega_0)}\Big)^{\frac12} \tag{4.2}$$

and with $T>0$ to be determined. We mention here that if $F(t,x,u,u')$ is not zero identically, the definitions for Λ_θ and $\mathcal{D}_\Lambda(v)$ are the same as in (4.1) and (4.2).

For any $v \in \Lambda_\theta$, let us define the series $\{\varphi^{(i)}\}$ and $\{\phi^{(i)}\}$ as follows. Let $\varphi^{(1)}$ be the solution of the linear initial-boundary value problem

$$\begin{cases} \Box\varphi^{(1)} = g_{i\alpha}(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\partial_{i\alpha}^2\varphi^{(1)}, \quad 0 \le t \le T, \quad x \in \mathbb{B}_1, \\ \varphi_t^{(1)} + \varphi_r^{(1)} = 0, \quad 0 \le t \le T, \quad x \in \mathbb{S}^{n-1}, \\ \varphi^{(1)}(0,x) = \widetilde{f}_0, \quad \varphi_t^{(1)}(0,x) = \widetilde{f}_1, \quad x \in \mathbb{B}_1, \end{cases} \tag{4.3}$$

and let $\phi^{(1)}$ be the solution of the inverted initial-boundary value problem

$$\begin{cases} \Box\phi^{(1)} = g_{i\alpha}(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\partial_{i\alpha}^2\phi^{(1)}, \quad 0 \le t \le T, \quad x \in \mathbb{B}_1, \\ \phi_t^{(1)} - \phi_r^{(1)} = 0, \quad 0 \le t \le T, \quad x \in \mathbb{S}^{n-1}, \\ \phi^{(1)}(T,x) = \widetilde{g}_0, \quad \phi_t^{(1)}(T,x) = \widetilde{g}_1, \quad x \in \mathbb{B}_1, \end{cases} \tag{4.4}$$

where we used the extension operator defined in (2.3) and (2.4). For $j \ge 2$, $\varphi^{(j)}$ is defined inductively as the solution of the linear initial-boundary value problem

$$\begin{cases} \Box\varphi^{(j)} = g_{i\alpha}(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\partial_{i\alpha}^2\varphi^{(j)}, \quad 0 \le t \le T, \quad x \in \mathbb{B}_1, \\ \varphi_t^{(j)} + \varphi_r^{(j)} = 0, \quad 0 \le t \le T, \quad x \in \mathbb{S}^{n-1}, \\ \varphi^{(j)}(0,x) = [\chi\phi^{(j-1)}]\tilde{}\,(0,x), \quad \varphi_t^{(j)}(0,x) = [\chi\phi^{(j-1)}]\tilde{}\,(0,x), \quad x \in \mathbb{B}_1, \end{cases} \tag{4.5}$$

and $\phi^{(j)}$ is defined as the solution of the inverted initial-boundary value problem

$$\begin{cases} \Box\phi^{(j)} = g_{i\alpha}(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\partial_{i\alpha}^2\phi^{(j)}, \quad 0 \le t \le T, \quad x \in \mathbb{B}_1, \\ \phi_t^{(j)} - \phi_r^{(j)} = 0, \quad 0 \le t \le T, \quad x \in \mathbb{S}^{n-1}, \\ \phi^{(j)}(T,x) = [\chi\varphi^{(j-1)}]\tilde{}\,(T,x), \quad \phi_t^{(j)}(T,x) = [\chi\varphi_t^{(j-1)}]\tilde{}\,(T,x), \quad x \in \mathbb{B}_1, \end{cases} \tag{4.6}$$

where χ is the characteristic function and $[\cdot]\tilde{}$ represents the extension operator defined in section 2.

Next, let us formally define a map $u = \Pi_\Lambda v$ by

$$\Pi_\Lambda v = \sum_{i=1}^{\infty}(-1)^{i-1}\big(\varphi^{(i)} + \phi^{(i)}\big). \tag{4.7}$$

We must show that the series is well-defined, and Π_Λ maps Λ_θ to itself and is a strict contraction with respect to the norm $\sup_{0\le t\le T}\sum_{j=0}^{s}(\|\partial_t^j v(t,\cdot)\|^2_{H^{s-j}(\Omega_0)})^{\frac{1}{2}}$, provided that ε_3 and θ are sufficiently small. Then, by the standard contraction mapping theorem, we can prove Theorem 1.6.

To verify this, we find that it is enough to prove the following theorem, and then all arguments are similar to those in sections 2 and 3.

THEOREM 4.1. *For $v \in \Lambda_\theta$, consider the following linear initial-boundary value problem:*

$$\begin{cases} \Box w = g_{i\alpha}(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\partial_{i\alpha}^2 w, \quad 0 \le t \le T, x \in \mathbb{B}_1, \\ w_t + w_r = 0, \quad 0 \le t \le T, \quad x \in \mathbb{S}^{n-1}, \\ w(0,x) = \widetilde{f}_0, \quad w_t(0,x) = \widetilde{f}_1, \quad x \in \mathbb{B}_1. \end{cases} \tag{4.8}$$

Suppose that $f_0 \in H^{s+1}(\Omega_0)$, $f_1 \in H^s(\Omega_0)$, $s \ge 3$. Then there exists a positive constant T_1, such that if $T \ge T_1$, then

$$\|\nabla w(T,\cdot)\|^2_{H^s(\Omega_0)} + \sum_{j=1}^{s+1}\|\partial_t^j w(T,\cdot)\|^2_{H^{s+1-j}(\Omega_0)} \le \lambda_2\big(\|\widetilde{f}_0\|^2_{H^{s+1}(\mathbb{B}_1)} + \|\widetilde{f}_1\|^2_{H^s(\mathbb{B}_1)}\big) \tag{4.9}$$

holds for some $\lambda_2 \in (0,1)$.

The well-posedness of classical solutions for linear system (4.8) can be established similarly by using semigroup methods as in Hughes, Kato, and Marsden [6] or energy methods based on differentiating the equation with respect to t as in Dafermos and Hrusa [3].

To complete the proof of the above theorem, we recall the following elliptic estimates involving the boundary condition of Neumann type [19].

LEMMA 4.2. *Suppose that* Ω_0 *is a bounded domain with smooth boundary* $\partial\Omega_0$ *and* $k = 0, 1, 2, \ldots$. *Given* $f \in H^k(\Omega_0)$, $g \in H^{k+\frac{1}{2}}(\partial\Omega_0)$, *and a solution* $h \in H^{k+2}(\Omega_0)$ *to the Neumann system*

$$\begin{cases} -\Delta h = f \quad on \quad \Omega_0, \\ \frac{\partial h}{\partial n} = g \quad on \quad \partial\Omega_0, \end{cases}$$

one has the estimate

$$\|h\|^2_{H^{k+2}(\Omega_0)} \le C_k\big(\|f\|^2_{H^k(\Omega_0)} + \|g\|^2_{H^{k+\frac{1}{2}}(\partial\Omega_0)} + \|h\|^2_{L^2(\Omega_0)}\big).$$

Next we return to prove Theorem 4.1.

Proof. The proof relies on the use of the underlying conservation law of the system, which is revealed by the shift technique presented below, and also on the energy estimates of Morawetz type. As the proof is rather long, we divide it into four steps.

Step 1. *A shift technique and estimates for* $\|\overline{w}\|_{L^2(\mathbb{B}_1)}$ *and* $\|\overline{w}\|_{L^2(\mathbb{S}^{n-1})}$. Let us integrate the linear wave equation in system (4.8) with respect to x on the unit ball $\mathbb{B}_1$ to yield

$$\frac{d}{dt}\int_{\mathbb{B}_1} w_t dx - \int_{\mathbb{S}^{n-1}} w_r d\sigma_y = \int_{\mathbb{B}_1} g_{i\alpha}\partial^2_{i\alpha} w dx.$$

Noting the dissipative boundary condition in system (4.8), we have

$$\frac{d}{dt}\Big[\int_{\mathbb{B}_1} w_t dx + \int_{\mathbb{S}^{n-1}} w d\sigma_y\Big] = \int_{\mathbb{B}_1} g_{i\alpha}\partial^2_{i\alpha} w dx.$$

Let

$$\mu_0 = \int_{\mathbb{B}_1} \widetilde{f}_1 dx + \int_{\mathbb{S}^{n-1}} \widetilde{f}_0 d\sigma_y = \int_{\mathbb{B}_1} \widetilde{f}_1 dx,$$

and then introduce a shift variable

$$\overline{w} = w - \frac{1}{|\mathbb{S}^{n-1}|}\Big[\mu_0 + \int_0^t \int_{\mathbb{B}_1} g_{i\alpha}\partial^2_{i\alpha} w dx d\tau\Big]. \tag{4.10}$$

We deduce that

$$\int_{\mathbb{B}_1} \overline{w}_t dx + \int_{\mathbb{S}^{n-1}} \overline{w} d\sigma_y = -\frac{|\mathbb{B}_1|}{|\mathbb{S}^{n-1}|}\int_{\mathbb{B}_1} g_{i\alpha}\partial^2_{i\alpha} w dx. \tag{4.11}$$

Moreover, a direct calculation shows that

$$\Big|\int_{\mathbb{B}_1} g_{i\alpha}\partial^2_{i\alpha} w dx\Big| \le C(T)\theta\big(\|\nabla w_t\|_{L^2(\mathbb{B}_1)} + \|\nabla^2 w\|_{L^2(\mathbb{B}_1)}\big). \tag{4.12}$$

Let us now go to estimate $\|\overline{w}\|_{L^2(\mathbb{B}_1)}$ and $\|\overline{w}\|_{L^2(\mathbb{S}^{n-1})}$. Combining (4.11) and (4.12), it is easy to get

$$\Big|\int_{\mathbb{B}_1}\overline{w}_t dx+\int_{\mathbb{S}^{n-1}}\overline{w}d\sigma_y\Big|\le C(T)\theta\big(\|\nabla w_t\|_{L^2(\mathbb{B}_1)}+\|\nabla^2 w\|_{L^2(\mathbb{B}_1)}\big). \tag{4.13}$$

Invoking the Poincaré inequality

$$\Big\|\overline{w}-\frac{1}{|\mathbb{S}^{n-1}|}\int_{\mathbb{S}^{n-1}}\overline{w}d\sigma_y\Big\|_{L^2(\mathbb{B}_1)}\le C\|\nabla\overline{w}\|_{L^2(\mathbb{B}_1)},$$

we can deduce from (4.13) that

$$\begin{aligned}&\|\overline{w}\|_{L^2(\mathbb{B}_1)}\\&\le C\|\nabla\overline{w}\|_{L^2(\mathbb{B}_1)}+C\Big|\int_{\mathbb{S}^{n-1}}\overline{w}d\sigma_y\Big|\\&\le C\big(\|\nabla\overline{w}\|_{L^2(\mathbb{B}_1)}+\|\overline{w}_t\|_{L^2(\mathbb{B}_1)}\big)\\&\quad+C(T)\theta\big(\|\nabla w_t\|_{L^2(\mathbb{B}_1)}+\|\nabla^2 w\|_{L^2(\mathbb{B}_1)}\big).\end{aligned} \tag{4.14}$$

Then, by the trace theorem, it follows that

$$\begin{aligned}&\|\overline{w}\|_{L^2(\mathbb{S}^{n-1})}\le C\|\overline{w}\|_{H^1(\mathbb{B}_1)}\\&\le C\big(\|\nabla\overline{w}\|_{L^2(\mathbb{B}_1)}+\|\overline{w}_t\|_{L^2(\mathbb{B}_1)}\big)\\&\quad+C(T)\theta\big(\|\nabla w_t\|_{L^2(\mathbb{B}_1)}+\|\nabla^2 w\|_{L^2(\mathbb{B}_1)}\big).\end{aligned} \tag{4.15}$$

Step 2. *Standard energy estimates.* The following energy estimates are standard. By taking the L^2 inner product of the linear wave equation in system (4.8) with w_t, and using Green's formula and integration by parts, we can compute as follows:

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\mathbb{B}_1}\big(|w_t|^2+|\nabla w|^2\big)dx-\int_{\mathbb{S}^{n-1}}w_t w_r dx\\&=\int_{\mathbb{B}_1}\Big(g_{i0}\partial_i\frac{|w_t|^2}{2}+\nabla_j(g_{ij}\nabla_i w w_t)\\&\qquad-\frac{d}{dx_j}g_{ij}\nabla_i w w_t-\frac{1}{2}g_{ij}\frac{d}{dt}(\nabla_i w\nabla_j w)\Big)dx\\&=-\frac{1}{2}\frac{d}{dt}\int_{\mathbb{B}_1}g_{ij}\nabla_i w\nabla_j w dx-\int_{\mathbb{B}_1}\Big(\frac{d}{dx_i}g_{i0}\frac{|w_t|^2}{2}+\frac{d}{dx_j}g_{ij}\nabla_i w w_t-\frac{1}{2}\frac{d}{dt}g_{ij}\nabla_i w\nabla_j w\Big)dx\\&\quad+\int_{\mathbb{S}^{n-1}}\Big(\frac{1}{2}g_{i0}x_i|w_t|^2+g_{ij}x_j\nabla_i w w_t\Big)d\sigma_y\\&\le-\frac{1}{2}\frac{d}{dt}\int_{\mathbb{B}_1}g_{ij}\nabla_i w\nabla_j w dx+\|\partial g_{i\alpha}\|_{L^\infty(\mathbb{B}_1)}\Big(\|w_t\|^2_{L^2(\mathbb{B}_1)}+\|\nabla w\|^2_{L^2(\mathbb{B}_1)}\Big)\\&\quad+\|g_{i\alpha}\|_{L^\infty(\mathbb{S}^{s-1})}\Big(\|w_t\|^2_{L^2(\mathbb{S}^{s-1})}+\|\Omega w\|^2_{L^2(\mathbb{S}^{s-1})}\Big).\end{aligned}$$

Keeping in mind (1.3), (1.5), and the boundary condition in (4.8), we deduce that

$$\begin{aligned}&\frac{d}{dt}\frac{1}{2}\Big[\|w_t\|^2_{L^2(\mathbb{B}_1)}+\|\nabla w\|^2_{L^2(\mathbb{B}_1)}+\int_{\mathbb{B}_1}g_{ij}\nabla_i w\nabla_j w dx\Big]+\|w_t\|^2_{L^2(\mathbb{S}^{n-1})}\\&\le C(T)\theta\big(\|w_t\|^2_{L^2(\mathbb{B}_1)}+\|\nabla w\|^2_{L^2(\mathbb{B}_1)}+\|w_t\|^2_{L^2(\mathbb{S}^{s-1})}+\|\Omega w\|^2_{L^2(\mathbb{S}^{s-1})}\big).\end{aligned} \tag{4.16}$$

Next, we do the higher-order energy estimates. The process is similar to the above. For $s \geq 3$ and $1 \leq k \leq s$, we apply the ∂_t^k derivative to the linear wave equation in system (4.8), and then take the L^2 inner product of the resulting equation with $\partial_t^{k+1} w$. Noting that $\partial_t^{k+1} w + \partial_t^k w_r = 0$ on the boundary, we finally arrive at

$$
\begin{aligned}
&\sum_{k=1}^{s}\Big\{\frac{1}{2}\frac{d}{dt}\Big(\|\partial_t^{k+1}w\|_{L^2(\mathbb{B}_1)}^2 + \|\nabla\partial_t^k w\|_{L^2(\mathbb{B}_1)}^2 \\
&\qquad + \int_{\mathbb{B}_1} g_{ij}\nabla_i\partial_t^k w\nabla_j\partial_t^k w dx\Big) + \|\partial_t^{k+1}w\|_{L^2(\mathbb{S}^{n-1})}^2\Big\} \\
&\leq C(T)\theta\sum_{k=0}^{s}\Big(\|\partial_t^{k+1}w\|_{L^2(\mathbb{S}^{s-1})}^2 + \|\Omega\partial_t^k w\|_{L^2(\mathbb{S}^{s-1})}^2\Big) \\
&\quad + C(T)\theta\sum_{k=0}^{s}\|\partial_t^{s+1-k}w\|_{H^k(\mathbb{B}_1)}^2 + C(T)\theta\|\nabla w\|_{H^s(\mathbb{B}_1)}^2.
\end{aligned}
\tag{4.17}
$$

Note that in the right-hand side of the above estimate $\|w\|_{H^{s+1}(\mathbb{B}_1)}^2$ doesn't appear.

By (4.8), we have

$$
\begin{cases}
\Delta(\partial_t^{s+1-k}w) = \partial_t^{s+3-k}w - \partial_t^{s+1-k}[g_{i\alpha}(t,x,\widetilde{v},\widetilde{v}_t,\nabla\widetilde{v})\partial_{i\alpha}^2 w], & x \in \mathbb{B}_1, \\
(\partial_t^{s+1-k}w)_r = -\partial_t^{s+2-k}w, & x \in \mathbb{S}^{n-1}.
\end{cases}
$$

Thus, proceeding from Lemma 4.2, we have

$$
\begin{aligned}
&\sum_{k=0}^{s}\|\partial_t^{s+1-k}w\|_{H^k(\mathbb{B}_1)}^2 = \|\partial_t^{s+1}w\|_{L^2(\mathbb{B}_1)}^2 \\
&\quad + \|\partial_t^s w\|_{H^1(\mathbb{B}_1)}^2 + \sum_{k=2}^{s}\|\partial_t^{s+1-k}w\|_{H^k(\mathbb{B}_1)}^2 \\
&\leq \|\partial_t^{s+1}w\|_{L^2(\mathbb{B}_1)}^2 + \|\partial_t^s w\|_{H^1(\mathbb{B}_1)}^2 + C\sum_{k=2}^{s}\Big(\|\partial_t^{s+1-k}w\|_{L^2(\mathbb{B}_1)}^2 \\
&\quad + \|\partial_t^{s+1-k}\big(w_{tt} - g_{i\alpha}\partial_{i\alpha}^2 w\big)\|_{H^{k-2}(\mathbb{B}_1)}^2 + \|\partial_t^{s+2-k}w\|_{H^{k-2+\frac{1}{2}}(\mathbb{S}^{n-1})}^2\Big).
\end{aligned}
$$

By the trace theorem, we obtain

$$
\begin{aligned}
&\sum_{k=0}^{s}\|\partial_t^{s+1-k}w\|_{H^k(\mathbb{B}_1)}^2 \\
&\leq \|\partial_t^{s+1}w\|_{L^2(\mathbb{B}_1)}^2 + \|\partial_t^s w\|_{H^1(\mathbb{B}_1)}^2 + C\sum_{k=2}^{s}\Big(\|\partial_t^{s+1-k}w\|_{L^2(\mathbb{B}_1)}^2 \\
&\quad + \|\partial_t^{s+3-k}w\|_{H^{k-2}(\mathbb{B}_1)}^2 + \|\partial_t^{s+1-k}\big(g_{i\alpha}\partial_{i\alpha}^2 w\big)\|_{H^{k-2}(\mathbb{B}_1)}^2 \\
&\quad + \|\partial_t^{s+2-k}w\|_{H^{k-1}(\mathbb{B}_1)}^2\Big) \\
&\leq C\|\partial_t^{s+1}w\|_{L^2(\mathbb{B}_1)}^2 + C\sum_{k=1}^{s}\|\partial_t^k w\|_{H^1(\mathbb{B}_1)}^2 \\
&\quad + C\sum_{k=0}^{s-1}\|\partial_t^{s+1-k}w\|_{H^k(\mathbb{B}_1)}^2 + C\|\partial_t^{s+1-k}\big(g_{i\alpha}\partial_{i\alpha}^2 w\big)\|_{H^{k-2}(\mathbb{B}_1)}^2.
\end{aligned}
$$

Thus, by (1.5), (4.1), and Lemma 3.5, we arrive at

$$\begin{aligned}
&\sum_{k=0}^{s}\|\partial_t^{s+1-k}w\|^2_{H^k(\mathbb{B}_1)}\\
&\le C\sum_{k=0}^{s-1}\|\partial_t^{s+1-k}w\|^2_{H^k(\mathbb{B}_1)}+C\|\partial_t^{s+1}w\|^2_{L^2(\mathbb{B}_1)}+C\sum_{k=1}^{s}\|\partial_t^{k}w\|^2_{H^1(\mathbb{B}_1)}\\
&\quad+C(T)\theta\Big(\sum_{k=0}^{s}\|\partial_t^{s+1-k}w\|^2_{H^k(\mathbb{B}_1)}+\|\nabla w\|^2_{H^s(\mathbb{B}_1)}\Big),
\end{aligned}$$

which leads to

$$\begin{aligned}
\sum_{k=0}^{s}\|\partial_t^{s+1-k}w\|^2_{H^k(\mathbb{B}_1)}&\le C\sum_{k=0}^{s-1}\|\partial_t^{s+1-k}w\|^2_{H^k(\mathbb{B}_1)}\\
&\quad+C\|\partial_t^{s+1}w\|^2_{L^2(\mathbb{B}_1)}+C\sum_{k=1}^{s}\|\partial_t^{k}w\|^2_{H^1(\mathbb{B}_1)}+\frac{1}{2}\|\nabla w\|^2_{H^s(\mathbb{B}_1)}
\end{aligned}$$

if we choose θ so small that $C(T)\theta<\frac{1}{2}$.

Repeating the above procedure for $s-1$ steps, we finally arrive at

$$\begin{aligned}
&\sum_{k=0}^{s}\|\partial_t^{s+1-k}w\|^2_{H^k(\mathbb{B}_1)}\\
&\le C\Big(\|\partial_t^{s+1}w\|^2_{L^2(\mathbb{B}_1)}+\sum_{k=1}^{s}\|\partial_t^{k}w\|^2_{H^1(\mathbb{B}_1)}+\|\nabla w\|^2_{H^s(\mathbb{B}_1)}\Big).
\end{aligned}\tag{4.18}$$

Similarly, by (4.14) and Lemma 4.2, we have

$$\begin{cases}
\|\nabla w\|^2_{H^1(\mathbb{B}_1)}\le C\Big(\|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)}+\|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)}\\
\quad+\|w_t\|^2_{L^2(\mathbb{B}_1)}+\|w_{tt}\|^2_{L^2(\mathbb{B}_1)}+\|\nabla w_t\|^2_{L^2(\mathbb{B}_1)}\Big),\\
\|\nabla w\|^2_{H^2(\mathbb{B}_1)}\le C\Big(\|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)}+\|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)}\\
\quad+\|w_t\|^2_{L^2(\mathbb{B}_1)}+\|w_{tt}\|^2_{H^1(\mathbb{B}_1)}+\|\nabla w_t\|^2_{H^1(\mathbb{B}_1)}\Big),\\
\quad\cdots\cdots,\\
\|\nabla w\|^2_{H^s(\mathbb{B}_1)}\le C\Big(\|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)}+\|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)}\\
\quad+\|w_t\|^2_{L^2(\mathbb{B}_1)}+\sum_{k=1}^{s}\|\partial_t^{s+1-k}w\|^2_{H^k(\mathbb{B}_1)}\Big),
\end{cases}\tag{4.19}$$

where the first estimate follows from the following calculation:

$$\begin{aligned}
&\|\nabla w\|^2_{H^1(\mathbb{B}_1)}=\|\nabla\overline{w}\|^2_{H^1(\mathbb{B}_1)}\le\|\overline{w}\|^2_{H^2(\mathbb{B}_1)}\\
&\le C\Big(\|\overline{w}\|^2_{L^2(\mathbb{B}_1)}+\|\overline{w}_r\|^2_{H^{\frac{1}{2}}(\mathbb{S}^{n-1})}+\|w_{tt}\|^2_{L^2(\mathbb{B}_1)}\Big)\\
&\quad+C(T)\theta\Big(\|\nabla w_t\|^2_{L^2(\mathbb{B}_1)}+\|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)}\Big)\\
&=C\Big(\|\overline{w}\|^2_{L^2(\mathbb{B}_1)}+\|w_r\|^2_{H^{\frac{1}{2}}(\mathbb{S}^{n-1})}+\|w_{tt}\|^2_{L^2(\mathbb{B}_1)}\Big)\\
&\quad+C(T)\theta\Big(\|\nabla w_t\|^2_{L^2(\mathbb{B}_1)}+\|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)}\Big)
\end{aligned}$$

$$\le C(T)\theta\|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)} + C\Big(\|w_{tt}\|^2_{L^2(\mathbb{B}_1)} + \|\nabla w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)} + \|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)} + \|w_t\|^2_{L^2(\mathbb{B}_1)}\Big),$$

which implies that

$$\|\nabla w\|^2_{H^1(\mathbb{B}_1)} + \|\overline{w}\|^2_{H^2(\mathbb{B}_1)} \le C\Big(\|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)} + \|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)} + \|w_{tt}\|^2_{L^2(\mathbb{B}_1)} + \|w_t\|^2_{H^1(\mathbb{B}_1)}\Big). \tag{4.20}$$

The second estimate in (4.19) can be obtained by similar argument:

$$\begin{aligned}
&\|\nabla w\|^2_{H^2(\mathbb{B}_1)} = \|\nabla\overline{w}\|^2_{H^2(\mathbb{B}_1)} \le \|\overline{w}\|^2_{H^3(\mathbb{B}_1)}\\
&\le C\Big(\|\overline{w}\|^2_{L^2(\mathbb{B}_1)} + \|\overline{w}_r\|^2_{H^{\frac{3}{2}}(\mathbb{S}^{n-1})} + \|w_{tt}\|^2_{H^1(\mathbb{B}_1)}\Big)\\
&\quad + C(T)\theta\Big(\|\nabla w_t\|^2_{H^1(\mathbb{B}_1)} + \|\nabla^2 w\|^2_{H^1(\mathbb{B}_1)}\Big)\\
&\le C\Big(\|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)} + \|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)} + \|w_t\|^2_{L^2(\mathbb{B}_1)}\\
&\qquad + \|w_{tt}\|^2_{H^1(\mathbb{B}_1)} + \|\nabla w_t\|^2_{H^1(\mathbb{B}_1)}\Big) + C(T)\theta\|\nabla w\|^2_{H^2(\mathbb{B}_1)},
\end{aligned}$$

which implies that

$$\|\nabla w\|^2_{H^2(\mathbb{B}_1)} + \|\overline{w}\|^2_{H^3(\mathbb{B}_1)} \le C\Big(\|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)} + \|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)} + \|w_{tt}\|^2_{H^1(\mathbb{B}_1)} + \|w_t\|^2_{H^2(\mathbb{B}_1)}\Big),$$

and similar arguments yield the remaining estimates in (4.19).

Combining (4.18) and (4.19), we have

$$\begin{aligned}
&\|\nabla w\|^2_{H^s(\mathbb{B}_1)} + \sum_{k=0}^{s}\|\partial_t^{s+1-k} w\|^2_{H^k(\mathbb{B}_1)}\\
&\le C\Big(\|\partial_t^{s+1} w\|^2_{L^2(\mathbb{B}_1)} + \sum_{k=1}^{s}\|\partial_t^k w\|^2_{H^1(\mathbb{B}_1)}\\
&\qquad + \|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)} + \|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)}\Big).
\end{aligned} \tag{4.21}$$

By (4.10) and (4.12), it is easy to get

$$\begin{cases}
\|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)} + \|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)} \le C\big(\|\nabla w\|^2_{L^2(\mathbb{B}_1)} + \|w_t\|^2_{L^2(\mathbb{B}_1)}\big)\\
\quad + C(T)\theta\big(\|\nabla w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)}\big),\\
\|\nabla w\|^2_{L^2(\mathbb{B}_1)} + \|w_t\|^2_{L^2(\mathbb{B}_1)} \le C\big(\|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)} + \|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)}\big)\\
\quad + C(T)\theta\big(\|\nabla w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)}\big),\\
\|\overline{w}_t\|^2_{L^2(\mathbb{S}^{n-1})} \le \|w_t\|^2_{L^2(\mathbb{S}^{n-1})} + C(T)\theta\big(\|\nabla w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)}\big),\\
\|w_t\|^2_{L^2(\mathbb{S}^{n-1})} \le \|\overline{w}_t\|^2_{L^2(\mathbb{S}^{n-1})} + C(T)\theta\big(\|\nabla w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)}\big).
\end{cases} \tag{4.22}$$

Finally, by using (4.21) and (4.22), we can improve (4.17) as follows:

$$
\begin{aligned}
(4.23)\qquad &\sum_{k=1}^{s}\Big\{\frac{1}{2}\frac{d}{dt}\Big(\|\partial_t^{k+1}w\|^2_{L^2(\mathbb{B}_1)}+\|\nabla\partial_t^k w\|^2_{L^2(\mathbb{B}_1)}\\
&\quad+\int_{\mathbb{B}_1} g_{ij}\nabla_i\partial_t^k w\nabla_j\partial_t^k w dx\Big)+\|\partial_t^{k+1}w\|^2_{L^2(\mathbb{S}^{n-1})}\Big\}\\
&\le C(T)\theta\sum_{k=0}^{s}\Big(\|\partial_t^{k+1}w\|^2_{L^2(\mathbb{S}^{s-1})}+\|\Omega\partial_t^k w\|^2_{L^2(\mathbb{S}^{s-1})}\Big)\\
&\quad+C(T)\theta\Big(\|\partial_t^{s+1}w\|^2_{L^2(\mathbb{B}_1)}+\sum_{k=1}^{s}\|\partial_t^k w\|^2_{H^1(\mathbb{B}_1)}+\|\nabla w\|^2_{L^2(\mathbb{B}_1)}\Big).
\end{aligned}
$$

Step 3. *Morawetz's energy estimates.* First of all, we rewrite the linear wave equation in system (4.8) in terms of the shift variable $\overline{w}$ as follows:

$$
(4.24)\qquad \Box\overline{w}=g_{i\alpha}(t,x,v,v_t,\nabla v)\partial^2_{i\alpha}\overline{w}-\frac{1}{|\mathbb{S}^{n-1}|}\frac{d}{dt}\int_{\mathbb{B}_1} g_{i\alpha}\partial^2_{i\alpha}\overline{w}dx.
$$

Next, we do the energy estimates of Morawetz type. By taking the L^2 inner product of (4.24) with $x\cdot\nabla\overline{w}$, we deduce that

$$
\begin{aligned}
&\frac{d}{dt}\Big[\int_{\mathbb{B}_1} x\cdot\nabla\overline{w}\ \overline{w}_t dx+\frac{1}{|\mathbb{S}^{n-1}|}\int_{\mathbb{B}_1} x\cdot\nabla\overline{w}dx\int_{\mathbb{B}_1} g_{i\alpha}\partial_{i\alpha}\overline{w}dx\Big]\\
&\quad-\int_{\mathbb{B}_1}\nabla_j(x\cdot\nabla\overline{w}\nabla_j\overline{w})dx+\int_{\mathbb{B}_1}|\nabla\overline{w}|^2dx\\
&=\frac{1}{2}\int_{\mathbb{B}_1} x\cdot\nabla\big(|\overline{w}_t|^2-|\nabla\overline{w}|^2\big)dx+\int_{\mathbb{B}_1} x\cdot\nabla\overline{w}g_{i\alpha}\nabla_i\overline{w}_\alpha dx\\
&\quad+\frac{1}{|\mathbb{S}^{n-1}|}\int_{\mathbb{B}_1} x\cdot\nabla\overline{w}_t dx\int_{\mathbb{B}_1} g_{i\alpha}\partial_{i\alpha}\overline{w}dx.
\end{aligned}
$$

Keeping in mind the boundary condition in (4.8) and $g_{i\alpha}=g_{\alpha i}$, and using Green's formula and integration by parts, we compute

$$
\begin{aligned}
&\frac{d}{dt}\Big[\int_{\mathbb{B}_1} x\cdot\nabla\overline{w}\ \overline{w}_t dx+\frac{1}{|\mathbb{S}^{n-1}|}\int_{\mathbb{B}_1} x\cdot\nabla\overline{w}dx\int_{\mathbb{B}_1} g_{i\alpha}\partial_{i\alpha}\overline{w}dx\Big]\\
&\quad-\int_{\mathbb{S}^{n-1}}|w_t|^2d\sigma_y+\frac{1}{2}\int_{\mathbb{B}_1}|\overline{w}_t|^2+|\nabla\overline{w}|^2dx\\
&=-\frac{n-1}{2}\int_{\mathbb{B}_1}\big(|\overline{w}_t|^2-|\nabla\overline{w}|^2\big)dx+\frac{1}{2}\int_{\mathbb{S}^{n-1}}\big(|\overline{w}_t|^2-|\nabla\overline{w}|^2\big)d\sigma_y\\
&\quad+\int_{\mathbb{S}^{n-1}} x\cdot\nabla\overline{w}g_{i\alpha}x_i\partial_\alpha\overline{w}d\sigma_y-\int_{\mathbb{B}_1}\Big(g_{i\alpha}\nabla_i\overline{w}\nabla_\alpha\overline{w}+x\cdot\nabla\overline{w}\frac{d}{dx_i}g_{i\alpha}\partial_\alpha\overline{w}\Big)dx\\
&\quad+\frac{1}{|\mathbb{S}^{n-1}|}\Big(\int_{\mathbb{S}^{n-1}}\overline{w}_t d\sigma_y-n\int_{\mathbb{B}_1}\overline{w}_t dx\Big)\int_{\mathbb{B}_1} g_{i\alpha}\partial_{i\alpha}\overline{w}dx\\
&\quad+\frac{1}{2}\int_{\mathbb{B}_1}\Big(ng_{i\alpha}\nabla_i\overline{w}\nabla_\alpha\overline{w}+x\cdot\nabla g_{i\alpha}\nabla_i\overline{w}\nabla_\alpha\overline{w}\Big)dx-\frac{1}{2}\int_{\mathbb{S}^{n-1}} g_{i\alpha}\nabla_i\overline{w}\nabla_\alpha\overline{w}d\sigma_y\\
&\le-\frac{n-1}{2}\int_{\mathbb{B}_1}\big(|\overline{w}_t|^2-|\nabla\overline{w}|^2\big)dx-\frac{1}{2}\int_{\mathbb{S}^{n-1}}|\Omega w|^2d\sigma_y
\end{aligned}
$$

$$+ C(T)\theta\big(\|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)} + \|w_t\|^2_{L^2(\mathbb{S}^{n-1})} + \|\Omega w\|^2_{L^2(\mathbb{S}^{n-1})}$$
$$+ \|\nabla w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)}\big),$$

where in the last inequality we used (2.20), (4.12), and (4.22). Consequently, we arrive at

$$\begin{aligned}(4.25)\qquad &\frac{d}{dt}\Big[\int_{\mathbb{B}_1} x\cdot\nabla\overline{w}\ \overline{w}_t dx + \frac{1}{|\mathbb{S}^{n-1}|}\int_{\mathbb{B}_1} x\cdot\nabla\overline{w}dx\int_{\mathbb{B}_1} g_{i\alpha}\partial_{i\alpha}\overline{w}dx\Big]\\ &\quad + \frac{2}{5}\big(\|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)} + \|\Omega w\|^2_{L^2(\mathbb{S}^{n-1})}\big)\\ &\le 2\|w_t\|^2_{L^2(\mathbb{S}^{n-1})} + C(T)\theta\big(\|\nabla w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)}\big)\\ &\quad - \frac{n-1}{2}\int_{\mathbb{B}_1}\big(|\overline{w}_t|^2 - |\nabla\overline{w}|^2\big)dx.\end{aligned}$$

At this stage, it is clear that we should estimate the last term of the above inequality. A straightforward calculation shows that

$$\begin{aligned}&\int_{\mathbb{B}_1}\big(|\overline{w}_t|^2 - |\nabla\overline{w}|^2\big)dx\\ &= \frac{d}{dt}\int_{\mathbb{B}^1}\overline{w}\ \overline{w}_t dx - \int_{\mathbb{B}_1}\Big\{|\nabla\overline{w}|^2 + \overline{w}\big(\Delta\overline{w}\\ &\quad + g_{i\alpha}\nabla_i\partial_\alpha\overline{w} - \frac{1}{|\mathbb{S}^{n-1}|}\frac{d}{dt}\int_{\mathbb{B}_1} g_{i\alpha}\partial_{i\alpha}\overline{w}dx\big)\Big\}dx\\ &= \frac{d}{dt}\Big[\int_{\mathbb{B}_1}\overline{w}\ \overline{w}_t dx + \frac{1}{|\mathbb{S}^{n-1}|}\int_{\mathbb{B}_1}\overline{w}dx\int_{\mathbb{B}_1} g_{i\alpha}\partial_{i\alpha}\overline{w}dx\Big]\\ &\quad - \int_{\mathbb{S}^{n-1}}\overline{w}\big(\partial_r\overline{w} + g_{i\alpha}x_i\partial_\alpha\overline{w}\big)d\sigma_y + \int_{\mathbb{B}_1}\big(g_{i\alpha}\nabla_i\overline{w}\partial_\alpha\overline{w}\\ &\quad + \nabla_i g_{i\alpha}\overline{w}\partial_\alpha\overline{w}\big)dx - \frac{1}{|\mathbb{S}^{n-1}|}\int_{\mathbb{B}_1}\overline{w}_t dx\int_{\mathbb{B}_1} g_{i\alpha}\partial_{i\alpha}\overline{w}dx.\end{aligned}$$

By (4.12), (4.14), (4.15), and (4.22), we arrive at

$$\begin{aligned}(4.26)\qquad &\int_{\mathbb{B}_1}\Big(|\overline{w}_t|^2 - |\nabla\overline{w}|^2\Big)dx\\ &\le \frac{d}{dt}\Big[\int_{\mathbb{B}_1}\overline{w}\ \overline{w}_t dx + \frac{1}{|\mathbb{S}^{n-1}|}\int_{\mathbb{B}_1}\overline{w}dx\int_{\mathbb{B}_1} g_{i\alpha}\partial_{i\alpha}\overline{w}dx + \frac{1}{2}\int_{\mathbb{S}^1}\overline{w}^2 d\sigma_y\Big]\\ &\quad + C(T)\theta\Big(\|\overline{w}_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla\overline{w}\|^2_{L^2(\mathbb{B}_1)} + \|w_t\|^2_{L^2(\mathbb{S}^{s-1})}\\ &\quad + \|\Omega w\|^2_{L^2(\mathbb{S}^{s-1})} + \|\nabla w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)}\Big).\end{aligned}$$

Combining (4.22), (4.25), and (4.26), we finally arrive at

$$\begin{aligned}(4.27)\qquad &\frac{d}{dt}\Big[\int_{\mathbb{B}_1}\Big(\frac{n-1}{2}\overline{w} + x\cdot\nabla\overline{w}\Big)\overline{w}_t dx + \frac{n-1}{4}\int_{\mathbb{S}^{n-1}}\overline{w}^2 d\sigma_y\\ &\qquad + \frac{n+1}{2|\mathbb{S}^{n-1}|}\int_{\mathbb{B}_1}\overline{w}dx\int_{\mathbb{B}_1} g_{i\alpha}\partial_{i\alpha}\overline{w}dx\Big]\\ &\quad + \frac{1}{4}\big(\|w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla w\|^2_{L^2(\mathbb{B}_1)} + \|\Omega w\|^2_{L^2(\mathbb{S}^{n-1})}\big)\\ &\le 3\|w_t\|^2_{L^2(\mathbb{S}^{s-1})} + C(T)\theta\big(\|\nabla w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)}\big).\end{aligned}$$

Noting that at the boundary

$$\partial_t^{k+1} w + \partial_t^k w_r = 0$$

holds for $1 \le k \le s$, $s \ge 3$, a similar argument yields that

$$\begin{aligned}(4.28)\quad &\sum_{k=1}^{s}\Big\{\frac{d}{dt}\Big[\int_{\mathbb{B}_1}\Big(\frac{n-1}{2}\partial_t^k w + x\cdot\nabla\partial_t^k w\Big)\partial_t^{k+1} w dx + \frac{n-1}{4}\|\partial_t^k w\|^2_{L^2(\mathbb{S}^{n-1})}\Big]\\ &\qquad + \frac{1}{4}\Big(\|\partial_t^{k+1} w\|^2_{L^2(\mathbb{B}_1)} + \|\nabla\partial_t^k w\|^2_{L^2(\mathbb{B}_1)} + \|\Omega\partial_t^k w\|^2_{L^2(\mathbb{S}^{n-1})}\Big)\Big\}\\ &\le 3\sum_{k=0}^{s}\|\partial_t^{k+1} w\|^2_{L^2(\mathbb{S}^{s-1})}\end{aligned}$$

for $s \ge 3$ and $1 \le k \le s$. Note that for the higher-order Morawetz energy estimates, we cannot use the shift variable $\overline{w}$, which will be made clear in the next step. We also point out that in the above estimate we used the same techniques as in (4.23).

Step 4. *Dissipative energy estimates.* Define

$$\begin{aligned}(4.29)\qquad X_0 &= \int_{\mathbb{B}_1}\Big(\frac{n-1}{2}\overline{w} + x\cdot\nabla\overline{w}\Big)\overline{w}_t dx + \frac{n-1}{4}\|\overline{w}\|^2_{L^2(\mathbb{S}^{n-1})}\\ &\quad + \frac{n-1}{2|\mathbb{S}^{n-1}|}\int_{\mathbb{B}_1}\overline{w}dx\int_{\mathbb{B}_1} g_{i\alpha}\partial_{i\alpha}\overline{w}dx\end{aligned}$$

and

$$(4.30)\qquad X_k = \int_{\mathbb{B}_1}\Big(\frac{n-1}{2}\partial_t^k w + x\cdot\nabla\partial_t^k w\Big)\partial_t^{k+1} w dx + \frac{n-1}{4}\|\partial_t^k w\|^2_{L^2(\mathbb{S}^{n-1})}$$

for $1 \le k \le s$, $s \ge 3$. By (2.23), (4.15), (4.12), and (4.22), it is obvious that

$$(4.31)\qquad \begin{cases}\Big|\int_{\mathbb{B}_1}\big(\frac{n-1}{2}\partial_t^k w + x\cdot\nabla\partial_t^k w\big)\partial_t^{k+1} w dx\Big|\\ \quad \le \frac{1}{2}\Big(\|\partial_t^{k+1} w\|^2_{L^2(\mathbb{B}_1)} + \|\nabla\partial_t^k w\|^2_{L^2(\mathbb{B}_1)}\Big) + \frac{n-1}{4}\|\partial_t^k w\|^2_{L^2(\mathbb{S}^{n-1})},\\ \Big|X_0 - \frac{n-1}{4}\|\overline{w}\|^2_{L^2(\mathbb{S}^{n-1})}\Big| \le C\Big(\|w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla w\|^2_{L^2(\mathbb{B}_1)}\Big)\\ \quad + C(T)\theta\Big(\|\nabla w_t\|^2_{L^2(\mathbb{B}_1)} + \|\nabla^2 w\|^2_{L^2(\mathbb{B}_1)}\Big),\end{cases}$$

where $1 \le k \le s$.

Noting (4.21), we multiply inequalities (4.16) and (4.17) by MC and then add the resulting inequalities to (4.27) and (4.28) to yield

$$(4.32)\qquad \frac{d}{dt}(A_s + MCE_s) + \frac{B_s}{8} \le 0,$$

where M is a big enough positive constant, A_s and B_s are defined as

$$\begin{cases} A_s = \sum_{k=0}^{s}\Big(MC\int_{\mathbb{B}_1} g_{ij}\nabla_i\partial_t^k w\nabla_i\partial_t^k w dx + X_k\Big),\\ B_s = E_s + \sum_{k=0}^{s}(\|\Omega\partial_t^k w\|^2_{L^2(\mathbb{S}^{n-1})} + \|\partial_t^{k+1} w\|^2_{L^2(\mathbb{S}^{n-1})}),\end{cases}$$

and E_s is given by

$$E_s = \|\nabla w\|^2_{L^2(\mathbb{B}_1)} + \|\partial_t^{s+1} w\|^2_{L^2(\mathbb{B}_1)} + \sum_{k=1}^{s}\|\partial_t^k w\|^2_{H^1(\mathbb{B}_1)}.$$

By (4.29), (4.30), and (4.31), it is easy to see that

$$-CE_s \le A_s \le CB_s,$$

which implies that

$$E_s \le A_s + CME_s \le \frac{1}{\gamma} B_s \tag{4.33}$$

for some small positive constant γ. Thus, we get

$$\frac{d}{dt}(A_s + MCE_s) + \gamma(A_s + MCE_s) \le 0.$$

Noting (4.30) and (4.31), it is easy to get

$$A_s(0) \le C\Big(\|\widetilde{f}_0\|^2_{H^{s+1}(\mathbb{B}_1)} + \|\widetilde{f}_1\|^2_{H^s(\mathbb{B}_1)}\Big).$$

By Gronwall's inequality, we have

$$E_s \le C\exp(-\gamma t)\Big(\|\widetilde{f}_0\|^2_{H^{s+1}(\mathbb{B}_1)} + \|\widetilde{f}_1\|^2_{H^s(\mathbb{B}_1)}\Big). \tag{4.34}$$

Finally, the estimate (4.9) follows (4.21), (4.22), (4.34), and the elliptic estimates in (2.28). □

In what follows, we will outline the steps to construct local exact boundary controllability for quasi-linear wave equations (1.2) with initial data (1.6) and final data (1.7), which are similar to those in sections 2 and 3.

Analogous to Lemma 2.8, we have the following.

LEMMA 4.3. *Let $\varphi^{(k)}$ and $\phi^{(k)}$ be defined as the solutions of systems* (4.3)–(4.6); *$s \ge 3$ is an integer. Then the estimates*

$$\begin{aligned}
&\sum_{j=0}^{s+1}\Big(\|\partial_t^j\varphi^{(k)}(T,\cdot)\|_{H^{s+1-j}(\Omega_0)} + \|\partial_t^j\phi^{(k)}(0,\cdot)\|_{H^{s+1-j}(\Omega_0)}\Big)\\
&\le \frac{1}{8}\sum_{j=0}^{s+1}\Big(\|\partial_t^j\varphi^{(k-1)}(T,x)\|_{H^{s+1-j}(\Omega_0)} + \|\partial_t^j\phi^{(k-1)}(0,x)\|_{H^{s+1-j}(\Omega_0)}\\
&\qquad + \|\partial_t^j\varphi^{(k-2)}(T,x)\|_{H^{s+1-j}(\Omega_0)} + \|\partial_t^j\phi^{(k-2)}(0,x)\|_{H^{s+1-j}(\Omega_0)}\Big)
\end{aligned}$$

hold, provided that θ is small enough and T is big enough.

By the above lemma, we can show that $u = \Pi_\Lambda v$ is well-defined and $u \in \Lambda_\theta$, similarly to Theorem 2.7.

Then, analogously to section 3, we can show that $\Pi_\Lambda : v \to u = \Pi_\Lambda v$ is a strict contraction from Λ_θ to itself. By the standard contraction mapping theorem, there exists a point $u \in \Lambda_\theta$, such that $u = \Pi_\Lambda u$. For $(t,x) \in [0,T] \times \partial\Omega_0$, let

$$h(t,x) = u(t,x);$$

then h is the desired control and the proof of Theorem 1.6 is complete.

Acknowledgments. The authors want to thank Professor Ta-tsien Li for many helpful discussions. The authors also want to thank the referees for their careful review and constructive comments on our paper.

REFERENCES

[1] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[2] M. M. Cavalcanti, *Exact controllability of the wave equation with mixed boundary condition and time-dependent coefficients*, Arch. Math. (Brno), 35 (1999), pp. 29–57.

[3] C. M. Dafermos and W. J. Hrusa, *Energy methods for quasilinear hyperbolic initial-boundary value problems. Applications to elastodynamics*, Arch. Rational Mech. Anal., 87 (1985), pp. 267–292.

[4] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.

[5] X. Fu, J. Yong, and X. Zhang, *Exact controllability for multidimensional semilinear hyperbolic equations*, SIAM J. Control Optim., to appear.

[6] T. J. R. Hughes, T. Kato, and J. E. Marsden, *Well-posed quasi-linear second-order hyperbolic systems with applications to nonlinear elastodynamics and general relativity*, Arch. Rational Mech. Anal., 63 (1976), pp. 273–294.

[7] J. Lagnese, *Exact boundary value controllability of a class of hyperbolic equations*, SIAM J. Control Optim., 16 (1978), pp. 1000–1017.

[8] I. Lasiecka and R. Triggiani, *Exact controllability of semilinear abstract systems with applications to waves and plates boundary control problems*, Appl. Math. Optim., 23 (1991), pp. 109–154.

[9] I. Lasiecka, R. Triggiani, and X. Zhang, *Nonconservative wave equations with unobserved Neumann B.C.: Global uniqueness and observability in one shot*, in Differential Geometric Methods in the Control of Partial Differential Equations (Boulder, CO, 1999), Contemp. Math. 268, AMS, Providence, RI, 2000, pp. 227–325.

[10] T. T. Li, *Exact boundary controllability for quasilinear wave equations*, in Trends in Partial Differential Equations of Mathematical Physics, Progr. Nonlinear Differential Equations Appl. 61, Birkhäuser, Basel, 2005, pp. 149–160.

[11] T. T. Li and Y. M. Chen, *Global Classical Solutions for Nonlinear Evolution Equations*, Pitman Monogr. Surveys Pure Appl. Math. 45, Longman Scientific & Technical, Harlow, John Wiley & Sons, New York, 1992.

[12] T.-T. Li and B.-P. Rao, *Exact boundary controllability for quasi-linear hyperbolic systems*, SIAM J. Control Optim., 41 (2003), pp. 1748–1755.

[13] T. T. Li and L. X. Yu, *Exact boundary controllability for* 1*-D quasilinear wave equations*, SIAM J. Control Optim., 45 (2006), pp. 1074–1083.

[14] J. L. Lions, *Contrôlabilité exacte des systèmes distribués* (Exact controllability of distributed systems), C. R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 471–475.

[15] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[16] D. L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–211.

[17] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[18] D. Tataru, *Carleman estimates and unique continuation for solutions to boundary value problems*, J. Math. Pures Appl. (9), 75 (1996), pp. 367–408.

[19] M. E. Taylor, *Partial Differential Equations.* I. *Basic Theory*, Appl. Math. Sci. 115, Springer-Verlag, New York, 1996.

[20] M. E. Taylor, *Partial Differential Equations.* III. *Nonlinear Equations*, Appl. Math. Sci. 115, Springer-Verlag, New York, 1996.

[21] P.-F. Yao, *On the observability inequalities for exact controllability of wave equations with variable coefficients*, SIAM J. Control Optim., 37 (1999), pp. 1568–1599.

[22] P. F. Yao, *Boundary Controllability for the Quasilinear Wave Equation*, manuscript, 2005.

[23] L. X. Yu, *Exact boundary controllability for higher order quasilinear hyperbolic equations*, Appl. Math. J. Chinese Univ. Ser. B, 20 (2005), pp. 127–141.

[24] E. Zuazua, *Exact controllability for the semilinear wave equation*, J. Math. Pures Appl. (9), 69 (1990), pp. 1–31.

[25] E. Zuazua, *Exact controllability for semilinear wave equations in one space dimension*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 109–129.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 1052–1079

ANTIANGIOGENIC THERAPY IN CANCER TREATMENT AS AN OPTIMAL CONTROL PROBLEM*

URSZULA LEDZEWICZ† AND HEINZ SCHÄTTLER‡

Abstract. Antiangiogenic therapy is a novel treatment approach in cancer therapy that aims at preventing a tumor from developing its own blood supply system that it needs for growth. In this paper a mathematical model for antiangiogenic treatments based on a biologically validated model by Hahnfeldt et al. is analyzed as an optimal control problem and a full solution of the problem is given. Geometric methods from optimal control theory are utilized to arrive at the solution.

Key words. optimal control, geometric methods, cancer treatment, antiangiogenic therapy

AMS subject classifications. 49K15, 92C50, 37N25

DOI. 10.1137/060665294

1. Introduction.

The most important limiting factor for the success of cancer chemotherapy treatments lies in both intrinsic and acquired drug resistance. Malignant cancer cell populations are highly heterogeneous—the number of genetic errors present within one cancer cell can lie in the thousands [16]—and fast duplications combined with genetic instabilities provide just one of several mechanisms which allow for quickly developing acquired resistance to anticancer drugs. In addition, intrinsic resistance (i.e., the specific drug's activation mechanism simply doesn't work) makes some cancer cells not susceptible to many cytotoxic agents. "... the truly surprising thing is that some malignancies can be cured even with current approaches" [8, p. 65]. Several mechanisms to circumvent the problem of drug resistance have been tried but so far without success, and currently no medical solution to the problem exists. In fact, it is acquired or intrinsic drug resistance which eventually makes most chemotherapy fail. At the same time, similar phenomena do not take place for the healthy proliferating cells. For example, regretfully, bone marrow does not develop drug resistance to the killing agent [10].

As of today, the search for therapy approaches that would avoid drug resistance still is of tantamount importance in medicine. Two such approaches that are currently being pursued in their experimental stages are immunotherapy and antiangiogenic treatments. While immunotherapy tries to coax the body's immune system to take action against the cancerous growth, tumor antiangiogenesis aims at depriving a tumor from developing the necessary blood cells and capillaries that it needs for further growth. Since the treatment does not target cancer cells but normal cells, no occurrence of drug resistance has been reported in lab studies. (These treatments, however, are only in the stage of experimental studies and initial clinical trials.) For

*Received by the editors July 17, 2006; accepted for publication (in revised form) March 20, 2007; published electronically July 11, 2007. This material is based upon research supported by the National Science Foundation under collaborative research grants DMS 0405827/0405848. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

http://www.siam.org/journals/sicon/46-3/66529.html

†Department of Mathematics and Statistics, Southern Illinois University at Edwardsville, Edwardsville, IL 62026-1653 (uledzew@siue.edu). The research of this author was partially supported by a SIUE 2006 Summer Research Fellowship.

‡Department of Electrical and Systems Engineering, Washington University, St. Louis, MO 63130-4899 (hms@wustl.edu).

this reason tumor antiangiogenesis has been called a therapy resistant to resistance which provides a new hope in treatment of tumor-type cancers [10].

There exist several mathematical models for the evolution of tumor antiangiogenesis as a dynamical system, with the one formulated by Hahnfeldt et al. in [9] probably being the most prominent one. This model was biologically validated in lab experiments and became the basis for several modifications and simplifications [5, 6] undertaken in an effort to both better understand the dynamical properties of the underlying mechanisms and to make the mathematical model easier and more tractable for analysis. For example, a dynamical systems analysis of the model by Hahnfeldt et al. and of several modifications (with more general growth models for the growth of cancer cells and slightly different dynamics for the evolution related to endothelial cells) is given in the paper by d'Onofrio and Gandolfi [5]; Ergun, Camphausen, and Wein [6] consider an optimal control problem for the scheduling of antiangiogenic inhibitors both as monotherapy and in combination with radiotherapy. While these models are variations of the specific dynamics proposed by Hahnfeldt et al. in [9], in the papers by Agur et al. [1] and Forys, Keifetz, and Kogan [7] more generally dynamical properties of models for angiogenesis are investigated under minimal assumptions on the form of the growth functions describing the dynamics.

In this paper we consider the original mathematical model for tumor antiangiogenesis formulated and validated by Hahnfeldt et al. in [9] and analyze it as an optimal control problem. Using geometric methods of optimal control theory, for this model we compute how to schedule a given amount of angiogenic inhibitors to achieve the maximum reduction in tumor volume possible. The key feature of the solution is an optimal singular arc whose geometric analysis forms the core of the mathematical argument. Optimal controls then are concatenations of bang controls (constant controls that give either a full or no dose of inhibitors) and the optimal singular control (a specific smooth control that administers the inhibitors using a time-varying feedback schedule at less than a maximum rate). The most general structure of optimal controls possible is a concatenation of the form "**0asa0**," where **a** and **0** denote trajectories with *full*, respectively, *no* antiangiogenic therapy, and **s** stands for a segment along the *singular* arc. However, depending on the initial condition not all of these pieces are present. Our theoretical analysis reduces the structure of optimal controls to at most this structure but for some initial conditions still allows for a one-parameter family of extremals of this form. Then, given any initial condition, the optimal solution is easily computed numerically based on our analysis. The most typical and medically most relevant scenarios are optimal protocols that take the simple form "**a0**" when all inhibitors are administered at the beginning or "**as0**" when the dosage is adjusted as the singular arc is reached and then all available inhibitors are being used up along the singular arc. If the optimal policy along the singular arc comes close to a point where the singular control saturates at the upper value a, then optimal trajectories actually leave the singular arc prior to saturation (and this is consistent with the behavior of optimal controls near saturation points; see, for example, [17] or [2]) and are of the type "**asa0**." The full structure "**0asa0**" arises only for initial conditions that are not significant for the underlying problem.

A preliminary announcement of some partial results presented in this paper has been given without proofs in [14]. Here the analysis is completed, and proofs are included.

2. Medical background and mathematical model [9]. A growing tumor, after it reaches just a few millimeters in diameter, no longer can rely on blood vessels of the host for its supply of nutrients, but it needs to develop its own vessels and

capillaries for blood supply. In this process, called *angiogenesis*, there is a reciprocal signaling between endothelial cells and tumor cells. Tumor cells produce vascular endothelial growth factor (VEGF) to stimulate endothelial cell growth; endothelial cells in turn provide the lining for the newly forming blood vessels that supply nutrients to the tumor and thus sustain tumor growth. But endothelial cells also have receptors which make them sensitive to inhibitors of inducers of angiogenesis such as, for example, endostatin, and pharmacologic therapies typically target the growth factor VEGF trying to impede the development of new blood vessels and capillaries. Overall, angiogenesis can be viewed as a complex balance of stimulatory and inhibitory mechanisms regulated through microenvironmental factors.

In the model developed by Hahnfeldt et al. in [9] these effects are summarized in a two-dimensional dynamical system with the *primary tumor volume* p and the *carrying capacity of the vasculature* q as variables. The latter is defined as the "maximal tumor volume potentially sustainable by the network" [9] and is implicitly assumed proportional to the number of endothelial cells. Thus the set $\mathcal{D}_0 = \{(p,q) \in \mathbb{R}^2_+ : p = q\}$ corresponds to points where the vasculature is adequate to support the tumor, while $\mathcal{D}_- = \{(p,q) \in \mathcal{D} : p < q\}$ corresponds to growing tumors and $\mathcal{D}_+ = \{(p,q) \in \mathcal{D} : p > q\}$ to shrinking tumors. A growth function describes the size of the tumor dependent on the carrying capacity q and is chosen as Gompertzian in the original model. Other models are equally realistic and are considered, for instance, in [5] or [7], but here we stay with the original choice. Thus the rate of change in the primary tumor volume is modeled as

$$\dot{p} = -\xi p \ln\left(\frac{p}{q}\right), \tag{2.1}$$

where ξ denotes a tumor growth parameter. The overall dynamics for the carrying capacity is a balance between stimulation and inhibition, and its basic structure is of the form

$$\dot{q} = -\mu q + S(p,q) - I(p,q) - Guq, \tag{2.2}$$

where μq describes the loss of endothelial cells due to natural causes (death, etc.), I and S denote endogenous inhibition and stimulation terms, respectively, and Guq represents a loss due to additional outside inhibition. The variable u represents the control in the system and corresponds to the angiogenic dose rate, while G is a constant that represents the antiangiogenic killing parameter. Generally μ is small, often this term is negligible compared to the other factors, and thus in the literature often μ is set to 0 in this equation.

In [9] a spatial analysis of the underlying consumption-diffusion model was carried out that led to the following two principal conclusions:

1. The inhibitor will impact endothelial cells in a way that grows like the volume of cancer cells to the power $\frac{2}{3}$.

The exponent $\frac{2}{3}$ arises since inhibitors need to be released through the surface of the tumor. Thus in [9] the inhibitor term is taken in the form

$$I(p,q) = dp^{\frac{2}{3}}q, \tag{2.3}$$

with d a constant, the "death" rate. The second implication of the analysis in [9] is that:

2. the inhibitor term will tend to grow at a rate of $q^\alpha p^\beta$ faster than the stimulator term where $\alpha + \beta = \frac{2}{3}$.

However, the choice of α and β is not imperative in their analysis and in fact is one of the main sources for the various other models also considered in the literature [5, 6]. In their original work [9] Hahnfeldt et al. select $\alpha = 1$ and $\beta = -\frac{1}{3}$ resulting in the simple stimulation term

$$S(p, q) = bp, \tag{2.4}$$

with b a constant, the "birth" rate. However, other choices are possible, and, for example, choosing $\alpha = 0$ and $\beta = \frac{2}{3}$ results in the equally simple form $S(p, q) = bq$ chosen in [5]. In that paper the dynamics for both models is analyzed, and it is shown for the uncontrolled system that there exists a unique globally asymptotically stable equilibrium (which, of course, is not viable biologically). Adding a control term, this equilibrium can be shifted to lower values or, depending on the parameter values, even eliminated altogether. In the latter case all trajectories converge to the origin in infinite time. This, in principle, would be the desired situation.

The problem then becomes how to administer a given amount of inhibitors to achieve the "best possible" effect. In this paper we use the dynamics of the original model from [9] and formulate this aim as the following optimal control problem.

[HPFH]. For a free terminal time T, minimize the value $p(T)$ subject to the dynamics

$$\dot{p} = -\xi p \ln\left(\frac{p}{q}\right), \qquad p(0) = p_0, \tag{2.5}$$

$$\dot{q} = bp - (\mu + dp^{\frac{2}{3}})q - Guq, \qquad q(0) = q_0, \tag{2.6}$$

$$\dot{y} = u, \qquad y(0) = 0, \tag{2.7}$$

over all measurable functions $u : [0, T] \to [0, a]$ for which the corresponding trajectory satisfies $y(T) \le A$.

As is customary in optimal control formulations, we adjoin the constraint as a third variable. Later on, for all of our numerical illustrations we use the following parameter values which are taken from [9]: The variables p and q are volumes measured in mm^3; $\xi = \frac{0.192}{\ln 10} = 0.084$ per day (adjusted to the natural logarithm), $b = 5.85$ per day, $d = 0.00873$ per mm^2 per day, $G = 0.15$ kg per mg of dose per day, and for illustrative purposes we chose a small positive value for μ: $\mu = 0.02$ per day. But we want to emphasize already that *our mathematical analysis and conclusions are valid independently of the specific parameter values* and lead to robust implications about the structure of optimal controls for this model.

3. The dynamical systems for constant controls. For the analysis of the optimal control problem it is of benefit to fully understand the dynamic properties of the systems for a constant control $u \equiv v$, with v some value in the control set $[0, a]$. Our statements in this section are only minor extensions of the analysis given in the paper by d'Onofrio and Gandolfi [5], and we refer the reader to that paper for the proofs about our claims of stability properties of the equilibria. All statements are for the natural domain $\mathbb{R}^2_+ = \{(p, q) : p > 0, q > 0\}$ of the system. The following fact about the dynamical behavior of the system is an easy corollary of the results proven in [5].

PROPOSITION 3.1. *For any admissible control u and arbitrary positive initial conditions p_0 and q_0, the corresponding solution (p, q) exists for all times $t \ge 0$, and both p and q remain positive.*

Assuming $b > \mu$, the uncontrolled system ($u = 0$) has a unique globally asymptotically stable equilibrium point at $(\bar{p}, \bar{q})$ given by $\bar{p} = \bar{q} = (\frac{b-\mu}{d})^{\frac{3}{2}}$. This value naturally is far too high to be acceptable, and the medically relevant region is contained in the domain

$$\mathcal{D} = \{(p, q) : 0 < p \leq \bar{p},\ 0 < q \leq \bar{q}\}. \tag{3.1}$$

In order to exclude irrelevant discussions about the structure of optimal controls in regions where the model does not represent the underlying medical problem to begin with, we henceforth restrict our discussions to this square domain $\mathcal{D}$.

PROPOSITION 3.2. *$\mathcal{D}$ is positively invariant for the flow of the control system; i.e., if $(p_0, q_0) \in \mathcal{D}$, then for any admissible control u defined over the interval $[0, \infty)$ the solution $(p(\cdot), q(\cdot))$ to the corresponding dynamics with initial condition $(p(0), q(0)) = (p_0, q_0)$ exists for all times $t \geq 0$ and lies in $\mathcal{D}$, $(p(t), q(t)) \in \mathcal{D}$.*

Proof. The positive invariance of the region $\mathcal{P} = \{(p, q) : 0 < p,\ 0 < q\}$ for any admissible control u directly follows from Proposition 3.1. The dynamics is clearly pointing into $\mathcal{D}$ on the boundary segment $\{(p, q) \in \mathcal{D} : p = \bar{p},\ 0 < q < \bar{q}\}$, since for $p > q$ we always have $\dot{p} < 0$. For a constant control v, the isoclines for $\dot{q} = 0$ are given by

$$q = \Xi_v(p) = \frac{bp}{\mu + Gv + dp^{\frac{2}{3}}}. \tag{3.2}$$

The functions Ξ_v are strictly increasing, $\Xi_v(0) = 0$, and at $\bar{p}$ take the value

$$\Xi_v(\bar{p}) = \frac{b}{b + Gv}\bar{p}. \tag{3.3}$$

In particular, the smallest value is given for $v = a$ and $\Xi_0(\bar{p}) = \bar{p}$. It thus follows that on the boundary segment $\{(p, q) \in \mathcal{D} : 0 < p < \bar{p},\ q = \bar{q}\}$ we have $\dot{q} < 0$ for all controls. The point $(\bar{p}, \bar{q})$ is the equilibrium point for $u = 0$, and the dynamics points into $\mathcal{D}$ for $u = a$ at this point. Thus, regardless of the control value v, trajectories can never leave the region $\mathcal{D}$. □

By increasing the value v of the control, the equilibrium can be shifted towards the origin along the diagonal and finally be eliminated altogether. As a function of v the equilibrium is the unique fixed point of the equation $p = \Xi_v(p)$ in $\{p > 0\}$ and is given by

$$\bar{p}(v) = \bar{q}(v) = \left(\frac{b - \mu - Gv}{d}\right)^{\frac{3}{2}} \tag{3.4}$$

provided $b - \mu > Gv$, and this equilibrium $(\bar{p}(v), \bar{q}(v))$ still is globally asymptotically stable. As $b - \mu \leq Gv$, the system no longer has an equilibrium point, and now all trajectories converge to the origin as $t \to \infty$ [5]. Thus, theoretically eradication of the tumor was possible in this case under the unrealistic scenario of constant treatment with an unlimited supply of inhibitors. Since this is the most desirable situation, for our analysis of the optimal control problem we also **assume** that

$$(\mathbf{A}) \qquad \mathbf{Ga} > \mathbf{b} - \mu > \mathbf{0}. \tag{3.5}$$

Figure 3.1 shows the phase portraits of the uncontrolled system on the left and for $u \equiv a$ on the right. In our figures we prefer to have the tumor volume as the vertical

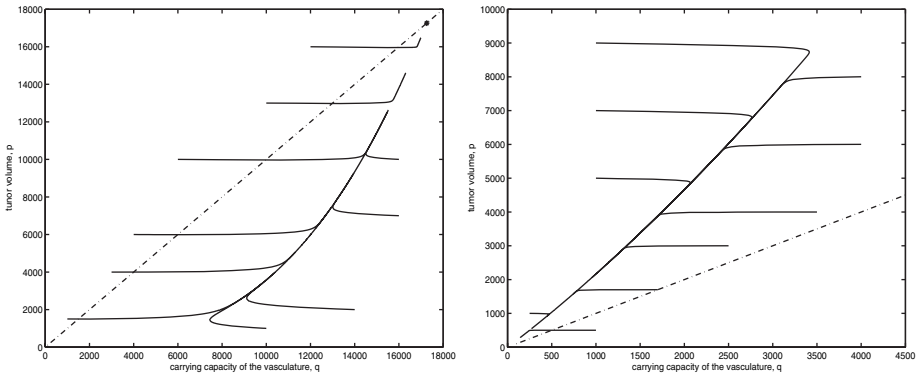


FIG. 3.1. *Phase portraits for* $u = 0$ *and* $u = a = 75$.

axis since this better visualizes the tumor reduction (respectively, increase). For comparison the diagonal is included in these figures as a dashed-dotted line. It is not difficult to extend all results of this paper to the case when the dynamics for $u \equiv a$ still has a positive equilibrium, but this will not be pursued here for reasons of space.

However, the domain $\mathcal{D}$ still contains initial conditions that give rise to degenerate cases that we want to exclude. Recall that $\mathcal{D}_+ = \{(p,q) \in \mathcal{D} : p > q\}$, $\mathcal{D}_0 = \{(p,q) \in \mathcal{D} : p = q\}$, and $\mathcal{D}_- = \{(p,q) \in \mathcal{D} : p < q\}$. Both of the trajectories for the constant controls $u = 0$ and $u = a$ cross the diagonal portion $\mathcal{D}_0$ transversally: For $u = 0$ trajectories cross from $\mathcal{D}_+$ into $\mathcal{D}_-$, while they cross in the opposite direction from $\mathcal{D}_-$ into $\mathcal{D}_+$ for $u = a$. Also, trajectories for $u = 0$ approach the stable equilibrium $(\bar{p}, \bar{q})$ from within the region $\mathcal{D}_-$, while trajectories for $u = a$ converge to the origin as $t \to \infty$ in the region $\mathcal{D}_+$. It follows from the dynamics for p, (2.5), that the p-value of trajectories is always decreasing in $\mathcal{D}_+$ and always increasing in $\mathcal{D}_-$. As a result, for some initial conditions (p_0, q_0, y_0), with $(p_0, q_0) \in \mathcal{D}_-$, it is possible that the (mathematically) optimal time T is $T = 0$. This situation arises when the amount of available inhibitors simply is not sufficient to reach a point in the region $\mathcal{D}_+$ that would have a lower p-value than p_0. In such a case it is not possible to decrease the tumor volume with the available amount of inhibitors. It is possible only to slow down the tumor's growth. Indeed, it is correct that the best way of doing this is to give the full dose $u = a$ until all inhibitors run out—this follows from the structure of optimal controls to be shown later—but this is not the mathematically "optimal" solution for problem [HPFH]. This one is simply to do nothing and take $T = 0$. Since this introduces a number of degeneracies into the analysis, we make the following definition.

DEFINITION 3.3. *We say an initial condition* $(p_0, q_0) \in \mathcal{D}_-$ *is* ill-posed *if for any admissible control it is not possible to reach a point* (p, q) *with* $p < p_0$. *In this case the optimal solution for the problem [HPFH] is given by* $T = 0$. *Otherwise* (p_0, q_0) *is* well-posed *and the optimal time* T *will be positive.*

It is clear that all initial conditions with $(p_0, q_0) \in \mathcal{D}_+ \cup \mathcal{D}_0$ are well-posed (since p decreases in $\mathcal{D}_+$ and trajectories with $u = a$ enter $\mathcal{D}_+$ from $\mathcal{D}_0$), and it is easily decided whether an initial condition $(p_0, q_0) \in \mathcal{D}_-$ is ill-posed. For our analysis of optimal controls, however, we *consider only well-posed initial conditions.*

4. The maximum principle and preliminary analysis of optimal controls. It follows from classical results that there exists an optimal solution to our

problem [4]. First-order necessary conditions for optimality of a control u are given by the *Pontryagin maximum principle* [18, 3, 4]: If u_* is an optimal control defined over an interval $[0,T]$ with corresponding trajectory $(p_*,q_*,y_*)^T$, then there exist a constant $\lambda_0 \geq 0$ and an absolutely continuous covector $\lambda : [0,T] \to (\mathbb{R}^3)^*$ (which we write as a row vector) such that (a) $(\lambda_0,\lambda(t)) \neq (0,0)$ for all $t \in [0,T]$, (b) the adjoint equations hold with transversality conditions

(4.1)
$$\dot{\lambda}_1 = \xi\lambda_1\left(\ln\left(\frac{p_*(t)}{q_*(t)}\right)+1\right)+\lambda_2\left(\frac{2}{3}d\frac{q_*(t)}{p_*^{\frac{1}{3}}(t)}-b\right), \quad \lambda_1(T)=\lambda_0,$$

(4.2)
$$\dot{\lambda}_2 = -\xi\lambda_1\frac{p_*(t)}{q_*(t)}+\lambda_2\left(\mu+dp_*^{\frac{2}{3}}(t)+Gu\right), \qquad \lambda_2(T)=0,$$

(4.3)
$$\dot{\lambda}_3 = 0, \qquad \lambda_3(T) = \begin{cases} 0 & \text{if } y(T)<A, \\ \text{free} & \text{if } y(T)=A, \end{cases}$$

and (c) the optimal control u_* minimizes the Hamiltonian H

$$H = -\lambda_1\xi p\ln\left(\frac{p}{q}\right)+\lambda_2\left(bp-\left(\mu+dp^{\frac{2}{3}}\right)q-Guq\right)+\lambda_3 u, \tag{4.4}$$

along $(\lambda(t),p_*(t),q_*(t))$ over the control set $[0,a]$ with the minimum value given by 0.

We call a pair $((p,q,y),u)$ consisting of an admissible control u with corresponding trajectory (p,q,y) an *extremal* (pair) if there exist multipliers (λ_0,λ) such that the conditions of the maximum principle are satisfied and the triple $((p,q,y),u,(\lambda_0,\lambda))$ is an extremal lift (to the cotangent bundle). Extremals with $\lambda_0 = 0$ are called *abnormal*, while those with a positive multiplier λ_0 are called *normal*. In this case it is possible to normalize $\lambda_0 = 1$. The following lemmas summarize some elementary properties of optimal controls and extremals for well-posed initial conditions.

LEMMA 4.1. *If u_* is an optimal control with corresponding trajectory $(p_*,q_*,y_*)^T$, then at the final time $p_*(T) = q_*(T)$ and $y_*(T) = A$; i.e., all available inhibitors have been used up.*

Proof. Since the p-dynamics is Gompertzian, (2.5), the cancer volume is growing for $p < q$ and is shrinking for $p > q$. This implies that optimal trajectories can terminate only at times where $p_*(T) = q_*(T)$. For, if $p_*(T) < q_*(T)$, then it would simply have been better to stop earlier since p was increasing over some interval $(T-\varepsilon,T]$. (Recall that we are assuming that the initial condition is well-posed so that the optimal final time T is positive.) On the other hand, if $p_*(T) > q_*(T)$, then we can always add another small interval $(T,T+\varepsilon]$ with the control $u = 0$ without violating any of the constraints and p will decrease along this interval if ε is small enough. Thus at the final time necessarily $p_*(T) = q_*(T)$. If now $y(T) < A$, then we can still add a small piece of a trajectory for $u = a$ over some interval $[0,\varepsilon]$. Since $\dot{q} < 0$ on the diagonal $\mathcal{D}_0$, the corresponding trajectory lies in $\mathcal{D}_+$, and thus the value of p is decreasing along this trajectory contradicting the optimality of T. ∎

LEMMA 4.2. *Extremals are normal. The multipliers λ_1 and λ_2 cannot vanish simultaneously; λ_2 has only simple zeros. The multiplier λ_3 is constant and nonnegative.*

Proof. The multipliers λ_1 and λ_2 satisfy the homogeneous linear system (4.1) and (4.2), and thus they vanish identically if they vanish at some time t. If $\lambda_0 = 0$, then

the nontriviality of $(\lambda_0, \lambda(t))$ implies that the multiplier λ_3, which is constant, is not zero. The condition $H \equiv 0$ on the Hamiltonian therefore gives $u \equiv 0$; i.e., the initial condition is ill-posed. Thus, without loss of generality we may assume that $\lambda_0 = 1$ and hence λ_1 and λ_2 cannot vanish simultaneously. In particular, whenever $\lambda_2(t) = 0$, then $\dot{\lambda}_2(t) \neq 0$, and thus λ_2 has only simple zeros.

For the final time T it follows from $p_*(T) = q_*(T)$, the transversality condition $\lambda_2(T) = 0$, and the condition $H(T) \equiv 0$ that $\lambda_3 u_*(T) = 0$. If $\lambda_3 < 0$, then the function $\Phi(t) = \lambda_3 - \lambda_2(t) G q_*(t)$ will be negative on some interval $(T - \varepsilon, T]$, and thus by the minimization condition (c) on the Hamiltonian the control must be given by $u_*(t) = a$ on this interval which is a contradiction. Hence $\lambda_3 \geq 0$. ∎

LEMMA 4.3. *If $\lambda_3 = 0$, then the corresponding optimal control is constant over the interval $[0, T]$ and given by the control $u \equiv a$.*

Proof. In this case the Hamiltonian function reduces to

$$H = -\lambda_1 \xi p \ln\left(\frac{p}{q}\right) + \lambda_2 \left(bp - \left(\mu + dp^{\frac{2}{3}}\right) q - Guq\right), \tag{4.5}$$

and thus the minimization condition (c) implies that

$$u_*(t) = \begin{cases} 0 & \text{if } \lambda_2(t) < 0, \\ a & \text{if } \lambda_2(t) > 0. \end{cases}$$

Since $\lambda_2(T) = 0$ and $\dot{\lambda}_2(T) = -\xi \lambda_1(T) \frac{p_*(t)}{q_*(t)} = -\xi < 0$, λ_2 is positive on some interval $(\tau, T]$, and here the control is given by $u_*(t) = a$. Since $p_*(T) = q_*(T)$, it follows that the trajectory entirely lies in $\mathcal{D}_-$ as long as the control is $u \equiv a$. But then λ_2 cannot have another zero τ since otherwise $H(\tau) = -\lambda_1(\tau) \xi p(\tau) \ln(\frac{p(\tau)}{q(\tau)}) \neq 0$. Thus the control must be constant $u \equiv a$. ∎

Except for this extremely degenerate case (the initial condition is such that with giving the full dose we reach the diagonal exactly when all inhibitors have been exhausted), we can, as we henceforth do, without loss of generality therefore assume that λ_3 is positive.

LEMMA 4.4. *If $\lambda_3 > 0$, then optimal controls end with an interval $(\tau, T]$ where $u_* \equiv 0$.*

The function

$$\Phi(t) = \lambda_3 - \lambda_2(t) G q_*(t), \tag{4.6}$$

which determines the structure of the optimal control u_* through the minimization property (c) on the Hamiltonian H, is called the *switching function* of the problem, and optimal controls satisfy

$$u_*(t) = \begin{cases} 0 & \text{if } \Phi(t) > 0, \\ a & \text{if } \Phi(t) < 0. \end{cases} \tag{4.7}$$

A priori the control is not determined by the minimum condition at times when $\Phi(t) = 0$. If $\Phi(\tau) = 0$, but $\dot{\Phi}(\tau) \neq 0$, then the control switches between $u = 0$ and $u = a$ depending on the sign of $\dot{\Phi}(\tau)$. On the other hand, if $\Phi(t)$ vanishes identically on an open interval, then the minimization property in itself gives no information about the control. However, in this case also all derivatives of $\Phi(t)$ must vanish, and this may and typically does determine the control. Controls of this kind are called *singular*, while we refer to the constant controls as *bang* controls. Optimal controls

then need to be synthesized from these candidates through an analysis of the switching function and its derivatives.

The computations of the derivatives of the switching function Φ can be expressed concisely within the framework of geometric optimal control theory, and we therefore now write the state as a 3-dimensional vector $z = (z_1, z_2, z_3)^T$, with $z_1 = p$, $z_2 = q$, and $z_3 = y$. In vector notation the dynamics takes the form

$$\dot{z} = f(z) + ug(z), \tag{4.8}$$

with

$$f(z) = \begin{pmatrix} -\xi p \ln\left(\frac{p}{q}\right) \\ bp - \left(\mu + dp^{\frac{2}{3}}\right) q \\ 0 \end{pmatrix} \tag{4.9}$$

and

$$g(z) = \begin{pmatrix} 0 \\ -Gq \\ 1 \end{pmatrix}. \tag{4.10}$$

The adjoint equation then simply becomes

$$\dot{\lambda}(t) = -\lambda(t) \left(Df(z(t)) + u_*(t) Dg(z(t))\right), \tag{4.11}$$

where Df and Dg denote the matrices of the partial derivatives of the vector fields which are evaluated along $z(t)$. The derivatives of the switching function can easily be computed using the following well known result that can be verified by an elementary direct calculation. Here $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^3$, i.e., for a covector $\lambda \in (\mathbb{R}^3)^*$ and a vector $z \in \mathbb{R}^3$, $\langle \lambda, z \rangle = \lambda z$.

PROPOSITION 4.5. *Let h be a continuously differentiable vector field, and define*

$$\Psi(t) = \langle \lambda(t), h(z(t)) \rangle. \tag{4.12}$$

Then the derivative of Ψ along a solution to the system equation (4.8) *for control u and a solution λ to the corresponding adjoint equation* (4.11) *is given by*

$$\dot{\Psi}(t) = \langle \lambda(t), [f + ug, h] z(t) \rangle, \tag{4.13}$$

where

$$[f, h](z) = Dh(z) f(z) - Df(z) h(z) \tag{4.14}$$

denotes the Lie bracket of the vector fields f and h.

5. Synthesis of optimal controlled trajectories. In this section we first give an overview of the structure of optimal controlled trajectories, but the proofs will be postponed to the remaining sections of the paper. We summarize the general structure of optimal controls and trajectories in the following theorem.

THEOREM 5.1. *Given $\tilde{z} = (\tilde{p}, \tilde{q}, 0)$, with $(\tilde{p}, \tilde{q}) \in \mathcal{D}$, optimal controls are* at most *concatenations of the form* **0asa0**, *with* **0** *denoting an arc along the constant control $u = 0$,* **a** *denoting an arc along the constant control $u = a$, and* **s** *denoting an arc along the singular curve $\mathcal{S}$.*

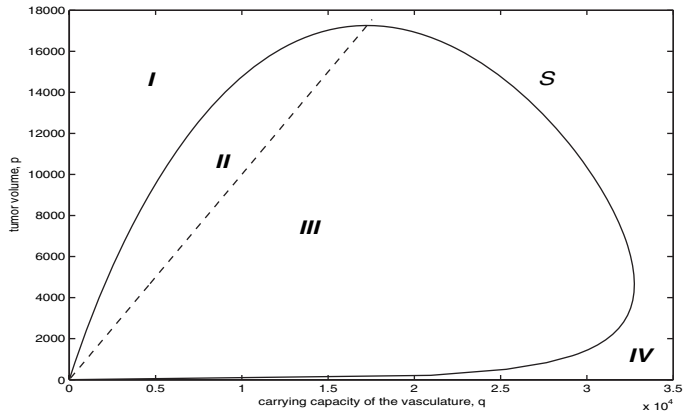


FIG. 5.1. *The regions* I, II, III, *and* IV.

This result limits the possible concatenations in the sense that it provides an upper bound. But for many initial conditions the concatenation structure is quite shorter (pieces are missing), and there exists a unique extremal of this type which is the optimal solution. However, there also are initial conditions for which there exists a one-parameter family of extremals of this type, and in these cases the optimal control needs to be computed numerically through minimizing a 1-dimensional function (see section 8). Once a simple maximal concatenation structure such as the one given in Theorem 5.1 has been determined, this is a straightforward argument.

Clearly, optimal trajectories lie in $\mathbb{R}^3$, and at every point the actions depend on the available amount of inhibitors. However, it is more illustrative to consider the projections of trajectories into the (p,q)-plane,[1] and it is convenient, and only a slight abuse of terminology, not to distinguish in our language between the trajectories in (p,q,y)-space and their projections onto the (p,q)-coordinates.

The anchor piece of the synthesis is an optimal singular arc. It will be shown in section 6 that singular trajectories can lie only on a looplike curve $\mathcal{S}$ in (p,q)-space where the vector fields f and the Lie bracket $[f,g]$ are linearly dependent and that there exists a unique arc Γ on $\mathcal{S}$ where the singular control is admissible, i.e., satisfies the control constraints $0 \le u \le a$. This curve $\mathcal{S}$ and the diagonal $\mathcal{D}_0 = \{(p,q) : p = q\}$ also form boundary curves between optimal bang-bang switchings in the order $a0$ and of the reverse order $0a$, and the concatenation structure of optimal controls is determined by the location of the initial condition relative to these curves. Denote by $\mathcal{S}_+$ the region outside of the singular loop $\mathcal{S}$ and by $\mathcal{S}_-$ the region inside this loop, and define the following regions (see Figure 5.1):

$$\mathrm{I} = \mathcal{D}_+ \cap \mathcal{S}_+, \quad \mathrm{II} = \mathcal{D}_+ \cap \mathcal{S}_-, \quad \mathrm{III} = \mathcal{D}_- \cap \mathcal{S}_-, \quad \mathrm{IV} = \mathcal{D}_- \cap \mathcal{S}_+.$$

In Figure 5.2 we indicate the structure of optimal controlled trajectories for a representative collection of initial conditions and have highlighted one example as a thick curve. (Pieces of trajectories corresponding to $u = a$ are shown as dashed curves, and pieces of $u = 0$ trajectories are shown as solid curves; pieces along the singular arc follow the curve $\mathcal{S}$.) The initial condition for the highlighted trajectory lies in region II, and the optimal control is of the form $as0$: The control initially is

[1] In our graphs we prefer to have q as the horizontal variable and p along the vertical axis. Visually this better corresponds to a decrease or increase in the primary cancer volume.

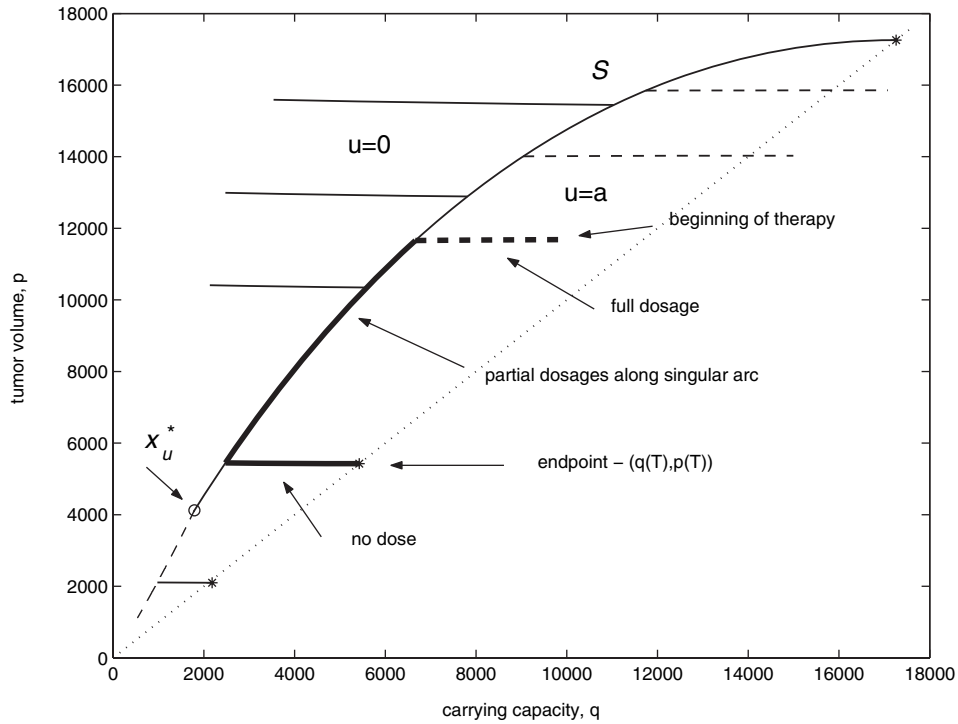


FIG. 5.2. *Synthesis of optimal controlled trajectories.*

given by $u = a$ until the singular arc Γ is reached; at this time the control switches to the singular control and follows the admissible singular arc Γ until all inhibitors have been exhausted; due to aftereffects the minimum value for the tumor volume then is reached when, after termination of therapy, the system crosses the diagonal along the trajectory for $u = 0$. This is the typical structure of optimal controlled trajectories for initial conditions in II, but it depends on two facts: The overall amount of inhibitors is large enough to reach the singular arc, but it is not so large that the singular control would saturate along the singular arc at a specific point x_u^* (computed in section 6.2). If there are not enough inhibitors to reach Γ, the optimal control is simple and is just of type $a0$, giving all inhibitors at maximum dosage from the beginning until they become exhausted. If the amount of inhibitors is large enough so that the singular control would saturate at the point x_u^* while following the singular arc, the structure of optimal controls is more complex, and in this case controls can be of type $a0$ or $asa0$. Typically, as the singular arc Γ is reached, now the control switches and follows the singular arc for some time period, but in this case optimal trajectories leave the singular arc before reaching x_u^*, and the remaining inhibitors are exhausted along a full dose segment with $u = a$. But for trajectories that meet the singular arc close to the saturation point x_u^* optimal controls do not switch to the singular control but simply follow $u = a$ until inhibitors are exhausted. The precise structure of optimal trajectories that come close to the saturation point is rather difficult but for a particular initial condition is easily resolved numerically.

For well-posed initial conditions in regions III and IV, optimal controlled trajectories will eventually enter region II along a trajectory for $u = a$ and then follow the pattern described above. The specific form depends on the amount of inhibitors

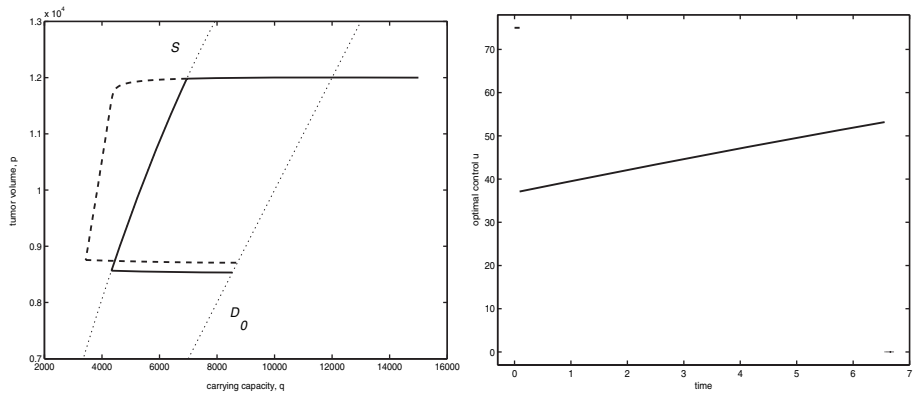


FIG. 5.3. *An optimal as*0 *trajectory with corresponding control.*

left over as the diagonal is crossed. For a typical initial condition, before crossing the diagonal $\mathcal{D}_0$, the control will simply be constant given by $u = a$ until region II is reached. Only for some initial conditions with a low p-value and a relatively high q-value can the control be $u = 0$ initially and then will switch to $u = a$. For some points in IV optimal controls could in principle have the full concatenation structure $0asa0$, but these are not of separate interest for the underlying problem. (Essentially, following $u = 0$ the trajectory enters the part of region III where the control switches to $u = a$ and then ends with the pattern described for region II.)

For initial conditions in region I the most typical structure of optimal controls is $0s0$. Since $p > q$, the tumor is shrinking already, but as the system reaches the singular arc, it is best to administer therapy according to the singular control until all inhibitors become exhausted. As above, if the trajectory comes close to the saturation point, this structure changes into $0sa0$, and if the initial condition actually already is close to this point, it may simply be $a0$ again.

A more precise description of all of these possibilities is given in section 8, where we prove these results. Also, the diagrams shown here are generated using the parameter values given in section 2 that are taken from [9], but the qualitative structure of the solutions described here is robust with respect to parameter changes. Only if the upper limit a on the dosage becomes too small will the singular arc disappear.

In Figure 5.3 we give one specific example of an optimal trajectory (on the left) and its corresponding control (on the right) of the type $as0$. The initial condition is given by $(\tilde{p}, \tilde{q}) = (12000, 15000)$ in region III. The optimal control takes the maximal value $u = a$ for the short interval from 0 to $t_1 = 0.0905$ when the trajectory reaches the singular arc. At this point the control switches to the time-varying singular control until all inhibitors are being exhausted at time $t_2 = 6.5579$. Then, due to aftereffects, the minimum value of the tumor volume is realized a short period later at the final time $T = 6.7221$ when the corresponding $u = 0$ trajectory reaches the diagonal. Note the extremely fast q-dynamics away from the singular arc. The optimal final value is given by $p_*(T) = 8533.4$. The optimal trajectory is shown as a solid curve in Figure 5.3, and the singular curve $\mathcal{S}$ and the diagonal $\mathcal{D}_0$ are shown as dotted curves. For comparison we also show the $a0$ trajectory that corresponds to the trajectory which applies all available inhibitors initially as a dashed curve. (Its initial segment agrees with the optimal trajectory and is not marked separately.) This strategy leads to a tumor reduction with value 8707.4 at time 4.1934.

6. Analysis of the singular arc. In this section we compute an explicit form for the singular control and the corresponding singular curve $\mathcal{S}$. We also show that there exists a unique connected arc of $\mathcal{S}$ where the singular control is admissible, i.e., satisfies the control constraints. Furthermore, the strengthened Legendre–Clebsch condition holds along $\mathcal{S}$, and thus the singular arc is locally optimal.

6.1. Computation of the singular control. Using Proposition 4.5 we get for the switching function $\Phi(t) = \langle \lambda(t), g(z(t)) \rangle$ that

$$\dot{\Phi}(t) = \langle \lambda(t), [f,g]z(t) \rangle \tag{6.1}$$

and

$$\ddot{\Phi}(t) = \langle \lambda(t), [f+ug,[f,g]]z(t) \rangle \,. \tag{6.2}$$

Direct calculations verify that

$$[f,g](z) = Gp \begin{pmatrix} \xi \\ -b \\ 0 \end{pmatrix} \tag{6.3}$$

and

$$[g,[f,g]](z) = -G^2 bp \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}. \tag{6.4}$$

If the switching function $\Phi(t) = \lambda_3 - \lambda_2(t) G q_*(t)$ vanishes at some time t, then $\lambda_2(t)$ is positive since $\lambda_3 > 0$, and thus we have

$$\langle \lambda(t), [g,[f,g]]z(t) \rangle = -\lambda_2(t) G^2 d p_*(t) < 0; \tag{6.5}$$

i.e., the so-called *strengthened Legendre–Clebsch condition* [11] is satisfied. Hence, and provided it is admissible, the singular control is of order 1, locally optimal, and given by

$$u_{\sin}(t) = -\frac{\langle \lambda(t), [f,[f,g]]z(t) \rangle}{\langle \lambda(t), [g,[f,g]]z(t) \rangle}. \tag{6.6}$$

Another direct computation verifies that

$$[f,[f,g]](z) = Gp \begin{pmatrix} \xi^2 + \xi b \frac{p}{q} \\ \xi b \ln\left(\frac{p}{q}\right) + \xi \left(\frac{2}{3} d \frac{q_*(t)}{\sqrt[3]{p_*(t)}} - b \right) - \left(\mu + d p_*^{\frac{2}{3}}(t) \right) b \\ 0 \end{pmatrix}. \tag{6.7}$$

The vector fields g, $[f,g]$, and $[g,[f,g]]$ are everywhere linearly independent, and thus $[f,[f,g]]$ can be expressed as a linear combination in this basis. In fact,

$$[f,[f,g]] = \left(\xi + b\frac{p}{q} \right) [f,g] - \psi [g,[f,g]], \tag{6.8}$$

with

$$\psi = \psi(p,q) = \frac{1}{G} \left(\xi \ln\left(\frac{p}{q}\right) + b\frac{p}{q} + \frac{2}{3}\xi \frac{d}{b} \frac{q}{p^{\frac{1}{3}}} - \left(\mu + d p^{\frac{2}{3}} \right) \right). \tag{6.9}$$

Along a singular arc we have that

$$\dot{\Phi}(t) = \langle \lambda(t), [f,g](z(t)) \rangle \equiv 0, \tag{6.10}$$

and therefore

$$\langle \lambda(t), [f,[f,g]](z(t)) \rangle = -\psi(z(t)) \langle \lambda(t), [g,[f,g]](z(t)) \rangle .$$

Since $\langle \lambda(t), [g,[f,g]](z(t)) \rangle \neq 0$, the singular control defined by (6.6) is simply given by the function ψ in feedback form, more precisely

$$u_{\sin}(t) = -\frac{\langle \lambda(t), [f,[f,g]]z(t) \rangle}{\langle \lambda(t), [g,[f,g]]z(t) \rangle} = \psi(p(t), q(t)). \tag{6.11}$$

Summarizing, we have so far the following.

PROPOSITION 6.1. *Let (z_*, u_*) be an extremal pair. If the control u_* is singular on an open interval (α, β), then u_* is of order 1, and the strengthened Legendre–Clebsch condition is satisfied. The singular control is given in feedback form as*

$$\begin{aligned} u_{\sin}(t) &= \psi(p_*(t), q_*(t)) \\ &= \frac{1}{G}\left(\xi \ln\left(\frac{p_*(t)}{q_*(t)} \right) + b\frac{p_*(t)}{q_*(t)} + \frac{2}{3}\xi \frac{d}{b} \frac{q_*(t)}{p_*^{\frac{1}{3}}(t)} - \left(\mu + d p_*^{\frac{2}{3}}(t) \right) \right). \end{aligned} \tag{6.12}$$

But note that the singular control is admissible only if this value lies in the interval $[0, a]$. Before addressing this issue, we first compute the singular curve itself.

6.2. Computation of the singular curve. For a trajectory to be an extremal, the singular curve also needs to satisfy the extra requirement that $H \equiv 0$ or, equivalently,

$$\langle \lambda(t), f(z(t)) \rangle \equiv 0. \tag{6.13}$$

Hence, along a singular arc, $\lambda(t)$ vanishes against the vector fields f, g, and $[f,g]$. Since $\lambda(t) \neq 0$, these vector fields must be linearly dependent. But g is always linearly independent of f and $[f,g]$, and thus the singular curve is precisely the locus where f and $[f,g]$ are linearly dependent. Both vector fields do not depend on y and have a y-coordinate equal to 0. With slight abuse of notation we can therefore also view them as vector fields on (p,q)-space and define the singular curve $\mathcal{S}$ as

$$\mathcal{S} = \{(p,q) : f(p,q) \wedge [f,g](p,q) = 0\}, \tag{6.14}$$

where

$$f(p,q) \wedge [f,g](p,q) = \begin{vmatrix} -\xi p \ln\left(\frac{p}{q}\right) & \xi \\ bp - \left(\mu + d p^{\frac{2}{3}}\right) q & -b \end{vmatrix}. \tag{6.15}$$

Thus the singular curve is given by the solutions of the equation

$$\mu + d p^{\frac{2}{3}} = b\frac{p}{q}\left(1 - \ln\left(\frac{p}{q} \right) \right). \tag{6.16}$$

The geometry of the singular curve becomes clear if we introduce a projective coordinate, i.e., make a blowup in the variables of the form

$$p = xq, \quad x > 0. \tag{6.17}$$

The quotient $\frac{q}{p}$ is proportional to the *endothelial density* which is used to replace the carrying capacity of the vasculature as a variable in some models like, for example, in [7]. As it turns out, the singular curve and its corresponding singular control can be expressed solely in terms of the variable x introduced here. This fact confirms mathematically the importance of this quantity. However, for the overall analysis, and in particular in view of a ready interpretation of the results, we preferred to keep the original variables p and q and use x only in the analysis of the singular arc. In these variables (6.16) simplifies to

$$\mu + dp^{\frac{2}{3}} = bx\,(1 - \ln x) \tag{6.18}$$

and can be rewritten in the form

$$p^2 + \varphi(x)^3 = 0, \tag{6.19}$$

with

$$\varphi(x) = \frac{bx(\ln x - 1) + \mu}{d}. \tag{6.20}$$

The function φ is strictly convex with a minimum at $x = 1$ and minimum value $\frac{\mu - b}{d}$. In particular, if $\mu \geq b$, then this equation has no positive solutions, and thus no admissible singular arc exists. The case $\mu < b$, which we assumed in (3.5), is the medically relevant case. For $\mu = 0$ the zeros of φ are given by $x_1^* = 0$ and $x_2^* = e$, and φ is negative on the interval $(0, e)$. In general, for $\mu > 0$, we have $\varphi(0) = \frac{\mu}{d} = \varphi(e)$, and thus now the zeros x_1^* and x_2^* satisfy $0 < x_1^* < 1 < x_2^* < e$. We thus have the following.

PROPOSITION 6.2. *The singular curve $\mathcal{S}$ entirely lies in the sector $\{(p, q) : x_1^* q < p < x_2^* q\}$, where x_1^* and x_2^* are the unique zeros of the equation $\varphi(x) = 0$ and satisfy $0 \leq x_1^* < 1 < x_2^* \leq e$. In the variables (p, x), with $x = \frac{p}{q}$, the singular curve can be parameterized in the form*

$$p^2 = \left(\frac{bx(1 - \ln x) - \mu}{d}\right)^3 \qquad \text{for} \quad x_1^* < x < x_2^*. \tag{6.21}$$

PROPOSITION 6.3. *Along the singular arc the singular control can be expressed solely as a function of x in the form*

$$\Psi(x) = \frac{1}{G}\left[\left(\frac{1}{3}\xi + bx\right)\ln x + \frac{2}{3}\xi\left(1 - \frac{\mu}{bx}\right)\right]. \tag{6.22}$$

There exists exactly one connected arc on the singular curve $\mathcal{S}$ along which the control is admissible, i.e., satisfies the bounds $0 \leq \Psi \leq a$. This arc is defined over an interval $[x_\ell^*, x_u^*]$, where x_ℓ^* and x_u^* are the unique solutions to the equations $\Psi(x_\ell^*) = 0$ and $\Psi(x_u^*) = a$, respectively, and these values satisfy $x_1^* < x_\ell^* < x_u^* < x_2^*$.

Figure 6.1 on the left gives a plot of the singular curve for the parameter values from [9] specified earlier and $\mu = 0.02$ and shows the admissible portion of the petallike

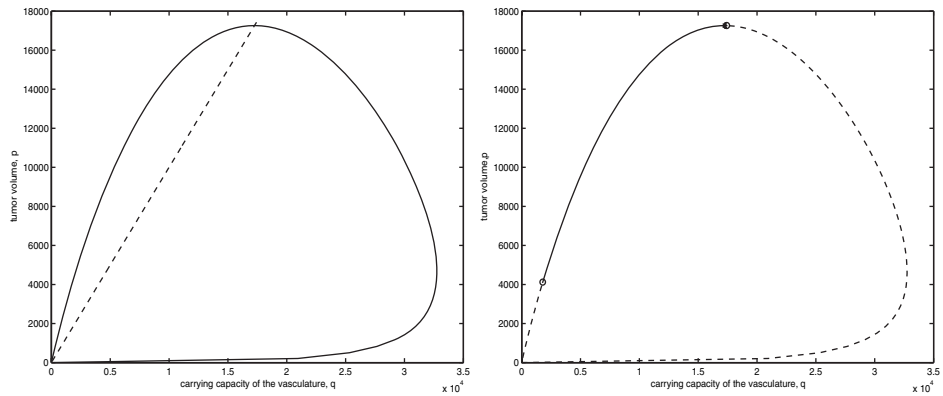


FIG. 6.1. *The singular curve and its admissible part.*

singular curve $\mathcal{S}$ for $a = 75$ marked as a solid curve on the right. The qualitative structure shown in Figure 6.1 is generally valid with the admissible portion shrinking for smaller values a.

Proof. In the variables p and x the singular control is given by

$$u_{\sin}(t) = \frac{1}{G}\left(\xi \ln x(t) + bx(t) + \frac{2}{3}\xi\frac{dp(t)^{\frac{2}{3}}}{bx(t)} - \left(\mu + dp(t)^{\frac{2}{3}}\right)\right). \tag{6.23}$$

But along the singular arc we have $p^{\frac{2}{3}} = -\varphi(x)$, and thus we obtain the singular control as a feedback function of x alone, $u_{\sin}(t) = \Psi(x(t))$, namely,

$$\begin{aligned}\Psi(x) &= \frac{1}{G}\left(\xi \ln x + bx + \frac{2}{3}\xi\frac{bx(1-\ln x)-\mu}{bx} - bx\,(1-\ln x)\right)\\ &= \frac{1}{G}\left[\left(\frac{1}{3}\xi + bx\right)\ln x + \frac{2}{3}\xi\left(1-\frac{\mu}{bx}\right)\right].\end{aligned}$$

Note that $\lim_{x\searrow 0}\Psi(x) = -\infty$ and $\lim_{x\to\infty}\Psi(x) = +\infty$. Now

$$\begin{aligned}\Psi'(x) &= \frac{1}{G}\left[b\,(\ln x + 1) + \frac{1}{3}\xi\left(\frac{1}{x} + 2\frac{\mu}{bx^2}\right)\right],\\ \Psi''(x) &= \frac{1}{Gx^3}\left(bx^2 - \frac{1}{3}\xi x - \frac{4}{3}\xi\frac{\mu}{b}\right),\end{aligned} \tag{6.24}$$

and the second derivative has a unique positive zero at

$$x_* = \frac{1}{6}\frac{\xi}{b}\left(1 + \sqrt{1 + 48\frac{\mu}{\xi}}\right).$$

It follows that Ψ is strictly concave for $0 < x < x_*$ and strictly convex for $x > x_*$. If the function Ψ has no stationary points, then Ψ is strictly increasing, and thus, as claimed, there exists a unique interval $[x_\ell^*, x_u^*]$ when Ψ takes values in $[0, a]$, and the limits are the unique solutions of the equations $\Psi(x) = 0$ and $\Psi(x) = a$, respectively. The same holds if Ψ has a unique stationary point at x_*. In the remaining case, it follows from the convexity properties that Ψ has a unique local maximum at $\tilde{x}_1 < x_*$ and a unique local minimum at $\tilde{x}_2 > x_*$. It suffices to show that Ψ is negative at

its local maximum. This, as before, implies that Ψ is strictly increasing when it is positive. Suppose now that $\Psi'(\tilde{x}) = 0$. Then

$$-b \ln \tilde{x} = b + \frac{1}{3}\xi \left(\frac{1}{\tilde{x}} + 2\frac{\mu}{b\tilde{x}^2} \right) > 0,$$

and thus

$$\begin{aligned} \Psi(\tilde{x}) &= \frac{1}{G}\left[\left(\frac{1}{3}\xi + b\tilde{x} \right) \left(-1 - \frac{1}{3}\frac{\xi}{b} \left(\frac{1}{\tilde{x}} + 2\frac{\mu}{b\tilde{x}^2} \right) \right) + \frac{2}{3}\xi \left(1 - \frac{\mu}{b\tilde{x}} \right) \right] \\ &= \frac{1}{G}\left[-b\tilde{x} - \frac{1}{9}\frac{\xi^2}{b} \left(\frac{1}{\tilde{x}} + 2\frac{\mu}{b\tilde{x}^2} \right) - \frac{4}{3}\frac{\xi\mu}{b\tilde{x}} \right] < 0. \end{aligned}$$

Hence Ψ is negative at any stationary point. □

7. Analysis of bang-bang junctions. Optimal controls are concatenations of the singular control with bang-bang structures, and in this section we analyze possible switchings among bang-bang pieces of an optimal trajectory. We start with a strictly local analysis of switchings that establishes the regions in (p, q)-space where switchings from $u = a$ to $u = 0$ or from $u = 0$ to $u = a$ are possible. We then proceed to analyze extremal bang-bang concatenation structures over the full interval. These results will then be used in section 8 to determine the overall concatenation structures of optimal controls.

The singular curve computed in section 6.2 also is a boundary curve between optimal switchings in the order $a0$ and of the reverse order $0a$. Recall that $\mathrm{I} = \mathcal{D}_+ \cap \mathcal{S}_+$, $\mathrm{II} = \mathcal{D}_+ \cap \mathcal{S}_-$, $\mathrm{III} = \mathcal{D}_- \cap \mathcal{S}_-$, and $\mathrm{IV} = \mathcal{D}_- \cap \mathcal{S}_+$.

PROPOSITION 7.1. *Along optimal trajectories there are no switchings from $u = a$ to $u = 0$ at points $(\tilde{p}, \tilde{q})$ in regions* I *and* III*, and there are no switchings from $u = 0$ to $u = a$ at points $(\tilde{p}, \tilde{q})$ in regions* II *and* IV.

Proof. It follows from (4.7) that the derivative of the switching function must be nonpositive at any time τ where the control switches from $u = 0$ to $u = a$ and nonnegative at every switching from $u = a$ to $u = 0$. Furthermore, since $H \equiv 0$ along extremal lifts, at any switching τ, the adjoint variable $\lambda(\tau)$ vanishes against both $f(z(\tau))$ and $g(z(\tau))$. Except for the points on the diagonal $\mathcal{D}_0 = \{(p, q) : p = q\}$, the vector fields f and g and the coordinate vector field $\frac{\partial}{\partial y} = (0, 0, 1)^T$ are linearly independent, and thus the Lie bracket $[f, g]$ can be written as a linear combination of these vector fields in the form

$$[f, g](z) = \alpha(z) f(z) + \beta(z) g(z) + \gamma(z) \frac{\partial}{\partial y};$$

i.e.,

$$Gp \begin{pmatrix} \xi \\ -b \\ 0 \end{pmatrix} = \alpha(z) \begin{pmatrix} -\xi p \ln\left(\frac{p}{q}\right) \\ bp - \left(\mu + dp^{\frac{2}{3}}\right) q \\ 0 \end{pmatrix} + \beta(z) \begin{pmatrix} 0 \\ -Gq \\ 1 \end{pmatrix} + \gamma(z) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Thus we have

$$\alpha(z) = -\frac{G}{\ln\left(\frac{p}{q}\right)}, \quad \beta(z) = \frac{b\left(\frac{p}{q}\right)\left(\ln\left(\frac{p}{q}\right) - 1\right) + \left(\mu + dp^{\frac{2}{3}}\right)}{\ln\left(\frac{p}{q}\right)},$$

and

$$\gamma(z) = -\beta(z). \tag{7.1}$$

At a switching time τ,

$$\begin{aligned}\dot{\Phi}(\tau) &= \langle \lambda(\tau), [f,g](z(\tau))\rangle \\ &= \alpha(z)\,\langle \lambda(\tau), f(z(\tau))\rangle + \beta(z)\,\langle \lambda(\tau), g(z(\tau))\rangle - \beta(z)\lambda_3 \\ &= -\beta(z)\lambda_3,\end{aligned}$$

and by Lemma 4.3 we may assume that λ_3 is positive. Thus the sign of $\dot{\Phi}(\tau)$ is the opposite of the sign of β. The denominator of β is positive in $\mathcal{D}_+$ and negative in $\mathcal{D}_-$. The zero set of the numerator of β is exactly the locus where the vector fields f and $[f,g]$ are linearly dependent, i.e., the singular curve $\mathcal{S}$ (see (6.16)). We have labeled the regions so that the numerator is positive in $\mathcal{S}_+$ and negative in $\mathcal{S}_-$ (recall that $b > \mu$). Hence $\dot{\Phi}(\tau)$ is negative in regions I and III and positive in regions II and IV. This proves the proposition. □

We now proceed to the analysis of bang-bang controls over the full interval. The admissible portion of the singular arc does not meet $\mathcal{D}_-$, and in $\mathcal{D}_-$ optimal controls will be bang-bang. We therefore begin with this analysis and assume as given a well-posed initial condition $\tilde{z} = (\tilde{p}, \tilde{q}, \tilde{y})$, with $(\tilde{p}, \tilde{q}) \in \mathcal{D}_-$. We first establish that optimal trajectories that start in $\mathcal{D}_-$ will enter $\mathcal{D}_+$ but then cannot return to $\mathcal{D}_-$ any more.

PROPOSITION 7.2. *Suppose $(p_*(\cdot), q_*(\cdot))$ is an optimal trajectory defined over the interval $[0,T]$ with a well-posed initial condition $(\tilde{p}, \tilde{q}) \in \mathcal{D}_-$. Then there exists a time $\tau \in (0,T)$ so that the trajectory lies in $\mathcal{D}_-$ for $t \in [0,\tau)$, crosses into $\mathcal{D}_+$ at time τ, and remains in $\mathcal{D}_+$ for times $t \in (\tau, T)$. Over the interval $[0,\tau)$ the control either is constant given by $u \equiv a$ or is of the form $0a$. In the latter case the junction must lie in the set $\mathcal{N}_- = \{(p,q) \in \mathcal{D}_- : bp < (\mu + dp^{\frac{2}{3}})q\}$.*

Proof. Recall once more that we consider only initial conditions that are well-posed; i.e., the corresponding optimal trajectory does cross over into $\mathcal{D}_+$. Define τ as the (possibly) first time when the trajectory lies on the diagonal $\mathcal{D}_0$. We first show that τ cannot be a switching time. For, if this were the case, then, since $p(\tau) = q(\tau)$,

$$H(\tau) = \lambda_2(\tau)p(\tau)\left(b - \left(\mu + dp(\tau)^{\frac{2}{3}}\right)\right) = 0. \tag{7.2}$$

But on $\mathcal{D}_0$ we have $b > \mu + dp(\tau)^{\frac{2}{3}}$, and thus $\lambda_2(\tau) = 0$. Hence the switching function at time τ is positive: $\Phi(\tau) = \lambda_3 > 0$. But then the control must be $u = 0$ in a neighborhood of τ, and the trajectory crosses from $\mathcal{D}_+$ into $\mathcal{D}_-$, which is a contradiction.

Of the bang controls only $u = a$ steers the system from $\mathcal{D}_-$ into $\mathcal{D}_+$, and nothing more needs to be shown about the interval $[0,\tau)$ if the control is constant and given by $u = a$ on this interval. If not, there exists a maximal interval (α,β), with $0 < \alpha < \tau < \beta < T$, so that the switching function is negative for $t \in (\alpha,\beta)$ and has zeros at α and β: $\Phi(\alpha) = \Phi(\beta) = 0$. The function Φ has a minimum over the interval $[\alpha,\beta]$ at some time $\sigma \in (\alpha,\beta)$, and by (6.1) and (6.3) we have

$$\dot{\Phi}(\sigma) = Gp(\sigma)\,(\xi\lambda_1(\sigma) - b\lambda_2(\sigma)) = 0. \tag{7.3}$$

Hence $\lambda_1(\sigma)$ and $\lambda_2(\sigma)$ have the same sign. But λ_2 is positive along trajectories for $u = a$ since $\Phi(t) = \lambda_3 - \lambda_2(t)Gq(t) < 0$ and $\lambda_3 > 0$. Hence $\lambda_1(\sigma) > 0$. Since there is a

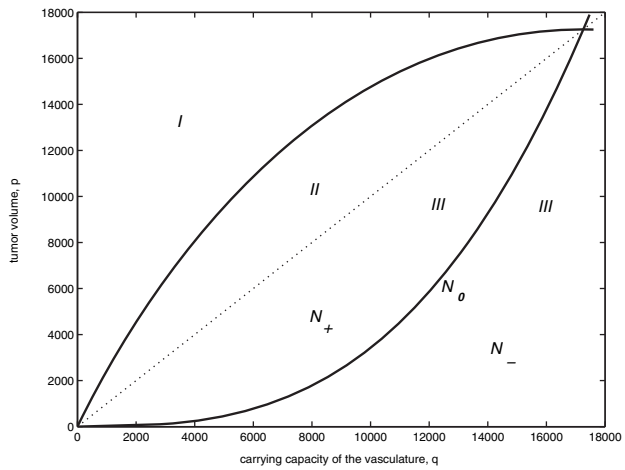


FIG. 7.1. *The curve* $\mathcal{N}_0$.

junction at time α, there also exists an interval $(\alpha - \varepsilon, \alpha)$ where the control is $u = 0$, and on this interval we have

$$H(t) = -\lambda_1(t)\xi p(t) \ln\left(\frac{p(t)}{q(t)}\right) + \lambda_2(t)\left(bp(t) - \left(\mu + dp(t)^{\frac{2}{3}}\right) q(t)\right) = 0, \tag{7.4}$$

while λ_2 is still positive for ε small. If we have

$$bp(\alpha) > (\mu + dp(\alpha)^{\frac{2}{3}})q(\alpha) \tag{7.5}$$

at the junction, then, by (7.4), λ_1 must be negative on this interval and also $\lambda_1(\alpha) < 0$. Hence there exists a last zero for λ_1 in the interval (α, σ), say, $\lambda_1(\rho) = 0$. At this zero the adjoint equation (4.2) reads

$$\dot{\lambda}_1(\rho) = \lambda_2(\rho)\left(\frac{2}{3} d \frac{q(\rho)}{\sqrt[3]{p(\rho)}} - b\right). \tag{7.6}$$

The curve $\mathcal{N}_0 = \{(p, q) : bp = (\mu + dp^{\frac{2}{3}})q\}$ is the $\dot{q}$-nullcline for $u = 0$, and since at the points on the $\dot{q}$-nullcline we have

$$f(z) = \begin{pmatrix} -\xi p \ln\left(\frac{p}{q}\right) \\ 0 \\ 0 \end{pmatrix} \tag{7.7}$$

with the first coordinate positive in $\mathcal{D}_-$, corresponding trajectories cross $\mathcal{N}_0$ from $\mathcal{N}_- = \{(p, q) : bp < (\mu + dp^{\frac{2}{3}})q\}$ into $\mathcal{N}_+ = \{(p, q) : bp > (\mu + dp^{\frac{2}{3}})q\}$ (Figure 7.1).

Because of the extra control term $-Gqa$ in the vector field g, this also holds for trajectories corresponding to the control $u = a$. Hence, if the junction at time α satisfies (7.5), then so does the arc forward in time; i.e., for all $t \in (\alpha, \sigma)$ we have

$$q(t) < \frac{bp(t)}{\mu + dp(t)^{\frac{2}{3}}}. \tag{7.8}$$

Hence

$$\frac{2}{3}d\frac{q(\rho)}{\sqrt[3]{p(\rho)}} - b < \frac{2}{3}d\frac{bp(\rho)^{\frac{2}{3}}}{\mu + dp(\rho)^{\frac{2}{3}}} - b = \frac{2}{3}b\frac{dp(\rho)^{\frac{2}{3}}}{\mu + dp(\rho)^{\frac{2}{3}}} - b < -\frac{1}{3}b < 0. \tag{7.9}$$

Since the multiplier λ_2 is positive along $u = a$, we therefore have $\dot{\lambda}_1(\rho) < 0$, which is a contradiction. Thus no $0a$-junction can lie in the interior of $\mathcal{N}_+$. The same reasoning also precludes that $0a$-junctions would lie on the $\dot{q}$-nullcline for $u = 0$, i.e., on $\mathcal{N}_0$. In this case it follows that $\lambda_1(\alpha) = 0$ and $\dot{\lambda}_1(\alpha) < 0$, and thus again there still exists a last zero of λ_1 in the interval (α, σ) where the same contradiction arises. Thus, any possible $0a$-junction in $\mathcal{D}_-$ must lie in $\mathcal{N}_-$.

We now show that if there is a junction at some time α, then the control must be constant $u = 0$ on the initial interval $[0, \alpha)$. Since trajectories for $u = 0$ cross $\mathcal{N}_0$ from $\mathcal{N}_-$ into $\mathcal{N}_+$, it follows that, as long as the control $u = 0$ is used for $t < \alpha$, the trajectory lies in $\mathcal{N}_-$. Since this trajectory also necessarily lies in $\mathcal{D}_-$ we have $p(t) < q(t)$. But then the identity

$$H(t) = -\lambda_1(t)\xi p(t)\ln\left(\frac{p(t)}{q(t)}\right) + \lambda_2(t)\left(bp(t) - \left(\mu + dp(t)^{\frac{2}{3}}\right)q(t)\right) = 0 \tag{7.10}$$

implies that neither λ_1 nor λ_2 can have any zeros (otherwise they would need to vanish simultaneously, contradicting Lemma 4.2). Hence λ_2 is positive, and λ_1 must be negative as long as the control is $u = 0$. This precludes any more switchings. For, if there is another switching $0 < \theta < \alpha$, then there also needs to be another zero for the derivative of the switching function in (θ, α), and at this derivative λ_1 and λ_2 must have the same sign.

Thus, if $(\tilde{p}, \tilde{q}) \in \mathcal{D}_-$ is a well-posed initial condition, then there exists a first time τ at which the trajectory crosses from $\mathcal{D}_-$ into $\mathcal{D}_+$ and the control over $[0, \tau)$ is either constant and given by $u \equiv a$ or it has exactly one switching from 0 to a with the junction in $\mathcal{N}_-$. It still remains to show that the trajectory cannot return from $\mathcal{D}_+$ into $\mathcal{D}_-$ for times $t > \tau$, i.e., that $(p(t), q(t))$ lies in $\mathcal{D}_+$ for $t \in (\tau, T)$. (It follows from Lemma 4.1 that the end point lies on $\mathcal{D}_0$.)

If the trajectory were to return to $\mathcal{D}_-$, then there would exist another time $\kappa > \beta$ where the trajectory again would cross $\mathcal{D}_0$ with control $u = 0$ in a neighborhood of κ. Now (7.10) implies that $\lambda_2(\kappa) = 0$, and thus the adjoint equation for λ_2 gives $\dot{\lambda}_2(\kappa) = -\xi\lambda_1(\kappa)$. Since λ_1 and λ_2 cannot vanish simultaneously, we have $\lambda_1(\kappa) \neq 0$, and λ_2 changes sign at τ. If $\lambda_1(\kappa) > 0$, then λ_2 is negative for $t > \tau$, t near τ. However, after crossing into $\mathcal{D}_-$, trajectories for $u = 0$ entirely lie in $\mathcal{D}_-$ and do not cross back into $\mathcal{D}_+$. Consequently it follows from (7.10) that λ_2 cannot have another zero along an $u = 0$ arc. (In $\mathcal{D}_-$ the expression $\lambda_1\xi p\ln(\frac{p}{q})$ can vanish only if $\lambda_1 = 0$, and this precludes λ_2 from having a zero.) But then the switching function $\Phi(t) = \lambda_3 - \lambda_2(t)Gq$ remains positive, and so there cannot be another switching in the control. But this structure clearly is not optimal since the value for p increases in $\mathcal{D}_-$, which is a contradiction. If $\lambda_1(\kappa) < 0$, we are back in the situation considered above: If all inhibitors have not been used up, there needs to exist another switching to $u = a$; but the entire forward orbit of the $u = 0$ trajectory lies in $\mathcal{N}_+$, and therefore this switching would lie in $\mathcal{N}_+$, violating the earlier statement. □

This proposition implies that, for well-posed initial conditions $(\tilde{p}, \tilde{q}, 0)$, with $(\tilde{p}, \tilde{q}) \in \mathcal{N}_+$, the optimal control is given in feedback form by $u \equiv a$ until the trajectory crosses over into $\mathcal{D}_+$ at time τ. The further structure of the optimal control depends on the

amount of inhibitors $y(\tau)$ that are still available at this time. If this amount is too small to reach the singular arc, then, since only junctions in the order $a0$ are optimal in region II $= \mathcal{D}_+ \cap \mathcal{S}_-$, the optimal control is simply given by $u = a$ until all inhibitors are exhausted and then follows $u = 0$ until the trajectory terminates on the diagonal $\mathcal{D}_0$. (If the control would switch prior to the time when all inhibitors are exhausted, by Proposition 7.1 it cannot switch back to the control $u = a$ in region II and thus would reach the diagonal with inhibitors still available, contradicting Lemma 4.1.) Hence, if enough inhibitors are available for the trajectory to reach the singular curve $\mathcal{S}$, then the trajectory will follow $u = a$ in region II.

We now consider the further structure of optimal controls for segments of the trajectory that lie in $\mathcal{D}_+$. We first show that segments corresponding to the control $u = 0$ can lie only at the beginning or at the end of the interval $[0, T]$.

PROPOSITION 7.3. *Let (α, β) be a maximal open interval where the optimal control is given by $u \equiv 0$ with corresponding trajectory (p_*, q_*) lying in $\mathcal{D}_+$. Then α and β cannot both be switching times. If α is a switching time, then $\beta = T$, the final time, and if β is a switching time, then $\alpha = 0$, the initial time.*

Proof. As before, on the interval (α, β), (7.10) holds, and on $\mathcal{D}_+$ we have $\xi p \ln(\frac{p}{q}) > 0$ and $bp - (\mu + dp^{\frac{2}{3}})q > 0$. As above, in this case neither λ_1 nor λ_2 can vanish, and therefore λ_1 and λ_2 have the same sign over (α, β). Since λ_2 is positive at switching times, it follows that both λ_1 and λ_2 are positive over $[\alpha, \beta]$ if at least one of the end points is a switching time. Along $u = 0$ the derivatives of the switching function are given by $\dot{\Phi}(t) = \langle \lambda(t), [f, g](z(t)) \rangle$ and $\ddot{\Phi}(t) = \langle \lambda(t), [f, [f, g]](z(t)) \rangle$. If there exists a time $\tau \in (\alpha, \beta)$ where $\dot{\Phi}(\tau) = 0$, then it follows from (6.3) and (6.8) that

$$\begin{aligned}
\ddot{\Phi}(\tau) &= \left(\xi + b\frac{p(\tau)}{q(\tau)}\right) \langle \lambda(\tau), [f, g](z(\tau)) \rangle - \psi(p(\tau), q(\tau)) \langle \lambda(\tau), [g, [f, g]](z(\tau)) \rangle \\
&= \left(\xi + b\frac{p(\tau)}{q(\tau)}\right) \dot{\Phi}(\tau) - \psi(p(\tau), q(\tau)) \langle \lambda(\tau), [g, [f, g]](z(\tau)) \rangle \\
&= \psi(p(\tau), q(\tau)) b G^2 p(\tau) \lambda_2(\tau) > 0. \qquad (7.11)
\end{aligned}$$

Here we use the fact that ψ is positive in $\mathcal{D}_+$.

Suppose α is a switching time. Then there exists an $\varepsilon > 0$ so that $\dot{\Phi}$ is positive in $(\alpha, \alpha + \varepsilon)$. (Since the control is $u = 0$, we have $\dot{\Phi}(\tau) \geq 0$, and even if $\dot{\Phi}(\alpha) = 0$, then this is implied by $\ddot{\Phi}(\alpha) > 0$.) Thus, if $\dot{\Phi}$ has zeros in (α, β), then there exists a smallest one; call it τ. But then $\dot{\Phi}(t) > 0$ on the interval (α, τ), and so Φ cannot have a local minimum at τ, contradicting $\ddot{\Phi}(\tau) > 0$. Hence Φ is strictly increasing over (α, β) as long as the control $u = 0$ is used, and there cannot be another zero at β. Similarly, if β is a switching time, then Φ is strictly decreasing over (α, β) as long as the control $u = 0$ is used, and again there cannot exist a previous zero at α. ∎

PROPOSITION 7.4. *Suppose (p_*, q_*) is an optimal trajectory corresponding to the constant control $u = a$ over some open interval (α, β) with switching times at α and β. Then $(p(\alpha), q(\alpha)) \notin$ II and $(p(\beta), q(\beta)) \in$ II. Furthermore, there exists a time $\tau \in (\alpha, \beta)$ where $\psi(p(t), q(t)) \geq a$.*

Proof. The statements about the junction points follow from Proposition 7.1. Along $u = a$ the switching function is negative over (α, β) and has a minimum at some time $\tau \in (\alpha, \beta)$ where $\dot{\Phi}(\tau) = 0$ and $\ddot{\Phi}(\tau) \geq 0$; the derivatives of the switching function are now given by $\dot{\Phi}(t) = \langle \lambda(t), [f, g](z(t)) \rangle$ and

$$\ddot{\Phi}(t) = \langle \lambda(t), [f + ag, [f, g]](z(t)) \rangle .$$

As above, it follows from (6.3) and (6.8) that

$$\begin{aligned}\ddot{\Phi}(\tau) &= \left(\xi + b\frac{p(\tau)}{q(\tau)}\right)\langle\lambda(\tau),[f,g](z(\tau))\rangle + \{a-\psi(p(\tau),q(\tau))\}\langle\lambda(\tau),[g,[f,g]](z(\tau))\rangle \\ &= \left(\xi + b\frac{p(\tau)}{q(\tau)}\right)\dot{\Phi}(\tau) + \{a-\psi(p(\tau),q(\tau))\}\langle\lambda(\tau),[g,[f,g]](z(\tau))\rangle \\ &= \{\psi(p(\tau),q(\tau))-a\}\,bG^2p(\tau)\lambda_2(\tau). \end{aligned} \tag{7.12}$$

Since $\Phi(\tau) = \lambda_3 - \lambda_2(\tau)Gaq(\tau) < 0$, we have $\lambda_2(\tau) > 0$, and thus we must have $\psi(p(\tau),q(\tau)) \geq a$. This proves the result. □

8. Synthesis of optimal controls. We now put together the results, consider concatenations between singular and bang arcs, and prove the results of section 5. We start with a strictly local analysis of singular junctions analogous to Proposition 7.1 for bang-bang junctions and show that all possible concatenations of bang controls with the singular arc are extremal. This classical result is included for the sake of completeness and is a direct consequence of the fact that the strengthened Legendre–Clebsch condition is satisfied. We then show that extremals which contain a saturating arc are not optimal and proceed to the analysis of the possible concatenations of bang and singular arcs.

PROPOSITION 8.1. *Let I be an open interval on which the optimal control u_* is singular and takes values in the interior of the control set. Then concatenations of both the forms* **bs** *and* **sb**, *where* **b** *stands for any of the two bang controls $u = 0$ or $u = a$, are extremal along I.*

Proof. Recall that, for any control u that is continuous from the left $(-)$ or right $(+)$, the second derivative of the switching function is given by

$$\ddot{\Phi}(t\pm) = \langle\lambda(t),[f,[f,g]](z(t))\rangle + u(t\pm)\langle\lambda(t),[g,[f,g]](z(t))\rangle,$$

and it vanishes identically on I along the singular control. Since the strengthened Legendre–Clebsch condition is satisfied, we have $\langle\lambda(t),[g,[f,g]](z(t))\rangle < 0$. By assumption the singular control takes values in the interior of the control set $[0,a]$, and thus $\langle\lambda(t),[f,[f,g]](z(t))\rangle > 0$. Hence, for $u = 0$ we get $\ddot{\Phi}(t) > 0$, and for $u = a$ we have $\ddot{\Phi}(t) < 0$. These signs are consistent with entry and exit from the singular arc for each control; i.e., for example, if $u = 0$ on an interval $(\tau-\varepsilon,\tau)$, then Φ is positive over this interval, consistent with the choice $u = 0$ as minimizing control. □

Thus, as long as the singular control has not saturated, it is possible to jump onto or off the singular arc with the constant controls $u = 0$ or $u = a$ at any point without violating the conditions of the maximum principle locally. Naturally, optimality over longer time intervals is not guaranteed and still needs to be analyzed. As an example, it follows from Proposition 7.3 that the singular arc can be left with the control $u = 0$ only when all inhibitors have been exhausted. On the other hand, trajectories for $u = a$ can (and sometimes must) leave the singular arc before all inhibitors have been exhausted. This follows from the result below. Recall that x_u^* is the point introduced in Proposition 6.3 where the singular control saturates at the upper control value a.

PROPOSITION 8.2. *At the saturation point x_u^* on the singular arc where the singular control saturates at the upper value $u = a$, it is not optimal to continue the control with $u = a$. Thus optimal trajectories need to leave the singular arc before saturation.*

This result may seem somewhat counterintuitive, but this is indeed the typical behavior at saturation in low dimensions (see, for example, [17] or [2]).

Proof. Consider the trajectory that follows the singular arc and at the saturation time τ continues with the control $u = a$. In general, we have

$$\ddot{\Phi}(t) = \left(\xi + b\frac{p(t)}{q(t)}\right)\dot{\Phi}(t) + (u(t) - \psi(p(t), q(t)))\,\langle\lambda(t), [g, [f, g]](z(t))\rangle\,.$$

Along the singular arc $\dot{\Phi}(\tau) = 0$ and at the saturation point we also have $\ddot{\Phi}(\tau) = 0$ for the control $u = a$ since $\psi = a$. Hence, along $u = a$ we get from the right that

$$\Phi^{(3)}(\tau+) = -\left(\frac{d}{dt}_{|t=\tau}\psi(p(t), q(t))\right)\langle\lambda(t), [g, [f, g]](z(t))\rangle \tag{8.1}$$

$$= \left(\frac{d}{dt}_{|t=\tau}\psi(p(t), q(t))\right) bG^2\lambda_2(\tau)p(\tau). \tag{8.2}$$

Recall that $\psi(p(t), q(t)) = \Psi(x(t))$, with $x = \frac{p}{q}$ and Ψ defined in (6.22) in Proposition 6.3. We thus have

$$\frac{d}{dt}_{|t=\tau}\psi(p(t), q(t)) = \Psi'(x_u^*)\dot{x}(\tau). \tag{8.3}$$

It follows from the proof of Proposition 6.3 that $\Psi'(x_u^*) > 0$, and in general we have

$$\dot{x} = \frac{\dot{p}q - p\dot{q}}{q^2} = -\xi x\ln x - bx^2 + (\mu + dp^{\frac{2}{3}})x + Gux.$$

Substituting $(\mu + dp^{\frac{2}{3}}) = bx(1 - \ln x)$ along the singular arc (cf. (6.18)), we get

$$\dot{x} = x\,(Gu - (\xi + bx)\ln x)\,.$$

But at the saturation point we also have

$$Gu(\tau) = Ga = \left(\frac{1}{3}\xi + bx(\tau)\right)\ln x(\tau) + \frac{2}{3}\xi\left(1 - \frac{\mu}{bx(\tau)}\right),$$

and thus

$$\dot{x}(\tau) = \frac{2}{3}\xi\left(x(\tau)\,(1 - \ln x(\tau)) - \frac{\mu}{b}\right) = \frac{2}{3}\frac{\xi}{b}dp(\tau)^{\frac{2}{3}} > 0.$$

Hence

$$\Phi^{(3)}(\tau+) > 0, \tag{8.4}$$

and Φ is positive for $t > \tau$, t near τ, contradicting the minimization property for $u = a$. □

Thus optimal trajectories cannot continue with the saturated control $u = a$ after the saturation point but instead must leave the singular arc prior to saturating. The analogous computation with $u = 0$ for $t > \tau$ shows that we can switch to $u = 0$ at saturation, but by Proposition 7.3 this is optimal only if all inhibitors have been exhausted. In general, if inhibitors are available to go beyond the saturation point, optimal trajectories must leave the singular arc before saturation occurs. When precisely this happens depends on the amount of inhibitors left. For example, it is

clear that if $\tilde{z} = (\tilde{p}, \tilde{q}, \tilde{y})$ is a point with $(\tilde{p}, \tilde{q}) \in \mathcal{S}$ before saturation for which $A - \tilde{y}$ is small, then it is not optimal to leave the singular arc simply because there are not enough inhibitors left so that the system could reach the region $\mathrm{II} = \mathcal{D}_+ \cap \mathcal{S}_-$, where a switching from $u = a$ to $u = 0$ again is optimal. In this case optimal trajectories follow the singular arc until inhibitors are exhausted. However, if enough inhibitors are available so that the singular arc would lose admissibility before they all are used up, then indeed by leaving the singular arc earlier trajectories can enter region II, and this will be the optimal strategy. But now the argument needs to become global. Since by now we have sufficiently reduced the possibilities in the structures of optimal controls and trajectories, we can reduce this problem to a 1-dimensional optimization problem that can easily be solved numerically. Based on our previous analysis of the structure of extremals, we now determine the optimal synthesis on $\mathcal{D}$. We not only establish the qualitative structure claimed in Theorem 5.1 but also show how to compute the optimal control for a given initial condition $\tilde{z} = (\tilde{p}, \tilde{q}, 0)$.

We will prove that, except for some initial conditions $(\tilde{p}, \tilde{q})$ in regions III and IV (which are less relevant for the underlying problem), our local analysis above only allows for at most a one-parameter family of extremals $\Gamma_\varepsilon(s) = (p_\varepsilon(s), q_\varepsilon(s), y_\varepsilon(s))$, $0 \le s \le T(\varepsilon)$, with the parameter ε ranging over a compact interval $I = [0, \theta]$. The corresponding value of the objective is given by $\upsilon(\varepsilon)$, $\upsilon(\varepsilon) = p_\varepsilon(T(\varepsilon))$, and it will be clear from the definition of the family Γ_ε that the value $\upsilon(\varepsilon)$ depends continuously on ε. Thus, for every initial condition there exists an optimal control that is determined by numerical minimization of $\upsilon(\varepsilon)$ over $[0, \theta]$. However, the structure of this one-parameter family of extremals depends on the location of $(\tilde{p}, \tilde{q})$, and we need to distinguish three cases.

Case 1. We start with an initial condition $(\tilde{p}, \tilde{q})$ in region I and will show that in this case optimal controls at most have the form **0sa0** (possibly with the initial **0s** sequence absent). Let ς_+ denote the reference trajectory that starts at $\tilde{z} = (\tilde{p}, \tilde{q}, 0)$, uses the control $u = 0$ until the singular arc is reached at some time τ (the existence of such a time is clear for initial conditions of this type), and then follows the singular arc for time σ until either all available inhibitors have been exhausted or the saturation point is reached. We now use the time ε along this trajectory as a parameter and construct the family Γ_ε over the compact parameter interval $[0, \theta]$, with $\theta = \tau + \sigma$ as follows: The trajectory $\Gamma_\varepsilon(\cdot)$ agrees with the reference trajectory ς_+ up to time ε and switches at time ε to the control $u = a$, which will then be followed until all remaining inhibitors have been exhausted, and then the control still is $u = 0$ until the trajectory terminates at time T as the diagonal is reached.

This family indeed contains all possible extremals starting at initial condition $\tilde{z}$: Initially the control can be only $u = 0$ or $u = a$. If the control is $u = a$, then it follows from the phase portrait for $u = a$ that this trajectory does not meet the admissible portion of the singular curve. Since any possible junction will lie in $\mathcal{D}_+$, it follows from Proposition 7.3 that the control can switch only to $u = 0$ as all inhibitors have been exhausted. At this point the optimal control then is still given by $u = 0$ until the diagonal $\mathcal{D}_0$ is reached at time $T(0)$ and the trajectory is terminated. This is the trajectory $\Gamma_0(\cdot)$ in our family. It actually follows from Proposition 7.1 that this trajectory would not be an extremal if the switching were to lie in region I, and in this case this trajectory could be excluded a priori. However, even if this is the case, for simplicity of argument we retain this curve anyway. Similarly, if initially the control is $u = 0$ on $[0, \varepsilon]$, with $\varepsilon \le \tau$, and then switches to $u = a$, the same reasoning applies, and the structure of the corresponding control is simply $0a0$. For $\varepsilon > \tau$, the trajectory $\Gamma_\varepsilon(\cdot)$ now follows the control $u = 0$ until time τ and then stays on the singular arc

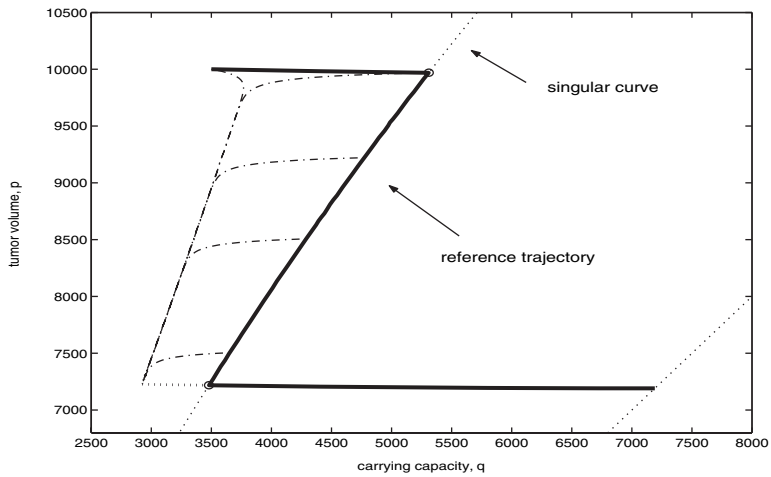


FIG. 8.1. *Parameterized extremals for an initial condition in region* I *without saturation.*

until time ε, at which time it leaves the singular arc to use $u = a$ until all inhibitors are being exhausted. The remaining time along the a trajectory is given by

$$\eta(\varepsilon) = \frac{1}{a}\left(A - \int_{\tau}^{\varepsilon} \psi(p_\varepsilon(s), q_\varepsilon(s))ds\right), \tag{8.5}$$

with $\eta(\tau + \theta) = 0$ if the inhibitors get exhausted along the singular arc. If the singular control saturates, we know that the trajectory $\Gamma_\theta(\cdot)$ is no longer optimal and therefore can terminate the construction of the one-parameter family of extremals with the trajectory that switches to $u = a$ at saturation. This simply provides us with a compact parameter interval, but in this case the minimum will be attained at a parameter $\varepsilon < \theta$. It follows from our local analysis given before that any possible extremal that could start at $\tilde{z}$ is part of this family $\Gamma_\varepsilon(\cdot)$. Essentially, trajectories cannot switch to $u = 0$ before all inhibitors have been exhausted, and thus once they switch to $u = a$ they need to use up all remaining inhibitors.

It is just a consequence of the continuous dependence of a solution to an ordinary differential equation on initial data and parameters that the end point $p_\varepsilon(T(\varepsilon))$ and thus the value $\upsilon(\varepsilon)$ depend continuously on ε. Hence, if $\hat{\varepsilon}$ is a parameter value where $\upsilon(\varepsilon)$ attains its minimum over $[0, \theta]$, then $\Gamma_{\hat{\varepsilon}}(\cdot)$ is the optimal trajectory starting at $\tilde{z}$ with a correspondingly defined optimal control.

The family $\Gamma_\varepsilon(s)$, $0 \le s \le T(\varepsilon)$, is illustrated in Figure 8.1 for the initial conditions $\tilde{p} = 10000$ and $\tilde{q} = 3500$. The reference trajectory is shown as a thick solid curve, and some sample trajectories of Γ_ε are shown as dashed-dotted curves. The heavily dotted curve is the curve of points when all inhibitors are being exhausted. Here the optimal control is of the type $0s0$ and given for $\varepsilon = \theta$; i.e., the optimal trajectory follows $u = 0$ until the singular curve $\mathcal{S}$ is reached, then follows the singular arc until all inhibitors have been exhausted, and finally uses $u = 0$ to reach the diagonal. This is always the case for initial conditions whose available inhibitors are too small to reach region II inside the loop $\mathcal{S}$ using the control $u = a$. This is the case for this example, and in fact the only extremal corresponding to this initial condition is the optimal trajectory Γ_θ. On the other hand, if initial conditions have an abundance of inhibitors so that the singular arc would saturate, then optimal trajectories exit the singular arc.

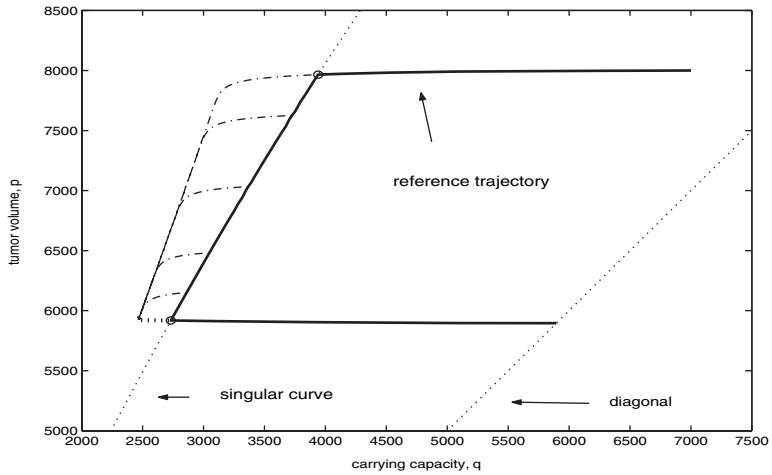


FIG. 8.2. *Parameterized extremals for an initial condition in region* II *without saturation.*

Case 2. Let $G = \text{II} \cup \mathcal{D}_0 \cup (\text{III} \cap \mathcal{N}_+)$, i.e., the region inside the loop $\mathcal{S}$, but "above" the curve $\mathcal{N}_0$, and consider a well-posed initial condition $\tilde{z} = (\tilde{p}, \tilde{q}, 0) \in G$. Then the optimal control is initially given by $u = a$. (If the control starts with $u = 0$, then the entire forward orbit lies in G, and in G switchings from 0 to a are not optimal (cf. Propositions 7.1 and 7.2)). If the a trajectory starting at $\tilde{z}$ does not intersect the admissible singular arc, the optimal control is simply $u = a$ until all inhibitors have been exhausted, and then $u = 0$ until the diagonal is reached. Nothing else needs to be done in this case. Otherwise let ς_- denote the reference trajectory that starts at $\tilde{z} = (\tilde{p}, \tilde{q}, 0)$, uses the control $u = a$ until the singular arc is reached at some time τ, and then, as above, follows the singular arc for time σ until either all available inhibitors have been exhausted or the saturation point is reached. Again we use the time along the singular arc of the reference trajectory as a parameter and construct the family Γ_ε over the compact parameter interval $[0, \theta]$, with $\theta = \sigma$ as follows: The trajectory $\Gamma_\varepsilon(\cdot)$ agrees with the reference trajectory ς_- along the initial segment for $u = a$ until the singular arc is reached at time τ and then $\Gamma_\varepsilon(\cdot)$ still follows the singular arc for time ε when it again switches to the control $u = a$. The end is as before: The control $u = a$ is used until all remaining inhibitors have been exhausted, at which time a final segment with $u = 0$ is added until the trajectory terminates at time T on the diagonal. Note that $\varepsilon = 0$ corresponds to the special case when the trajectory does not follow the singular arc but continues straight with $u = a$. As above, this family contains all possible extremals that start at $\tilde{z}$ but also may have some members that are not extremals (for example, the second arc with $u = 0$ may violate Proposition 7.4). The optimal trajectory is given by $\Gamma_{\hat{\varepsilon}}(\cdot)$, where $\hat{\varepsilon}$ is a minimizer of $\upsilon(\varepsilon)$ over $[0, \theta]$. In particular, in this case optimal controls at most have the form **asa0** with possibly some of the pieces absent.

Figure 8.2 shows an example of this family $\Gamma_\varepsilon(s)$, $0 \le s \le T(\varepsilon)$, for initial conditions $p_0 = 8000$ and $q_0 = 7000$. Like in Figure 8.1 the reference trajectory ς_- is shown as a thick solid curve, sample trajectories from the family are shown as dashed-dotted curves, and the heavily dotted curve of points is the curve when all inhibitors have been exhausted. As above, no saturation occurs, and it is optimal to follow the singular arc until all inhibitors have been exhausted (the trajectory corresponding to

the rightmost parameter value θ gives the minimum value). Also for this example this is in fact the only extremal.

This, however, is no longer true when there are still inhibitors available at the saturation point. An example of this scenario is given for the initial condition $p_0 = 5000$ and $q_0 = 2500$ in region III. In this case, however, the numerical differences between the values of the objectives are minute. In fact, no difference in the trajectories of the family is discernable, and the cost varies only between 3761.65 and 3761.98. It is clear that, although present mathematically, these differences are of no significance, and for all practical purposes one may simply continue the singular arc at saturation with $u = a$ without any noticeable loss.

It is furthermore clear that the case of initial conditions with $(\tilde{p}, \tilde{q})$ on the admissible portion of the singular arc can be analyzed in exactly the same way by setting $\tau = 0$. If the initial condition lies on the inadmissible portion, optimal controls are of the form $a0$ with all inhibitors being exhausted along $u = a$.

Case 3. The last case corresponds to initial conditions $(\tilde{p}, \tilde{q})$ in the region $F = (\mathrm{III} \cap \mathcal{N}_-) \cup (\mathcal{S} \cap \mathcal{D}_-) \cup \mathrm{IV}$, i.e., points that lie in $\mathcal{D}_-$ "below" $\mathcal{N}_0$. These initial conditions all have relatively very large q-values in contrast to small p-values. In these cases, in principle, the control can start with $u = 0$ and switch to $u = a$, while still in F. Once the control is $u = a$, the trajectory enters the region G and the construction of Case 2 applies. Thus, and if a large amount of inhibitors is available, here the full structure **0asa0** can arise. Since trajectories eventually enter the region G, and this leads to a repetition of the construction, we skip a precise description of what now would be a two-parameter family of extremals over a compact rectangle.

9. Conclusion. We presented a complete solution for a mathematical model for tumor antiangiogenesis for the problem of optimally scheduling a given number of inhibitors in order to minimize the primary tumor volume. Based on our theoretical analysis of the problem, for any specific initial condition the optimal solution can easily be computed numerically and as such provides a benchmark value to which other strategies should be compared. From a practical point of view, it is not realistic to employ the singular control. It is a feedback control, and the required information certainly is not available, although it could be predetermined offline from the initial condition. Naturally, strategies of the type $a0$ which give all available inhibitors in one session are the easiest to implement in practice. It follows from our analysis that for some initial conditions these are indeed the optimal ones. This is certainly the case for initial conditions for which a trajectories do not meet the admissible singular arc but also for initial conditions when this intersection point is close to the saturation point. Indeed, the dynamics for $u \equiv a$ very much has a differential algebraic structure with the q-dynamics fast and the p-dynamics slow. As a result, after a brief transient phase in steady state the system essentially follows the $\dot{q}$-nullcline. This nullcline is very close to the singular curve near the saturation point, and thus there the differences in the objective are almost unnoticeable. For initial conditions far away from this point the singular arc and the $\dot{q}$-nullcline are separated, and then the singular control is noticeably better. Of course, only knowing the optimal solution allows one to make such an analysis. However, this, and also comparisons with other models, will be pursued elsewhere.

Here we only conclude with the statement that it is shown in [12, 13] that the qualitative structure of optimal solutions as concatenations of the form $0asa0$ for the model by Hahnfeldt et al. [9] analyzed in this paper is exactly the same for the modified model considered by Ergun et al. [6] while optimal controls are bang-bang with at

most two switchings of the form $0a0$ for the modification considered by d'Onofrio and Gandolfi [5, 15].

Acknowledgments. We thank two anonymous referees for their careful reading of the paper and valuable suggestions regarding the exposition of the material. Thanks are especially due to a referee who clarified some of the underlying medical aspects and whose comments we incorporated into section 2.

REFERENCES

[1] Z. Agur, L. Arakelyan, P. Daugulis, and Y. Ginosar, *Hopf point analysis for angiogenesis models*, Discrete Contin. Dyn. Sys. Series B, 4 (2004), pp. 29–38.
[2] B. Bonnard and J. de Morant, *Toward a geometric theory in the time-minimal control of chemical batch reactors*, SIAM J. Control Optim., 33 (1995), pp. 1279–1311.
[3] A.E. Bryson and Y.C. Ho, *Applied Optimal Control*, Hemisphere, Washington, DC, 1975.
[4] L. Cesari, *Optimization - Theory and Applications*, Springer-Verlag, New York, 1983.
[5] A. d'Onofrio and A. Gandolfi, *Tumour eradication by antiangiogenic therapy: Analysis and extensions of the model, by Hahnfeldt et al.* (1999), Math. Biosci., 191 (2004), pp. 159–184.
[6] A. Ergun, K. Camphausen, and L.M. Wein, *Optimal scheduling of radiotherapy and angiogenic inhibitors*, Bull. Math. Bio., 65 (2003), pp. 407–424.
[7] U. Forys, Y. Keifetz, and Y. Kogan, *Critical-point analysis for three-variable cancer angiogenesis models*, Math. Biosci. Eng., 2 (2005), pp. 511–525.
[8] J.H. Goldie, *Drug resistance in cancer: A perspective*, Cancer Metastasis Rev., 20 (2001), pp. 63–68.
[9] P. Hahnfeldt, D. Panigrahy, J. Folkman, and L. Hlatky, *Tumor development under angiogenic signaling: A dynamical theory of tumor growth, treatment response, and postvascular dormancy*, Cancer Res., 59 (1999), pp. 4770–4775.
[10] R.S. Kerbel, *A cancer therapy resistant to resistance*, Nature, 390 (1997), pp. 335–336.
[11] A.J. Krener, *The high order maximal principle and its application to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256–293.
[12] U. Ledzewicz and H. Schättler, *A synthesis of optimal controls for a model of tumor growth under angiogenic inhibitors*, in Proceedings of the 44th IEEE Conference on Decision and Control, Sevilla, Spain, 2005, pp. 934–939.
[13] U. Ledzewicz and H. Schättler, *Optimal control for a system modelling tumor anti-angiogenesis*, ICGST-ACSE J., 6 (2006), pp. 33–39.
[14] U. Ledzewicz and H. Schättler, *Application of optimal control to a system describing tumor anti-angiogenesis*, in Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems (MTNS), Kyoto, Japan, 2006, pp. 478–484.
[15] U. Ledzewicz and H. Schättler, *Analysis of optimal controls for a mathematical model of tumor anti-angiogenesis*, Optimal Control Appl. Methods, to appear.
[16] L.A. Loeb, *A mutator phenotype in cancer*, Cancer Res., 61 (2001), pp. 3230–3239.
[17] H. Schättler and M. Jankovic, *A synthesis of time-optimal controls in the presence of saturated singular arcs*, Forum Math., 5 (1993), pp. 203–241.
[18] L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Macmillan, New York, 1964.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 1080–1097

BOUNDARY OBSERVATION AND EXACT CONTROL OF A QUASI-ELECTROSTATIC PIEZOELECTRIC SYSTEM IN MULTILAYERED MEDIA*

B. KAPITONOV†, B. MIARA‡, AND G. PERLA MENZALA§

Abstract. We study the evolution of a layered quasi-electrostatic piezoelectric system. Under suitable assumptions on the geometry of a region and the interfaces as well as a monotonicity condition on the coefficients, we prove a boundary observation inequality which together with the Hilbert uniqueness method introduced by Lions give us a solution of the exact controllability problem for the model under study.

Key words. distributed systems, boundary observation, transmission conditions, exact controllability

AMS subject classifications. 35Q99, 74F99, 35B40

DOI. 10.1137/050629884

1. Introduction. Piezoelectricity is an electromechanical interaction. Piezoelectric materials have the property that an electric field creates stress and that a deformation creates polarization. These very special properties explain why such materials are of great use in industry: They are both actuators and sensors. The constitutive equations relate the stress tensor to the electric field and the polarization vector to the strain tensor (see [5] or [6]). In this work we consider the evolution problem of a piezoelectric structure whose three-dimensional (3-D) mechanical displacement vector field $u = u(x,t) = (u^1, u^2, u^3)$ and electric field E are acting on a bounded domain Ω of $\mathbb{R}^3$ with smooth boundary $\partial\Omega = S_0 \cup S_1$. The coupled system which models the phenomenon is

$$\begin{cases} u_{tt} - \operatorname{Div}\sigma = 0 \\ \operatorname{div}\mathcal{D} = 0 \end{cases} \qquad \text{in } \Omega \times (0, +\infty), \tag{1.1}$$

where

$$\sigma = C\,\mathcal{E}(u) - PE,$$
$$\mathcal{D} = P^t\,\mathcal{E}(u) + DE.$$

Here σ is the mechanical stress and $\mathcal{D}$ is the electric displacement. The linearized strain tensor $\mathcal{E}(u) = \left[\mathcal{E}_{k\ell}(u)\right]$ has components $\mathcal{E}_{k\ell}(u) = \frac{1}{2}\left(\frac{\partial u^\ell}{\partial x_k} + \frac{\partial u^k}{\partial x_\ell}\right)$, $1 \le \ell, k \le 3$.

*Received by the editors April 27, 2005; accepted for publication (in revised form) January 25, 2007; published electronically July 20, 2007. The first and third authors were partially supported in this work by Brazilian grants, the first author by a grant from CNPq Project 303981/03-2 and the third author by CNPq and PRONEX.

http://www.siam.org/journals/sicon/46-3/62988.html

†Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences, Russia, and the National Laboratory of Scientific Computation (LNCC/MCT), Brazil (borisvk@lncc.br).

‡Laboratoire de Modélisation et Simulation numérique, École Supérieure d'Ingénieurs en Électrotechnique et Électronique, 2 Boulevard Blaise Pascal, 93160 Noisy-le-Grand, France (miarab@esiee.fr). This author was supported in part by European Community's Human Potential Programme under contract "Smart systems," HPRN-CT-2002-00284.

§National Laboratory of Scientific Computation LNCC/MCT, Rua Getulio Vargas 333, Quitandinha, Petropolis 25651-070, RJ, Brazil, and Institute of Mathematics Federal University of Rio de Janeiro, RJ, P.O. 68530, Rio de Janeiro, RJ, Brazil (perla@lncc.br).

We assume that the electric field E is such that $E = -\nabla q$, where q is the so-called electric potential. In (1.1), $C = (c_{ijk\ell})$ is the fourth order elasticity tensor, which is symmetric and positive, $P = (e_{kij})$ is a third order piezoelectric symmetric tensor, and $D = (d_{ij})$ is a second order symmetric and positive dielectric tensor. We assume that the material is piecewise homogeneous.

From now on, a summation convention with respect to repeated indices will be used. With the above considerations the quasi-electrostatic piezoelectric system (1.1) can be rewritten as $(1 \le i \le 3)$

$$\begin{cases} u^i_{tt} - \dfrac{\partial}{\partial x_j}\sigma_{ij}(u,q) = 0 \\ \dfrac{\partial}{\partial x_i} D_i(u,q) = 0 \end{cases} \qquad \text{in } \Omega \times (0,+\infty), \tag{1.2}$$

where $\sigma_{ij}(u,q) = c_{ijk\ell}\mathcal{E}_{k\ell}(u) + e_{kij}\,\frac{\partial q}{\partial x_k}$ and $D_i(u,q) = e_{ik\ell}\,\mathcal{E}_{k\ell}(u) - d_{ij}\,\frac{\partial q}{\partial x_j}$. We assume that the domain Ω with smooth boundary S has the form $\Omega = \mathcal{O}_0 \backslash \overline{\mathcal{O}}_1$, where $\mathcal{O}_0$ and $\mathcal{O}_1$ are open bounded domains with $\overline{\mathcal{O}}_1 \subset \mathcal{O}_0$, where $\overline{\mathcal{O}}_1$ denotes the closure of $\mathcal{O}_1$, $\partial\mathcal{O}_0 = S_0$, and $\partial\mathcal{O}_1 = S_1$. Thus $S = S_0 \cup S_1$. Let $n > 1$ be a given integer. For each m with $1 \le m \le n$ let B_m be an open subset with smooth boundary Γ_m and such that $\overline{\mathcal{O}}_1 \subset B_m \subset \mathcal{O}_0$, $\overline{B}_m \subset B_{m+1}$. We set $\Omega_0 = B_1 \backslash \overline{\mathcal{O}}_1$, $\Omega_m = B_{m+1} \backslash \overline{B}_m$ for $1 \le m \le n-1$ and $\Omega_n = \mathcal{O}_0 \backslash \overline{B}_n$.

We associate with system (1.2) the following given initial conditions,

$$u(x,0) = f_1(x), \quad u_t(x,0) = f_2(x), \tag{1.3}$$

and boundary conditions

$$\begin{cases} \sigma_{ij}(u,q)\eta_j = 0 \\ q = 0 \end{cases} \qquad \text{on } S_0 \times (0,+\infty), \tag{1.4}$$

$$\begin{cases} D_i(u,q)\eta_i = 0 \\ u = 0 \end{cases} \qquad \text{on } S_1 \times (0,+\infty), \tag{1.5}$$

where $\eta = \eta(x) = (\eta_1, \eta_2, \eta_3)$ is the unit normal vector pointing toward the exterior of Ω.

Our main purpose in this work will be to prove a special uniqueness theorem (boundary observation) for the transmission problem associated with (1.2)–(1.5). We will assume that the transmission conditions are the following:

$$\begin{cases} \sigma_{ij}\big(u^{(m-1)}, q^{(m-1)}\big)\eta_j = \sigma_{ij}\big(u^{(m)}, q^{(m)}\big)\eta_j, \\ D_i\big(u^{(m-1)}, q^{(m-1)}\big)\eta_i = D_i\big(u^{(m)}, q^{(m)}\big)\eta_i, \\ u^{(m-1)} = u^{(m)}, \quad q^{(m-1)} = q^{(m)}, \end{cases} \tag{1.6}$$

for any $(x,t) \in \Gamma_m \times (0,+\infty)$, $m = 1, 2, \ldots, n$, and all $i, j \in \{1,2,3\}$. Here $\eta = \eta(x)$ is the unit normal vector pointing toward the exterior of B_m.

In (1.6), $u^{(m)}$ and $q^{(m)}$ denote the restrictions of the corresponding functions on Ω_m, $1 \le m \le n$. Figures 1 and 2 illustrate simple such situations when $n = 0$ or $n = 2$.

FIG. 1. $n = 0$

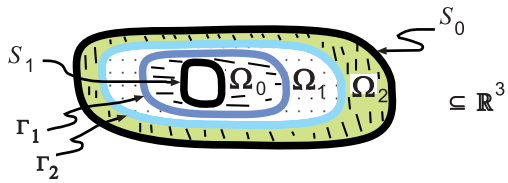


FIG. 2. $n = 2$

From now on we will assume that the coefficients $c_{ijk\ell}$, d_{ij}, and $e_{ik\ell}$ satisfy the symmetries

$$c_{ijk\ell} = c_{k\ell ij} = c_{jik\ell},$$
$$d_{ij} = d_{ji}, \quad e_{ik\ell} = e_{i\ell k}.$$

Also, the tensors $(c_{ijk\ell})$ and (d_{ij}) are elliptic in the sense that there exist $c_0 > 0$ and $d_0 > 0$ such that

$$c_{ijk\ell}\,\lambda_{k\ell}\,\lambda_{ij} \geq c_0 \sum_{i,j=1}^{3} (\lambda_{ij})^2, \quad d_{ij}\xi_j\xi_i \geq d_0|\xi|^2 \tag{1.7}$$

for any real symmetric matrix $[\lambda_{ij}]$ of order 3 and any vector $\xi = (\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3$.

In order to mention the main result of this paper we describe the assumptions on the coefficients for problem (1.2)–(1.6).

HYPOTHESIS I. *The coefficients $c_{ijk\ell} = c_{ijk\ell}(x)$ and $d_{ij} = d_{ij}(x)$ are piecewise constant functions on $\overline{\Omega}$, which lose continuity only on $\Gamma_1, \Gamma_2, \dots$ and Γ_n. All the coefficients e_{kij} are constant on $\overline{\Omega}$.*

We will obtain an estimate of the form

$$(T - T_0) \sum_{m=0}^{n} \int_{\Omega_m} \left\{ |u_t^{(m)}|^2 + c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)}) \mathcal{E}_{ij}(u^{(m)}) + d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i} \right\} dx \leq C \int_0^T \int_{S_0} |u_t|^2\, d\Gamma\, dt \tag{1.8}$$

for some $T_0 > 0$, $C > 0$, and any $T > T_0$. Here we use the notation $|u_t|^2 = \frac{\partial u}{\partial t} \cdot \frac{\partial u}{\partial t}$, where the dot $\cdot$ denotes the usual inner product in $\mathbb{R}^3$ and $\int_{S_0} |u_t|^2\, d\Gamma$ means the surface integral of $|u_t|^2$ over the surface S_0.

This is a so-called boundary observation inequality, which will be stated and proved in section 3 (see Theorem 3.3), provided that we assume geometric properties on the region Ω and the interfaces Γ_i. Such assumptions are basically of "star-shaped" type. In addition, to prove (1.8) we will assume monotonicity conditions on the coefficients of system (1.2).

The need for the above requirements was already noticed by Lions in [16] while treating transmission problems. Later on, Lagnese [13] also used those types of assumptions to prove controllability results for a large class of hyperbolic problems.

Results on controllability of physical systems are quite important, especially in the case of systems driven by coupled equations such as those found in thermoelasticity [14], magnetoelasticity [4], or for flexible multistructures [3], among many others.

In a simple case we may check that our monotonicity conditions are in fact optimal (see Hypothesis II, (e) and (f), in section 3): Suppose that the tensor (e_{kij}) vanishes in Ω. In this case the electric potential does not interact with the displacement vector, and u satisfies a wave-like equation that may be chosen to be a scalar wave equation with the corresponding boundary conditions. It follows from [13] that in this situation the above monotonicity assumptions are optimal. In the general case, that is, when the tensor (e_{kij}) does not vanish, the optimality would require further study.

Using (1.8) and the Hilbert uniqueness method (HUM) introduced by Lions [15], [16], we study the following exact controllability problem: Given, as in (1.3), an initial distribution $f = (f_1, f_2)$ and a desired terminal state $g = (g_1, g_2)$, with f and g belonging to appropriate function spaces, we want to find a vector-valued function $Q = Q(x,t)$ (the control function) in a suitable function space and a time $T > 0$ such that the solution u of (1.2), (1.3), (1.6) with boundary conditions

$$\begin{cases} \sigma_{ij}(u,q)\eta_j = Q^i(x,t) \\ q = 0 \end{cases} \qquad \text{on } S_0 \times (0,+\infty),$$

$$\begin{cases} D_i(u,q)\eta_i = 0 \\ u = 0 \end{cases} \qquad \text{on } S_1 \times (0,+\infty) \tag{1.9}$$

satisfies

$$u(x,T) = g_1(x), \qquad u_t(x,T) = g_2(x). \tag{1.10}$$

Several authors have considered the well-posedness of the initial-boundary value problem for quasi-electrostatic equations; see, for instance, [1], [2], [11], and [17] (and the references therein). The exact controllability problem for system (1.2) has been studied by Miara [18]. Boundary controllability in transmission problems for the wave equation was considered by Lions [16] and Nicaise [19], [20]. Uniform stabilization and exact control for the Maxwell system in multilayered media was studied by Kapitonov [7] (see also [21]). Boundary controllability in transmission problems for a class of second order hyperbolic systems has been studied by Lagnese [13]. Stabilization and exact boundary controllability for a system of electromagneto-elasticity was studied recently by Kapitonov and Perla Menzala [8], [10] and by Kapitonov and Raupp [9].

The sections of this paper are as follows. Solvability of (1.2)–(1.6) for the appropriate class of functions is shown in section 2. This is done via semigroup theory and with the help of known results for transmission problems for elliptic equations (see (2.1)–(2.2) below). Actually, for such a problem we could also give the variational form and the expression of the energy and conclude that the solution is a saddle point of this energy (hence, the Lax–Milgram theorem applies as well). In section 3 we prove the boundary observation result via the multiplier method "slightly" modified in order to take into account additional terms after spatial integration of the fundamental identity. In the last section, the exact controllability problem is solved using the boundary observability inequality and HUM.

2. Well-posedness. The well-posedness of the initial-boundary value problem with transmission conditions (1.2)–(1.6) is proved by standard semigroup methods. Thus, we just outline the proof of well-posedness. Let Ω be a bounded region of $\mathbb{R}^3$ with Lipschitz boundary as considered in section 1. Let us assume Hypothesis I on the coefficients and consider $F = [F_{k\ell}]$ to be a real symmetric matrix of order 3 such that $F_{k\ell} \in H^1(\Omega_m)$, $m = 0, 1, 2, \ldots, n$. In Ω we consider the elliptic problem

$$\frac{\partial}{\partial x_i}\left(d_{ij}\,\frac{\partial q}{\partial x_j}\right) = \frac{\partial}{\partial x_i}\big(e_{ik\ell}\,F_{k\ell}\big) \quad \text{in} \quad \Omega_m, \quad m = 0, 1, 2, \ldots, n, \tag{2.1}$$

with boundary conditions

$$\begin{cases} q = 0 \qquad \text{on} \qquad S_0, \\ d_{ij}\,\dfrac{\partial q}{\partial x_j}\,\eta_i = e_{ik\ell}\,F_{k\ell}\,\eta_i \quad \text{on} \quad S_1, \end{cases} \tag{2.2}$$

and transmission conditions on Γ_m, $m = 1, 2, \ldots, n$,

$$\begin{cases} q^{(m-1)} = q^{(m)}, \\ d_{ij}^{(m-1)}\,\dfrac{\partial q^{(m-1)}}{\partial x_j}\,\eta_i - d_{ij}^{(m)}\,\dfrac{\partial q^{(m)}}{\partial x_j}\,\eta_i \\ \qquad = e_{ik\ell}\,F_{k\ell}^{(m-1)}\,\eta_i - e_{ik\ell}\,F_{k\ell}^{(m)}\,\eta_i. \end{cases} \tag{2.3}$$

Using well-known elliptic results, we deduce that there exists a unique solution $q \in H^2(\Omega_m)$ of problem (2.1)–(2.3), which we will denote by $q = \beta(F)$. (Observe that the existence of a solution $q \in H^2(\Omega_m)$ of problem (2.1)–(2.3) depends on the regularity of the boundary and the boundary loading.) Additionally, the restriction of q to the subset Ω_m will be written as $q^{(m)} = \beta^{(m)}(F)$. Let X be the real Hilbert space of pairs (u, v) of three-component vector-valued functions u and v such that $v^{(m)} \in [L^2(\Omega_m)]^3$, $u^{(m)} \in [H^2(\Omega_m)]^3$, and $u = 0$ on S_1. Here $H^s(\Omega)$ denotes the Sobolev space of order s.

The inner product in X is given as follows:

$$\langle (u, v), (\tilde{u}, \tilde{v})\rangle_X = \sum_{m=0}^{n} \int_{\Omega_m} \Big\{ v \cdot \tilde{v} + c_{ijk\ell}\,\mathcal{E}_{k\ell}(u)\mathcal{E}_{ij}(\tilde{u}) + d_{ij}\,\frac{\partial}{\partial x_j}\left(\beta(\mathcal{E}(u))\right)\frac{\partial}{\partial x_i}\left(\beta(\mathcal{E}(\tilde{u}))\right)\Big\} dx,$$

where $\mathcal{E}(u) = [\mathcal{E}_{k\ell}(u)]$. In X we define the unbounded operator $\mathcal{A}$ with domain $\mathcal{D}(\mathcal{A})$, which consists of all elements (u, v) belonging to X such that

$$v^{(m)} \in \left[H^1(\Omega_m)\right]^3, \quad v = 0 \quad \text{on } S_1,$$

$$\left[c_{ijk\ell}\,\mathcal{E}_{k\ell}(u) + e_{kij}\,\frac{\partial}{\partial x_k}\left(\beta(\mathcal{E}(u))\right)\right]\eta_j = 0 \quad \text{on } S_0,$$

and

$$\begin{cases} u^{(m)} = u^{(m-1)}, \\ \left[c_{ijk\ell}^{(m-1)}\,\mathcal{E}_{k\ell}(u^{(m-1)}) + e_{kij}\,\dfrac{\partial}{\partial x_k}\left(\beta^{(m-1)}(\mathcal{E}(u))\right)\right]\eta_j \\ \qquad = \left[c_{ijk\ell}^{(m)}\,\mathcal{E}_{k\ell}(u^{(m)}) + e_{kij}\,\dfrac{\partial}{\partial x_k}\left(\beta^{(m)}(\mathcal{E}(u))\right)\right]\eta_j \end{cases}$$

on Γ_m, $m = 1, 2, \ldots, n$. In $\mathcal{D}(\mathcal{A})$ the operator $\mathcal{A}$ is given by

$$\mathcal{A}(u,v) = \left(v, \frac{\partial}{\partial x_i} \left\{ c_{ijk\ell}\, \mathcal{E}_{k\ell}(u) + e_{kij}\, \frac{\partial}{\partial x_k} \left(\beta(\mathcal{E}(u)) \right) \right\} \right).$$

The skew-self-adjointness of $\mathcal{A}$ can be verified in the standard way. Consequently, the operator $\mathcal{A}$ generates a one-parameter group of unitary operators $\{U(t)\}_{t\in\mathbb{R}}$ on X, $U(t)$ is strongly continuous with respect to t, and $U(t)f$ is strongly differentiable with respect to t whenever $f = (f_1, f_2) \in \mathcal{D}(\mathcal{A})$. Furthermore, if $f = (f_1, f_2) \in X$, we consider $f^n = (f_1^n, f_2^n) \in \mathcal{D}(\mathcal{A})$ such that $||f - f^n||_X \to 0$ as $n \to +\infty$. Let $\Phi \in L^2(0,T;\mathcal{D}(\mathcal{A}^*))$ such that $\frac{d}{dt}\Phi \in L^2(0,T;X)$ and $\Phi(T) = 0$. For any such Φ we have that $U(t)f^n$ satisfies the identity

$$\int_0^T \left\{ \left\langle U(t)f^n, \frac{d}{dt}\Phi \right\rangle_X + \langle U(t)f^n, \mathcal{A}^*\Phi \rangle_X \right\} dt = -\langle f^n, \Phi(0) \rangle_X. \tag{2.4}$$

Passing to the limit in (2.4) as $n \to +\infty$, we obtain that

$$\int_0^T \left\{ \left\langle U(t)f, \frac{d}{dt}\Phi \right\rangle_X + \langle U(t)f, \mathcal{A}^*\Phi \rangle_X \right\} dt = -\langle f, \Phi(0) \rangle_X; \tag{2.5}$$

that is, $U(t)f$ is the weak solution of the abstract Cauchy problem

$$\frac{dV}{dt} = \mathcal{A}\, V, \quad V(0) = f$$

associated with (1.2)–(1.6).

3. Boundary observation. In this section we prove the boundary observation result (1.8). The proof is based on the theory of multipliers and is motivated by the invariance of system (1.2) relative to the one-parameter group of dilations in all variables. A good reference for the use of this technique is Komornik's book [12]. The multipliers have to be conveniently modified in such a way that we can handle the extra boundary terms appearing in the identities. Let $h = h(x)$ be an auxiliary scalar smooth function on $\overline{\Omega}$ (which we will choose later on). Consider the multipliers M_1 and M_2 (which is actually an operator) given by

$$M_1\, u^i = t\, \frac{\partial u^i}{\partial t} + \nabla h \cdot \nabla u^i + u^i, \quad i = 1, 2, 3,$$

and

$$M_2\, q = tq\, \frac{\partial}{\partial t} - \nabla h \cdot \nabla q.$$

We will use the standard notation

$$\nabla h \cdot \nabla = \sum_{j=1}^{3} \frac{\partial h}{\partial x_j}\, \frac{\partial}{\partial x_j}.$$

Let us multiply the first equation of (1.2) by $2M_1 u^i$ and apply the operator $2M_2 h$ to the second equation in (1.2). Here the derivatives of the components of the tensors

$C = (c_{ijk\ell})$, $D = (d_{ij})$, and $P^t = (e_{kij})$ have to be understood in the distributional sense. Adding and rearranging terms, we obtain

$$\begin{aligned} 0 &= 2M_1 u^i \left[u^i_{tt} - \frac{\partial}{\partial x_j} \left(c_{ijk\ell}\, \mathcal{E}_{k\ell}(u) + e_{kij}\, \frac{\partial q}{\partial x_k} \right) \right] \\ &\quad + 2M_2 q \left[\frac{\partial}{\partial x_i} \left\{ e_{ik\ell}\, \mathcal{E}_{k\ell}(u) - d_{ij}\, \frac{\partial q}{\partial x_j} \right\} \right] \\ &= \frac{\partial}{\partial t}\, F - \frac{\partial}{\partial x_j}\, G_j - J, \end{aligned} \tag{3.1}$$

where

$$\begin{aligned} F &= t \left[|u_t|^2 + c_{ijk\ell}\, \mathcal{E}_{k\ell}(u) \mathcal{E}_{ij}(u) + d_{ij}\, \frac{\partial q}{\partial x_j}\, \frac{\partial q}{\partial x_i} \right] \\ &\quad + 2u_t \cdot \big[(\nabla h \cdot \nabla) u + u \big], \end{aligned} \tag{3.2}$$

$$\begin{aligned} G_j &= 2\big[t\, u^i_t + \nabla h \cdot \nabla u^i + u^i \big] \left[c_{ijk\ell}\, \mathcal{E}_{k\ell}(u) + e_{kij}\, \frac{\partial q}{\partial x_k} \right] \\ &\quad + 2 \left(tq\, \frac{\partial}{\partial t} - \nabla h \cdot \nabla q \right) \left(d_{ij}\, \frac{\partial q}{\partial x_i} - e_{jk\ell}\, \mathcal{E}_{jk\ell}(u) \right) \\ &\quad + \frac{\partial h}{\partial x_j} \left[|u_t|^2 + d_{k\ell}\, \frac{\partial q}{\partial x_\ell}\, \frac{\partial q}{\partial x_k} - c_{pqk\ell}\, \mathcal{E}_{k\ell}(u) \mathcal{E}_{pq}(u) \right. \\ &\quad \left. - 2e_{ki\ell}\, \mathcal{E}_{i\ell}(u)\, \frac{\partial q}{\partial x_k} \right], \end{aligned} \tag{3.3}$$

$$\begin{aligned} J &= (\Delta h - 1) c_{ijk\ell}\, \mathcal{E}_{k\ell}(u) \mathcal{E}_{ij}(u) \\ &\quad - 2\, \frac{\partial^2 h}{\partial x_p \partial x_j}\, c_{ijk\ell}\, \mathcal{E}_{k\ell}(u)\, \frac{\partial u^i}{\partial x_p} + (3 - \Delta h) |u_t|^2 \\ &\quad + 2\, \frac{\partial^2 h}{\partial x_i \partial x_k}\, d_{ij}\, \frac{\partial q}{\partial x_j}\, \frac{\partial q}{\partial x_k} + (1 - \Delta h) d_{ij}\, \frac{\partial q}{\partial x_j}\, \frac{\partial q}{\partial x_i} \\ &\quad + 2(\Delta h - 1) e_{kij}\, \mathcal{E}_{ij}(u)\, \frac{\partial q}{\partial x_k} \\ &\quad - 2\, \frac{\partial^2 h}{\partial x_j \partial x_\ell}\, e_{kij}\, \frac{\partial u^i}{\partial x_\ell}\, \frac{\partial q}{\partial x_k} - 2\, \frac{\partial^2 h}{\partial x_i \partial x_j}\, e_{ik\ell}\, \mathcal{E}_{k\ell}(u)\, \frac{\partial q}{\partial x_j}. \end{aligned} \tag{3.4}$$

Observation. If we consider $h(x) = \frac{1}{2} |x - x_0|^2$ for some $x_0 \in \mathbb{R}^3$, then we can verify that $J \equiv 0$. In that case identity (3.1) represents a true conservation law. However, due to the terms appearing in G_j (see (3.4)), we would require a definite sign for $\frac{\partial h}{\partial \eta} \cdot$ That is why we will choose h as a "small" perturbation of $\frac{1}{2} |x - x_0|^2$.

Let $\{u, q\}$ be a smooth solution of (1.2)–(1.6). Integration over $\Omega_m \times (0, T)$ of

identity (3.1) and summation over m implies that

$$
\begin{aligned}
&T\sum_{m=0}^{n}\int_{\Omega_m}\left\{|u_t|^2+c_{ijk\ell}\,\mathcal{E}_{k\ell}(u)\mathcal{E}_{ij}(u)+d_{ij}\,\frac{\partial q}{\partial x_j}\,\frac{\partial q}{\partial x_i}\right\}dx\\
&\quad+\left[2\sum_{m=0}^{n}\int_{\Omega_m}u_t\cdot\big\{(\nabla h\cdot\nabla)u+u\big\}dx\right]_{t=0}^{t=T}\\
&\quad=\sum_{m=1}^{n}\int_0^T\int_{\Gamma_m}\big(V_{m-1}-V_m\big)d\Gamma\,dt\\
&\quad+\int_0^T\int_{S_0}V_n\,d\Gamma\,dt+\int_0^T\int_{S_1}V_0\,d\Gamma\,dt\\
&\quad+\sum_{m=0}^{n}\int_0^T\int_{\Omega_m}J_m(x,t)dx\,dt,
\end{aligned}
\tag{3.5}
$$

where $J_m=J_m(u,q,h)$ denotes the restriction of J (given by (3.4)) to the region Ω_m and

$$
\begin{aligned}
V_m=\;&2\big[t\,u_t^{(m)}+(\nabla h\cdot\nabla)u^{(m)}+u^{(m)}\big]^i\left[c_{ijk\ell}^{(m)}\,\mathcal{E}_{k\ell}(u^{(m)})+e_{kij}\,\frac{\partial q^{(m)}}{\partial x_k}\right]\eta_j\\
&+\frac{\partial h}{\partial\eta}\left[|u_t^{(m)}|^2+d_{k\ell}^{(m)}\,\frac{\partial q^{(m)}}{\partial x_\ell}\,\frac{\partial q^{(m)}}{\partial x_k}-c_{pqk\ell}^{(m)}\,\mathcal{E}_{k\ell}(u^{(m)})\mathcal{E}_{pq}(u^{(m)})\right.\\
&\left.-\,2e_{ki\ell}\,\mathcal{E}_{i\ell}(u^{(m)})\frac{\partial q^{(m)}}{\partial x_k}\right]\\
&+2tq^{(m)}\left[d_{ij}^{(m)}\,\frac{\partial^2 q^{(m)}}{\partial x_i\partial t}-e_{jk\ell}\,\mathcal{E}_{k\ell}(u_t^{(m)})\right]\eta_j\\
&-2\,\nabla h\cdot\nabla q^{(m)}\left[d_{ij}^{(m)}\,\frac{\partial q^{(m)}}{\partial x_i}-e_{jk\ell}\,\mathcal{E}_{k\ell}(u^{(m)})\right]\eta_j
\end{aligned}
\tag{3.6}
$$

$(0\le m\le n)$.

In (3.6), $\frac{\partial h}{\partial\eta}$ denotes the normal derivative of h at $x\in\Gamma_m$ (or S_0, S_1). The lemma that we prove below shows that the differences $V_{m-1}-V_m$ will have the "good" sign if we choose h conveniently and assume a monotonicity condition on the coefficients, as follows.

LEMMA 3.1. *Let $\{u,q\}$ be a smooth solution of problem* (1.2)–(1.6). *Then, the identity*

$$
\begin{aligned}
V_{m-1}-V_m=-\frac{\partial h}{\partial\eta}\Big[&\big(c_{ijk\ell}^{(m-1)}-c_{ijk\ell}^{(m)}\big)\mathcal{E}_{k\ell}(u^{(m-1)})\mathcal{E}_{ij}(u^{(m-1)})\\
&+c_{ijk\ell}^{(m)}\,\mathcal{E}_{k\ell}(u^{(m)}-u^{(m-1)})\mathcal{E}_{ij}(u^{(m)}-u^{(m-1)})\\
&+\left(d_{ij}^{(m)}-d_{ij}^{(m-1)}\right)\frac{\partial q^{(m)}}{\partial x_j}\,\frac{\partial q^{(m)}}{\partial x_i}\\
&+d_{ij}^{(m-1)}\left(\frac{\partial q^{(m-1)}}{\partial x_j}-\frac{\partial q^{(m)}}{\partial x_j}\right)\left(\frac{\partial q^{(m-1)}}{\partial x_i}-\frac{\partial q^{(m)}}{\partial x_i}\right)\Big]
\end{aligned}
\tag{3.7}
$$

holds on Γ_m for $m=1,2,\dots,n$.

Proof. We use interface conditions (1.6). In fact, direct calculations using (3.6) and the interface conditions imply that

$$
\begin{aligned}
V_{m-1} - V_m &= 2(\nabla h \cdot \nabla)(u^{i(m-1)} - u^{i(m)}) \left(c^{(m)}_{ijk\ell}\, \mathcal{E}_{k\ell}(u^{(m)}) + e_{kij}\, \frac{\partial q^{(m)}}{\partial x_k} \right) \eta_j \\
&\quad - \frac{\partial h}{\partial \eta} \left[c^{(m-1)}_{pqk\ell}\, \mathcal{E}_{k\ell}(u^{(m-1)}) \mathcal{E}_{pq}(u^{(m-1)}) - c^{(m)}_{pqk\ell}\, \mathcal{E}_{k\ell}(u^{(m)}) \mathcal{E}_{pq}(u^{(m)} \right] \\
&\quad + \frac{\partial h}{\partial \eta} \left[d^{(m-1)}_{k\ell}\, \frac{\partial q^{(m-1)}}{\partial x_\ell}\, \frac{\partial q^{(m-1)}}{\partial x_k} - d^{(m)}_{k\ell}\, \frac{\partial q^{(m)}}{\partial x_\ell}\, \frac{\partial q^{(m)}}{\partial x_k} \right] \\
&\quad - 2\, \frac{\partial h}{\partial \eta} \left[e_{ki\ell}\, \mathcal{E}_{i\ell}(u^{(m-1)}) \frac{\partial q^{(m-1)}}{\partial x_k} - e_{ki\ell}\, \mathcal{E}_{i\ell}(u^{(m)})\, \frac{\partial q^{(m)}}{\partial x_k} \right] \\
&\quad - 2 \left[d^{(m)}_{ij}\, \frac{\partial q^{(m)}}{\partial x_i} - e_{jk\ell}\, \mathcal{E}_{k\ell}(u^{(m)}) \right] \eta_j \left[\nabla h \cdot \nabla q^{(m-1)} - \nabla h \cdot \nabla q^{(m)} \right]. \qquad (3.8)
\end{aligned}
$$

We use the identity

$$
\begin{aligned}
(\nabla h \cdot \nabla)(u^{(m-1)} - u^{(m)}) &= \frac{\partial h}{\partial x_k} \left(\frac{\partial u^{(m-1)}}{\partial x_k} - \frac{\partial u^{(m)}}{\partial x_k} \right) \\
= \frac{\partial h}{\partial x_k}\, \eta_k \left(\frac{\partial u^{(m-1)}}{\partial \eta} - \frac{\partial u^{(m)}}{\partial \eta} \right) &= \frac{\partial h}{\partial \eta} \left(\frac{\partial u^{(m-1)}}{\partial \eta} - \frac{\partial u^{(m)}}{\partial \eta} \right)
\end{aligned}
$$

in order to obtain

$$
\begin{aligned}
&2(\nabla h \cdot \nabla)(u^{i(m-1)} - u^{i(m)}) \left(c^{(m)}_{ijk\ell}\, \mathcal{E}_{k\ell}(u^{(m)}) + e_{kij}\, \frac{\partial q^{(m)}}{\partial x_k} \right) \eta_j \\
&= 2\, \frac{\partial h}{\partial \eta} \left[\left(\frac{\partial u^{i(m-1)}}{\partial \eta} - \frac{\partial u^{i(m)}}{\partial \eta} \right) \eta_j \right] \left[c^{(m)}_{ijk\ell}\, \mathcal{E}_{k\ell}(u^{(m)}) + e_{kij}\, \frac{\partial q^{(m)}}{\partial x_k} \right] \\
&= 2\, \frac{\partial h}{\partial \eta} \left[\left(\frac{\partial u^{i(m-1)}}{\partial x_j} - \frac{\partial u^{i(m)}}{\partial x_j} \right) \right] \left[c^{(m)}_{ijk\ell}\, \mathcal{E}_{k\ell}(u^{(m)}) + e_{kij}\, \frac{\partial q^{(m)}}{\partial x_k} \right] \\
&= 2\, \frac{\partial h}{\partial \eta}\, c^{(m)}_{ijk\ell}\, \mathcal{E}_{k\ell}(u^{(m)})\, \mathcal{E}_{ij}(u^{(m-1)}) - 2\, \frac{\partial h}{\partial \eta}\, c^{(m)}_{ijk\ell}\, \mathcal{E}_{k\ell}(u^{(m)})\, \mathcal{E}_{ij}(u^{(m)}) \\
&\quad + 2\, \frac{\partial h}{\partial \eta}\, e_{kij}\, \mathcal{E}_{ij}(u^{(m-1)})\, \frac{\partial q^{(m)}}{\partial x_k} - 2\, \frac{\partial h}{\partial \eta}\, e_{kij}\, \mathcal{E}_{ij}(u^{(m)})\, \frac{\partial q^{(m)}}{\partial x_k}. \qquad (3.9)
\end{aligned}
$$

In a similar way we get

$$
\begin{aligned}
&-2 \left[d^{(m)}_{ij}\, \frac{\partial q^{(m)}}{\partial x_i} - e_{jk\ell}\, \mathcal{E}_{k\ell}(u^{(m)}) \right] \eta_j \left[\left(\nabla h \cdot \nabla q^{(m-1)} \right) - \left(\nabla h \cdot \nabla q^{(m)} \right) \right] \\
&= -2 \left[\nabla h \cdot \eta \left(\frac{\partial q^{(m-1)}}{\partial \eta} - \frac{\partial q^{(m)}}{\partial \eta} \right) \right] \left[d^{(m)}_{ij}\, \frac{\partial q^{(m)}}{\partial x_i} - e_{jk\ell}\, \mathcal{E}_{k\ell}(u^{(m)}) \right] \eta_j \\
&= -2\, \frac{\partial h}{\partial \eta} \left[\left(d^{(m)}_{ij}\, \frac{\partial q^{(m)}}{\partial x_i} - e_{jk\ell}\, \mathcal{E}_{k\ell}(u^{(m)}) \right) \eta_j \left(\frac{\partial q^{(m-1)}}{\partial \eta} - \frac{\partial q^{(m)}}{\partial \eta} \right) \right] \\
&= 2\, \frac{\partial h}{\partial \eta} \left[d^{(m)}_{ij}\, \frac{\partial q^{(m)}}{\partial x_i} - e_{jk\ell}\, \mathcal{E}_{k\ell}(u^{(m)}) \right] \left[\frac{\partial q^{(m)}}{\partial x_j} - \frac{\partial q^{(m-1)}}{\partial x_j} \right]
\end{aligned}
$$

$$
\begin{aligned}
&= 2\,\frac{\partial h}{\partial \eta}\, d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i} - 2\,\frac{\partial h}{\partial \eta}\, d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m-1)}}{\partial x_i} \\
&\quad + 2\,\frac{\partial h}{\partial \eta}\, e_{jk\ell}\, \mathcal{E}_{k\ell}(u^{(m)})\, \frac{\partial q^{(m-1)}}{\partial \eta} - 2\,\frac{\partial h}{\partial \eta}\, e_{jk\ell}\, \mathcal{E}_{k\ell}(u^{(m)})\, \frac{\partial q^{(m)}}{\partial x_j}. \qquad (3.10)
\end{aligned}
$$

From (3.8)–(3.10) we obtain the identity

$$
\begin{aligned}
V_{m-1} - V_m &= 2\,\frac{\partial h}{\partial \eta}\, c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)}) \mathcal{E}_{ij}(u^{(m-1)}) \\
&\quad - \frac{\partial h}{\partial \eta}\, c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)}) \mathcal{E}_{ij}(u^{(m)}) \\
&\quad - \frac{\partial h}{\partial \eta}\, c_{ijk\ell}^{(m-1)}\, \mathcal{E}_{k\ell}(u^{(m-1)}) \mathcal{E}_{ij}(u^{(m-1)}) \\
&\quad - 2\,\frac{\partial h}{\partial \eta}\, d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m-1)}}{\partial x_j} + \frac{\partial h}{\partial \eta}\, d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i} \\
&\quad + \frac{\partial h}{\partial \eta}\, d_{ij}^{(m-1)}\, \frac{\partial q^{(m-1)}}{\partial x_j}\, \frac{\partial q^{(m-1)}}{\partial x_i} \\
&\quad + 2\,\frac{\partial h}{\partial \eta}\, e_{kij}\, \mathcal{E}_{ij}(u^{(m-1)})\, \frac{\partial q^{(m)}}{\partial x_k} - 2\,\frac{\partial h}{\partial \eta}\, e_{kij}\, \mathcal{E}_{ij}(u^{(m)})\, \frac{\partial q^{(m)}}{\partial x_k} \\
&\quad + 2\,\frac{\partial h}{\partial \eta}\, e_{kij}\, \mathcal{E}_{ij}(u^{(m)})\, \frac{\partial q^{(m-1)}}{\partial x_k} - 2\,\frac{\partial h}{\partial \eta}\, e_{kij}\, \mathcal{E}_{ij}(u^{(m-1)})\, \frac{\partial q^{(m-1)}}{\partial x_k}. \qquad (3.11)
\end{aligned}
$$

Using the interface conditions

$$
D_k\big(u^{(m-1)}, q^{(m-1)}\big)\eta_k = D_k\big(u^{(m)}, q^{(m)}\big)\eta_k,
$$

we obtain

$$
\begin{aligned}
&\big[e_{kij}\mathcal{E}_{ij}(u^{(m)}) - e_{ikj}\mathcal{E}_{ij}(u^{(m-1)})\big]\eta_k \\
&= \left(d_{k\ell}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_\ell} - d_{k\ell}^{(m-1)}\, \frac{\partial q^{(m-1)}}{\partial x_\ell}\right)\eta_k. \qquad (3.12)
\end{aligned}
$$

Using (3.12), we deduce

$$
\begin{aligned}
&2\,\frac{\partial h}{\partial \eta}\ e_{kij}\, \mathcal{E}_{ij}(u^{(m-1)})\, \frac{\partial q^{(m)}}{\partial x_k} - 2\,\frac{\partial h}{\partial \eta}\, e_{kij}\, \mathcal{E}_{ij}(u^{(m)})\, \frac{\partial q^{(m)}}{\partial x_k} \\
&\quad + 2\,\frac{\partial h}{\partial \eta}\, e_{kij}\, \mathcal{E}_{ij}(u^{(m)})\, \frac{\partial q^{(m-1)}}{\partial x_k} - 2\,\frac{\partial h}{\partial \eta}\, e_{kij}\, \mathcal{E}_{ij}(u^{(m-1)})\, \frac{\partial q^{(m-1)}}{\partial x_k} \\
&= 2\,\frac{\partial h}{\partial \eta}\left[e_{kij}\, \mathcal{E}_{ij}(u^{(m)}) - e_{kij}\, \mathcal{E}_{ij}(u^{(m-1)})\right]\left[\frac{\partial q^{(m-1)}}{\partial x_k} - \frac{\partial q^{(m)}}{\partial x_k}\right] \\
&= 2\,\frac{\partial h}{\partial \eta}\left[e_{kij}\, \mathcal{E}_{ij}(u^{(m)}) - e_{kij}\, \mathcal{E}_{ij}(u^{(m-1)})\right]\eta_k\left[\frac{\partial q^{(m-1)}}{\partial \eta} - \frac{\partial q^{(m)}}{\partial \eta}\right] \\
&= 2\,\frac{\partial h}{\partial \eta}\left[d_{k\ell}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_\ell} - d_{k\ell}^{(m-1)}\, \frac{\partial q^{(m-1)}}{\partial x_\ell}\right]\eta_k\left[\frac{\partial q^{(m-1)}}{\partial \eta} - \frac{\partial q^{(m)}}{\partial \eta}\right] \\
&= -2\,\frac{\partial h}{\partial \eta}\, d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i} + 2\,\frac{\partial h}{\partial \eta}\, d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m-1)}}{\partial x_i} \\
&\quad - 2\,\frac{\partial h}{\partial \eta}\, d_{ij}^{(m-1)}\, \frac{\partial q^{(m-1)}}{\partial x_j}\, \frac{\partial q^{(m-1)}}{\partial x_i} + 2\,\frac{\partial h}{\partial \eta}\, d_{ij}^{(m-1)}\, \frac{\partial q^{(m-1)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i}. \qquad (3.13)
\end{aligned}
$$

From (3.9)–(3.13) it follows that

$$
\begin{aligned}
V_{m-1} - V_m = {} & 2\,\frac{\partial h}{\partial \eta}\, c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)})\mathcal{E}_{ij}(u^{(m-1)}) \\
& - \frac{\partial h}{\partial \eta}\, c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)})\mathcal{E}_{ij}(u^{(m)}) \\
& - \frac{\partial h}{\partial \eta}\, c_{ijk\ell}^{(m-1)}\, \mathcal{E}_{k\ell}(u^{(m-1)})\mathcal{E}_{ij}(u^{(m-1)}) \\
& + 2\,\frac{\partial h}{\partial \eta}\, d_{ij}^{(m-1)}\, \frac{\partial q^{(m-1)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i} - \frac{\partial h}{\partial \eta}\, d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i} \\
& - \frac{\partial h}{\partial \eta}\, d_{ij}^{(m-1)}\, \frac{\partial h^{(m-1)}}{\partial x_j}\, \frac{\partial q^{(m-1)}}{\partial x_i}.
\end{aligned}
\tag{3.14}
$$

The conclusion of Lemma 3.1 follows from (3.14), observing the validity of the identities

$$
\begin{aligned}
& c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)})\, \mathcal{E}_{ij}(u^{(m)}) + c_{ijk\ell}^{(m-1)}\, \mathcal{E}_{k\ell}(u^{(m-1)})\, \mathcal{E}_{ij}(u^{(m-1)}) \\
& \quad - 2\, c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)})\, \mathcal{E}_{ij}(u^{(m-1)}) \\
& \quad = \big(c_{ijk\ell}^{(m-1)} - c_{ijk\ell}^{(m)}\big)\mathcal{E}_{k\ell}(u^{(m-1)})\mathcal{E}_{ij}(u^{(m-1)}) \\
& \quad\quad + \, c_{ijk\ell}^{(m)} \Big[\mathcal{E}_{k\ell}(u^{(m)}) - \mathcal{E}_{k\ell}(u^{(m-1)})\Big] \Big[\mathcal{E}_{ij}(u^{(m)}) - \mathcal{E}_{ij}(u^{(m-1)})\Big]
\end{aligned}
$$

and

$$
\begin{aligned}
& d_{ij}^{(m)} \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_k} + d_{ij}^{(m-1)}\, \frac{\partial q^{(m-1)}}{\partial x_j}\, \frac{\partial q^{(m-1)}}{\partial x_i} - 2\, d_{ij}^{(m-1)}\, \frac{\partial q^{(m-1)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i} \\
& \quad = \Big(d_{ij}^{(m)} - d_{ij}^{(m-1)}\Big)\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m-1)}}{\partial x_i} \\
& \quad\quad + d_{ij}^{(m-1)} \left(\frac{\partial q^{(m-1)}}{\partial x_j} - \frac{\partial q^{(m)}}{\partial x_j}\right) \left(\frac{\partial q^{(m-1)}}{\partial x_i} - \frac{\partial q^{(m)}}{\partial x_i}\right).
\end{aligned}
$$

Both identities together with (3.14) imply the conclusion of Lemma 3.1. □

Next, we would like to get a bound for the last term on the right-hand side of (3.5), as follows. Let us choose a convenient function $h(x)$: Let $\Phi(x)$ be the solution of the elliptic problem

$$
\begin{cases}
\Delta\, \Phi = 1 \quad \text{in} \quad \Omega, \\
\dfrac{\partial \Phi}{\partial \eta} = 2\, \dfrac{\mathrm{Vol}(\Omega)}{\mathrm{Area}(S_0)} \quad \text{on} \quad S_0, \\
\dfrac{\partial \Phi}{\partial \eta} = -\, \dfrac{\mathrm{Vol}(\Omega)}{\mathrm{Area}(S_0)} \quad \text{on} \quad S_1,
\end{cases}
\tag{3.15}
$$

where $\mathrm{Area}(S_j)$ means the surface area of S_j .

Clearly, problem (3.15) admits a solution $\Phi(x)$ (depending on the boundary regularity) such that $\Phi \in C^2(\Omega) \cap C^1(\overline{\Omega})$. Let $\delta > 0$ and $x_0 \in \mathbb{R}^3$ (to be chosen later) and define

$$
h(x) = \delta\, \Phi(x) + \frac{1}{2}\, |x - x_0|^2. \tag{3.16}
$$

Now, we will estimate the term $\sum_{m=0}^{n}\int_0^T\int_{\Omega_m} J_m\,dxdt$ in (3.5).

LEMMA 3.2. *Under the assumption of Lemma* 3.1, *Hypothesis* I, *and choosing* h *as in* (3.16), *we have that*

$$\sum_{m=0}^{n}\int_0^T\int_{\Omega_m} J_m\,dxdt \tag{3.17}$$
$$\leq \delta\,C\,T\sum_{m=0}^{n}\int_{\Omega_m}\left\{|u_t^{(m)}|^2+c_{ijk\ell}^{(m)}\,\mathcal{E}_{k\ell}(u^{(m)})\mathcal{E}_{ij}(u^{(m)})+d_{ij}^{(m)}\,\frac{\partial q^{(m)}}{\partial x_j}\,\frac{\partial q^{(m)}}{\partial x_i}\right\}dx$$

for any $\delta>0$ *and some positive constant* C *which depends only on* Φ *and the coefficients of system* (1.2).

Proof. The index m will be omitted to simplify notation. With our choice of $h(x)$, straightforward calculations show that J given by (3.4) while restricted to Ω can be written as

$$\begin{aligned}J=&-\delta\left[|u_t|^2+c_{ijk\ell}\,\mathcal{E}_{k\ell}(u)\mathcal{E}_{ij}(u)+d_{ij}\,\frac{\partial q}{\partial x_j}\,\frac{\partial q}{\partial x_i}\right]\\&+2\delta\Bigg[c_{ijk\ell}\,\mathcal{E}_{k\ell}(u)\mathcal{E}_{ij}(u)-\frac{\partial^2\Phi}{\partial x_p\partial x_j}\,c_{ijk\ell}\,\mathcal{E}_{k\ell}(u)\,\frac{\partial u^i}{\partial x_p}\\&\quad+\frac{\partial^2\Phi}{\partial x_i\partial x_k}\,d_{ij}\,\frac{\partial q}{\partial x_j}\,\frac{\partial q}{\partial x_k}+e_{kij}\,\mathcal{E}_{ij}(u)\,\frac{\partial q}{\partial x_k}\\&\quad-\frac{\partial^2\Phi}{\partial x_j\partial x_\ell}\,e_{kij}\,\frac{\partial u^i}{\partial x_\ell}\,\frac{\partial q}{\partial x_k}-\frac{\partial^2\Phi}{\partial x_i\partial x_j}\,e_{ik\ell}\,\mathcal{E}_{k\ell}(u)\,\frac{\partial q}{\partial x_j}\Bigg].\end{aligned} \tag{3.18}$$

Let us estimate each term on the right-hand side of (3.18).

Let us note $\psi=(\psi_1,\psi_2,\psi_3)$, where $\psi_i=\frac{\partial^2\Phi}{\partial x_i\partial x_j}\,\frac{\partial q}{\partial x_k}\cdot$ Thus, for any $\delta_1>0$ we have the inequality

$$2\delta\,\frac{\partial^2\Phi}{\partial x_i\partial x_j}\,d_{ij}\,\frac{\partial q}{\partial x_j}\,\frac{\partial q}{\partial x_k}=2\delta\,D\nabla q\cdot\psi\leq\delta\,\delta_1\,D\nabla q\cdot\nabla q+\delta\,\delta_1^{-1}D\psi\cdot\psi, \tag{3.19}$$

where $D=(d_{ij})$. Letting

$$C(\Phi)=\max_{\substack{x\in\overline{\Omega}\\ i,j=1,2,3}}\left|\frac{\partial^2\Phi}{\partial x_i\partial x_j}\right|,$$

the use of Hypothesis I with (1.7) yields the inequality

$$|D\psi\cdot\psi|\leq 9||D||C^2(\Phi)d_0^{-1}\,D\nabla q\cdot\nabla q,$$

which together with (3.19) gives us the estimate

$$2\delta\,\frac{\partial^2\Phi}{\partial x_i\partial x_k}\,d_{ij}\,\frac{\partial q}{\partial x_j}\,\frac{\partial q}{\partial x_k}\leq\delta\left[\delta_1+9d_0^{-1}\delta_1^{-1}||D||C^2(\Phi)\right]d_{ij}\,\frac{\partial q}{\partial x_j}\,\frac{\partial q}{\partial x_i}. \tag{3.20}$$

Next, for any $\delta_2>0$ we have the inequality

$$2\delta\,e_{kij}\,\mathcal{E}_{ij}(u)\,\frac{\partial q}{\partial x_k}\leq 3\delta\,\delta_2||P||\mathcal{E}_{ij}^2(u)+9\delta\delta_2^{-1}||P||\,|\nabla q|^2,$$

where $P = (e_{kij})$. Again, we use Hypothesis I and (1.7) to obtain from the above inequality the estimate

(3.21)
$$2\delta\, e_{kij}\, \mathcal{E}_{ij}(u)\, \frac{\partial q}{\partial x_k} \leq 3\delta\delta_2 ||P|| c_0^{-1}\, c_{jk\ell}\, \mathcal{E}_{k\ell}(u)\mathcal{E}_{ij}(u) + 9\delta\delta_2^{-1}||P||d_0^{-1} d_{ij}\, \frac{\partial q}{\partial x_j}\, \frac{\partial q}{\partial x_i}.$$

In a similar way we can deduce that for any $\delta_3 > 0$

$$\begin{aligned} -2\delta\, \frac{\partial^2 \Phi}{\partial x_i \partial x_j}\, e_{ik\ell}\, \mathcal{E}_{k\ell}(u)\, \frac{\partial q}{\partial x_j} &\leq 3\delta\delta_3 ||P|| c_0^{-1}\, c_{ijk\ell}\, \mathcal{E}_{k\ell}(u)\mathcal{E}_{ij}(u) \\ &\quad + 81\, \delta\delta_3^{-1}||P||d_0^{-1}\, d_{ij}\, \frac{\partial q}{\partial x_j}\, \frac{\partial q}{\partial x_i}. \end{aligned} \tag{3.22}$$

Now, let us note $v_{ij} = \frac{\partial^2 \Phi}{\partial x_j \partial x_\ell}\, \frac{\partial u^i}{\partial x_\ell}\cdot$ Clearly, we have that $v_{ij}^2 \leq 9C^2(\Phi)\big(\frac{\partial u^k}{\partial x_\ell}\big)^2$. For any $\delta_4 > 0$, we have the estimate

$$\begin{aligned} &-2\delta\, \frac{\partial^2 \Phi}{\partial x_j \partial x_\ell}\, e_{kij}\, \frac{\partial u^i}{\partial x_\ell}\, \frac{\partial q}{\partial x_k} = -2\,\delta\, e_{kij}\, v_{ij}\, \frac{\partial q}{\partial x_k} \\ &\leq 3\,\delta\delta_4 ||P|| v_{ij}^2 + 9\delta\delta_4^{-1}||P||\, |\nabla q|^2 \\ &\leq 27\,\delta\delta_4 ||P|| C^2(\Phi) \sum_{i=1}^{3} |\nabla u^i|^2 + 9\,\delta\delta_4^{-1}||P||d_0^{-1}\, d_{ij}\, \frac{\partial q}{\partial x_j}\, \frac{\partial q}{\partial x_i}. \end{aligned} \tag{3.23}$$

Finally, we estimate the term $-2\delta\, \frac{\partial^2 \Phi}{\partial x_p \partial x_j}\, c_{ijk\ell}\, \mathcal{E}_{k\ell}(u)\, \frac{\partial u^i}{\partial x_p}$. Denoting by $\widetilde{C} = \max_{1\leq i,j,k,\ell\leq 3} |c_{ijk\ell}|$, and for any $\delta_5 > 0$, we have the inequalities

$$\begin{aligned} &-2\delta\, \frac{\partial^2 \Phi}{\partial x_p \partial x_j}\, c_{ijk\ell}\, \mathcal{E}_{k\ell}(u)\, \frac{\partial u^i}{\partial x_p} \leq 6\,\delta\,\tilde{c}|\mathcal{E}_{k\ell}(u)|c(\Phi) \sum_{j=1}^{3} \left|\frac{\partial u^i}{\partial x_j}\right| \\ &\leq 3\,\delta\delta_3^{-1}\,\tilde{c}c(\Phi)c_0^{-1}\, c_{ijk\ell}\, \mathcal{E}_{k\ell}(u)\mathcal{E}_{ij}(u) + 9\,\delta\delta_5\tilde{c}\, c(\Phi)|\nabla u^i|^2, \end{aligned} \tag{3.24}$$

where we again used Hypothesis I. Integration of identity (3.18) on $\Omega_m \times (0,T)$, adding from $m = 0$ up to $m = n$, and estimates (3.19)–(3.24) give us the inequality

$$\begin{aligned} \sum_{m=0}^{n} \int_0^T \int_{\Omega_m} J_m\, dxdt &\leq \delta\, A \sum_{m=0}^{n} \int_0^T \int_{\Omega_m} c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)})\mathcal{E}_{ij}(u^{(m)}) dxdt \\ &\quad + \delta\, B \sum_{m=0}^{n} \int_0^T \int_{\Omega_m} d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i}\, dxdt \end{aligned} \tag{3.25}$$

for some positive constants A and B. According to our estimates (3.19)–(3.24) we can take

$$A = 1 + 3\tilde{c}(\Phi)\delta_5^{-1}c_0^{-1} + 3\tilde{c}\, c_0^{-1}(\delta_2 + \delta_3) + 9\alpha_0^{-1}\, \delta_5\tilde{c}\, c(\Phi) + 27\,\delta_4\alpha_0^{-1}\, \tilde{c}\, c^2(\Phi)$$

and

$$B = \delta_1 + 9\, d_0^{-1}\delta_1^{-1}||D||c^2(\Phi) + 9\, d_0^{-1}\delta_2^{-1}\tilde{c} + 81\, \tilde{c}\, d_0^{-1}\delta_3^{-1} + 9\, \tilde{c}\, d_0^{-1}\delta_4^{-1} - 1,$$

where $\alpha_0 > 0$ is chosen such that

$$\sum_{m=0}^{n} \int_{\Omega_m} c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)})\mathcal{E}_{ij}(u^{(m)}) dx \geq \alpha_0 \sum_{m=0}^{n} ||u^{(m)}||^2_{[H^1(\Omega_m)]^3} \tag{3.26}$$

for any $u^{(m)} \in [H^1(\Omega_m)]^3$ with $u^{(m-1)} = u^{(m)}$ on Γ_m, $m = 1, 2, \ldots, n$, and $u\big|_{S_1} = 0$. Choosing $\delta_1 > 1$, we conclude from (3.25) the existence of a positive constant C such that

$$\begin{aligned}
&\sum_{m=0}^{n} \int_0^T \int_{\Omega_m} J_m \, dxdt \\
&\quad \le \delta\, C \sum_{m=0}^{n} \int_0^T \int_{\Omega_m} \left\{ c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)})\, \mathcal{E}_{ij}(u^{(m)}) + d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i} \right\} dxdt \\
&\quad \le \delta\, C\, T \sum_{m=0}^{n} \int_{\Omega_m} \left\{ |u_t^{(m)}|^2 + c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)}) \mathcal{E}_{ij}(u^{(m)}) + d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i} \right\} dx
\end{aligned}$$

because we can easily verify that if $\{u, q\}$ is the solution of problem (1.2)–(1.6) as constructed in section 2, then the quantity

$$E(t) = \sum_{n=0}^{n} \int_\Omega \left\{ |u_t^{(m)}|^2 + c_{ijk\ell}^{(m)}\, \mathcal{E}_{k\ell}(u^{(m)}) \mathcal{E}_{ij}(u^{(m)}) + d_{ij}^{(m)}\, \frac{\partial q^{(m)}}{\partial x_j}\, \frac{\partial q^{(m)}}{\partial x_i} \right\} dx$$

is independent of t. This proves Lemma 3.2. □

HYPOTHESIS II. There exist $\delta_0 \ge 0$ and $x_0 \in \mathcal{O}_1$ ($\mathcal{O}_1$ is as in section 1) such that

(a) $\delta_0 C < 1$, where $C > 0$ is as in the conclusion of Lemma 3.2,

(b) $\delta_0 \frac{\partial \Phi}{\partial \eta} + (x - x_0) \cdot \eta \ge 0$ for any $x \in \Gamma_m$, $m = 1, 2, \ldots, n$,

(c) $(x - x_0) \cdot \eta \ge -2\delta_0 \frac{\mathrm{Vol}(\Omega)}{\mathrm{Area}(S_0)}$ for any $x \in S_0$,

(d) $(x - x_0) \cdot \eta \le \delta_0 \frac{\mathrm{Vol}(\Omega)}{\mathrm{Area}(S_1)}$ for any $x \in S_1$.

The coefficients of system (1.2) satisfy

(e) $\big(c_{ijk\ell}^{(m-1)} - c_{ijk\ell}^{(m)}\big)\lambda_{k\ell}\lambda_{ij} \ge 0$, $m = 1, 2, \ldots, n$, for any real symmetric matrix (λ_{ij}) of order 3,

(f) $\big(d_{ij}^{(m)} - d_{ij}^{(m-1)}\big)\xi_j\xi_i \ge 0$, $m = 1, 2, \ldots, n$, for any (real) vector $\xi = (\xi_1, \xi_2, \xi_3)$.

Remark. We note that above assumptions (a)–(d) in Hypothesis II are valid when $\delta_0 = 0$ for star-shaped surfaces $\Gamma_1, \Gamma_2, \ldots, \Gamma_n$, S_0, S_1. Moreover, if $\Gamma_1, \ldots, \Gamma_n$ are strongly star-shaped with respect to a point x_0, then the above conditions hold with $\delta_0 > 0$ for a class of domains Ω which includes star-shaped domains. In this sense condition (b) could be considered as a relaxation of the star-shape condition that we usually find in piezoelectric problems.

From now on we fix $\delta_0 > 0$ and $x_0 \in \mathcal{O}_1$ satisfying Hypothesis II. Thus, we will work with the auxiliary function

$$h(x) = \delta_0\, \Phi(x) + \frac{1}{2}\, |x - x_0|^2.$$

The following inequality can be proved by standard arguments:

$$\begin{aligned}
&\left| 2 \sum_{m=0}^{n} \int_{\Omega_m} u_t^{(m)} \cdot \big[(\nabla h \cdot \nabla) u^{(m)} + u^{(m)} \big] dx \right| \\
&\qquad \le C_2 \sum_{m=0}^{n} \int_{\Omega_m} \left\{ |u_t^{(m)}|^2 + c_{ijk\ell}^{(m)}\, \mathcal{E}_{i\ell}(u^{(m)}) \mathcal{E}_{ij}(u^{(m)}) \right\} dx
\end{aligned} \tag{3.27}$$

for some positive constant C_2 which actually could be chosen to be $C_2 = \mathrm{Max}\ \{4, \alpha_0^{-1} \max\{1, C(\Omega)\}\}$, where $\alpha_0 > 0$ is as in (3.26) and

$$C(\Omega) = 2 \max_{x \in \overline{\Omega}} \left\{ \delta_0^2 |\nabla \Phi(x)|^2 + |x - x_0|^2 \right\}.$$

THEOREM 3.3. *Let $\{u,q\}$ be the unique solution of problem* (1.2)–(1.6) *obtained in section* 2. *Then, the inequality*

$$[(1-\delta_0 C)T - C_2]\sum_{m=0}^{n}\int_{\Omega_m}\Big\{|u_t^{(m)}|^2 + c_{ijk\ell}^{(m)}\,\mathcal{E}_{k\ell}(u^{(m)})\mathcal{E}_{ij}(u^{(m)}) + d_{ij}^{(m)}\,\frac{\partial q^{(m)}}{\partial x_j}\,\frac{\partial q^{(m)}}{\partial x_i}\Big\}dx \le \int_0^T\int_{S_0}\frac{\partial h}{\partial \eta}\,|u_t|^2\,d\Gamma\,dt$$

holds, where δ_0 is the fixed positive constant in the definition of $h(x)$, $C>0$, and C_2 is as in (3.27).

Proof. Using assumptions (b), (e), and (f) of Hypothesis II together with Lemma 3.1, we deduce that

$$V_{m-1} - V_m \le 0 \quad \text{on} \quad \Gamma_m \quad \text{for} \quad m = 1,2,\dots,n.$$

Thus, from (3.5), Lemma 3.2, and (3.27) we obtain that

$$[(1-\delta_0 C)T - C_2]\sum_{m=0}^{n}\int_{\Omega_m}\Big\{|u_t^{(m)}|^2 + c_{ijk\ell}^{(m)}\,\mathcal{E}_{k\ell}(u^{(m)})\mathcal{E}_{ij}(u^{(m)}) + d_{ij}^{(m)}\,\frac{\partial q^{(m)}}{\partial x_j}\,\frac{\partial q^{(m)}}{\partial x_i}\Big\}dx \le \int_0^T\int_{S_0} V_n\,d\Gamma\,dt + \int_0^T\int_{S_1} V_0\,d\Gamma\,dt. \tag{3.28}$$

Using the boundary conditions (1.4)–(1.5) and the expressions of V_m and V_0 in (3.6), we deduce that

$$V_n\Big|_{S_0} = \frac{\partial h}{\partial \eta}\,|u_t|^2 - \frac{\partial h}{\partial \eta}\left\{c_{ijk\ell}\,\mathcal{E}_{k\ell}(u)\mathcal{E}_{ij}(u) + d_{ij}\,\frac{\partial q}{\partial x_j}\,\frac{\partial q}{\partial x_i}\right\},$$
$$V_0\Big|_{S_1} = \frac{\partial h}{\partial \eta}\left\{c_{ijk\ell}\,\mathcal{E}_{k\ell}(u)\mathcal{E}_{ij}(u) + d_{ij}\,\frac{\partial q}{\partial x_j}\,\frac{\partial q}{\partial x_i}\right\}.$$

Substitution of the above expression into (3.28) proves the conclusion of Theorem 3.3. □

COROLLARY 3.4. *Under the assumptions of Theorem* 3.3, *let $(f_1,f_2)\in\mathcal{D}(\mathcal{A})$. Suppose that $\{u,q\}$ is the solution of problem* (1.2)–(1.6) *satisfying the condition*

$$u(x,t) = 0 \quad \textit{for any} \quad (x,t)\in S_0\times(0,T).$$

Then, $u(x,t)\equiv 0$, $q(x,t)\equiv 0$ $\forall\,(x,t)\in\Omega\times(0,T)$ for any $T>T_0=\frac{C_2}{1-\delta_0 C}$.

4. Exact controllability. As a consequence of Corollary 3.4 it follows that for $T>T_0=\frac{C_2}{1-\delta_0 C}$ the expression

$$||(f_1,f_2)||_{\mathcal{F}} = \left(\int_0^T\int_{S_0}|u_t|^2\,d\Gamma\,dt\right)^{1/2} \tag{4.1}$$

defines a norm on the set of initial data (f_1,f_2) (where $(u,u_t)=U(t)(f_1,f_2)$ and $\mathcal{F}$ denotes the Hilbert space obtained by completing $\mathcal{D}(\mathcal{A})$ with respect to the norm (4.1)). We also have

$$||(f_1,f_2)|| \le C||(f_1,f_2)||_{\mathcal{F}}.$$

Let us denote by $\mathcal{F}'$ the dual space of $\mathcal{F}$ with respect to X. In the cylinder $\Omega \times (0,T)$ we consider the mixed problem

$$\begin{cases} \dfrac{\partial^2 w_i}{\partial t^2} - \dfrac{\partial}{\partial x_j}\,\sigma_{ij}(w,\psi) = 0 \\ \dfrac{\partial}{\partial x_i}\,D_i(w,\psi) = 0 \end{cases} \qquad \text{in } \Omega_m \times (0,+\infty), \tag{4.2}$$

$m = 0, 1, 2, \ldots, n,$

$$w(x,0) = g_1(x), \quad w_t(x,0) = g_2(x) \quad \text{in} \quad \Omega_m, \tag{4.3}$$

$$\begin{cases} \sigma_{ij}(w,\psi)\eta_j = Q^i(x,t) \\ \psi = 0 \end{cases} \qquad \text{on } S_0 \times (0,+\infty), \tag{4.4}$$

$$\begin{cases} D_i(w,\psi)\eta_i = 0 \\ w = 0 \end{cases} \qquad \text{on } S_1 \times (0,+\infty),$$

$$\begin{cases} \sigma_{ij}\big(w^{(m-1)},\psi^{(m-1)}\big)\eta_j = \sigma_{ij}\big(w^{(m)},\psi^{(m)}\big)\eta_j, \\ D_i\big(w^{(m-1)},\psi^{(m-1)}\big)\eta_i = D_i\big(w^{(m)},\psi^{(m)}\big)\eta_i, \\ \quad w^{(m-1)} = w^{(m)}, \quad \psi^{(m-1)} = \psi^{(m)}, \\ \text{on} \quad \Gamma_m \times (0,T), \quad m = 1, 2, \ldots, n, \end{cases} \tag{4.5}$$

where $Q = Q(x,t) \in [L^2(S_0 \times (0,T))]^3$ and $g = (g_1, g_2) \in \mathcal{F}'$.

In order to give a definition of a solution of problem (4.2)–(4.5) we observe that, in the case when we have a smooth (vector) function w and scalar function ψ, the identity

$$\begin{aligned} \sum_{m=0}^{n} \int_{\Omega_m} \Big\{ & u_t \cdot w_t + c_{ijk\ell}\,\mathcal{E}_{k\ell}(u)\mathcal{E}_{ij}(w) \\ & + d_{ij}\,\frac{\partial}{\partial x_j}\,(\beta(\mathcal{E}(u)))\,\frac{\partial}{\partial x_i}\,(\beta(\mathcal{E}(w))\Big\} dx \Big|_{t=T} \\ = \sum_{m=0}^{n} \int_{\Omega_m} \Big\{ & f_2 \cdot g_2 + c_{ijk\ell}\,\mathcal{E}_{k\ell}(f_1)\mathcal{E}_{ij}(g_1) \\ & + d_{ij}\,\frac{\partial}{\partial x_j}\,(\beta(\mathcal{E}(f_1))\,\frac{\partial}{\partial x_i}\,(\beta(\mathcal{E}(g_1))\Big\} dx \\ & + \int_0^T \int_{S_0} u_t \cdot Q\, d\Gamma\, dt \end{aligned} \tag{4.6}$$

holds for all $f = (f_1, f_2) \in \mathcal{D}(\mathcal{A})$. Here, we note $(u, u_t) = U(t)f$. Given $\mathcal{E}(y)$ (or $\mathcal{E}(w)$), we note $\beta(\mathcal{E}(u))$ (or $\beta(\mathcal{E}(w))$), the solution of problem (2.1)–(2.3) as described in section 2.

We rewrite (4.2) in the form $\frac{d}{dt}(w, \tilde{w}) = \mathcal{A}(w, \tilde{w})$. By definition, $(w(t), \tilde{w}(t)) \in L^\infty(0,T;\mathcal{F}')$ is a solution of (4.2)–(4.5) if

$$\langle (w(t), \tilde{w}(t)), U(t)\varphi \rangle_X = \langle g, \varphi \rangle_X + \int_0^t \int_{S_0} Q \cdot v_t\, d\Gamma\, ds \tag{4.7}$$

for all $\varphi \in \mathcal{F}$ and $0 < t < T$. Let $(v, v_t) = U(t)\varphi$ and $\langle , \rangle$ be the duality $\mathcal{F}'\mathcal{F}$. In a similar way we define a solution of (4.2), (4.4), and (4.5) with zero data at $t = T$ as an element $(w(t), \tilde{w}(t)) \in L^\infty(0, T, \mathcal{F}')$ such that

$$\langle (w(t), \tilde{w}(t)), U(t)\varphi \rangle_X = -\int_t^T \int_{S_0} Q \cdot v_t \, d\Gamma \, ds \tag{4.8}$$

for all $\varphi \in \mathcal{F}$ and $0 < t < T$.

Let f be an arbitrary element of $\mathcal{F}$ and $(w(t), \tilde{w}(t))$ be a solution of (4.2), (4.4), (4.5) with zero data at $t = T > T_0 = \frac{C_2}{1-\delta_0 C}$ and boundary function $Q = -u_t$, where $(u, u_t) = U(t)f$. Let us define $Mf = (w(x, 0), \tilde{w}(x, 0))$. From (4.8) we deduce that

$$\langle Mf, \varphi \rangle = \int_0^T \int_{S_0} u_t \cdot v_t \, d\Gamma \, dt = \langle f, \varphi \rangle_{\mathcal{F}} \, . \tag{4.9}$$

From this it follows that M is an isomorphism from $\mathcal{F}$ onto $\mathcal{F}'$.

Finally, we consider problem (4.2)–(4.5) and suppose that the initial data $g = (g_1, g_2)$ belongs to $\mathcal{F}'$. We set

$$f = M^{-1}g, \quad (u, u_t) = U(t)f, \quad Q = -u_t \, .$$

Using identity (4.7) with $t = T > T_0$, we find that

$$\langle (w(T), \tilde{w}(T)), U(T)\varphi \rangle = \langle Mf, \varphi \rangle - \langle f, \varphi \rangle_{\mathcal{F}} = 0 \tag{4.10}$$

for any $\varphi \in \mathcal{F}$, due to (4.9). This means that $(w(T), \tilde{w}(T))$ "generates" the zero functional on $\mathcal{F}$. In conclusion, we have proved the following theorem.

THEOREM 4.1. *Let us assume Hypotheses* I *and* II. *If* $T > T_0 = \frac{C_2}{1-\delta_0 C}$, *then for any initial data* $f = (f_1, f_2) \in \mathcal{F}'$ *of problem* (1.2), (1.3), (1.6) *with boundary conditions*

$$\begin{cases} \sigma_{ij}(u, q)\eta_j = Q^i(x, t) \\ q = 0 \end{cases} \qquad \text{on } S_0 \times (0, +\infty), \tag{4.11}$$

$$\begin{cases} D_i(u, q)\eta_i = 0 \\ u = 0 \end{cases} \qquad \text{on } S_1 \times (0, +\infty), \tag{4.12}$$

we can find a vector-valued function $Q(x, t) \in \left[L^2(0, T; L^2(S_0))\right]^3$ *such that the corresponding solution of* (1.2), (1.3), (1.6), (4.11), (4.12) *satisfies*

$$u(x, T) = 0, \qquad u_t(x, T) = 0.$$

Acknowledgments. Stimulating discussions concerning this work with members of the Department of Mathematics of the Universidad Autonoma de Madrid (Spain) happened while the third author was a visiting Professor there during September, 2005, partly with financial support from the project "Ecuaciones en Derivadas Parciales: Homogenizacion, control y Aplicaciones" (CEAL) coordinated by Professor E. Zuazua, to whom we express our sincere acknowledgements. The authors would also like to express their thanks to the referees of this journal for their suggestions which improved the final version of this manuscript.

REFERENCES

[1] M. Akamatsu and G. Nakamura, *Well-posedness of initial-boundary value problems for piezoelectric equations*, Appl. Anal., 81 (2002), pp. 129–141.

[2] G. Cimatti, *The piezoelectric continuum*, Ann. Mat., 183 (2004), pp. 495–514.

[3] R. Dager and E. Zuazua, *Wave Propagation, Observation, and Control in* 1*-D Flexible Multistructures*, Math. Appl., 50, Springer, Paris, 2006.

[4] T. Duyckaerts, *Stabilisation haute fréquence d'équations aux dérivées partielles linéaires*, Thèse de Doctorat, Universit Paris XI, Orsay, 2004.

[5] J. N. Eringen and G. A. Maugin, *Electrodynamics of Continua*, Springer, Berlin, 1990.

[6] T. Ikeda, *Fundamentals of Piezoelectricity*, Oxford University Press, London, 1996.

[7] B. V. Kapitonov, *Stabilization and exact boundary controllability for Maxwell's equations*, SIAM J. Control Optim., 32 (1994), pp. 408–420.

[8] B. V. Kapitonov and G. Perla Menzala, *Energy decay and a transmission problem in electromagneto-elasticity*, Adv. Differential Equations, 7 (2002), pp. 819–846.

[9] B. Kapitonov and G. Perla Menzala, *Uniform stabilization and exact control of a multilayered piezoelectric body*, Portugal. Math., 60 (2003), pp. 411–454.

[10] B. Kapitonov and M. A. Raupp, *Boundary observation and exact control of multilayered piezoelectric body*, Math. Methods Appl. Sci., 26 (2003), pp. 431–452.

[11] B. Kapitonov, B. Miara, and G. Perla Menzala, *Stabilization of a layered piezoelectric* 3*-D body by boundary dissipation*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 198–215.

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

[13] J. E. Lagnese, *Boundary controllability in problems of transmission for a class of second order hyperbolic systems*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 343–357.

[14] G. Lebeau and E. Zuazua, *Decay rates for the three-dimensional linear system of themoelasticity*, Arch. Ration. Mech. Anal., 148 (1999), pp. 179–231.

[15] J. L. Lions, *Exact controllability, stabilization, and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[16] J. L. Lions, *Controlabilité Exacte, Perturbations et Stabilisation de Systémes Distribués*, Tome 1, *Controlabilité exacte*, Coll. RMA, Vol. 8, Masson, Paris, 1988.

[17] R. V. N. Melnik and K. N. Melnik, *A note on the class of weakly coupled problems of nonstationary piezoelectricity*, Comm. Numer. Methods Engrg., 14 (1998), pp. 839–847.

[18] B. Miara, *Contrôlabilité d'un corps piézoélectrique*, C.R. Acad. Sci. Paris, 333 (2001), pp. 267–270.

[19] S. Nicaise, *Boundary exact controllability of interface problems with singularities* I: *Addition of the coefficients of singularities*, SIAM J. Control Optim., 34 (1996), pp. 1512–1532.

[20] S. Nicaise, *Boundary exact controllability of interface problems with singularities* II: *Addition of internal controls*, SIAM J. Control Optim., 35 (1997), pp. 585–603.

[21] D. L. Russell, *The Dirichlet–Neumann boundary control problem associated with Maxwell's equations in a cylindrical region*, SIAM J. Control Optim., 24 (1986), pp. 199–229.

[22] P. Russell, *Exact boundary value controllability theorems for waves and heat processes in star-complemented regions*, in Differential Games and Control Theory, E. O. Roxin, P.-T. Lin, and R. Sternberg, eds., Marcel-Dekker, New York, 1974.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 1098–1115

ON REGULARITY OF SOLUTIONS AND LAGRANGE MULTIPLIERS OF OPTIMAL CONTROL PROBLEMS FOR SEMILINEAR EQUATIONS WITH MIXED POINTWISE CONTROL-STATE CONSTRAINTS*

A. RÖSCH† AND F. TRÖLTZSCH‡

Abstract. A class of nonlinear elliptic and parabolic optimal control problems with mixed control-state constraints is considered. Extending a method known for the control of ordinary differential equations to the case of PDEs, the Yosida–Hewitt theorem is applied to show that the Lagrange multipliers are functions of certain L^p-spaces. By bootstrapping arguments, under natural assumptions, optimal controls are shown to be Lipschitz continuous in the elliptic case and Hölder continuous for parabolic problems.

Key words. optimal control, semilinear elliptic equation, semilinear parabolic equation, mixed control-state constraints, multiplier regularity, regularity of optimal controls, Yosida–Hewitt theorem

AMS subject classifications. 49K20, 49N10, 49N15, 90C45

DOI. 10.1137/060671565

1. Introduction.

The solutions of optimal control problems with mixed control-state constraints exhibit better regularity properties than those with pure pointwise state constraints. This fact about the control of ordinary differential equations has been known for a long time. We refer the reader, for instance, to early contributions to linear programming problems related to control problems with constraints of bottleneck type in [22] or [11] and to the more recent exposition by Dmitruk [8]. A first extension to an optimal control problem for the heat equation was presented in [19].

More recently, associated results were shown for more general parabolic equations in Bergounioux and Tröltzsch [4] and Arada and Raymond [3], and for elliptic problems in Tröltzsch [21] and Rösch and Tröltzsch [17]. In all of these papers on the control of PDEs, it was shown that Lagrange multipliers exist in certain L^p-spaces. Different techniques were applied to prove these results. While [4], [17], and [21] used duality theorems, in [3] it was shown that multipliers in $(L^\infty)^*$ are more regular by exploiting the smoothing property of the state equation and using some compactification approach for parabolic equations.

Here, assuming a natural regularity condition, we show the regularity of Lagrange multipliers by the Yosida–Hewitt theorem [23], following an idea explained for ordinary differential equations by Dmitruk [8]. This approach is close to the one suggested by Arada and Raymond but still simplifies and unifies the proof, since compactification arguments are not needed. We also deal with the elliptic case that needs slightly different techniques than the parabolic problems discussed in [3].

Moreover, our paper differs from our former ones by deriving higher regularity of multipliers and optimal controls up to Lipschitz continuity. We extend ideas presented

*Received by the editors September 25, 2006; accepted for publication (in revised form) February 19, 2007; published electronically July 27, 2007.
http://www.siam.org/journals/sicon/46-3/67156.html

†Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria (arnd.roesch@oeaw.ac.at).

‡Technische Universität Berlin, Fakultät II—Mathematik und Naturwissenschaften, Str. des 17. Juni 136, D-10623 Berlin, Germany (troeltzsch@math.tu-berlin.de).

by Rösch and Wachsmuth [18] for a simplified class of elliptic problems. This is the main contribution of this paper.

2. Elliptic optimal control problem and main assumptions. We consider first the following elliptic optimal control problem:

$$\min J(y,u) = \int_\Omega \varphi(x,y,u)\,dx + \int_\Gamma \psi(x,y)\,ds \tag{2.1}$$

subject to

$$\begin{aligned} A\,y + d(x,y) &= u \quad \text{in } \Omega, \\ \frac{\partial y}{\partial \nu_A} + b(x,y) &= 0 \quad \text{on } \Gamma \end{aligned} \tag{2.2}$$

and to

$$g_i(x,y(x),u(x)) \le 0 \quad \text{a.e. on } \Omega, \quad i = 1,\dots,k. \tag{2.3}$$

The inequalities (2.3) are our mixed control-state constraints, which are the main issue of this paper.

Our theory is based upon the following assumptions:

(A1) $\Omega \subset \mathbb{R}^N$, $N \in \mathbb{N}$, is a bounded domain with Lipschitz boundary in the sense of Nečas [13].

(A2) A is a uniformly elliptic differential operator of the form

$$Ay(x) = -\sum_{i,j=1}^N \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}y(x)\right) + c_0(x)y(x)$$

with coefficients $a_{ij} \in C^{0,1}(\bar\Omega)$, $i,j = 1,\dots,N$, that satisfy the condition of uniform ellipticity

$$\sum_{i,j=1}^N a_{ij}(x)\xi_i\xi_j \ge m_0|\xi|^2 \quad \forall x \in \bar\Omega,\ \forall \xi \in \mathbb{R}^N$$

with some $m_0 > 0$. Moreover, c_0 belongs to $L^\infty(\Omega)$ and satisfies $c_0 \ge 0$ a.e. on Ω and $c_0(x) > 0$ on a set of positive measure.

(A3) $\varphi = \varphi(x,y,u) : \Omega \times \mathbb{R}^2 \to \mathbb{R}$ and $g_i = g_i(x,y,u) : \Omega \times \mathbb{R}^2 \to \mathbb{R}$ are given functions enjoying the following properties:
For all fixed y, u, they are Lipschitz with respect to $x \in \Omega$. They are partially differentiable with respect to y and u for all fixed $x \in \bar\Omega$. The derivatives are uniformly Lipschitz on bounded sets, i.e.,
For all $M > 0$ there exists $L(M) > 0$ such that

$$\begin{aligned} &|\varphi(x,y_1,u_1) - \varphi(x,y_2,u_2)| + \left|\frac{\partial\varphi}{\partial y}(x,y_1,u_1) - \frac{\partial\varphi}{\partial y}(x,y_2,u_2)\right| \\ &+ \left|\frac{\partial\varphi}{\partial u}(x,y_1,u_1) - \frac{\partial\varphi}{\partial u}(x,y_2,u_2)\right| \\ &\qquad \le L(M)(|y_1 - y_2| + |u_1 - u_2|), \end{aligned} \tag{2.4}$$

$$\begin{aligned} &|g_i(x,y_1,u_1) - g_i(x,y_2,u_2)| + \left|\frac{\partial g_i}{\partial y}(x,y_1,u_1) - \frac{\partial g_i}{\partial y}(x,y_2,u_2)\right| \\ &+ \left|\frac{\partial g_i}{\partial u}(x,y_1,u_1) - \frac{\partial g_i}{\partial u}(x,y_2,u_2)\right| \\ &\qquad \le L(M)(|y_1 - y_2| + |u_1 - u_2|) \end{aligned} \tag{2.5}$$

hold for a.e. $x \in \Omega$, for all real y_j, u_j with $\max(|y_j|, |u_j|) \le M$, $j = 1, 2$, and for $i = 1, \dots, k$. Moreover, we require

$$|\varphi(x,0,0)| + \left|\frac{\partial \varphi}{\partial y}(x,0,0)\right| + \left|\frac{\partial \varphi}{\partial u}(x,0,0)\right| \le C \quad \text{a.e. on } \Omega,$$
$$|g_i(x,0,0)| + \left|\frac{\partial g_i}{\partial y}(x,0,0)\right| + \left|\frac{\partial g_i}{\partial u}(x,0,0)\right| \le C \quad \text{a.e. on } \Omega.$$

(A4) The functions $\psi = \psi(x,y) : \Gamma \times \mathbb{R} \to \mathbb{R}$, $d = d(x,y) : \Omega \times \mathbb{R} \to \mathbb{R}$, and $b = b(x,y) : \Gamma \times \mathbb{R} \to \mathbb{R}$ are measurable with respect to $x \in \Gamma$ or $x \in \Omega$, respectively, for all fixed $y \in \mathbb{R}$, and differentiable with respect to y for all x. For $y = 0$, they are bounded with respect to x; i.e.,

$$\|\psi(\cdot,0)\|_{L^\infty(\Omega)} + \left\|\frac{\partial \psi}{\partial y}(\cdot,0)\right\|_{L^\infty(\Omega)} + \|b(\cdot,0)\|_{L^\infty(\Gamma)} + \left\|\frac{\partial b}{\partial y}(\cdot,0)\right\|_{L^\infty(\Gamma)} + \|d(\cdot,0)\|_{L^\infty(\Omega)} + \left\|\frac{\partial d}{\partial y}(\cdot,0)\right\|_{L^\infty(\Omega)} \le C.$$

Moreover, they are uniformly Lipschitz on bounded sets; i.e., ψ, b, d, and their derivatives $\partial\psi/\partial y$, $\partial b/\partial y$, $\partial d/\partial y$ satisfy (2.4) or (2.5) with respect to y for almost all $x \in \Omega$ or $x \in \Gamma$, respectively.

(A5) It holds that

$$\frac{\partial d}{\partial y}(x,y) \ge 0 \qquad \forall y \in \mathbb{R}, \ \text{a.e. on } \Omega,$$
$$\frac{\partial b}{\partial y}(x,y) \ge 0 \qquad \forall y \in \mathbb{R}, \ \text{a.e. on } \Gamma.$$

We should mention that the Lipschitz continuity with respect to x of φ and g_i, $i = 1, \dots, k$, is needed only for the results of sections 5 and 6. To have Lagrange multipliers in L^p-spaces, measurability and boundedness with respect to x are sufficient.

3. L^1-regularity of Lagrange multipliers. We consider the controls in the space $U = L^\infty(\Omega)$ and the states y in $Y = H^1(\Omega) \cap C(\bar\Omega)$. Then, thanks to the assumptions (A1), (A2), and (A4), for all $u \in U$ a unique state $y_u \in Y$ exists that solves (2.2) in the weak sense. We refer the reader to Alibert and Raymond [2], who consider the nonlinear system (2.2) including distributed and boundary control and certain unbounded coefficients. Due to their more general setting, their assumptions differ slightly from ours. We also mention Casas [5], who presented a similar technique for the case of boundary control under assumptions that are analogous to ours. The boundedness of the solution y was proven in [2], [5] by the Stampacchia truncation method. For (2.2) and our assumptions, this method can be found in [21, Thm. 7.3].

The control-to-state mapping $G : u \mapsto y$ is continuously Fréchet differentiable from U to Y; cf. again the technique of [2], [5] that can be directly transferred to our problem.

We assume now once and for all that $\bar u \in U$ is a locally optimal control with associated state $\bar y = G(\bar u)$. Local optimality means that there is an $\varepsilon > 0$ such that

$$J(y,u) \ge J(\bar y, \bar u)$$

is satisfied for all (y,u) that satisfy (2.2)–(2.3) and $\|u - \bar u\|_{L^\infty(\Omega)} < \varepsilon$.

We do not discuss the existence of global solutions of the optimal control problem. If the constraints (2.3) include, in particular, $\alpha \le u \le \beta$ with $\alpha, \beta \in L^\infty(\Omega)$, the admissible set is nonempty, and suitable assumptions on the behavior of φ and g_i with respect to u are required, then the existence of a global solution can be shown. This is, however, not the issue of this paper.

We begin our analysis with the existence of Lagrange multipliers in $(L^\infty(\Omega))^*$, the dual space to $L^\infty(\Omega)$. The elements of $(L^\infty(\Omega))^*$ can be represented by finitely additive set functions on $\bar\Omega$ that are also called *finitely additive measures*. We shall use the latter terminology.

To derive necessary optimality conditions, we need a standard constraint qualification and assume the following *linearized Slater condition:*

(A6) There exist $\hat u \in L^\infty(\Omega)$ and $\sigma > 0$ such that

$$\begin{aligned} g_i(x,\bar y(x),\bar u(x)) + \frac{\partial g_i}{\partial y}(x,\bar y(x),\bar u(x))\hat y(x) & \\ + \frac{\partial g_i}{\partial u}(x,\bar y(x),\bar u(x))\hat u(x) &\le -\sigma \quad \text{a.e. in } \Omega, \end{aligned} \tag{3.1}$$

where $\hat y \in Y$ is the solution of the linearized equation

$$\begin{aligned} A\hat y + \frac{\partial d}{\partial y}(x,\bar y(x))\hat y &= \hat u \quad \text{in } \Omega \\ \frac{\partial \hat y}{\partial \nu_A} + \frac{\partial b}{\partial y}(x,\bar y(x))\hat y &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{3.2}$$

Remark 3.1. It holds that $\hat y = G'(\bar u)\hat u$.

Invoking this assumption, the following first-order necessary conditions of Karush–Kuhn–Tucker type can be shown.

THEOREM 3.2. *Suppose that $\bar u$ is locally optimal for* (2.1)–(2.3) *with associated state $\bar y = G(\bar u)$. If the assumptions* (A1)–(A6) *are satisfied, then there exist nonnegative finitely additive measures $\mu_i \in L^\infty(\Omega)^*$, $i = 1,\dots,k$, and an adjoint state $p \in W^{1,s}(\Omega)$ for all $1 \le s < \frac{N}{N-1}$, such that the conditions*

$$\int_\Omega \left(\frac{\partial \varphi}{\partial u}(x,\bar y,\bar u) + p\right) h\,dx + \int_\Omega \sum_{i=1}^k \frac{\partial g_i}{\partial u}(x,\bar y,\bar u) h\,d\mu_i = 0 \quad \forall h \in L^\infty(\Omega), \tag{3.3}$$

$$\int_\Omega g_i(x,\bar y,\bar u)\,d\mu_i = 0, \quad i = 1,\dots,k, \tag{3.4}$$

and the adjoint equation

$$\begin{aligned} A^* p + \frac{\partial d}{\partial y}(x,\bar y)p &= \frac{\partial \varphi}{\partial y}(x,\bar y) + \sum_{i=1}^k \left(\frac{\partial g_i}{\partial y}(x,\bar y,\bar u)^*\mu_i\right)\Big|_\Omega, \\ \frac{\partial p}{\partial \nu_{A^*}} + \frac{\partial b}{\partial y}(x,\bar y)p &= \frac{\partial \psi}{\partial y}(x,\bar y) + \sum_{i=1}^k \left(\frac{\partial g_i}{\partial y}(x,\bar y,\bar u)^*\mu_i\right)\Big|_\Gamma \end{aligned} \tag{3.5}$$

are satisfied.

The proof of the theorem can be performed analogously to Alibert and Raymond [2] or Casas [5], where also the definition and the proof of existence and uniqueness of a weak solution of (3.5) are presented. Notice that the multiplication operators $y \mapsto$

$\frac{\partial g_i}{\partial y}(x,\bar y,\bar u)\,y$ are continuous from $C(\bar\Omega)$ to $L^\infty(\Omega)$. Therefore, the adjoint mappings $\mu_i \mapsto \frac{\partial g_i}{\partial y}(x,\bar y,\bar u)^*\,\mu_i$ are continuous from $L^\infty(\Omega)^*$ to $C(\bar\Omega)^*$ so that their images are regular Borel measures, and the restrictions of them to Ω and Γ are well defined.

As linear continuous functionals on $L^\infty(\Omega)$, the finitely additive measures μ_i must vanish on sets of Lebesgue measure zero. Thanks to Theorem 1.24 of Yosida and Hewitt [23], each $\mu \in L^\infty(\Omega)^*$ can be uniquely written in the form

$$\mu = \mu_c + \mu_p,$$

where μ_c is countably additive and μ_p is purely finitely additive. Moreover, if $\mu \ge 0$, then μ_c and μ_p are nonnegative, too [23, Thm. 1.23].

Let us briefly comment on the associated definitions. *Countable additivity* is equivalent to the following property: For every sequence $\{E_n\}_{n=1}^\infty$ of Lebesgue-measurable sets with $\bar\Omega \supset E_1 \supset E_2 \ldots \supset E_n \ldots$ and $\cap_{n=1}^\infty E_n = \emptyset$, it holds that

$$\lim_{n\to\infty} \mu_c(E_n) = 0. \tag{3.6}$$

Pure finite additivity is defined as follows [23, Def. 1.13]: A nonnegative finitely additive measure μ is said to be purely finitely additive if every countably additive measure λ with $0 \le \lambda \le \mu$ is identically zero. An arbitrary finitely additive measure is purely finitely additive if its nonnegative and its nonpositive parts are purely finitely additive.

Every nonnegative purely finitely additive measure μ_p can be characterized by the following behavior [23, Thm. 1.22]: If λ is nonnegative and countably additive, then there exists a decreasing sequence $\bar\Omega \supset E_1 \supset E_2 \ldots \supset E_n \ldots$ of Lebesgue-measurable sets such that $\lim_{n\to\infty} \lambda(E_n) = 0$ and $\mu_p(E_n) = \mu_p(\Omega)$ for all n. We refer the reader also to Ioffe and Tikhomirov [10, Chap. 8.3.3].

We shall apply this theorem with the Lebesgue measure λ. This means that $\lambda(E_n) = meas(E_n) \to 0$, $n \to \infty$, but

$$\int_{E_n} d\mu_p = \|\mu_p\|_{L^\infty(\Omega)^*} \quad \forall n. \tag{3.7}$$

Our next goal is to show that, under an additional constraint qualification, the singular (i.e., purely finitely additive) parts of all Lagrange multipliers vanish. In this case, we will have at least $\mu_i \in L^1(\Omega)$ for all $i \in \{1,\ldots,k\}$. This property is a consequence of the Radon–Nikodym theorem, since the measures vanish on sets of Lebesgue measure zero.

The following assumption is needed for this purpose.

(A7) Define, for $\delta > 0$, the δ-active sets

$$M_i^\delta := \{x \in \Omega :\ g_i(x,\bar y(x),\bar u(x)) \ge -\delta\}.$$

Assume that there exist $\delta > 0$ and $\tilde u \in L^\infty(\Omega)$ such that there holds

$$\frac{\partial g_i}{\partial u}(x,\bar y(x),\bar u(x))\tilde u(x) \ge 1 \quad \text{a.e. on } M_i^\delta \tag{3.8}$$

for all $i \in \{1,\ldots,k\}$.

We shall discuss the consequences of this assumption later. It is equivalent to a "uniformly positive linear independence condition"; cf. Dmitruk [8]. For some types

of constraints, this assumption is automatically satisfied. In other cases, the optimal solution must fulfill a separation condition.

THEOREM 3.3. *Suppose that $\bar{u} \in U$, $\bar{y} \in Y$, and $\mu_i \in L^\infty(\Omega)^*$, $\mu_i \geq 0$, $i \in \{1, \dots, k\}$, satisfy the first-order necessary optimality conditions of Theorem* 3.2*, and assume that* (A7) *is satisfied. Then the purely finitely additive parts of all μ_i are vanishing so that all μ_i, $i = 1, \dots, k$, can be represented by densities in $L^1(\Omega)$.*

Proof. The proof follows the one given by Dmitruk [8] for the case of ordinary differential equations. We mention first that

$$\int_{\Omega \setminus M_i^\delta} d\mu_i = 0$$

holds true for all $i \in \{1, \dots, k\}$. Otherwise the complementarity condition (3.4) cannot be satisfied, since $g_i < -\delta$ on $\Omega \setminus M_i^\delta$.

Consider, for arbitrary $j \in \{1, \dots, k\}$, the singular part $\mu_{p,j}$ of μ_j. Thanks to Theorem 1.22 of Yosida and Hewitt, there exists a decreasing sequence $\{E_n\}_{n=1}^\infty$ with the properties mentioned above such that

$$\int_{E_n} d\mu_{p,j} = \int_\Omega d\mu_{p,j} \quad \forall n. \tag{3.9}$$

Without limitation of generality, we can assume $E_n \subset M_j^\delta$. We define now

$$h_n = \chi_{E_n} \tilde{u},$$

where $\tilde{u}$ is taken from (3.8) and χ_{E_n} denotes the characteristic function of E_n. Inserting h_n into the gradient equation (3.3), we find that

$$\begin{aligned}
&- \int_\Omega \left(\frac{\partial \varphi}{\partial u}(x, \bar{y}, \bar{u}) + p \right) h_n \, dx = \int_\Omega \sum_{i=1}^k \frac{\partial g_i}{\partial u}(x, \bar{y}, \bar{u}) \, h_n \, d\mu_i \\
&= \sum_{i=1}^k \int_{M_i^\delta} \frac{\partial g_i}{\partial u}(x, \bar{y}, \bar{u}) \, \tilde{u} \, \chi_{E_n} \, d\mu_i \geq \int_{M_j^s} \frac{\partial g_i}{\partial u}(x, \bar{y}, \bar{u}) \, \tilde{u} \, \chi_{E_n} \, d\mu_j \\
&\geq \int_{M_j^\delta} \frac{\partial g_j}{\partial u}(x, \bar{y}, \bar{u}) \, \tilde{u} \, \chi_{E_n} \, d\mu_{p,j} \geq \int_{M_j^\delta} \chi_{E_n} \, d\mu_{p,j} \\
&= \int_{E_n} \chi_{E_n} \, d\mu_{p,j} = \int_\Omega \chi_{E_n} \, d\mu_{p,j} = \|\mu_{p,j}\|_{L^\infty(\Omega)^*}.
\end{aligned}$$

The last inequality was obtained by (3.8). In view of (3.6), the left-hand side tends to zero as $n \to \infty$. Therefore, $\|\mu_{p,j}\|_{L^\infty(\Omega)^*} = 0$. ∎

Remark 3.4. Thanks to the regularity $\mu_i \in L^1(\Omega)$, the adjoint equation admits the simpler form

$$\begin{aligned}
A^* p + \frac{\partial d}{\partial y}(x, \bar{y}) p &= \frac{\partial \varphi}{\partial y}(x, \bar{y}) + \sum_{i=1}^k \frac{\partial g_i}{\partial y}(x, \bar{y}, \bar{u}) \mu_i, \\
\frac{\partial p}{\partial \nu_{A^*}} + \frac{\partial b}{\partial y}(x, \bar{y}) p &= \frac{\partial \psi}{\partial y}(x, \bar{y}).
\end{aligned} \tag{3.10}$$

Moreover, the optimality condition (3.3) and the complementarity condition (3.4) read now

$$\frac{\partial \varphi}{\partial u}(x,\bar{y},\bar{u}) + p + \sum_{i=1}^{k} \frac{\partial g_i}{\partial u}(x,\bar{y},\bar{u})\mu_i = 0 \quad \text{a.e. in } \Omega, \tag{3.11}$$

$$\int_\Omega g_i(x,\bar{y},\bar{u})\mu_i(x)\,dx = 0 \quad \forall i \in \{1,\dots,k\}. \tag{3.12}$$

4. Some examples of constraints. Next, we discuss the regularity condition (3.8) for some examples that might be of interest in the applications.

Example 1 (control constraints). Consider the constraints

$$u_a(x) \le u(x) \le u_b(x) \quad \text{a.e. on } \Omega.$$

We define

$$g_1(x,y,u) = u - u_b(x),$$
$$g_2(x,y,u) = u_a(x) - u.$$

Assume $u_b(x) - u_a(x) \ge \alpha > 0$ a.e. on Ω, and take $\delta = \alpha/3$. Then $M_1(\delta) \cap M_2(\delta) = \emptyset$. Therefore, we can define

$$\tilde{u}(x) = \begin{cases} 1 & \text{on } M_1^\delta, \\ -1 & \text{on } M_2^\delta, \\ 0 & \text{else.} \end{cases}$$

Then

$$\begin{aligned} \frac{\partial g_1}{\partial u}\tilde{u} &= 1 \quad \text{on } M_1^\delta, \\ \frac{\partial g_2}{\partial u}\tilde{u} &= 1 \quad \text{on } M_2^\delta. \end{aligned}$$

In this case, the assumption (A7) is automatically satisfied. However, the existence of regular Lagrange multipliers can here be obtained in an easier and even better way, without assuming $u_b(x) - u_a(x) \ge \alpha > 0$, since

$$\mu_1(x) = \left(\frac{\partial \varphi}{\partial u}(x) + p(x)\right)^+,$$
$$\mu_2(x) = \left(\frac{\partial \varphi}{\partial u}(x) + p(x)\right)^-$$

are Lagrange multipliers; see [20, Thm. 2.29, (2.58), or sect. 6.1, (6.8)].

Example 2 (pure mixed control-state constraints of bottleneck type). Consider the constraint

$$y_a(x) \le \lambda u(x) + y(x) \le y_b(x)$$

with $\lambda \ne 0$ and assume again $y_b(x) - y_a(x) \ge \alpha > 0$ a.e. on Ω. We define $g_1(x,y,u) = \lambda u + y - y_b(x)$, $g_2(x,y,u) = -\lambda u - y + y_a(x)$, and

$$\tilde{u}(x) = \begin{cases} \frac{1}{\lambda} & \text{on } M_1^\delta, \\ -\frac{1}{\lambda} & \text{on } M_2^\delta, \\ 0 & \text{else.} \end{cases}$$

Again, condition (3.8) is automatically satisfied. Also here, the regularity of Lagrange multipliers can be obtained without assuming $y_b - y_a \geq \alpha$ by a transformation to a control constrained problem; cf. [12].

Example 3 (control constraints and unilateral mixed constraint). Let the following constraints be given:

$$\begin{array}{rcccl} u_a(x) & \leq & u(x) & \leq & u_b(x), \\ & & \lambda u(x) - y(x) & \leq & y_b(x), \end{array}$$

with $\lambda > 0$. We define

$$\begin{aligned} g_1(x,y,u) &= u - u_b(x), \\ g_2(x,y,u) &= u_a(x) - u, \\ g_3(x,y,u) &= \lambda u - y - y_b(x) \end{aligned}$$

and assume, for some $\delta > 0$, the separation condition $M_2^\delta \cap M_3^\delta = \emptyset$. Moreover, we assume again $u_b(x) - u_a(x) \geq \alpha > 0$. Then, if δ is sufficiently small, $M_1^\delta \cap M_2^\delta = \emptyset$ is automatically satisfied. We set

$$\tilde{u}(x) = \begin{cases} \max(1/\lambda, 1) & \text{on } M_1^\delta \cup M_3^\delta, \\ -1 & \text{on } M_2^\delta, \\ 0 & \text{else.} \end{cases}$$

Then (3.8) is satisfied. However, we had to assume a separation condition that depends on the unknown solution $(\bar{u}, \bar{y})$. If we have, for example, $u_a(x) \equiv 0$ and we know from maximum principle arguments that $u \geq 0 \Rightarrow y_u \geq 0$ a.e. on Ω, then obviously $y_b(x) \geq \beta > 0$ yields $y(x) + y_b(x) \geq \beta > 0$. In this case, $M_2^\delta \cap M_3^\delta = \emptyset$ is automatically satisfied; we have obtained a result of [17]. We should mention that Arada and Raymond [3] also introduced a separation condition of this type.

Example 4 (equidirected mixed constraints). Consider the general constraints (2.3) and assume that condition (5.2) below is satisfied. Here we can define

$$\tilde{u}(x) \equiv \frac{1}{m} \quad \forall x \in \Omega,$$

and (3.8) is automatically satisfied.

Example 5 (bilateral control and mixed control-state constraints). For the following constraints, a separation condition is needed again:

$$\begin{array}{rcccl} u_a & \leq & u & \leq & u_b, \\ y_a & \leq & u + y & \leq & y_b. \end{array}$$

We define g_1, g_2, g_3, and M_i^δ, $i = 1, 2, 3$, analogously to Example 3. Additionally, we introduce

$$g_4(x,y,u) = y_a(x) - u - y$$

and $M_4^\delta = \{x \in \Omega : y_a(x) - \bar{u}(x) - \bar{y}(x) \geq -\delta\}$. We require, for some $\delta > 0$,

$$(M_2^\delta \cup M_4^\delta) \cap (M_1^\delta \cup M_3^\delta) = \emptyset. \tag{4.1}$$

Then, by the same arguments as before, we see that (A7) is fulfilled. Again, we have to assume (4.1), an additional separation condition.

5. Higher regularity of local solutions. In this section we show how the regularity $\mu_i \in L^1(\Omega)$ can be improved by bootstrapping arguments to finally obtain Lipschitz regularity of $\bar{u}$. To this aim, we have to impose stronger conditions on φ and on the g_i:

(A8) The function φ possesses the second derivative $\partial^2\varphi/\partial u^2(x,y,u)$ on $\bar{\Omega}\times\mathbb{R}^2$. All functions g_i, $i=1,\dots,k$, are defined on $D\times\mathbb{R}^2$, where $D\subset\mathbb{R}^N$ is an open set containing $\bar{\Omega}$. They satisfy (A3) on this extended set.
Moreover, there is a constant $m>0$ such that the monotonicity properties

$$\frac{\partial^2\varphi}{\partial u^2}(x,y,u)\ge m \quad \forall x\in\bar{\Omega},\ \forall (y,u)\in\mathbb{R}^2, \tag{5.1}$$

$$\frac{\partial g_i}{\partial u}(x,y,u)\ge m \quad \forall x\in D,\ \forall (y,u)\in\mathbb{R}^2 \tag{5.2}$$

are satisfied.

Remark 5.1. The extension of the g_i from $\bar{\Omega}$ to a larger open set D is needed in the proof of the next theorem. We apply the Robinson implicit function theorem in an open covering of $\bar{\Omega}\times[-M,M]$. In our examples, the dependence of the g_i on x comes with that of the functions u_a, u_b or y_a, y_b defining the bounds. The extension of these functions to a neighborhood around $\bar{\Omega}$ should not cause difficulties.

We will also consider bilateral constraints of the form

$$\alpha_i(x)\le\gamma_i(x,y(x),u(x))\le\beta_i(x),\quad i=1,\dots,l, \tag{5.3}$$

where the γ_i, $i=1,\dots,l$, satisfy (A8) and $\alpha_i\le\beta_i$ are Lipschitz functions.

LEMMA 5.2. *Suppose that $g_1,\dots,g_k$ satisfy assumption* (A8). *Then there exist functions $\phi_i:\bar{\Omega}\times\mathbb{R}^2\to\mathbb{R}$ with the following properties: All $\phi_i(x,y)$ are Lipschitz with respect to x for all $y\in\mathbb{R}$,*

$$|\phi_i(x,y_1)-\phi_i(x,y_2)|\le L(M)|y_1-y_2| \tag{5.4}$$

is satisfied for all $x\in\bar{\Omega}$ and all $|y_j|\le M$, and there holds

$$g_i(x,y,u)\begin{cases} =0 & \Leftrightarrow \quad u=\phi_i(x,y),\\ <0 & \Leftrightarrow \quad u<\phi_i(x,y),\\ >0 & \Leftrightarrow \quad u>\phi_i(x,y).\end{cases} \tag{5.5}$$

Proof. Consider, for fixed i, the equation

$$g_i(x,y,u)=0. \tag{5.6}$$

By (5.2) we have $\lim_{u\to\pm\infty} g_i(x,y,u)=\pm\infty$, and hence, for each $(x,y)\in D\times\mathbb{R}$, (5.6) has a unique solution $u=\phi_i(x,y)$. To show the Lipschitz property of ϕ_i, we invoke the implicit function theorem of Robinson [16, Thm. 2.1]. It ensures that, for each pair $(x_0,y_0)\in D\times\mathbb{R}$ and each $\varepsilon>0$, there is an (open) neighborhood $N_\varepsilon(x_0,y_0)\subset D\times\mathbb{R}$ such that

$$|\phi_i(x,y)-\phi_i(\xi,\eta)|\le(\lambda+\varepsilon)|g_i(x,y,\phi_i(x,y))-g_i(\xi,\eta,\phi_i(x,y))| \tag{5.7}$$

holds for all (x,y) and (ξ,η) in $N_\varepsilon(x_0,y_0)$, where $\lambda=1/m$ with m defined by (5.2).

The collection of all neighborhoods $N_\varepsilon(x_0,y_0)$, $(x_0,y_0)\in D\times[-M,M]$, defines an open covering of the compact set $\bar{\Omega}\times[-M,M]$. Selecting a finite covering, an easy

application of the triangle inequality shows that (5.4) holds everywhere in $\bar{\Omega} \times \mathbb{R}$ with a suitable constant $L(M)$.

In view of the strong monotonicity of g with respect to u for all fixed (x, y), the reader may now readily verify the relations (5.5). □

LEMMA 5.3. *Assume that the optimality system* (3.10)–(3.12) *is fulfilled with Lagrange multipliers* $\mu_i \in L^1(\Omega)$. *If* (A8) *is satisfied, then the Lagrange multipliers* μ_i *satisfy a.e. on* Ω *the equation*

(5.8)
$$\sum_{i=1}^{k} \frac{\partial g_i}{\partial u}(x, \bar{y}(x), \bar{u}(x))\mu_i(x) = \max\left(0, -\left(\frac{\partial \varphi}{\partial u}\big(x, \bar{y}(x), \min_{i=1,\dots,k} \phi_i(x, \bar{y}(x))\big) + p(x)\right)\right).$$

Proof. We extend an idea introduced in [18] and consider two cases for $x \in \Omega$.

(i) $x \in M^+ = \{x \in \Omega : \sum_{i=1}^{k} \frac{\partial g_i}{\partial u}(x, \bar{y}(x), \bar{u}(x))\mu_i(x) > 0\}$.

Assumption (A8) ensures in particular that $\partial g_i/\partial u \geq 0$ so that, for each $x \in M^+$, at least one multiplier $\mu_i(x)$ must be positive. In view of the complementary slackness condition (3.12), a.e. in this set, at least one inequality constraint is active. Therefore, in view of (5.5), we have

$$\bar{u}(x) = \min_i \phi_i(x, \bar{y}(x)) \quad \text{a.e. on } M^+. \tag{5.9}$$

Moreover, from $\sum_{i=1}^{k} \frac{\partial g_i}{\partial u}(x, \bar{y}(x), \bar{u}(x))\mu_i(x) > 0$ and the gradient equation (3.11) we deduce

$$\frac{\partial \varphi}{\partial u}(x, \bar{y}(x), \bar{u}(x)) + p(x) < 0 \qquad \text{a.e. on } M^+.$$

Inserting the expression (5.9) for $\bar{u}$ in this inequality, it follows that

$$0 < -\left(\frac{\partial \varphi}{\partial u}(x, \bar{y}(x), \min_i \phi_i(x, \bar{y}(x))) + p(x)\right) \text{ a.e. on } M^+.$$

Therefore, again in view of (3.11), we obtain

$$\sum_{i=1}^{k} \frac{\partial g_i}{\partial u}(x, \bar{y}(x), \bar{u}(x))\mu_i(x) = \max\left(0, -\left(\frac{\partial \varphi}{\partial u}(x, \bar{y}(x), \min_i \phi_i(x, \bar{y}(x))) + p(x)\right)\right),$$

since the left-hand side is positive.

(ii) $x \in \Omega \setminus M^+ = \{x \in \Omega : \sum_{i=1}^{k} \frac{\partial g_i}{\partial u}(x, \bar{y}(x), \bar{u}(x))\mu_i(x) = 0\}$.

Here, the gradient equation (3.11) shows

$$-\left(\frac{\partial \varphi}{\partial u}(x, \bar{y}(x), \bar{u}(x)) + p(x)\right) = 0. \tag{5.10}$$

Moreover, we have

$$\bar{u}(x) \leq \min_i \phi_i(x, \bar{y}(x)).$$

From the monotonicity condition (5.1), it follows that

$$\frac{\partial \varphi}{\partial u}(x, \bar{y}(x), \bar{u}(x)) \leq \frac{\partial \varphi}{\partial u}(x, \bar{y}(x), \min_i \phi_i(x, \bar{y}(x))).$$

Together with (5.10), this implies

$$-\left(\frac{\partial\varphi}{\partial u}(x,\bar y(x),\min_i \phi_i(x,\bar y(x))) + p(x)\right) \le 0,$$

and hence

$$\sum_{i=1}^k \frac{\partial g_i}{\partial u}(x,\bar y(x),\bar u(x))\mu_i(x) = 0 = \max\left(0, -\left(\frac{\partial\varphi}{\partial u}(x,\bar y(x),\min_i \phi_i(x,\bar y(x))) + p(x)\right)\right)$$

holds also a.e. in $\Omega\setminus M^+$, too. □

THEOREM 5.4. *Suppose that $(\bar y,\bar u)\in H^1(\Omega)\cap C(\bar\Omega)\times L^\infty(\Omega)$ satisfy, together with $p\in W^{1,s}(\Omega)$, $1\le s<\frac{N}{N-1}$, and $\mu_1,\dots,\mu_k\in L^1(\Omega)$, the optimality conditions of Theorem 3.2. If the assumptions* (A3) *and* (A8) *are satisfied, then all multipliers μ_i, $i=1,\dots,k$, are bounded and measurable functions. If Γ is of class $C^{1,1}$, then $\bar u$ and $\sum_{i=1}^k \frac{\partial g_i}{\partial u}(x,\bar y,\bar u)\mu_i$ are Lipschitz functions on $\bar\Omega$.*

Proof. We show this result by a bootstrapping argument. At the beginning, we know that $\bar u\in L^\infty(\Omega)$ and $\bar y\in C(\bar\Omega)$.

Thanks to $p\in W^{1,s}(\Omega)$, by Sobolev embedding theorems there is a $\sigma>0$ such that $p\in L^{s_1}(\Omega)$ with $s_1=1+\sigma$ (see also our arguments at the end of the proof). From the gradient equation (3.11), we deduce

$$\sum_{i=1}^k \frac{\partial g_i}{\partial u}(x,\bar y,\bar u)\mu_i = -\frac{\partial\varphi}{\partial u}(x,\bar y,\bar u) - p \in L^{s_1}(\Omega). \tag{5.11}$$

Because of (5.2) and by the nonnegativity of the multipliers μ_i, this implies $\mu_i\in L^{s_1}(\Omega)$ for all $i\in\{1,\dots,k\}$ and hence

$$\sum_{i=1}^k \frac{\partial g_i}{\partial y}(x,\bar y,\bar u)\,\mu_i \in L^{s_1}(\Omega).$$

Inserting this into (3.10), the right-hand side is seen to belong to $L^{s_1}(\Omega)$. Therefore,

$$p\in W^{1,s_1}(\Omega)\hookrightarrow L^{s_2}(\Omega),\ \text{where } s_2=s_1+\sigma \text{ and } \sigma>0.$$

We explain below why the same σ can be taken. By (5.11), we find

$$\sum_{i=1}^k \frac{\partial g_i}{\partial u}(x,\bar y,\bar u)\mu_i \in L^{s_2}(\Omega).$$

Repeating this bootstrapping method, we get numbers s_i with $s_{i+1}\ge s_i+\sigma$. We can take the same $\sigma>0$ for all i for the following reason: If $p\in W^{1,s}(\Omega)$, then $p\in L^r(\Omega)$ for all r given by

$$\frac{1}{r}=\frac{1}{s}-\frac{1}{N}, \tag{5.12}$$

provided that $1<\frac{N}{s}$; cf. Adams [1]. Let us assume $1<\frac{N}{s}$. Then (5.12) implies

$$r-s=\frac{s^2}{N-s}>\frac{s^2}{N}>1/N$$

by $s \geq 1$, and we are justified to take $\sigma = 1/N$.

After finitely many steps, in any case we arrive at a situation, where $N/s_{i+1} < 1$ while $N/s_i > 1$ (notice that we have some freedom in the choice of σ to avoid the equality sign in both the equations).

In this case, it holds that $p \in W^{1,s_{i+1}}(\Omega) \hookrightarrow C(\bar{\Omega})$. This implies

$$\mu_i \in L^\infty(\Omega) \quad \forall i \in \{1, \dots, k\}.$$

Now we need the higher smoothness $C^{1,1}$ of Γ. Exploiting again (3.10), we obtain $p \in W^{2,s}(\Omega)$ for all $s < \infty$. This regularity result follows from Grisvard [9]. Therefore, p is continuously differentiable (Adams [1]) and hence Lipschitz.

Now, we invoke formula (5.8). Since $\bar{u}$ is bounded and measurable, $\bar{y}$ is also Lipschitz. The same holds true for the function

$$\min_{i \in \{1,\dots,k\}} \phi_i(x, \bar{y}(x)),$$

since all ϕ_i are Lipschitz. Thanks to this, the right-hand side of (5.8) is Lipschitz so that the left-hand side must have this property, too.

From the gradient equation (5.11), we now obtain

$$\frac{\partial \varphi}{\partial u}(\cdot, \bar{y}, \bar{u}) \in C^{0,1}(\bar{\Omega}). \tag{5.13}$$

Next we make use of the assumption (A8) and (5.1), i.e., $\frac{\partial^2 \varphi}{\partial u^2} \geq m > 0$. Invoking the implicit function theorem again, we arrive at the Lipschitz continuity of $\bar{u}$. □

Bilateral nonlinear mixed constraints. Finally, we consider the constraints (5.3), where we need an additional separation assumption to prove the Lipschitz continuity of $\bar{u}$. We assume the following.

(A9) The functions γ_i satisfy Assumption (A8) on the g_i. Moreover φ satisfies (A8), too, and there is a $\delta > 0$ such that the sets

$$M^\alpha_{i,\delta} := \{x : \gamma_i(x, \bar{u}(x), \bar{y}(x)) \leq \alpha_i(x) + \delta\},$$
$$M^\beta_{i,\delta} := \{x : \beta_i(x) - \delta \leq \gamma_i(x, \bar{u}(x), \bar{y}(x))\}$$

satisfy the condition

$$\bigcup_{i=1}^k M^\alpha_{i,\delta} \cap \bigcup_{i=1}^k M^\beta_{i,\delta} = \emptyset.$$

THEOREM 5.5. *Consider the optimal control problem* (2.1)–(2.3) *for constraints of the form* (5.3)*, i.e., for*

$$g_i = \begin{cases} \gamma_i - \beta_i, & i \in \{1, \dots, l\}, \\ \alpha_{i-l} - \gamma_{i-l}, & i \in \{l+1, \dots, 2l\}. \end{cases}$$

Suppose that $\bar{y} \in H^1(\Omega) \cap C(\bar{\Omega})$ *and* $\bar{u} \in L^\infty(\Omega)$ *satisfy together the first-order necessary optimality conditions. Assume that* (A9) *is satisfied and that* Γ *is of class* $C^{1,1}$*. Then the functions*

$$\sum_{i=1}^{l} \frac{\partial g_i}{\partial u}(x, \bar{y}, \bar{u})\mu_i, \qquad \sum_{i=l+1}^{2l} \frac{\partial g_i}{\partial u}(x, \bar{y}, \bar{u})\mu_i$$

and the optimal control $\bar u$ are Lipschitz.

Proof. Let us recall first that we have assumed $\partial\gamma_i/\partial u \ge m$ for all $i \in \{1,\dots,l\}$. Therefore, in view of the definition of the g_i, it holds that

$$\frac{\partial g_i}{\partial u} \ge m \ \text{ if } \ 1 \le i \le l, \qquad \frac{\partial g_i}{\partial u} \le -m \ \text{ if } \ l+1 \le i \le 2l.$$

Now we proceed similarly to the proof of Theorem 5.4 and distinguish four cases with respect to $x \in \Omega$:

$$\sum_{i=1}^{l} \frac{\partial g_i}{\partial u}\mu_i > 0,\ \sum_{i=1}^{l} \frac{\partial g_i}{\partial u}\mu_i = 0,\ \sum_{i=l+1}^{2l} \frac{\partial g_i}{\partial u}\mu_i < 0,\ \sum_{i=l+1}^{2l} \frac{\partial g_i}{\partial u}\mu_i = 0.$$

Here and in what follows we suppress the arguments $(x,\bar y,\bar u)$ in $\partial g_i/\partial u$ for convenience. The first two cases concern the upper bounds; hence they are of the type considered in Theorem 5.4. Let us therefore concentrate on the remaining two cases.

(i) $\sum_{i=l+1}^{2l}(\frac{\partial g_i}{\partial u}\mu_i)(x) < 0$. At least one of the multipliers μ_i, $i \in \{l+1,\dots,2l\}$, must be positive; thus one of the associated lower constraints is active. Hence, by the separation assumption (A9), no one of the upper constraints can be almost active. This implies that all multipliers μ_i with $i \in \{1,\dots,l\}$ must vanish a.e. on this set, i.e.,

$$\sum_{i=1}^{l} \left(\frac{\partial g_i}{\partial u}\mu_i\right)(x) = 0.$$

Invoking the gradient equation (3.3), we find

$$0 < -\sum_{i=l+1}^{2l} \left(\frac{\partial g_i}{\partial u}\mu_i\right)(x) = \frac{\partial \varphi}{\partial u}(x,\bar y(x),\bar u(x)) + p(x)$$

and hence

$$-\sum_{i=l+1}^{2l} \left(\frac{\partial g_i}{\partial u}\mu_i\right)(x) = \max\left(0, \frac{\partial \varphi}{\partial u}(x,\bar y(x),\bar u(x)) + p(x)\right).$$

Moreover, we have in this case that

$$\bar u(x) = \max_{i\in\{1,\dots,l\}} \phi_i^\alpha(x,\bar y(x))$$

with Lipschitz functions ϕ_i^α, which are associated to the lower bounds and defined by

$$g_i(x,y,u) = \alpha_i(x) \quad \Leftrightarrow \quad u = \phi_i^\alpha(x,y).$$

This follows by the arguments of Lemma 5.2. Consequently,

$$-\sum_{i=l+1}^{2l} \left(\frac{\partial g_i}{\partial u}\mu_i\right)(x) = \max\left(0, \frac{\partial \varphi}{\partial u}(x,\bar y(x), \max_{i\in\{1,\dots,l\}} \phi_i^\alpha(x,\bar y(x))) + p(x)\right)$$

holds on this set.

(ii) $\sum_{i=l+1}^{2l}(\frac{\partial g_i}{\partial u}\mu_i)(x) = 0$. The gradient equation implies then

$$\frac{\partial \varphi}{\partial u} + p = -\sum_{i=1}^{l}\left(\frac{\partial g_i}{\partial u}\mu_i\right)(x) \le 0$$

and

$$\bar{u}(x) \ge \max_{i\in\{1,\dots,l\}} \phi_i^\alpha(x, \bar{y}(x)).$$

In view of the monotonicity property (5.1), we obtain

$$\frac{\partial \varphi}{\partial u}(x, \bar{y}(x), \max_{i\in\{1,\dots,l\}} \phi_i^\alpha(x, \bar{y}(x))) + p(x) \le \frac{\partial \varphi}{\partial u}(x, \bar{y}(x), \bar{u}(x)) + p(x) \le 0.$$

Obviously, it therefore holds that

(5.14)
$$0 = -\sum_{i=l+1}^{2l}\left(\frac{\partial g_i}{\partial u}\mu_i\right)(x) = \max\left(0, \frac{\partial \varphi}{\partial u}(x, \bar{y}(x), \max_{i\in\{1,\dots,l\}} \phi_i^\alpha(x, \bar{y}(x))) + p(x)\right),$$

so that (5.14) is satisfied a.e. on Ω. Invoking the same bootstrapping arguments as in the proof of Theorem 5.4, we deduce the desired Lipschitz properties. ∎

6. The parabolic case. It is fairly obvious that the method of the preceding sections can be extended to problems with parabolic state equation. There are some differences in the regularity results of the equation, but the main ideas are analogous. Here, we briefly sketch the arguments to show Hölder regularity of the optimal control.

In [3], the L^1-regularity of Lagrange multipliers has already been investigated for parabolic equations. Therefore, we prove Hölder continuity on the assumption that the Lagrange multipliers belong to L^1. In [3], sufficient conditions can be found that ensure this property.

We consider the following parabolic counterpart to the elliptic optimal control problem (2.1)–(2.3):

$$\min J(y,u) := \int_\Omega \int_0^T \varphi(x,t,y,u)\,dxdt + \int_\Gamma \int_0^T \psi(x,t,y)\,dsdt \tag{6.1}$$

subject to

$$\begin{aligned} \frac{\partial y}{\partial t} + A\,y + d(x,t,y) &= u && \text{in } Q := \Omega\times(0,T),\\ \frac{\partial y}{\partial \nu_A} + b(x,t,y) &= 0 && \text{in } \Sigma := \Gamma\times(0,T),\\ y(\cdot,0) &= y_0(\cdot) && \text{in } \Omega \end{aligned} \tag{6.2}$$

and to

$$g_i(x,t,y(x,t),u(x,t)) \le 0 \quad \text{a.e. in } Q,\ i = 1,\dots,k. \tag{6.3}$$

We rely on the following general assumptions:

(A10) The given data have to satisfy direct extensions of (A1)–(A5) to the parabolic case that are obtained as follows: In (A1), we additionally assume that Γ is of class $C^{1,1}$. (A2) remains unchanged except that c_0 is now a function of $L^\infty(Q)$ not restricted in sign. In (A3)–(A5), the sets Ω and Γ are replaced by Q and Σ, respectively, and $\tilde{x} := (x,t)$ replaces x in these assumptions. Moreover, we assume that y_0 is Hölder continuous in Ω.

In particular, d, b are monotone nondecreasing with respect to y and $d(\cdot,\cdot,0)$, $b(\cdot,\cdot,0)$ belong to $L^\infty(Q)$ and $L^\infty(\Sigma)$, respectively.

Under these assumptions, for all $u \in L^r(Q)$ with $r > N/2+1$, the parabolic equation (6.2) has a unique solution $y \in W(0,T) \cap C(\bar{Q})$; cf. Casas [6] or Raymond and Zidani [15]. The space $W(0,T)$ is defined by

$$W(0,T) = \left\{ y \in L^2(0,T;H^1(\Omega)) : \frac{dy}{dt} \in L^2(0,T;H^1(\Omega)') \right\}.$$

For the remainder of this section, let $\bar{u} \in L^\infty(Q)$ be (locally) optimal for (6.1)–(6.3). We assume that nonnegative Lagrange multipliers $\mu_i \in L^1(Q)$ and an adjoint state p exist such that the following first-order necessary optimality conditions are satisfied:

$$\begin{aligned} -\frac{\partial p}{\partial t} + A^*p + \frac{\partial d}{\partial y}(x,t,\bar{y})\,p &= \frac{\partial \varphi}{\partial y}(x,t,\bar{y}) + \sum_{i=1}^k \frac{\partial g_i}{\partial y}(x,t,\bar{y},\bar{u})\,\mu_i && \text{in } Q, \\ \frac{\partial p}{\partial \nu_{A^*}} + \frac{\partial b}{\partial y}(x,t,\bar{y})\,p &= \frac{\partial \psi}{\partial y}(x,t,\bar{y}) && \text{in } \Sigma, \\ p(\cdot,T) &= 0 && \text{in } \Omega, \end{aligned} \tag{6.4}$$

$$\frac{\partial \varphi}{\partial u}(x,t,\bar{y},\bar{u}) + p + \sum_{i=1}^k \frac{\partial g_i}{\partial u}(x,t,\bar{y},\bar{u})\,\mu_i = 0 \quad \text{a.e. in } Q, \tag{6.5}$$

$$\iint_Q g_i(x,t,\bar{y},\bar{u})\,\mu_i\,dxdt = 0 \quad \forall i \in \{1,\dots,k\}. \tag{6.6}$$

The adjoint state p is the weak solution of (6.4) and belongs to $L^{\tilde{r}}(0,T,W^{1,r}(\Omega))$ for all $\tilde{r} > 1$, $r > 1$ satisfying

$$\frac{N}{2} + \frac{1}{2} < \frac{N}{2r} + \frac{1}{\tilde{r}};$$

cf. [14, Thm. 4.3]. Now we are going to show Hölder continuity of $\bar{u}$. To this end, we additionally assume the following:

(A11) The function φ possesses the second-order derivative $\partial^2\varphi/\partial u^2(x,t,y,u)$ on $\bar{Q}\times\mathbb{R}^2$. All functions g_i, $i=1,\dots,k$, are defined on $D\times\mathbb{R}^2$, where $D \subset \mathbb{R}^{N+1}$ is an open set containing $\bar{Q}$. They satisfy (A3) on this extended set. There is a constant $m > 0$ such that the monotonicity properties

$$\frac{\partial^2 \varphi}{\partial u^2}(x,t,y,u) \ge m \quad \forall (x,t) \in \bar{Q},\ \forall (y,u) \in \mathbb{R}^2, \tag{6.7}$$

$$\frac{\partial g_i}{\partial u}(x,t,y,u) \ge m \quad \forall (x,t) \in D,\ \forall (y,u) \in \mathbb{R}^2 \tag{6.8}$$

are satisfied.

The assertions of the Lemmas 5.2 and 5.3 do not depend on the special structure of the underlying PDE. Obviously, they can be directly transferred to the parabolic case. Therefore, the following extension of (5.8) is satisfied a.e. in Q:

$$\begin{aligned} &\sum_{i=1}^k \frac{\partial g_i}{\partial u}(x,t,\bar{y}(x,t),\bar{u}(x,t))\mu_i(x,t) \\ &\qquad = \max\left(0, -\left(\frac{\partial \varphi}{\partial u}\big(x,t,\bar{y}(x,t), \min_{i=1,\dots,k} \phi_i(x,t,\bar{y}(x,t))\big) + p(x,t)\right)\right). \end{aligned} \tag{6.9}$$

The functions ϕ_i are constructed again by the Robinson implicit function theorem that ensures, in particular, an estimate of the type (5.7). Now, the functions g_i in this estimate are only locally Hölder continuous so that all $\phi_i(x,t,y)$ are locally Hölder continuous: There is a constant $\lambda \in (0,1)$ and, for all $M > 0$, a constant $H(M) > 0$ depending on M such that

$$|\phi_i(x_1,t_1,y_1) - \phi_i(x_2,t_2,y_2)| \le H(M)|(x_1,t_1,y_1) - (x_2,t_2,y_2)|^\lambda \tag{6.10}$$

holds for all $(x_i,t_i) \in \bar{Q}$ and for all $y_i \in [-M,M]$.

THEOREM 6.1. *Suppose that $(\bar{y},\bar{u}) \in W(0,T) \cap C(\bar{Q}) \times L^\infty(Q)$ satisfy, together with $p \in L^{\tilde{r}}(0,T,W^{1,r}(\Omega))$ for all $\tilde{r} > 1$, $r > 1$, and $\mu_1,\dots,\mu_k \in L^1(Q)$, the optimality conditions* (6.4)–(6.6). *If the assumptions* (A10) *and* (A11) *are satisfied, then all multipliers μ_i, $i = 1,\dots,k$, belong to $L^\infty(Q)$. Moreover, the optimal control $\bar{u}$ and the expression $\sum_{i=1}^k \frac{\partial g_i}{\partial u}(x,t,\bar{y},\bar{u})\mu_i$ are Hölder continuous on $\bar{Q}$.*

Proof. We proceed by bootstrapping arguments following the proof of Theorem 5.4. By our assumptions, we know $\bar{u} \in L^\infty(Q)$ and $\bar{y} \in C(\bar{Q})$.

Consider now the adjoint equation (6.4). Thanks to Theorem 4.2, (i), in [14], the right-hand sides of the adjoint equation in $L^s(Q)$ are transformed into solutions in $L^\alpha(Q)$ with $\alpha \ge s$, if

$$\frac{1}{s}\left(\frac{N}{2}+1\right) < \frac{1}{\alpha}\left(\frac{N}{2}+1\right) + 1,$$

and hence the right-hand sides from $L^s(Q)$ are transformed into $L^\alpha(Q)$ for all $\alpha \ge 1$ with

$$\alpha < \frac{s(N/2+1)}{N/2+1-s}$$

provided that $s < N/2+1$. For $s > N/2+1$, the transformation is from $L^s(Q)$ to $C(\bar{Q})$. The gain of smoothness $\alpha - s$ is

$$\alpha - s = \frac{s^2}{N/2+1-s} - \varepsilon,$$

where $\varepsilon > 0$ can be taken arbitrarily small. Therefore, by $s \ge 1$, at least the gain

$$\alpha - s \ge \frac{s^2}{N/2+1} \ge \frac{1}{N/2+1} =: \sigma$$

is obtained, and hence $p \in L^{s+\sigma}(Q)$.

We start a bootstrapping procedure at $s := 1$. From the gradient equation (6.5), we deduce

$$\sum_{i=1}^k \frac{\partial g_i}{\partial u}(x,t,\bar{y},\bar{u})\mu_i = -\frac{\partial \varphi}{\partial u}(x,t,\bar{y},\bar{u}) - p \ \in L^{s+\sigma}(Q). \tag{6.11}$$

Because of (6.8) and by the nonnegativity of the multipliers μ_i, this implies

$$\mu_i \in L^{s+\sigma}(Q) \quad \forall i \in \{1,\dots,k\}.$$

Inserting this into (6.4), the right-hand sides of the adjoint equation are seen to belong to $L^{s+\sigma}(Q)$. Therefore, we obtain by the same arguments as before

$$p \in L^{s+2\sigma}(Q).$$

By (6.11) and the boundedness of the functions $\partial g_i/\partial u(x,t,\bar{y},\bar{u})$, we find

$$\sum_{i=1}^{k} \frac{\partial g_i}{\partial u}(x,t,\bar{y},\bar{u})\mu_i \in L^{s+2\sigma}(Q).$$

Repeating this bootstrapping method, after finitely many steps, we arrive at the situation that $N/2+1 < 1+(j+1)\sigma$ while $N/2+1 > 1+j\sigma$. In this case, it holds that $p \in C(\bar{Q})$, and (6.11) implies

$$\mu_i \in L^\infty(Q) \quad \forall i \in \{1,\dots,k\}.$$

We know that p is bounded on $\bar{Q}$ and its terminal value is zero and hence Hölder continuous on $\bar{\Omega}$. Therefore, Theorem 4 in Di Benedetto [7] yields Hölder continuity of p. (For our case of variational boundary data, this theorem ensures Hölder continuity of the solution on $\bar{\Omega}\times[0,T-\varepsilon]$ for all $\varepsilon>0$. Moreover, it states Hölder continuity on $\bar{Q}$ if the prescribed terminal data are Hölder.)

Now, we invoke formula (6.9). Since $\bar{y}$ bounded and y_0 is Hölder continuous, $\bar{y}$ exhibits this property, too. The same holds true for the function

$$\min_{i\in\{1,\dots,k\}} \phi_i(x,t,\bar{y}(x,t)),$$

since, by (6.10), all ϕ_i are Hölder continuous. Thanks to this, the right-hand side of (6.9) is Hölder continuous so that the left-hand side has this property, too.

From the gradient equation (6.11), we now obtain

$$\frac{\partial \varphi}{\partial u}(\cdot,\bar{y},\bar{u}) \in C^{0,\kappa}(\bar{Q}) \tag{6.12}$$

with some $\kappa \in (0,1)$. Next we make use of the assumption (A11) and (6.7), i.e., $\frac{\partial^2 \varphi}{\partial u^2} \geq m > 0$. Invoking the implicit function theorem again, we deduce the Hölder continuity of $\bar{u}$. □

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, Boston, 1978.

[2] J.-J. Alibert and J.-P. Raymond, *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls*, Numer. Funct. Anal. Optim., 3–4 (1997), pp. 235–250.

[3] N. Arada and J. P. Raymond, *Optimal control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1391–1407.

[4] M. Bergounioux and F. Tröltzsch, *Optimal control of linear bottleneck problems*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 235–250.

[5] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[6] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.

[7] E. Di Benedetto, *On the local behaviour of solutions of degenerate parabolic equations with measurable coefficients*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 13 (1986), pp. 487–535.

[8] A. Dmitruk, *Maximum principle for the general optimal control problem with phase and regular mixed constraints*, Comput. Math. Model., 4 (1993), pp. 364–377.

[9] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[10] A. D. Ioffe and V. M. Tichomirov, *Theorie der Extremalaufgaben*, Deutscher Verlag der Wissenschaften, Berlin, 1979.

[11] N. Levinson, *A class of continuous linear programming problems*, J. Math. Anal. Appl., 16 (1966), pp. 73–83.

[12] C. Meyer and F. Tröltzsch, *On an elliptic optimal control problem with pointwise mixed control-state constraints*, in Recent Advances in Optimization. Proceedings of the 12th French-German-Spanish Conference on Optimization (Avignon, 2004), A. Seeger, ed., Lecture Notes in Econom. and Math. Systems 563, Springer-Verlag, New York, 2006, pp. 187–204.

[13] J. Nečas, *Les méthodes directes en théorie des equations elliptiques*, Academia, Prague, 1967.

[14] J.-P. Raymond and F. Tröltzsch, *Second order sufficient optimality conditions for nonlinear parabolic control problems with state constraints*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 431–450.

[15] J.-P. Raymond and H. Zidani, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations*, Appl. Math. Optim., 39 (1999), pp. 143–177.

[16] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.

[17] A. Rösch and F. Tröltzsch, *Existence of regular Lagrange multipliers for a nonlinear elliptic optimal control problem with pointwise control-state constraints*, SIAM J. Control Optim., 45 (2006), pp. 548–564.

[18] A. Rösch and D. Wachsmuth, *Regularity of solutions for an optimal control problem with mixed control-state constraints*, Top, 14 (2006), pp. 263–278.

[19] F. Tröltzsch, *A minimum principle and a generalized bang-bang-principle for a distributed optimal control problem with constraints on control and state*, Z. Angew. Math. Mech., 59 (1979), pp. 737–739.

[20] F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen—Theorie, Verfahren und Anwendungen*, Vieweg, Wiesbaden, Germany, 2005.

[21] F. Tröltzsch, *Regular Lagrange multipliers for control problems with mixed pointwise control-state constraints*, SIAM J. Optim., 15 (2005), pp. 616–634.

[22] W. F. Tyndall, *A duality theorem for a class of continuous linear programming problems*, J. Soc. Indust. Appl. Math., 13 (1965), pp. 644–666.

[23] K. Yosida and E. Hewitt, *Finitely additive measures*, Trans. Amer. Math. Soc., 72 (1952), pp. 46–66.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 1116–1132

PATHWISE STOCHASTIC OPTIMAL CONTROL*

L. C. G. ROGERS†

Abstract. This paper approaches optimal control problems for discrete-time controlled Markov processes by representing the value of the problem in a dual Lagrangian form; the value is expressed as an infimum over a family of Lagrangian martingales of an expectation of a pathwise supremum of the objective adjusted by the Lagrangian martingale term. This representation opens up the possibility of numerical methods based on Monte Carlo simulation, which may be advantageous in high-dimensional problems or in problems with complicated constraints.

Key words. deterministic, stochastic, dynamic programming, pathwise optimization, dual

AMS subject classifications. 90C40, 90C46, 90C39, 93E20, 93E25

DOI. 10.1137/050642885

1. Introduction.

The title of this paper refers to this: we intend to show that the solution of a stochastic optimal control problem can be characterized in terms of a *pathwise* optimization. In simple terms, this means that we can repeatedly generate sample paths, solving a *deterministic* optimization for each sample path separately, to obtain an approximation to the solution of the problem.

This approach is in contrast to the more familiar method of trying to find the value function of the problem, and the associated optimal control; the more familiar approach requires consideration of all possible future evolutions of the process at each time that a control choice is to be made. This method is well developed and generally effective, but there are certainly problems (such as the optimal control of a diffusion in high dimensions) where the approach is impractical.

The approach we follow is foreshadowed by various papers in the control literature, where the relationship between deterministic and stochastic optimal control is explored. There is, for example, the paper of Davis and Burstein [4], where the theme of optimal control of a diffusion process is considered. The tools applied, notably the use of the stochastic flow of a "null" solution to the optimal control problem, are strongly specific to that particular context, but the form of the solution, involving a pathwise optimization of the original objective modified by a Lagrangian term, invites extension. Other interesting papers around this theme are those of Rockafellar and Wets [11] and Wets [13], and that of Back and Pliska [2], who present the maximization of some concave path functional over a family of adapted processes in terms of the maximization of the same functional modified by a linear (Lagrangian) functional over the larger family of *measurable* processes. The linear functional is of course the gradient of the objective at the optimum, in some suitable sense.

Contributions [11], [13], and [2] represent the Lagrangian form of the solution in quite abstract terms. In contrast, the approach to be followed in this paper derives simple and quite explicit representations which may be the basis for effective numerical techniques. This approach does not require any convexity assumptions on the objective, unlike [11], [13], [2], and the proofs are simple and completely elementary.

*Received by the editors October 17, 2005; accepted for publication (in revised form) January 27, 2007; published electronically August 29, 2007.
http://www.siam.org/journals/sicon/46-3/64288.html

†Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK (l.c.g.rogers@statslab.cam.ac.uk).

Although our first result has the appearance of the "Lagrangian form" of the problem studied in [11], [13], [2], the subsequent results do not.

The approach of this paper develops the recent result of Rogers [12], proved independently by Haugh and Kogan [6], on Monte Carlo pricing of American options.[1] This result states the following. Given an adapted process[2] $(Z_t)_{0\le t\le T}$, the value Y_0^* at time 0 of the optimal stopping problem satisfies

$$\begin{aligned} Y_0^* &\equiv \sup_{\tau\in\mathcal{T}} EZ_\tau \\ &= \inf_{M\in\mathcal{M}_0} E\left[\sup_{0\le t\le T}(Z_t - M_t)\right], \end{aligned} \tag{1}$$

where $\mathcal{T}$ is the family of stopping times, and $\mathcal{M}_0$ is the space of uniformly integrable martingales started at 0. The importance of this result is that it gives a way to find the value of an American option via Monte Carlo simulation; given the sample path of $Z - M$, we simply stop at the best place without considering what might be happening on any other path, and in particular without considering what the value function might be at any time. The numerical methods presented in [12] are crude but good enough to get upper and lower bounds in a number of interesting examples which were different by about 0.5%–2%. Andersen and Broadie [1] present a more systematic way to search out "good" martingales and achieve bounds that are generally better. Jamshidian [7] proposes a "multiplicative" version of the result of [12], [6].

Now the optimal stopping problem is a particularly simple class of optimal control problems; *could any variant of result* (1) *be used for more general stochastic control problems?* Passing to complete generality introduces a couple of major complications: the first is that the space of possible controls is no longer a two-point set but can be very large; and the second is that the choice of controls now affects the law of the process, and there is no canonical choice. However, the main message of this paper is that we *can* extend the dual methodology that worked so well for optimal stopping problems; we present a number of different forms of the main idea. We present results only in a discrete-time setting; there are doubtless continuous-time analogues, but we prefer to present the main ideas in the technically simplest form. Our main focus is on the development of Monte Carlo methodologies that use the main ideas of this paper to solve optimal control problems. Existing techniques for solving Hamilton–Jacobi–Bellman equations by PDE methods are reasonably satisfactory provided the problem is not too involved, but it does not take much imagination to come up with examples that are so complicated that only a simulation methodology could possibly work. The different forms of the main result that we derive suggest different techniques for approaching the problem of Monte Carlo approximation of the solution. There are also links to the "occupation measure" approach to optimal control of a Markov process (which Kurtz and Stockbridge [8] trace back to Manne [10]); this we discuss in an appendix.

2. The problem and its solution. We shall consider the optimal control of a discrete-time Markov process with a finite time horizon T. The Markov process X takes values in some measurable space $(\mathcal{X}, \mathcal{G})$, and the control process $\mathbf{a} \equiv$

[1] See Davis and Karatzas [5] for a weaker partial result.

[2] The process is also required to satisfy a mild integrability condition.

$(a_0, a_1, \ldots, a_{T-1})$ belongs to the class $\mathcal{A}$ of adapted processes with values in some measurable space $(A, \mathcal{B})$ of permitted controls. The objective is

$$E\left[\sum_{j=0}^{T-1} f_j(X_j, a_j) + F(X_T)\right], \tag{2}$$

which is to be maximized over $\mathbf{a} \in \mathcal{A}$. For simplicity, we shall assume that the functions f_j and F are *bounded* measurable to avoid having to worry over finiteness of objectives and other such inessential issues; this restriction is made solely for ease of exposition. We shall suppose that there is some reference measure m over $(\mathcal{X}, \mathcal{G})$ such that for each $a \in A$ the transition under control a has density $p(x, x'; a)$ with respect to m, and that there is some reference Markovian transition density $p^*(x, x')$. We write

$$\varphi(x, x'; a) = \frac{p(x, x'; a)}{p^*(x, x')}$$

for the controlled transition density with respect to the reference Markovian transition p^*. We write $V_j(x)$ for the value function of the problem starting from state x at time j:

$$V_j(x) = \sup_{\mathbf{a}\in\mathcal{A}} E\left[\sum_{r=j}^{T-1} f_r(X_r, a_r) + F(X_T) \,\middle|\, X_j = x\right]. \tag{3}$$

We may view the effect of control as being an alteration of the law of the underlying process X. If we do this, then by introducing the notation $(0 \le k \le t < T)$

$$\Lambda_{k,t}(\mathbf{a}) \equiv \prod_{r=k}^{t-1} \varphi(X_r, X_{r+1}; a_r), \quad \Lambda_t(\mathbf{a}) \equiv \Lambda_{0,t}(\mathbf{a}), \tag{4}$$

we may recast the optimization problem in the form

$$V_0(X_0) = \sup_{\mathbf{a}\in\mathcal{A}} E^*\left[\sum_{j=0}^{T-1} \Lambda_j(\mathbf{a}) f_j(X_j, a_j) + \Lambda_T(\mathbf{a}) F(X_T)\right], \tag{5}$$

where the expectation is now taken with respect to the fixed reference probability P^*. We shall need the following notation (for (bounded) measurable $g, h_j : \mathcal{X} \mapsto \mathbb{R}$):

$$Pg(x, a) = E^*[\, g(X_1)\varphi(x, X_1; a) \mid X_0 = x], \quad (x \in \mathcal{X},\ a \in A), \tag{6}$$

$$(\mathcal{L}h)_j(x) = \sup_a\big[\, f_j(x, a) + Ph_{j+1}(x, a)\,\big], \quad (x \in \mathcal{X}, j = 0, \ldots, T-1). \tag{7}$$

The first notation is just the expectation of $g(X_1)$ if at time 0 we are in state x and use action a; the second defines the one-step Bellman operator.

The first result is the following.

THEOREM 1.

$$V_0(X_0) = \min_{(h_j)\in\mathcal{H}} E^*\left[\sup_{\mathbf{a}} \left\{ \sum_{j=0}^{T-1} \Lambda_j(\mathbf{a})\{f_j(X_j,a_j) + E_j^*(\eta_{j+1}) - \eta_{j+1}\} + \Lambda_T(\mathbf{a})F(X_T) \right\}\right] \tag{8}$$

$$= \min_{(h_j)\in\mathcal{H}} \left[h_0(X_0) + \sum_{j=0}^{T-1} E^* \sup_{\mathbf{a}} \Lambda_j(\mathbf{a}) \{ (\mathcal{L}h)_j(X_j) - h_j(X_j) \}^+ \right], \tag{9}$$

where the random variables η_j are defined in terms of the functions (h_j) via

$$\eta_{j+1} \equiv h_{j+1}(X_{j+1})\varphi(X_j, X_{j+1}; a_j), \tag{10}$$

and the set $\mathcal{H}$ is the set of sequences $(h_j)_{j=0}^T$ of (bounded) measurable functions from $\mathcal{X}$ to $\mathcal{X}$, satisfying the terminal condition

$$h_T = F.$$

Remarks. (i) To get from the form (5) to (8), we add a martingale-difference sequence $E_j^*(\eta_{j+1}) - \eta_{j+1}$ to the objective, then do a *pathwise* optimization over the controls, take expectations, and finally minimize over our choice of the martingale difference sequence. This is formally similar to what we did in (1); as there, the martingale-difference sequence can be interpreted as a Lagrangian process to account for the adapted constraint on the controls $\mathbf{a}$. Once we have included this term in the objective, we optimize pathwise, allowing ourselves to see the entire path and pick controls in an anticipative way. Notice that because of the form (10) of η_{j+1}, the conditional expectation appearing in (8) can as well be expressed as

$$E_j^*(\eta_{j+1}) = Ph_{j+1}(X_j, a_j). \tag{11}$$

(ii) As we shall see, the minimum is attained when we take $h_j = V_j$. This fact is of little practical value, however, since we cannot assume that we know V—it is, after all, the solution we seek! Nevertheless, the result allows us to obtain *upper* bounds on the value function.

(iii) The choice of reference measure must be expected to be critical in practice. We cannot expect a simulation method to work well if most of the paths simulated are quite unlike the paths of the optimally controlled process.

(iv) The form (8) is well suited to Monte Carlo, since it involves an expectation of a pathwise supremum. The second form (9) can be evaluated with no backward recursion. It can be reworked in the situation where

$$\psi(x) \equiv \int \sup_{a\in A} p(x,x';a)\, m(dx) < \infty \tag{12}$$

for all x. This allows us to define a new transition density

$$\bar{p}(x,x') = \frac{\sup_{a\in A} p(x,x';a)}{\psi(x)}$$

with a corresponding path probability $\bar{P}$. Writing $\Psi_j \equiv \prod_{i=0}^{j-1} \psi(X_i)$, the final form (9) becomes simply

$$\min_{(h_j)} \left[h_0(X_0) + \sum_{j=0}^{T-1} \bar{E}\Psi_j \{ (\mathcal{L}h)_j(X_j) - h_j(X_j) \}^+ \right]. \tag{13}$$

This is of interest because it expresses the solution in terms of a *fixed* measure, which we could call the *maximum-likelihood measure*, together with a reweighting factor which is independent of any choice of controls.

Proof. The problem is to find

$$V_0(X_0) = \sup_{\mathbf{a}\in\mathcal{A}} v_0(X_0; \mathbf{a}),$$

where of course we define

$$v_0(X_0; \mathbf{a}) \equiv E^* \left[\sum_{j=0}^{T-1} \Lambda_j(\mathbf{a}) f_j(X_j, a_j) + \Lambda_T(\mathbf{a}) F(X_T) \right].$$

Now fixing $\mathbf{a} \in \mathcal{A}$, for any P^*-martingale M,

$$\begin{aligned} v_0(X_0; \mathbf{a}) &\equiv E^* \left[\sum_{j=0}^{T-1} \Lambda_j(\mathbf{a}) f_j(X_j, a_j) + \Lambda_T(\mathbf{a}) F(X_T) \right] \\ &= E^* \left[\sum_{j=0}^{T-1} \Lambda_j(\mathbf{a}) \{ f_j(X_j, a_j) + \Delta M_{j+1} \} + \Lambda_T(\mathbf{a}) F(X_T) \right], \end{aligned}$$

since for $\mathbf{a} \in \mathcal{A}$ the process $\Lambda(\mathbf{a})$ is adapted. We shall specialize the martingale slightly by expressing the martingale-differences as

$$\Delta M_{j+1} = E^*_j(\eta_{j+1}) - \eta_{j+1}, \quad \eta_{j+1} \equiv h_{j+1}(X_{j+1}) \varphi(X_j, X_{j+1}; a_j). \tag{14}$$

Notice that

$$\Lambda_j(\mathbf{a}) \eta_{j+1} = \Lambda_{j+1}(\mathbf{a}) h_{j+1}(X_{j+1}); \tag{15}$$

this fact is used in the following reworking. The first inequality comes by relaxing the constraint that $\mathbf{a} \in \mathcal{A}$:

$$\begin{aligned} V_0(X_0) &= \sup_{\mathbf{a}\in\mathcal{A}} v_0(X_0; \mathbf{a}) \\ &= \sup_{\mathbf{a}\in\mathcal{A}} E^* \left[\sum_{j=0}^{T-1} \Lambda_j(\mathbf{a}) \{ f_j(X_j, a_j) + E^*_j(\eta_{j+1}) - \eta_{j+1} \} + \Lambda_T(\mathbf{a}) F(X_T) \right] \\ &= \sup_{\mathbf{a}\in\mathcal{A}} E^* \left[\sum_{j=0}^{T-1} \Lambda_j(\mathbf{a}) \{ f_j(X_j, a_j) + P h_{j+1}(X_j, a_j) - \eta_{j+1} \} + \Lambda_T(\mathbf{a}) F(X_T) \right] \end{aligned}$$

$$\leq E^*\left[\sup_{\mathbf{a}}\left\{\sum_{j=0}^{T-1}\Lambda_j(\mathbf{a})\{f_j(X_j,a_j)+Ph_{j+1}(X_j,a_j)-\eta_{j+1}\}\right.\right.$$
$$\left.\left.+\Lambda_T(\mathbf{a})F(X_T)\right\}\right]$$
$$= E^*\left[\sup_{\mathbf{a}}\left\{\sum_{j=0}^{T-1}\left\{\Lambda_j(\mathbf{a})\{f_j(X_j,a_j)+Ph_{j+1}(X_j,a_j)\}-\Lambda_{j+1}(\mathbf{a})h_{j+1}(X_{j+1})\right\}\right.\right.$$
$$\left.\left.+\Lambda_T(\mathbf{a})F(X_T)\right\}\right]$$
$$= E^*\left[\sup_{\mathbf{a}}\left\{h_0(X_0)+\sum_{j=0}^{T-1}\Lambda_j(\mathbf{a})\{f_j(X_j,a_j)+Ph_{j+1}(X_j,a_j)-h_j(X_j)\}\right\}\right]$$
$$\leq E^*\left[h_0(X_0)+\sum_{j=0}^{T-1}\sup_{\mathbf{a}}\Lambda_j(\mathbf{a})\{f_j(X_j,a_j)+Ph_{j+1}(X_j,a_j)-h_j(X_j)\}\right]$$
$$= h_0(X_0)+\sum_{j=0}^{T-1}E^*\left[\sup_{\mathbf{a}}\Lambda_j(\mathbf{a})\{\,(\mathcal{L}h)_j(X_j)-h_j(X_j)\,\}\right]$$
$$\leq h_0(X_0)+\sum_{j=0}^{T-1}E^*\left[\sup_{\mathbf{a}}\Lambda_j(\mathbf{a})\{\,(\mathcal{L}h)_j(X_j)-h_j(X_j)\,\}^+\right].$$

Taking the infimum over the functions $(h_j)\in\mathcal{H}$, we get

$$V_0(X_0)\leq\inf_{(h_j)\in\mathcal{H}}E^*\left[\sup_{\mathbf{a}}\left\{\sum_{j=0}^{T-1}\Lambda_j(\mathbf{a})\{f_j(X_j,a_j)-\eta_{j+1}+E_j^*(\eta_{j+1})\}\right.\right.$$
$$\left.\left.+\Lambda_T(\mathbf{a})F(X_T)\right\}\right]$$
$$\leq\inf_{(h_j)\in\mathcal{H}}\left[h_0(X_0)+\sum_{j=0}^{T-1}E^*\left[\sup_{\mathbf{a}}\Lambda_j(\mathbf{a})\{\,(\mathcal{L}h)_j(X_j)-h_j(X_j)\,\}\right]\right]$$
$$\leq\inf_{(h_j)\in\mathcal{H}}\left[h_0(X_0)+\sum_{j=0}^{T-1}E^*\left[\sup_{\mathbf{a}}\Lambda_j(\mathbf{a})\{\,(\mathcal{L}h)_j(X_j)-h_j(X_j)\,\}^+\right]\right]. \tag{16}$$

In fact, there is equality throughout. To see this, we use the Bellman equation for the value function

$$V_j=(\mathcal{L}V)_j,$$

so if we take $h_j=V_j$, the sum in (16) vanishes and leaves only $h_0(X_0)=V_0(X_0)$. □

Remark. The proof also shows that

$$V_0(x_0) = \min_{(h_j)\in\mathcal{H}} \left[h_0(X_0) + \sum_{j=0}^{T-1} E^* \sup_{\mathbf{a}} \Lambda_j(\mathbf{a}) \{ (\mathcal{L}h)_j(X_j) - h_j(X_j) \} \right], \tag{17}$$

which is a fact that we will refer back to later.

Theorem 1 gives us a way to approach a stochastic optimal control problem by Monte Carlo methods, by simulating paths repeatedly and computing the expressions inside the expectations (8). However, it is important that this optimization, over the sequence $\mathbf{a}$, can be done efficiently; otherwise the method will be too slow. Fortunately, it turns out that the optimization required may be performed *recursively*, so we have a sequence of optimization problems over the choice of only one a_j at a time.

To explain this in more detail, let us focus on the form (8). We can rewrite the expression inside the expectation on the right-hand side as

$$\begin{aligned} &\sum_{j=0}^{T-1} \Lambda_j(\mathbf{a})\{f_j(X_j,a_j) + Ph_{j+1}(X_j,a_j) - \eta_{j+1}\} + \Lambda_T(\mathbf{a})F(X_T) \\ &= \sum_{j=0}^{m-1} \Lambda_j(\mathbf{a})\{f_j(X_j,a_j) + Ph_{j+1}(X_j,a_j) - \eta_{j+1}\} + \Lambda_m(\mathbf{a})Z_m, \end{aligned}$$

where

$$Z_m \equiv \sum_{j=m}^{T-1} \Lambda_{m,j}(\mathbf{a})\{f_j(X_j,a_j) + Ph_{j+1}(X_j,a_j) - \eta_{j+1}\} + \Lambda_{m,T}(\mathbf{a})F(X_T)$$

contains all dependence on $a_m, \dots, a_{T-1}$. Recursively,

$$\begin{aligned} Z_m &= f_m(X_m,a_m) + Ph_{m+1}(X_m,a_m) - \eta_{m+1} + \Lambda_{m,m+1}(\mathbf{a})Z_{m+1} \\ &= f_m(X_m,a_m) + Ph_{m+1}(X_m,a_m) + \varphi(X_m,X_{m+1};a_m)\big[\, Z_{m+1} - h_{m+1}(X_{m+1})\,\big]. \end{aligned}$$

Assuming we already have the maximizing values of $a_{m+1}, \dots, a_{T-1}$, this is a maximization over a_m only!

3. Towards an algorithm. It is clear from the statement of Theorem 1 that the choice of the Lagrangian functions (h_j) is critical. The following little result offers a possible approach to finding good choices.

PROPOSITION 1. *Suppose that*

$$B \equiv \sup_{a,x,x'} \varphi(x,x';a) < \infty,$$

and suppose we are given a sequence $(V_j^{(0)})_{j=0}^T$ *of functions from* $\mathcal{X}$ *to* $\mathcal{X}$*, with* $V_T^{(0)} = F$*. Define recursively the functions* $(V_k^{(n)})_{k=0}^T$ *for* $n = 1, 2, \dots$ *by*

$$V_k^{(n+1)}(x) = E^* \left[\sup_{\mathbf{a}} \left\{ \sum_{j=k}^{T-1} \Lambda_{k,j}(\mathbf{a})\{f_j(X_j,a_j) - V_{j+1}^{(n)}(X_{j+1})\varphi(X_j,X_{j+1};a_j) + PV_{j+1}^{(n)}(X_j,a_j)\} + \Lambda_{k,T}(\mathbf{a})F(X_T) \right\} \,\middle|\, X_k = x \right] \tag{18}$$

for $x \in \mathcal{X}$, $k = 0, \ldots, T$. Defining

$$\Delta_k^{(n)} \equiv \sup_x |V_k^{(n)}(x) - V_k^{(n-1)}(x)|,$$

$k = 0, \ldots, T$, $n \geq 1$, we have

$$\Delta_k^{(n)} \leq (1+B) \sum_{r=k+1}^{T} \Delta_r^{(n-1)}. \tag{19}$$

Remarks. The impact of Proposition 1 lies in the fact that $V_T^{(n)} = F$ for all n, so $\Delta_T^{(n)} = 0$ for all n. Hence from (19) we conclude that (provided that the $\Delta_k^{(n-1)}$ are finite)

$$\Delta_k^{(n)} = 0 \quad \forall n \geq T - k.$$

Thus by applying the recursive construction of Proposition 1 we compute the true value function by going backwards step by step from the end. Now in one sense all we have done is re-express the familiar backward recursion of the Bellman equation in a more complicated form, but there is nevertheless something gained; if we are not able to compute the recursive recipe (18) exactly (as would be the case were we using Monte Carlo in a high-dimensional problem, for example), we can still use the *approximate* output of the nth stage to begin on the $(n+1)$th.

Proof. Clearly,

$$\begin{aligned} -V_{j+1}^{(n)}(X_{j+1})\varphi(X_j, X_{j+1}; a_j) &\leq -V_{j+1}^{(n-1)}(X_{j+1})\varphi(X_j, X_{j+1}; a_j) \\ &\quad + \Delta_{j+1}^{(n)}\varphi(X_j, X_{j+1}; a_j) \\ &\leq -V_{j+1}^{(n-1)}(X_{j+1})\varphi(X_j, X_{j+1}; a_j) + B\Delta_{j+1}^{(n)} \end{aligned}$$

and

$$PV_{j+1}^{(n)}(X_j, a_j) \leq PV_{j+1}^{(n-1)}(X_j, a_j) + \Delta_{j+1}^{(n)},$$

so using this in (18) gives us

$$\begin{aligned} V_k^{(n+1)}(x) &\equiv E^* \Bigg[\sup_{\mathbf{a}} \Bigg\{ \sum_{j=k}^{T-1} \Lambda_{k,j}(\mathbf{a}) \{ f_j(X_j, a_j) - V_{j+1}^{(n)}(X_{j+1})\varphi(X_j, X_{j+1}; a_j) \\ &\qquad + PV_{j+1}^{(n)}(X_j, a_j) \} + \Lambda_{k,T}(\mathbf{a}) F(X_T) \Bigg\} \,\Bigg|\, X_k = x \Bigg] \\ &\leq E^* \Bigg[\sup_{\mathbf{a}} \Bigg\{ \sum_{j=k}^{T-1} \Lambda_{k,j}(\mathbf{a}) \{ f_j(X_j, a_j) - V_{j+1}^{(n-1)}(X_{j+1})\varphi(X_j, X_{j+1}; a_j) \\ &\qquad + PV_{j+1}^{(n-1)}(X_j, a_j) \} + \Lambda_{k,T}(\mathbf{a}) F(X_T) \Bigg\} \,\Bigg|\, X_k = x \Bigg] \\ &\qquad + (1+B) \sum_{r=k+1}^{T} \Delta_r^{(n)} \\ &= V_k^{(n)}(x) + (1+B) \sum_{r=k+1}^{T} \Delta_r^{(n)}. \end{aligned}$$

Thus

$$V_k^{(n+1)}(x) - V_k^{(n)}(x) \le (1+B) \sum_{r=k+1}^{T} \Delta_r^{(n)},$$

and a similar bound on the other side establishes the result. □

Discussion. For the purposes of this discussion, we assume for ease of exposition that $f_j = f$ for all j, and that there exists a sequence of functions ψ_k such that the integral $P\psi_k(x,a)$ is known in closed form. The reason for this is to permit approximation of the value function as linear combinations of the ψ_k; this is similar to what Longstaff and Schwartz [9] do.

When might we use this approach? When the steps of the dynamic programming algorithm are numerically intensive as, for example, in a situation where $\mathcal{X}$ is very high dimensional and the required integrations are difficult to do, or when the pointwise optimization over $a \in A$ is hard, then the simulation-based approach of Theorem 1 may be of value. One advantage of this approach is that it seeks only the solution starting from a particular x_0, whereas the dynamic programming approach is calculating the solution from *all* starting points.

The first thing to do will be to simulate some paths of the process.

What law should we use for the initial simulation? Probably we should not use the reference Markovian law P^*, as the paths of X under P^* can't be expected to look very much like the paths of the optimally controlled process, and so we will get little relevant information about the objective if we just simulate from P^*. This problem becomes more acute the larger T is, so it may be worth simulating initially only out to some $T_1 < T$, and gradually increasing T_1 as the algorithm proceeds. Since the intermediate rewards f_j could all be zero (or very small), we should not forget to include a term $F(X_{T_1})$ in the objective, as we will ultimately be steering towards this. The *maximum likelihood measure* $\bar{P}$ is also not a very promising candidate, as the law does not depend in any way on f_j, F, but it suggests something we might try instead. Since the objective is

$$v_0(X_0;\mathbf{a}) \equiv E^*\left[\sum_{j=0}^{T-1} \Lambda_j(\mathbf{a}) f(X_j,a_j) + \Lambda_T(\mathbf{a}) F(X_T)\right]$$

$$= E^*\left[\Lambda_T(\mathbf{a})\ \left\{\sum_{j=0}^{T-1} f(X_j,a_j) + F(X_T)\right\}\right]$$

$$\simeq \varepsilon^{-1} E^*\left[\Lambda_T(\mathbf{a}) \exp\left\{\varepsilon \sum_{j=0}^{T-1} f(X_j,a_j) + \varepsilon F(X_T)\right\} - 1\right]$$

$$= \varepsilon^{-1} E^*\left[\prod_{j=0}^{T-1} \varphi(X_j, X_{j+1}; a_j) e^{\varepsilon f(X_j,a_j)}\ .e^{\varepsilon F(X_T)} - 1\right],$$

this suggests we might modify the definition of $\bar{P}$ by defining

$$\psi(x) \equiv \int \sup_{a \in A} p(x, x'; a) e^{\varepsilon f(x,a)} \, m(dx),$$

$$\bar{p}(x, x') = \frac{\sup_{a \in A} p(x, x'; a) e^{\varepsilon f(x,a)}}{\psi(x)}.$$

The effect of this is to lead the process in directions where the running reward is higher. The choice of ε will need to be tuned a bit.

How do we move from one simulation to the next? We suppose that our current estimate $V_t^{(n)}$ of the value is expressed as a linear combination of the ψ_k:

$$V_t^{(n)} = \sum_k c_{t,k}^{(n)} \psi_k,$$

which allows us to write expressions for $PV_t^n(x, a)$. Once we have simulated sample paths $(X_0^{(i)}, X_1^{(i)}, \ldots, X_T^{(i)})$ for $i = 1, \ldots, N$, we perform the pathwise optimization in (8) and then have some estimate of the value at the points $X_t^{(i)}$ at time t. We now regress these values onto the functions ψ_k to get the next approximation to the value. The next simulation should follow according to what we now think is an approximation to the optimal path law, and one way to achieve this would be as follows. Suppose that at time t on the simulated path we have reached x; first choose $x' \in \{X_t^{(i)}, i = 1, \ldots, N\}$ at random, points "nearer" to x being chosen with higher probability, and jump to that point, $x' = X_t^{(q)}$, say. Then make the move to y at time $t + 1$ according to the density $p(x', \cdot; a_t^{(q)})$, where $a_t^{(q)}$ was the control optimally chosen at $X_t^{(q)}$.

4. Variants of the main result.

4.1. Least-squares characterization. The study [12] of Monte Carlo valuation of American options showed that the optimal policy was in some sense a "minimum-variance" policy, and there is an analogue in this setting too. Writing

$$Y(X; h) \equiv \sup_{\mathbf{a}} \left\{ \sum_{j=0}^{T-1} \Lambda_j(\mathbf{a}) \{ f_j(X_j, a_j) - \eta_{j+1} + E_j^*(\eta_{j+1}) \} + \Lambda_T(\mathbf{a}) F(X_T) \right\}$$

(where the η_j are as in (10)), Theorem 1 says that $V(X_0) = \inf_{(h_j)} E^*[\, Y(X; h)\,]$. Moreover, the infimum is attained by taking $h_j = V_j$, and in that case the proof of Theorem 1 shows that the random variable $Y(X; V)$ is almost surely constant. We therefore have the following alternative characterization of the optimal solution.

COROLLARY 1. *Assuming that V_0 is nonnegative,*[3] *the problem*

$$\inf_{(h_j) \in \mathcal{H}} E^*[\, Y(X; h)^2\,]$$

is solved by taking $h_j = V_j$.

[3]Nonnegativity is needed only because we use the reasoning $E^* Y(X; h)^2 = \operatorname{var}(Y(X; h)) + E^*(Y(X; h))^2 \geq E^*(Y(X; h))^2 \geq (\min E^* Y(X; h))^2$, and the final step is not true unless we have $E^* Y(X; h) \geq 0$ for all h.

4.2. Multiplicative form of the main result. As in the case of Jamshidian's version of the optimal stopping result, we have a multiplicative form of Theorem 1.

THEOREM 2.

$$V_0(X_0) \le \inf_{\eta>0} E^* \left[\sup_{\mathbf{a}} \left\{ \sum_{j=0}^{T-1} \Lambda_j(\mathbf{a})\, f_j(X_j, a_j)\, \frac{\eta_{j+1}}{E_j^*[\eta_{j+1}]} + \Lambda_T(\mathbf{a}) F(X_T) \right\} \right], \tag{20}$$

where the random variables η_j are positive. Provided

$$g_j^*(X_j, X_{j+1}, a_j) \equiv V_j(X_j) - V_{j+1}(X_{j+1})\varphi(X_j, X_{j+1}; a) > 0, \tag{21}$$

result (20) *can be strengthened into the statement*

$$V_0(X_0) = \min_{\eta>0} E^* \left[\sup_{\mathbf{a}} \left\{ \sum_{j=0}^{T-1} \Lambda_j(\mathbf{a})\, f_j(X_j, a_j)\, \frac{\eta_{j+1}}{E_j^*[\eta_{j+1}]} + \Lambda_T(\mathbf{a}) F(X_T) \right\} \right], \tag{22}$$

with the minimizing choice of η_{j+1} being $\eta_{j+1} = g_j^*(X_j, X_{j+1}, a_j)$.

Remark. Condition (21) could be weakened to nonnegativity; we simply need to change f_j to $f_j - j$ and apply the theorem to this modified problem (whose value is $T(T-1)/2$ less than the value of the original problem).

Proof. The proof follows similarly to the proof of Theorem 1. Fixing $\mathbf{a} \in \mathcal{A}$, and letting η be any strictly positive adapted process, we have

$$\begin{aligned} v_0(X_0; \mathbf{a}) &= E^* \left[\sum_{j=0}^{T-1} \Lambda_j(\mathbf{a}) f_j(X_j, a_j) + \Lambda_T(\mathbf{a}) F(X_T) \right] \\ &= E^* \left[\sum_{j=0}^{T-1} \Lambda_j(\mathbf{a})\, f_j(X_j, a_j)\, \frac{\eta_{j+1}}{E_j^*[\eta_{j+1}]} + \Lambda_T(\mathbf{a}) F(X_T) \right]. \end{aligned}$$

Just as before,

$$\begin{aligned} V_0(X_0) &= \sup_{\mathbf{a}\in\mathcal{A}} v_0(X_0; a) \\ &= \sup_{\mathbf{a}\in\mathcal{A}} E^* \left[\sum_{j=0}^{T-1} \Lambda_j(\mathbf{a})\, f_j(X_j, a_j)\, \frac{\eta_{j+1}}{E_j^*[\eta_{j+1}]} + \Lambda_T(\mathbf{a}) F(X_T) \right] \\ &\le E^* \left[\sup_{\mathbf{a}} \left\{ \sum_{j=0}^{T-1} \Lambda_j(\mathbf{a})\, f_j(X_j, a_j)\, \frac{\eta_{j+1}}{E_j^*[\eta_{j+1}]} + \Lambda_T(\mathbf{a}) F(X_T) \right\} \right]. \end{aligned}$$

Taking the infimum over all choices of η leads to the first statement, (20).

For the second statement, (22), we again use the Bellman equation; positivity of g_j^* allows us to conclude that

$$\frac{f_j(X_j, a_j)}{E_j^*[\eta_{j+1}]}\, \eta_{j+1} \le \eta_{j+1},$$

and once again the sum telescopes to $V_0(X_0)$. □

4.3. "Strong" form of the main result. In Theorems 1 and 2, the effect of the controls is to modify the measure; if we simulate paths according to the measure P^*, then the controls applied do not affect the path of X—they simply affect the value assigned to the path. It may sometimes be more helpful to be able to allow the controls to act on the path directly, for which we need to formulate the problem slightly differently.

We shall suppose that if some control sequence $(a_j)_{j=0}^{T-1}$ is chosen, and the initial value X_0 for the process is given, then the trajectory X is determined by the relations

$$X_{j+1} = \xi(j, X_j, a_j, \varepsilon_{j+1}), \qquad (j = 0, \dots, T-1), \tag{23}$$

where the ε_j are independent random variables with common distribution, which we could take to be uniform on $[0,1]$ if we wish. The function ξ expresses the Markovian evolution; from a theoretical point of view it may be a little unusual to specify a Markov process in this way, rather than through the transition kernel, but from the point of view of simulating the paths of the process, this is *exactly* the way we think of the controlled Markov process! The difference is exactly the difference between a strong solution of a stochastic differential equation, constructed over a given driving process, and a weak solution, constructed in law on some probability space (as in Theorems 1 and 2).

Given a sequence (h_j) of functions of the Markovian state variable, we define

$$Ph_{j+1}(x,a) = E\, h_{j+1}(\xi(j,x,a,\varepsilon_{j+1})).$$

Then we have the following result.

THEOREM 3.

$$V_0(X_0) = \min_{(h_j)\in\mathcal{H}} E\left[\sup_{\mathbf{a}} \left\{ \sum_{j=0}^{T-1} (f_j(X_j,a_j) - h_{j+1}(X_{j+1}) + Ph_{j+1}(X_j,a_j)) + F(X_T) \right\} \right], \tag{24}$$

where the X_j and a_j are related through (23). *The minimum is attained by taking $h_j = V_j$.*

Remarks. The Monte Carlo approach to evaluating the right-hand side of (24) would generate a sequence of ε values and then find the optimal controls. In effect, what this means is that we have to solve a deterministic optimization problem along each path, where the choice of control will now affect where the path goes, and doing this is arguably no easier than solving the Bellman equation for the original stochastic control problem. However, in situations where this deterministic control problem can be dealt with more simply, there may be value in this result.

Proof. This closely follows along the lines of the proof of Theorem 1; we leave this to the reader to check. □

4.4. Infinite horizon. So far we have been considering only finite-horizon problems, but it is at least as important to develop methods for infinite-horizon discounted problems, as these will generate time-independent strategies that are easier to interpret and implement. Throughout this section, we will assume that f is uniformly bounded, and that we aim to find the value function $V : \mathcal{X} \to \mathcal{X}$ solving

$$V(x) = \sup_a E^* \left[f(x,a) + \beta\varphi(x, X_1; a)V(X_1) \;\middle|\; X_0 = x \right]. \tag{25}$$

Under the assumptions that $0 < \beta < 1$ and that f is uniformly bounded, it is well known that the Bellman operator $\mathcal{L} : L^\infty(\mathcal{X}) \to L^\infty(\mathcal{X})$ defined by

$$\mathcal{L}g(x) \equiv \sup_{a \in \mathcal{A}} E^* \left[f(x,a) + \beta\varphi(x, X_1; a) g(X_1) \;\middle|\; X_0 = x \right] \tag{26}$$

is a monotone contraction whose unique fixed point is the value function V solving (25).

To see where the dual method leads in this infinite-horizon setting, we need to introduce for each $h \in L^\infty(\mathcal{X})$ the operator $\mathcal{L}_h : L^\infty(\mathcal{X}) \to L^\infty(\mathcal{X})$ defined by

$$\begin{aligned} \mathcal{L}_h g(x) \equiv E^* \Big[\sup_a \big\{ & f(x,a) + Ph(x,a) - h(X_1)\varphi(x, X_1; a) \\ & + \beta\varphi(x, X_1; a) g(X_1) \big\} \;\Big|\; X_0 = x \Big]. \end{aligned} \tag{27}$$

Just as for $\mathcal{L}$, the operator $\mathcal{L}_h$ is a monotone contraction with a unique fixed point, which we denote by g_h^*. The analogue of Theorem 1 for the infinite-horizon setting is the following.

THEOREM 4. *Assuming that f is uniformly bounded, the value function V is characterized as*

$$V = \inf_h \; g_h^* = \min_h \; g_h^*, \tag{28}$$

where the infimum is attained by taking $h = \beta V$.

Proof. Evidently, the supremum in the definition of $\mathcal{L}_h g$ will be reduced if we insist that a must be a function of only X_0 and not of X_1; therefore

$$\begin{aligned} \mathcal{L}_h g(x) &\geq \sup_a E^* \Big[f(x,a) + Ph(x,a) - h(X_1)\varphi(x, X_1; a) \\ &\qquad + \beta\varphi(x, X_1; a) g(X_1) \;\Big|\; X_0 = x \Big] \\ &= \sup_a E^* \Big[f(x,a) + \beta\varphi(x, X_1; a) g(X_1) \;\Big|\; X_0 = x \Big] \\ &\equiv \mathcal{L}g(x). \end{aligned}$$

Since $\mathcal{L}V = V$, we deduce immediately that whatever h is, we shall have $\mathcal{L}_h V \geq V$, and by induction we conclude that for all n,

$$\mathcal{L}_h^n V \geq V.$$

By the contraction mapping principle, $\mathcal{L}_h^n V \to g_h^*$ as $n \to \infty$, and so for any h we have $g_h^* \geq V$, and hence $V \leq \inf_h g_h^*$.

To conclude, we observe that taking $h = \beta V$ gives for any x, a,

$$f(x,a) + Ph(x,a) \leq \sup_{a'} \{ f(x,a') + Ph(x,a') \} = V(x).$$

Hence,

$$\mathcal{L}_h V(x) \equiv E^*\Bigg[\sup_a \big\{ f(x,a) - h(X_1)\varphi(x,X_1;a) + Ph(x,a) + \beta\varphi(x,X_1;a)V(X_1)\big\} \;\Big|\; X_0 = x \Bigg]$$

$$\leq V(x) + E^*\Bigg[\sup_a \big\{ -h(X_1)\varphi(x,X_1;a) + \beta\varphi(x,X_1;a)V(X_1)\big\} \;\Big|\; X_0 = x \Bigg]$$

$$= V(x).$$

By induction, $\mathcal{L}_h^n V \leq V$, and so taking the limit as $n \to \infty$ leads to the conclusion that $g_h^* \leq V$. □

As in the finite-horizon case, we can ask about possible recursive methods for generating a better approximation to the solution from an existing one. The following result, proved only under rather restrictive conditions, shows that something can be done.

PROPOSITION 2. *Suppose that f is uniformly bounded, that*

$$B \equiv \sup_{x,x',a} \varphi(x,x';a) < \infty,$$

and that β is so small that

$$\frac{\beta(1+B)}{1-\beta B} < 1.$$

Then the sequence $(g_n)_{n=0}^{\infty}$ generated by taking an arbitrary $g_0 \in L^\infty(\mathcal{X})$ and letting g_{n+1} be the unique fixed point of $\mathcal{L}_{\beta g_n}$ converges to the value function.

Proof. The relation linking g_{n+1} and g_n can be expressed as

$$g_{n+1}(x) = E^*\Bigg[\sup_a \big\{ f(x,a) - \beta g_n(X_1)(X_1)\varphi(x,X_1;a) + \beta P g_n(x,a) + \beta\varphi(x,X_1;a)g_{n+1}(X_1)\big\} \;\Big|\; X_0 = x \Bigg].$$

If we set $\Delta_n \equiv \sup_x |g_n(x) - g_{n-1}(x)|$, then this leads to

$$g_{n+1}(x) \leq E^*\Bigg[\sup_a \big\{ f(x,a) - \beta g_{n-1}(X_1)(X_1)\varphi(x,X_1;a) + \beta P g_{n-1}(x,a) + \beta(1+B)\Delta_n + \beta\varphi(x,X_1;a)g_{n+1}(X_1)\big\} \;\Big|\; X_0 = x \Bigg],$$

so if we set $\tilde{g}_{n+1} \equiv g_{n+1} + A$, we have

$$\tilde{g}_{n+1}(x) + A \le E^*\Bigg[\sup_a \big\{ f(x,a) - \beta g_{n-1}(X_1)(X_1)\varphi(x,X_1;a) + \beta P g_{n-1}(x,a) + \beta(1+B)\Delta_n + \beta\varphi(x,X_1;a)(\tilde{g}_{n+1}(X_1) + A)\big\} \;\Big|\; X_0 = x\Bigg]$$

$$\le E^*\Bigg[\sup_a \big\{ f(x,a) - \beta g_{n-1}(X_1)(X_1)\varphi(x,X_1;a) + \beta P g_{n-1}(x,a) + \beta(1+B)\Delta_n + \beta BA + \beta\varphi(x,X_1;a)\tilde{g}_{n+1}(X_1)\big\} \;\Big|\; X_0 = x\Bigg].$$

Taking

$$A \equiv \frac{\beta(1+B)\Delta_n}{1-\beta B}$$

gives us

$$\tilde{g}_{n+1}(x) \le E^*\Bigg[\sup_a \big\{ f(x,a) - \beta g_{n-1}(X_1)(X_1)\varphi(x,X_1;a) + \beta P g_{n-1}(x,a) + \beta\varphi(x,X_1;a)\tilde{g}_{n+1}(X_1)\big\} \;\Big|\; X_0 = x\Bigg],$$

from which we conclude that $\tilde{g}_{n+1} \equiv g_{n+1} - A \le g_n$. A similar argument for the lower bound gives

$$\Delta_{n+1} \le \frac{\beta(1+B)}{1-\beta B}\,\Delta_n,$$

and the result follows. □

Remarks. Proposition 2 shows how we may recursively construct approximations to the solution using this methodology, provided the discount factor β is small enough. The assumptions of Proposition 2 are unlikely to be satisfied in most applications, but at least the methodology can be tried; the conditions are sufficient but not necessary!

5. Conclusions. This paper has presented a novel strategy for solving stochastic optimal control problems by using duality ideas. This approach is completely general, but is particularly well suited to problems where the state space is so large that it is hard to determine where the value function should be closely approximated. The methodology involves modifying the objective by adding in appropriate martingale differences and then carrying out a *pathwise* optimization, an approach that is well suited to Monte Carlo evaluation. We have shown that under suitable regularity conditions, a recursive method for improving the martingale difference sequence converges to the true solution.

Choosing the martingale difference sequence well is of course key to the success of the method, but there remain important issues in performing the simulations and related calculations in an efficient manner. The whole study of numerical implementation has barely begun.

Appendix A. Links to the occupation measure approach. The approach of Theorem 1 is in some sense a dual approach, but how is it related to another dual approach, the occupation measure approach, as explained and studied in [10], [3], [8]? In this approach, the original optimization problem is re-expressed as

$$\sup_{(\mu_t),(\kappa_t)} \sum_{t=0}^{T} \int_{\mathcal{X}} \mu_t(dx) \int_A \kappa_t(x,da) f_t(x,a) \tag{A1}$$

subject to the constraints

$$\mu_0(dx) = \delta_{x_0}(dx), \tag{A2}$$

$$\mu_{t+1}(dx) = \int_{\mathcal{X}} \mu_t(dx') \int_A \kappa_t(x',da) p(x',x;a) m(dx), \quad (t = 0,\dots,T-1), \tag{A3}$$

where we write $f_T(x,a) \equiv F(x)$, and each of the measures μ_t is a probability measure, and κ_t is a Markov kernel from $\mathcal{X}$ into A for each t.

The interpretation of this is that μ_t is the law of the controlled process at time t under controls given by the Markov kernels κ_t; frequently, the Markov kernels will be degenerate, in the sense that $\kappa_t(x,da) = \delta_{\alpha(t,x)}(da)$ for all t,x, but this formulation allows randomized decision rules also.

Introducing Lagrangian multiplier functions $v_t : \mathcal{X} \to \mathbb{R}$ for each $t = 0,\dots,T$ changes the optimization problem into the Lagrangian form

$$\begin{aligned}
&\sup_{\mu_t,\kappa_t \geq 0} \Bigg[v_0(x_0) + \sum_{t=0}^{T-1} \int_{\mathcal{X}} \mu_t(dx) \Bigg\{ -v_t(x) + \int_A \kappa_t(x,da) f_t(x,a) \\
&\qquad + \int_A \kappa_t(x,da) P v_{t+1}(x,a) \Bigg\} + \int_{\mathcal{X}} \mu_T(dx) \int_A \kappa_T(x,da) \{F(x) - v_T(x)\} \Bigg] \\
&= \sup_{\mu_t,\kappa_t \geq 0} \Bigg[v_0(x_0) + \sum_{t=0}^{T-1} \int_{\mathcal{X}} \mu_t(dx) \int_A \kappa_t(x,da) \big\{ -v_t(x) + f_t(x,a) + P v_{t+1}(x,a) \big\} \\
&\qquad + \int_{\mathcal{X}} \mu_T(dx) \{F(x) - v_T(x)\} \Bigg].
\end{aligned}$$

We deduce the dual-feasibility conditions

$$v_t(x) \geq f_t(x,a) + P v_{t+1}(x,a) \quad (x \in \mathcal{X}, a \in A, t = 0,\dots,T-1), \tag{A4}$$

$$v_T(x) \geq F(x), \tag{A5}$$

and the dual problem is now to minimize $v_0(x_0)$ subject to (A4), (A5). These conditions are obviously equivalent to

$$v_t(x) \geq \sup_{a \in A} \big\{ f_t(x,a) + P v_{t+1}(x,a) \big\} \equiv (\mathcal{L}v)_t(x) \quad (x \in \mathcal{X}, t = 0,\dots,T-1), \tag{A6}$$

$$v_T(x) \geq F(x), \tag{A7}$$

which are solved by taking $v_T = F$, and $v_t = (\mathcal{L}v)_t$ for $0 \leq t < T$—the Bellman equations. The value of the dual problem is also evidently equal to the value of the

primal problem. However, it will often be the case that the operations involved in the Bellman equations (taking expectations; pointwise maximization) will be hard to do numerically, so casting the dual problem in Lagrangian form gives us

$$L(\{g_t\}) \equiv \inf_{(v_t)} \left\{ v_0(x_0) + \sum_{t=0}^{T-1} \int g_t(x)\{(\mathcal{L}v)_t(x) - v_t(x)\}\, m(dx) \right\}$$

for nonnegative multiplier functions (g_t). The dual form of this programming problem is

$$\sup_{g_t \geq 0} L(\{g_t\}) \leq V_0(x_0),$$

and (9) is the same expression, for a particular choice of the multipliers (g_t), attaining the value.

Acknowledgments. The author thanks participants at the Isaac Newton Institute program Developments in Quantitative Finance 2005, and at the Cambridge Finance Seminar, for helpful discussions and comments; in particular, Mark Broadie, Mark Davis, Michael Dempster, Paul Glasserman, David Hodge, Stan Pliska, Jose Scheinkman, Nizar Touzi, Richard Weber, and Peter Whittle.

REFERENCES

[1] L. Andersen and M. Broadie, *A primal-dual simulation algorithm for pricing multi-dimensional American options*, Management Sci., 50 (2004), pp. 1222–1234.

[2] K. Back and S. R. Pliska, *The shadow price of information in continuous time decision problems*, Stochastics, 22 (1987), pp. 151–186.

[3] A. G. Bhatt and V. S. Borkar, *Occupation measures for controlled Markov processes: Characterization and optimality*, Ann. Probab., 24 (1996), pp. 1531–1562.

[4] M. H. A. Davis and G. Burstein, *A deterministic approach to stochastic optimal control with application to anticipative optimal control*, Stochastics Stochastics Rep., 40 (1992), pp. 203–256.

[5] M. H. A. Davis and I. Karatzas, *A deterministic approach to optimal stopping, with applications*, in Probability, Statistics and Optimisation: A Tribute to Peter Whittle, F. P. Kelly, ed., Wiley, New York, Chichester, 1994, pp. 455–466.

[6] M. Haugh and L. Kogan, *Pricing American options: A duality approach*, Oper. Res., 52 (2004), pp. 258–270.

[7] F. Jamshidian, *Numeraire-Invariant Option P and American, Bermudan and Trigger Stream Rollover*, Tech. report, University of Twente, Twente, The Netherlands, 2004.

[8] T. G. Kurtz and R. H. Stockbridge, *Existence of Markov controls and characterization of optimal Markov controls*, SIAM J. Control Optim., 36 (1998), pp. 609–653.

[9] F. A. Longstaff and E. A. Schwartz, *Valuing American options by simulation: A simple least-squares approach*, Rev. Financial Stud., 14 (2001), pp. 113–147.

[10] A. S. Manne, *Linear programming and sequential decisions*, Management Sci., 6 (1960), pp. 259–267.

[11] R. T. Rockafellar and R. J. B. Wets, *Nonanticipativity and L^1 martingales in stochastic optimization problems*, Math. Programming Stud., 6 (1976), pp. 170–187.

[12] L. C. G. Rogers, *Monte Carlo valuation of American options*, Math. Finance, 12 (2002), pp. 271–286.

[13] R. J. B. Wets, *On the relation between stochastic and deterministic optimization*, in Control Theory, Numerical Methods and Computer Systems Modelling, Lecture Notes in Econom. and Math. Systems 107, Springer, Berlin, 1975, pp. 350–361.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 3, pp. 1133–1153

ANALYSIS OF THE SQP-METHOD FOR OPTIMAL CONTROL PROBLEMS GOVERNED BY THE NONSTATIONARY NAVIER–STOKES EQUATIONS BASED ON L^p-THEORY*

DANIEL WACHSMUTH†

Abstract. The aim of this article is to present a refined convergence theory of the SQP-method applied to optimal control problems for the nonstationary Navier–Stokes equations. We will employ a second-order sufficient optimality condition, which requires that the second derivative of the Lagrangian is positive definite on a subspace of inactive constraints. Therefore, we have to use the L^p-theory of optimal controls of the nonstationary Navier–Stokes equations rather than Hilbert space methods. Estimates of state and adjoint equations with respect to L^p-norms are provided. Finally, the local convergence of the SQP-method is confirmed by numerical tests.

Key words. optimal control, Navier–Stokes equations, control constraints, Lipschitz stability, SQP-method

AMS subject classifications. Primary, 49M37; Secondary, 49N60

DOI. 10.1137/S0363012904443506

1. Introduction.

We are considering the optimal control of the nonstationary Navier–Stokes equations. As a model problem we minimize the following quadratic objective functional J:

$$J(y,u) = \frac{\alpha_T}{2}\int_\Omega |y(x,T)-y_T(x)|^2dx + \frac{\alpha_Q}{2}\int_Q |y(x,t)-y_Q(x,t)|^2dxdt + \frac{\alpha_R}{2}\int_Q |\operatorname{curl} y(x,t)|^2dxdt + \frac{\gamma}{2}\int_Q |u(x,t)|^2dxdt \tag{1.1}$$

subject to the nonstationary Navier–Stokes equations

$$\begin{aligned} y_t - \nu\Delta y + (y\cdot\nabla)y + \nabla p &= u && \text{in } Q,\\ \operatorname{div} y &= 0 && \text{in } Q,\\ y(0) &= y_0 && \text{in } \Omega, \end{aligned} \tag{1.2}$$

and the control constraints $u \in U_{ad}$ with a set of admissible controls defined by

$$U_{ad} = \{u \in L^2(Q)^2 :\ u_{a,i}(x,t) \le u_i(x,t) \le u_{b,i}(x,t) \text{ a.e. on } Q,\ i=1,2\}.$$

Here, Ω is an open bounded subset of $\mathbb{R}^2$ with a C^3-boundary Γ such that Ω is locally on one side of Γ, and Q is defined by $Q = \Omega\times(0,T)$. Further, functions $y_T \in L^2(\Omega)^2$, $y_Q \in L^2(Q)^2$, and $y_0 \in H \subset L^2(\Omega)^2$ are given. The parameters γ and ν are positive real numbers. The bounds u_a, u_b are required to be in $L^2(Q)^2$ with $u_{a,i}(x,t) \le u_{b,i}(x,t)$ a.e. on Q, $i=1,2$.

Control of the nonstationary Navier–Stokes flow has been studied very intensively since the pioneering work [1]. Necessary as well as sufficient optimality conditions

*Received by the editors May 4, 2004; accepted for publication (in revised form) February 15, 2007; published electronically August 31, 2007. This work was supported by DFG SFB 557 "Control of complex turbulent shear flows" at TU Berlin.

http://www.siam.org/journals/sicon/46-3/44350.html

†Institut für Mathematik, Technische Universität Berlin, Str. des 17. Juni 136, D-10623 Berlin, Germany (wachsmut@math.tu-berlin.de).

were established; cf. [8, 9, 16, 17, 27]. The optimality system can be used to derive regularity properties of optimal controls. It was proved that under certain regularity assumptions a locally optimal control of the problem (1.1) is a continuous function in space and time; cf. [29].

The aim of this article is the presentation of a convergence theory of the SQP-method to solve the optimization problem (1.1). This method is widely applied to solve finite dimensional as well as function space optimization problems. If a stability result is available, one can prove local convergence of the SQP-method using the concept of strongly regular generalized equations [21]. The first convergence result in the context of optimal control of partial differential equations was given in [26]. A convergence proof for the SQP-method applied to constrained optimal control problems of the Navier–Stokes equations was given in [13]. However, the convergence was proved under a strong second-order sufficient optimality condition: the second derivative of the Lagrangian was required to be coercive for all control test functions.

We will prove quadratic convergence of the SQP-method in an L^∞-neighborhood of a reference control, which has to fulfill a second-order sufficient optimality condition. In contrast to the approaches in [13, 17], we require that the second derivative of the Lagrangian is positive definite only on a subspace associated with inactive constraints. Furthermore, we get convergence of the control iterates with respect to the L^∞-norm, whereas the convergence theory of [13, 17] gives convergence in weaker, say, L^q-norms, with $q < 7/2$.

The semismooth Newton method is another method to tackle nonlinear optimal control problems. It was applied to optimal control problems for nonstationary Navier–Stokes equations in [28]. Under the assumption of a strong sufficient optimality condition, locally superlinear convergence was proved. Applying the regularity results of our article, it is possible to prove convergence of that method under the weaker coercivity assumptions involving strongly active constraints.

The outline of the paper is as follows. In section 2, we will introduce some notation and state common results concerning solvability of the nonstationary Navier–Stokes system (1.2). Sections 3 and 4 contain a brief overview of optimality conditions including first-order necessary and second-order sufficient conditions. In section 5, the SQP-method is considered. The main result of the article—local convergence of SQP—is stated and proved in section 5.3. Numerical results confirming the convergence theory are presented in section 6. The required regularity results for the linearized Navier–Stokes equation and the adjoint equation can be found in sections 2.2 and 3.1. Throughout the article, we investigate the theory of optimal controls of the nonstationary Navier–Stokes equations in the L^p-space context.

2. Notation and preliminary results. Here, we will restrict ourselves to the two-dimensional case, $n = 2$. First, we introduce some notation and provide some results that we will need later on.

To begin with, we define the spaces of solenoidal functions

$$H_p := \{v \in L^p(\Omega)^2 : \operatorname{div} v = 0\}, \qquad V_p := \{v \in W_0^{1,p}(\Omega)^2 : \operatorname{div} v = 0\}.$$

Here, p denotes an arbitrary exponent $p \geq 2$. These spaces are Banach spaces with their norms denoted by $|\cdot|_p$ and $|\cdot|_{1,p}$, respectively. For $p = 2$, we get the frequently used solenoidal spaces $H := H_2$ and $V := V_2$, which are Hilbert spaces with scalar products $(\cdot,\cdot)_H$ and $(\cdot,\cdot)_V$, respectively. The dual of V with respect to the scalar product of H is denoted by V' with the duality pairing $\langle\cdot,\cdot\rangle_{V',V}$.

We shall work in the standard space of abstract functions from $[0,T]$ to a real Banach space X, $L^p(0,T;X)$, endowed with its natural norm,

$$\|y\|_{L^p(X)} := \|y\|_{L^p(0,T;X)} = \left(\int_0^T |y(t)|_X^p dt\right)^{1/p}, \quad 1 \le p < \infty,$$
$$\|y\|_{L^\infty(X)} := \operatorname*{vrai\,max}_{t\in(0,T)} |y(t)|_X.$$

In what follows, we will identify the spaces $L^p(0,T;L^p(\Omega)^2)$ and $L^p(Q)^2$ for $1 < p < \infty$, and denote their norm by $\|u\|_p := |u|_{L^p(Q)^2}$. We denote by $(\cdot,\cdot)_Q$ the usual $L^2(Q)^2$-scalar product to avoid ambiguity.

In all what follows, $\|\cdot\|$ stands for norms of abstract functions, while $|\cdot|$ denotes norms of "stationary" spaces such as H and V.

To deal with the time derivative in (1.2), we introduce the common spaces of functions y, whose time derivatives y_t exist as abstract functions

$$W^\alpha(0,T;V) := \{y \in L^2(0,T;V) :\ y_t \in L^\alpha(0,T;V')\}, \quad W(0,T) := W^2(0,T;V),$$

where $1 \le \alpha \le 2$. Endowed with the norm

$$\|y\|_{W^\alpha} := \|y\|_{W^\alpha(0,T;V)} = \|y\|_{L^2(V)} + \|y_t\|_{L^\alpha(V')},$$

these spaces are Banach spaces (respectively, Hilbert spaces) in the case of $W(0,T)$. Every function of $W(0,T)$ is, up to changes on sets of zero measure, equivalent to a function of $C([0,T],H)$, and the imbedding $W(0,T) \hookrightarrow C([0,T],H)$ is continuous; cf. [2, 20].

Furthermore, we introduce the following space of abstract functions in the L^p-context:

$$W_p^{2,1} := \{y \in L^p(0,T;W^{2,p}(\Omega)^2 \cap V_p) :\ y_t \in L^p(0,T;L^p(\Omega)^2)\},$$

which is continuously imbedded in $C([0,T], W_0^{2-2/p,\,p}(\Omega)^2)$ (see [19]). Here, $W_0^{2-2/p,\,p}(\Omega)^2$ denotes the space of solenoidal $W^{2-2/p,\,p}$-functions, where zero boundary values are prescribed if $p \ge 4/3$. We abbreviate $H^{2,1} = W_2^{2,1}$ for $p = 2$. Note that in this case we have $W_0^{2-2/2,\,2}(\Omega)^2 = V$.

We define the trilinear form $b : V \times V \times V \mapsto \mathbb{R}$ by

$$b(y,v,w) = ((y\cdot\nabla)v, w)_2 = \int_\Omega \sum_{i,j=1}^2 y_i \frac{\partial v_j}{\partial x_i}\, w_j\, dx.$$

To specify the problem setting, we introduce a linear operator $A : L^2(0,T;V) \mapsto L^2(0,T;V')$ by

$$\int_0^T \langle (Ay)(t),\, v(t)\rangle_{V',V} dt := \int_0^T (y(t), v(t))_V dt,$$

and a nonlinear operator B by

$$\int_0^T \big\langle \big(B(y)\big)(t),\, v(t)\big\rangle_{V',V} dt := \int_0^T b(y(t), y(t), v(t)) dt.$$

B is continuous, for instance, as an operator from $W(0,T)$ to $L^2(0,T;V')$. For convenience, we will use the notation

$$b_Q(y,v,w) = \int_0^T b(y(t), v(t), w(t)) dt.$$

2.1. The state equation. We begin with the investigation of weak solutions for the nonstationary Navier–Stokes equations (1.2) in the Hilbert space setting.

DEFINITION 2.1 (weak solution). *Let $f \in L^2(0,T;V')$ and $y_0 \in H$ be given. A function $y \in L^2(0,T;V)$ with $y_t \in L^2(0,T;V')$ is called a weak solution of* (1.2) *if*

$$\begin{aligned} y_t + \nu A y + B(y) &= f, \\ y(0) &= y_0. \end{aligned} \tag{2.1}$$

Results concerning the solvability of (2.1) are standard; cf. [24] for proofs and further details.

THEOREM 2.2 (existence and uniqueness of solutions). *For every $f \in L^2(0,T;V')$ and $y_0 \in H$,* (2.1) *has a unique solution $y \in W(0,T)$. Moreover, the mapping $(y_0, f) \mapsto y$ is locally Lipschitz continuous from $H \times L^2(0,T;V')$ to $W(0,T)$.*

For more regular data, one expects more regular solutions.

THEOREM 2.3 (regularity). *For the higher regularity of the weak solutions of* (2.1) *the following holds. Let $y_0 \in V$ and $f \in L^2(Q)^2$ be given. Then the weak solution of* (2.1) *fulfills $y \in H^{2,1}$. The solution mapping $(y_0, f) \mapsto y$ is locally Lipschitz continuous between $L^2(Q)^2 \times V$ and $H^{2,1}$.*

For the proof, we refer again to Temam [24]. However, the result was proved there under the additional assumption $f \in L^2(0,T;H)$. A closer inspection of the proof shows that $f \in L^2(Q)^2$ suffices to get $y \in L^2(0,T;H^2(\Omega)^2) \cap L^\infty(0,T;V)$ and $y_t \in L^2(Q)^2$. Then the regularity $y_t \in L^2(0,T;H)$ follows by density arguments.

Now, we want to specify the notation of a solution of (2.1) in the L^p-context.

DEFINITION 2.4 (strong solution in L^p). *Let $f \in L^p(Q)^2$ and $y_0 \in W_0^{2-2/p,\,p}(\Omega)^2$ be given. A function $y \in W_p^{2,1}$ is called a strong solution to the exponent $p > 2$ of* (1.2) *if there holds*

$$-\int_0^T (y, \phi')dt + \nu \int_0^T (\nabla y, \nabla \phi)dt + \int_0^T b(y,y,\phi) = \int_0^T (f,\phi)dt + (y_0, \phi(0)) \tag{2.2}$$

for all test functions $\phi \in L^q(0,T;V_q)$ with $\phi_t \in L^q(0,T;L^q(\Omega)^2)$ and $\phi(T) = 0$, where q is the dual exponent to p, $1/q + 1/p = 1$.

Here the space $W_0^{2-2/p,\,p}(\Omega)^2$ is the natural trace space. Every abstract function of $L^p(0,T;W^{2,p}(\Omega)^2)$ with a time derivative in $L^p(0,T;L^p(\Omega)^2)$ is—after changes on a zero measure set—continuous with values in this space [19]. Obviously, every strong L^p-solution is a weak solution. For existence of L^p-solutions we have the following theorem.

THEOREM 2.5 (L^p-solutions). *Let $f \in L^p(Q)^2$ and $y_0 \in W_0^{2-2/p,\,p}(\Omega)^2$ be given with $p \geq 2$. Then the weak solution y of* (2.1) *in the sense of Definition* 2.1 *is a strong solution and satisfies $y \in W_p^{2,1}$. There exists a constant $c > 0$ such that*

$$\|y\|_{W_p^{2,1}} \leq c\left\{|y_0|_{W^{2-2/p,\,p}} + \|f\|_p\right\}.$$

Moreover, the mapping $(f, y_0) \mapsto y$ is locally Lipschitz continuous, and hence the strong solution y is unique.

If $p = 2$, this result reduces to Theorem 2.3. For the non-Hilbert space case $p > 2$, existence and regularity was proved in [30]. Lipschitz continuity for this case is a conclusion of the estimates in the next section.

2.2. Linearized equation. In the course of the article, we will need a similar existence and regularity result for the linearized Navier–Stokes equations. Let a function $\bar{y} \in L^p(0,T;W^{2,p}(\Omega)^2) \cap L^\infty(0,T;W_0^{2-2/p,\,p}(\Omega)^2)$ be given. Then we are looking for solutions of the linearized system

$$\begin{aligned} y_t + \nu Ay + B'(\bar{y})y &= f, \\ y(0) &= y_0. \end{aligned} \tag{2.3}$$

We will show that this system admits a unique solution, which belongs to the space $W_p^{2,1}$ and depends continuously on the data.

To this end, we will rely on a regularity result for the nonstationary Stokes equations,

$$\begin{aligned} y_t + \nu Ay &= f, \\ y(0) &= y_0. \end{aligned} \tag{2.4}$$

Concerning L^p-solutions, the following result is due to Solonnikov [22] for the two- and three-dimensional cases; in [30] it was generalized to arbitrary spatial dimensions.

THEOREM 2.6. *Let $p > 1$, $p \neq 3/2$, $y_0 \in W_0^{2-2/p,\,p}(\Omega)^2$, $f \in L^p(Q)^2$. Then there exists a unique weak solution y of (2.4) satisfying $y \in W_p^{2,1}$. Furthermore, there exists a constant $c > 0$ such that the estimate*

$$\|y_t\|_p + \|y\|_{L^p(W^{2,p})} \leq c\left\{\|f\|_p + |y_0|_{W^{2-2/p,\,p}}\right\}$$

is satisfied.

Now, we can prove the regularity result for the linearized system. A similar result can be found in [23], where equations with the linear term $(\bar{y}\cdot\nabla)y$ instead of $B'(\bar{y})y = (\bar{y}\cdot\nabla)y + (y\cdot\nabla)\bar{y}$ are studied.

THEOREM 2.7. *Let $\bar{y} \in L^p(0,T;W^{2,p}(\Omega)^2) \cap L^\infty(0,T;W_0^{2-2/p,\,p}(\Omega)^2)$, $f \in L^p(Q)^2$, and $y_0 \in W_0^{2-2/p,\,p}(\Omega)^2$ be given with $2 \leq p < \infty$. Then the system (2.3) has a unique solution $y \in W_p^{2,1}$. Moreover, there is a constant $c > 0$ independent of f and y_0 such that the following estimate holds:*

$$\|y\|_{W_p^{2,1}} \leq c\left\{\|f\|_p + |y_0|_{W^{2-2/p,\,p}}\right\}. \tag{2.5}$$

Proof. Step 1. $p = 2$. The existence of a unique weak solution together with the estimate in the case $p = 2$ was proved in [17]. Let us write the system (2.3) in a slightly modified form,

$$\begin{aligned} y_t + \nu Ay &= f - B'(\bar{y})y, \\ y(0) &= y_0, \end{aligned} \tag{2.6}$$

to estimate y in terms of f, $B'(\bar{y})y$, and y_0. Here, we want to apply the regularity result of Theorem 2.6. To this end, we have to estimate the L^p-norm of the right-hand side of (2.6) for different values of p. The proof is then carried out using bootstrapping arguments.

Step 2. $2 < p < 4$. From the previous step, we know the existence of a unique weak solution $y \in W_2^{2,1}$ of (2.3). Let us investigate $B'(\bar{y})y = (y\cdot\nabla)\bar{y} + (\bar{y}\cdot\nabla)y$.

By assumption, we have $\bar{y} \in L^\infty(0,T;W_0^{2-2/p,\,p}(\Omega)^2)$. The space $W_0^{2-2/p,\,p}(\Omega)^2$ is continuously imbedded in $W^{1,q}(\Omega)^2$ for $q = \frac{2p}{4-p}$, $p < 4$; cf. [2, 25]. Furthermore,

the space V is continuously imbedded in $L^{q'}(\Omega)^2$ for $q' < \infty$. Applying Hölder's inequality with $\frac{1}{p} = \frac{1}{q} + \frac{1}{q'}$, we obtain

$$\|(y \cdot \nabla)\bar{y}\|_p \le c\|y\|_{L^\infty(L^{q'})}\|\bar{y}\|_{L^\infty(W^{1,q})} \le c\|y\|_{L^\infty(V)}\|\bar{y}\|_{L^\infty(W^{2-2/p,\,p})}. \tag{2.7}$$

The estimation of the second addend of $B'(\bar{y})y$ needs a bit more effort. Using the interpolation identity

$$[W^{2,2}(\Omega)^2, W^{1,2}(\Omega)^2]_\theta = W^{2-2/p,\,2}(\Omega)^2, \quad \theta = 1 - 2/p,$$

and the imbedding $W^{2-2/p,\,2}(\Omega)^2 \hookrightarrow W^{1,p}(\Omega)^2$, we find a.e. on $[0,T]$

$$|y(t)|^p_{W^{1,p}} \le c|y(t)|^p_{W^{2-2/p,\,2}} \le c|y(t)|^{p-2}_{W^{2,2}}|y(t)|^2_{W^{1,2}}.$$

Integrating with respect to the time variable yields

$$\|y\|^p_{L^p(W^{1,p})} \le c\|y\|^{p-2}_{L^{p-2}(W^{2,2})}\|y\|^2_{L^\infty(V)} \le c\|y\|^{p-2}_{L^2(W^{2,2})}\|y\|^2_{L^\infty(V)} \le c\|y\|^p_{W_2^{2,1}} \tag{2.8}$$

provided $p \le 4$. Also, we can derive

$$\|(\bar{y} \cdot \nabla)y\|_p \le c\|\bar{y}\|_\infty\|y\|_{L^p(W^{1,p})}. \tag{2.9}$$

Collecting (2.7)–(2.9), we find

$$\|B'(\bar{y})y\|_p \le c\|\bar{y}\|_{W_p^{2,1}}\|y\|_{W_2^{2,1}} \le c\|\bar{y}\|_{W_p^{2,1}}\left\{\|f\|_2 + |y_0|_V\right\}.$$

Now, we can utilize Theorem 2.6 to obtain the solution estimate

$$\begin{aligned}\|y\|_{W_p^{2,1}} &\le c\left\{\|f\|_p + |y_0|_{W^{2-2/p,\,p}} + \|B'(\bar{y})y\|_p\right\}\\ &\le c\left\{\|f\|_p + |y_0|_{W^{2-2/p,\,p}}\right\} + c\|\bar{y}\|_{W_p^{2,1}}\left\{\|f\|_2 + |y_0|_V\right\}\\ &\le c(1 + \|\bar{y}\|_{W_p^{2,1}})\left\{\|f\|_p + |y_0|_{W^{2-2/p,\,p}}\right\}.\end{aligned}$$

Thus, the (weak) solution y is of class $W_p^{2,1}$. Using density arguments, it is easy to verify that y is also a strong solution. Since every strong solution is a weak solution, and weak solutions are unique, it follows that y is the unique strong solution of the linearized system. This completes the proof for exponents $p \in (2,4)$.

Step 3. $4 \le p < \infty$. By Step 2, the solution y of (2.3) is in $W^{2,1}_{4-\varepsilon}$, $0 < \varepsilon \le 2$. It is—after changes on a set of zero measure—continuous with values in the space $W_0^{2-2/(4-\varepsilon),4-\varepsilon}(\Omega)^2$, which is itself continuously imbedded in $L^\infty(\Omega)^2$. Hence, the imbedding of $W^{2,1}_{4-\varepsilon}$ in $L^\infty(Q)^2$ is continuous.

Again, we have to estimate the L^p-norm of $B'(\bar{y})y$. We begin with its first addend, which can be treated as

$$\|(y \cdot \nabla)\bar{y}\|_p \le c\|y\|_\infty\|\nabla\bar{y}\|_p \le c\|y\|_{W^{2,1}_{4-\varepsilon}}\|\bar{y}\|_{W_p^{2,1}}. \tag{2.10}$$

To estimate the second addend of $B'(\bar{y})y$, we observe that for $\varepsilon = \frac{8}{p+2}$ the imbedding

$$W_0^{2-\frac{2}{4-\varepsilon},4-\varepsilon}(\Omega)^2 = W_0^{\frac{3}{2}-\frac{1}{p},\frac{4p}{p+2}}(\Omega)^2 \hookrightarrow W_0^{1,p}(\Omega)^2$$

is continuous. Consequently, we obtain for this choice of ε,

$$y \in W^{2,1}_{4-\varepsilon} \hookrightarrow L^\infty(0,T; W_0^{1,p}(\Omega)^2).$$

Hence, we arrive at

$$\|(\bar{y}\cdot\nabla)y\|_p \le c\|\bar{y}\|_\infty \|y\|_{L^\infty(W^{1,p})} \le c\|\bar{y}\|_{W_p^{2,1}} \|y\|_{W_{4-\varepsilon}^{2,1}}, \tag{2.11}$$

which allows us to conclude by Theorem 2.6,

$$\|y\|_{W_p^{2,1}} \le c(1+\|\bar{y}\|_{W_p^{2,1}})\left\{\|f\|_p + |y_0|_{W^{2-2/p,\,p}}\right\},$$

and the claim is proved for all p in $[2,\infty)$. □

Remark 2.8. The Lipschitz continuity of the solution mapping of the nonstationary Navier–Stokes equations can be proved using the previous lemma. Let data $f_i \in L^p(Q)^2$ and $y_{0,i} \in W_0^{2-2/p,\,p}(\Omega)^2$ be given, with $i = 1, 2$. Denote the associated strong solutions by y_i, $i = 1, 2$. Then the difference $d := y_1 - y_2$ satisfies

$$\begin{aligned} d_t + \nu A d + (y_1\cdot\nabla)d + (d\cdot\nabla)y_2 &= f_1 - f_2, \\ d(0) &= y_{0,1} - y_{0,2}. \end{aligned}$$

With analogous arguments as above, one can verify

$$\|y_1 - y_2\|_{W_p^{2,1}} \le c(1+\|y_1\|_{W_p^{2,1}} + \|y_2\|_{W_p^{2,1}})\left\{\|f_1 - f_2\|_p + |y_{0,1} - y_{0,2}|_{W^{2-2/p,\,p}}\right\},$$

which is the claimed Lipschitz continuity of the solution mapping associated with the nonlinear system.

3. First-order necessary optimality conditions. Now let us return to our optimal control problem. We briefly recall the necessary conditions for local optimality. For the proofs and further discussion see [1, 5, 9, 16, 27] and the references cited therein.

Definition 3.1 (locally optimal control). *A control $\bar{u} \in U_{ad}$ is said to be locally optimal in $L^2(Q)^2$ if there exists a constant $\rho > 0$ such that*

$$J(\bar{y},\bar{u}) \le J(y_\rho, u_\rho)$$

holds for all $u_\rho \in U_{ad}$ with $\|\bar{u} - u_\rho\|_2 \le \rho$. Here, $\bar{y}$ and y_ρ denote the states associated with $\bar{u}$ and u_ρ, respectively.

In the following, we denote by $B'(\bar{y})^*$ the adjoint of $B'(\bar{y})$, given by

$$[B'(\bar{y})^*\lambda]v = \int_Q b(\bar{y}, v, \lambda) + b(v, \bar{y}, \lambda)dt.$$

Theorem 3.2 (necessary condition). *Let $\bar{u}$ be a locally optimal control with associated state $\bar{y} = y(\bar{u})$. Then there exists a unique solution $\bar{\lambda} \in W^{4/3}(0,T;V)$ of the adjoint equation*

$$\begin{aligned} -\bar{\lambda}_t + \nu A\bar{\lambda} + B'(\bar{y})^*\bar{\lambda} &= \alpha_Q(\bar{y} - y_Q) + \alpha_R \vec{\operatorname{curl}}\operatorname{curl}\bar{y}, \\ \bar{\lambda}(T) &= \alpha_T(\bar{y}(T) - y_T). \end{aligned} \tag{3.1}$$

Moreover, the variational inequality

$$(\gamma\bar{u} + \bar{\lambda}, u - \bar{u})_{L^2(Q)^2} \ge 0 \quad \forall u \in U_{ad} \tag{3.2}$$

is satisfied.

Proofs can be found in [9, 10, 27]. The regularity of $\bar{\lambda}$ is proved in [17].

The variational inequality (3.2) can be reformulated equivalently in different ways. First, let us introduce the normal cone $N_{U_{ad}}(\bar u)$ of the set of admissible controls at a given control $\bar u$, which is defined by

$$N_{U_{ad}}(\bar u) = \begin{cases} \{ z \in L^2(Q)^2 : \ (z, u - \bar u)_2 \le 0 \ \forall u \in U_{ad} \} & \text{if } \bar u \in U_{ad}, \\ \emptyset & \text{otherwise.} \end{cases} \tag{3.3}$$

Then the variational inequality (3.2) can be written equivalently as the inclusion

$$\gamma \bar u + \bar\lambda + N_{U_{ad}}(\bar u) \ni 0. \tag{3.4}$$

This representation fits into the context of generalized equations (see, for instance, [13, 29]), and will be utilized in the course of the article.

Second, the variational inequality

$$(\gamma \bar u_i(x,t) + \bar\lambda_i(x,t)) \cdot (u - \bar u_i(x,t)) \ge 0 \quad \forall u \in [u_{a.i}(x,t), u_{b.i}(x,t)]$$

has to be fulfilled pointwise a.e. on Q for $i = 1, 2$. This is in turn equivalent to the projection representation of the optimal control

$$\bar u_i(x,t) = \mathrm{Proj}_{[u_{a,i}(x,t), u_{b,i}(x,t)]} \left(-\frac{1}{\gamma} \bar\lambda_i(x,t) \right) \qquad \text{a.e. on } Q,\ i = 1, 2. \tag{3.5}$$

With this formula, one can see that a locally optimal control inherits some regularity from the associated adjoint state. This form is also used in connection with Lipschitz stability of optimal controls, see section 5.2 below.

3.1. Adjoint equation. The adjoint state λ is the solution of a linearized adjoint equation backward in time. So it is natural to look for its dependence on the given data. The existence of L^p-solutions of the adjoint equation is the topic of the next theorem.

THEOREM 3.3. *Let $y_Q \in L^p(Q)^2$ and $y_T \in W_0^{2-2/p,\,p}(\Omega)^2$ be given with $p \ge 2$. If $\bar y \in L^p(0,T; W^{2,p}(\Omega)^2) \cap L^\infty(0,T; W_0^{2-2/p,\,p}(\Omega)^2)$, then the weak solution λ of (3.1) is a strong solution and satisfies $\lambda \in W_p^{2,1}$. Moreover, the adjoint λ depends continuously on the given data y_Q, y_T, $\bar y$.*

Proof. The result in the case $p = 2$ was proved in [17, Prop. 2.4] for the homogeneous final value $\lambda(T) = 0$. It can be extended to the inhomogeneous case using known results for the nonstationary Stokes system.

Let us sketch the proof for the case $p > 2$. For simplicity, we define $f := \alpha_Q(\bar y - y_Q) + \alpha_R \vec{\mathrm{curl}}\, \mathrm{curl}\, \bar y$ and $\lambda_T := \alpha_T(\bar y(T) - y_T)$. Under the assumptions of the theorem, we know that $f \in L^p(Q)^2$ and $\lambda_T \in W_0^{2-2/p,\,p}(\Omega)^2$. Now, we will investigate the system

$$\begin{aligned} -\lambda_t + \nu A\lambda + B'(\bar y)^* \lambda &= f, \\ \lambda(T) &= \lambda_T. \end{aligned} \tag{3.6}$$

Via the transformations $w(t) = \lambda(T - t)$, $\hat y(t) = \bar y(T - t)$, $g(t) = f(T - t)$, $w_0 = \lambda_T$, this system is carried over in the forward-in-time equation

$$\begin{aligned} w_t + \nu A w + B'(\hat y)^* w &= g, \\ w(0) &= w_0. \end{aligned} \tag{3.7}$$

Obviously, $\hat{y}, g, w_0$ inherits their regularity from $\bar{y}, f, \lambda_T$. Hence, the adjoint state λ has the same regularity as the auxiliary state w. The proof is finished if the next lemma is verified. ∎

LEMMA 3.4. *Let $\hat{y} \in L^p(0,T;W^{2,p}(\Omega)^2) \cap L^\infty(0,T;W_0^{2-2/p,\,p}(\Omega)^2)$, $g \in L^p(Q)^2$, and $w_0 \in W_0^{2-2/p,\,p}(\Omega)^2$ be given with $2 \le p < \infty$. Then the system (3.6) has a unique solution $w \in W_p^{2,1}$. Moreover, there is a constant $c > 0$ independent of g and w_0 such that the following estimate is true:*

$$\|w\|_{W_p^{2,1}} \le c\left\{\|g\|_p + |w_0|_{W^{2-2/p,\,p}}\right\}.$$

Proof. The proof is very similar to the proof of Theorem 2.7. Therefore, we will briefly repeat its steps. First, let us investigate the action of $B'(\hat{y})^* w$ on a test function $v \in W(0,T)$:

$$\begin{aligned}[B'(\hat{y})^* w]v &= b_Q(\hat{y}, v, w) + b_Q(v, \hat{y}, w) = -b_Q(\hat{y}, w, v) + b_Q(v, \hat{y}, w)\\ &= \sum_{i,j=1}^{2} \int_Q \left(-\hat{y}_i(x,t)\frac{\partial w_j(x,t)}{\partial x_i} v_j(x,t) + v_i(x,t)\frac{\partial \hat{y}_j(x,t)}{\partial x_i} w_j(x,t)\right) dxdt\\ &= \int_0^T [-(\hat{y}\cdot\nabla)w + (\nabla\hat{y})^T w]\cdot v\, dt.\end{aligned}$$

Here, we used the identity $b_Q(y,v,w) = -b_Q(y,w,v)$, which holds for functions $y, v, w \in L^2(0,T;V)$ [24]. Consequently, we are allowed to identify the functional $B(\hat{y})^* w$ with the function $-(\hat{y}\cdot\nabla)w + (\nabla\hat{y})^T w$. For $\hat{y}, w \in H^{2,1}$ we find $-(\hat{y}\cdot\nabla)w + (\nabla\hat{y})^T w \in L^2(Q)^2$.

Step 1. $p = 2$. The result for $p = 2$ was proved, for instance, in [17].

Step 2. $2 < p < 4$. With the help of (2.7)–(2.9), we conclude that

$$\|B'(\hat{y})^* w\|_p \le c\left(\|\hat{y}\|_\infty \|w\|_{W_2^{2,1}} + \|\hat{y}\|_{L^\infty(W^{2-2/p,\,p})}\|w\|_{L^\infty(V)}\right).$$

Then Theorem 2.6 gives us the boundedness of the solution w in $W_p^{2,1}$.

Step 3. $4 \le p < \infty$. Let $w \in W_{4-\varepsilon}^{2,1}$ be the strong solution of Step 2, $0 < \varepsilon < 2$. Analogously as in (2.10) and (2.11), we find

$$\|B'(\hat{y})^* w\|_p \le c\|\hat{y}\|_{W_p^{2,1}} \|w\|_{W_{4-\varepsilon}^{2,1}},$$

and the claim follows immediately. ∎

3.2. Lagrange functional. Let us introduce the Lagrange function

$$\mathcal{L} : W(0,T) \times L^2(Q)^2 \times W^{4/3}(0,T) \mapsto \mathbb{R}$$

for the optimal control problem as follows:

$$\mathcal{L}(y,u,\lambda) = J(u,y) - \left\{\langle y_t,\, \lambda\rangle_{L^2(V'),L^2(V)} + \nu(y,\lambda)_{L^2(V)} + b_Q(y,y,\lambda) - (u,\lambda)_Q\right\}.$$

This function is twice Fréchet-differentiable with respect to $(y,u) \in W(0,T)\times L^2(Q)^2$; cf. [27]. The reader can readily verify that the necessary conditions can be expressed equivalently by

$$\begin{gathered}\mathcal{L}_y(\bar{y},\bar{u},\bar{\lambda})\, h = 0 \quad \forall h \in W(0,T) \text{ with } h(0) = 0,\\ \mathcal{L}_u(\bar{y},\bar{u},\bar{\lambda})(u-\bar{u}) \ge 0 \quad \forall u \in U_{ad}.\end{gathered}$$

Here, $\mathcal{L}_y$, $\mathcal{L}_u$ denote the partial Fréchet-derivative of $\mathcal{L}$ with respect to y and u.

In what follows we denote the pair of state and control (y, u) by v for convenience. The second derivative of the Lagrangian $\mathcal{L}$ at $y \in W(0,T)$ with associated adjoint state λ in the directions $v_1 = (w_1, h_1)$, $v_2 = (w_2, h_2) \in W(0,T) \times L^2(Q)^2$ is given by

$$\mathcal{L}_{vv}(y,u,\lambda)[v_1,v_2] = \mathcal{L}_{yy}(y,u,\lambda)[w_1,w_2] + \mathcal{L}_{uu}(y,u,\lambda)[h_1,h_2] \tag{3.8}$$

with

$$\begin{aligned}\mathcal{L}_{yy}(y,u,\lambda)[w_1,w_2] = \alpha_T(w_1(T),w_2(T))_H + \alpha_Q(w_1,w_2)_Q + \alpha_R(\operatorname{curl} w_1, \operatorname{curl} w_2)_Q \\ - b_Q(w_1,w_2,\lambda) - b_Q(w_2,w_1,\lambda)\end{aligned}$$

and

$$\mathcal{L}_{uu}(y,u,\lambda)[h_1,h_2] = \gamma(h_1,h_2)_2.$$

It can be verified that it satisfies the estimate

$$|\mathcal{L}_{yy}(y,u,\lambda)[w_1,w_2]| \le c\left(1 + \|\lambda\|_{L^2(V)}\right)\|w_1\|_{W(0,T)}\|w_2\|_{W(0,T)} \tag{3.9}$$

for all $w_1, w_2 \in W(0,T)$. To shorten notation, we abbreviate $[v,v]$ by $[v]^2$, i.e.,

$$\mathcal{L}_{vv}(\bar v,\bar\lambda)[(w,h)]^2 := \mathcal{L}_{vv}(\bar v,\bar\lambda)[(w,h),(w,h)].$$

4. Second-order sufficient optimality condition. Let $\bar v := (\bar y, \bar u)$ be an admissible reference pair satisfying the first-order necessary optimality conditions.

Definition 4.1 (strongly active sets). *Let $\varepsilon > 0$ and $i \in \{1,2\}$ be given. Define sets $Q_{\varepsilon,i} \subseteq Q = \Omega \times [0,T]$ by*

$$Q_{\varepsilon,i} = \{(x,t) \in Q : |\gamma \bar u_i(x,t) + \bar\lambda_i(x,t)| > \varepsilon\}.$$

We assume further that the reference pair $\bar v = (\bar y, \bar u)$ satisfies the following coercivity assumption on $\mathcal{L}''(\bar v, \bar\lambda)$; in what follows, this is called the second-order sufficient condition:

(SSC)
There exist $\varepsilon > 0$ and $\delta > 0$ such that

$$\mathcal{L}_{vv}(\bar v,\bar\lambda)[(w,h)]^2 \ge \delta\,\|h\|_2^2$$

holds for all pairs $(w,h) \in W(0,T) \times L^2(Q)^2$ with

$$h = u - \bar u,\ u \in U_{ad},\ h_i = 0 \text{ on } Q_{\varepsilon,i} \text{ for } i = 1,2,$$

and $w \in W(0,T)$ being the weak solution of the linearized equation

$$\begin{aligned} w_t + Aw + B'(\bar y)w &= h, \\ w(0) &= 0. \end{aligned}$$

The sufficiency of (SSC) was proved in [27]: An admissible control that satisfies the first-order necessary conditions together with (SSC) is locally optimal in $L^\infty(Q)^2$. Observe that $\mathcal{L}''$ has to be positive definite only on the subspace of control variations that are zero on the set of strongly active constraints. We will show that for a control $\bar u$ satisfying (SSC), the SQP-method started in a neighborhood of $\bar u$ will indeed converge to that local solution.

5. SQP-method. In this section, we consider the SQP-method to compute a local optimum of the control problem (1.2). It is a well-known method, applied very often to optimal control problems of partial differential equations; see, e.g., [13, 26]. For the analysis of other local methods such as (quasi-) Newton and semismooth Newton methods in connection with nonstationary Navier–Stokes equations, we refer to [17, 28].

The SQP-method solves in every step a linear-quadratic optimal control problem (P^n). Given starting values y_n, u_n, λ_n, it computes the next iterates $y_{n+1}, u_{n+1}, \lambda_{n+1}$ as the minimizers of

$$J^n(y,u) = J(y_n,u_n) + \nabla J(y_n,u_n)(y-y_n,u-u_n) + \frac{1}{2}\mathcal{L}_{vv}(y_n,u_n,\lambda_n)[(y-y_n,u-u_n)]^2$$

subject to the linearized state equation

$$\begin{aligned} y_t + \nu Ay + B'(y_n)(y-y_n) &= u - B(y_n), \\ y(0) &= y_0, \end{aligned}$$

and the control constraint $u \in U_{ad}$. For convenience we write the functional as

$$\begin{aligned} J^n(y,u) = {} & \frac{\alpha_T}{2}\int_\Omega |y(x,T)-y_T(x)|^2 dx + \frac{\alpha_Q}{2}\int_Q |y(x,t)-y_Q(x,t)|^2 dxdt \\ & + \frac{\alpha_R}{2}\int_Q |\operatorname{curl} y(x,t)|^2 dxdt + \frac{\gamma}{2}\int_Q |u(x,t)|^2 dxdt - b_Q(y-y_n, y-y_n, \lambda_n). \end{aligned}$$

In what follows, we investigate local convergence of this method. Here, the sufficient condition (SSC) plays an essential role. As one expects, we get quadratic convergence as soon as the iterates lie in a neighborhood of a local solution.

5.1. Generalized Newton method. The SQP-method can be interpreted as a Newton method for a generalized equation of the form

$$0 \in F(x) + N(x), \tag{5.1}$$

where F is a $C^{1,1}$-mapping between two Banach spaces X and Z, while $N : X \mapsto 2^Z$ is a set-valued mapping with closed graph. One can write the Newton method formally as follows: Given iterate x_n, compute the next iterate x_{n+1} by solving

$$0 \in F(x_n) + F'(x_n)(x-x_n) + N(x).$$

Before we state an abstract result concerning the convergence of the generalized Newton method, we will introduce the notation of strong regularity in the sense of Robinson [21]. Let $\bar{x}$ be a solution of (5.1). The generalized equation is said to be *strongly regular* at the point $\bar{x}$ if there are open balls $B_X(\bar{x},\rho_x)$ and $B_Z(0,\rho_z)$ such that for all $z \in B_Z(0,\rho_z)$ the linearized and perturbed equation

$$z \in F(\bar{x}) + F'(\bar{x})(x-\bar{x}) + N(x)$$

admits a unique solution $x = x(z)$ in $B_X(\bar{x},\rho_x)$, and the mapping $z \mapsto x$ is Lipschitz continuous $B_Z(0,\rho_z)$ from to $B_X(\bar{x},\rho_x)$. The following theorem [3, 7] gives the mentioned convergence result.

THEOREM 5.1. *Let $\bar{x}$ be a solution of* (5.1) *and assume that* (5.1) *is strongly regular at $\bar{x}$. Then there exists an open ball $B_X(\bar{x},\rho_x')$ such that for every starting*

element $x_1 \in B_X(\bar{x}, \rho_x')$ the generalized Newton method generates a unique sequence $\{x_n\}_{n=1}^{\infty}$. The iterates x_n remain in $B_X(\bar{x}, \rho_x')$, and it holds that

$$\|x_{n+1} - \bar{x}\|_X \le c_N \|x_n - \bar{x}\|_X^2 \quad \forall n \in \mathbb{N}, \tag{5.2}$$

where c_N is independent of n.

5.2. Strong regularity: L^∞-stability of optimal controls. Let $(\bar{y}, \bar{u}, \bar{\lambda})$ satisfy the first-order necessary optimality conditions (see Theorem 3.2), together with the second-order sufficient optimality condition (SSC). The optimality system consisting of state equation (1.2), adjoint equation (3.1), and the inclusion (3.4) can be written in the condensed form

$$F(\bar{y}, \bar{u}, \bar{\lambda}) + \big(0, 0, 0, 0, N_{U_{ad}}(\bar{u})\big)^T \ni 0, \tag{5.3}$$

where the function F,

$$\begin{aligned} F : W_p^{2,1} \times L^\infty(Q)^2 \times W_p^{2,1} \to \\ L^p(Q)^2 \times W_0^{2-2/p,\,p}(\Omega)^2 \times L^p(Q)^2 \times W_0^{2-2/p,\,p}(\Omega)^2 \times L^\infty(Q)^2, \end{aligned} \tag{5.4}$$

is given by

$$F(y,u,\lambda) = \begin{pmatrix} y_t + \nu A y + B(y) \\ y(0) \\ -\lambda_t + \nu A \lambda + B'(y)^* \lambda \\ \lambda(T) \\ \gamma u + \lambda \end{pmatrix} - \begin{pmatrix} u \\ y_0 \\ \alpha_Q(y - y_Q) + \alpha_R \vec{\operatorname{curl}} \operatorname{curl} y \\ \alpha_T(y(T) - y_T) \\ 0 \end{pmatrix}. \tag{5.5}$$

Further, we have to redefine the normal cone $N_{U_{ad}}$ to be a subset of $L^\infty(Q)^2$,

$$N_{U_{ad}} = \begin{cases} \{ z \in L^\infty(Q)^2 : \, (z, u - \bar{u})_2 \le 0 \ \forall u \in U_{ad} \} & \text{if } \bar{u} \in U_{ad}, \\ \emptyset & \text{otherwise.} \end{cases}$$

We will apply Theorem 5.1 to the generalized equation (5.3). To do so, we have to show strong regularity of this equation at the reference triple $(\bar{y}, \bar{u}, \bar{\lambda})$. To this end, let us investigate the linearized and perturbed inclusion

$$z \in F(\bar{y}, \bar{u}, \bar{\lambda}) + F'(\bar{y}, \bar{u}, \bar{\lambda})(y - \bar{y}, u - \bar{u}, \lambda - \bar{\lambda}) + \big(0, 0, 0, 0, N_{U_{ad}}(\bar{u})\big)^T. \tag{5.6}$$

Here, the perturbation vector $z = (z_y, z_0, z_Q, z_T, z_u)$ is restricted to be in the space Z given by

$$Z := L^p(Q)^2 \times W_0^{2-2/p,\,p}(\Omega)^2 \times L^p(Q)^2 \times W_0^{2-2/p,\,p}(\Omega)^2 \times L^\infty(Q)^2. \tag{5.7}$$

We equip Z with the natural norm

$$\|z\|_Z = \|(z_y, z_0, z_Q, z_T, z_u)\|_Z := \|z_y\|_p + |z_0|_{W^{2-2/p,\,p}} + \|z_Q\|_p + |z_T|_{W^{2-2/p,\,p}} + \|z_u\|_\infty.$$

To prove strong regularity of (5.3), we have to consider the linearized and perturbed generalized equation (5.6). It represents a system that can be written in a more

convenient way. The addends coming from the Navier–Stokes nonlinearity can be equivalently transformed, due to its quadratic character, into

$$B(\bar y) + B'(\bar y)(y - \bar y) = B'(\bar y)y + B(\bar y) - B'(\bar y)\bar y = B'(\bar y)y - B(\bar y).$$

Then, (5.6) builds up the optimality system of the following perturbed linear-quadratic optimization problem, henceforth called (P_z): Find the minimizer of $J^{(z)}$ given by

$$\begin{aligned}(5.8)\quad J^{(z)}(y,u) = {}& \frac{\alpha_R}{2}\int_Q |\operatorname{curl} y|^2 + \frac{\gamma}{2}\int_Q |u|^2 \\ & + \frac{\alpha_T}{2}|y(T) - y_d|_H^2 + \frac{\alpha_Q}{2}\|y - y_Q\|_2^2 + \frac{\alpha_R}{2}\|\operatorname{curl} y\|_2^2 + \frac{\gamma}{2}\|u\|_2^2 \\ & + (z_Q, y)_Q + (z_T, y(T))_\Omega - (z_u, u)_Q - b_Q(y - \bar y, y - \bar y, \bar\lambda)\end{aligned}$$

subject to the linearized state equation

$$\begin{aligned}(5.9)\qquad y_t + \nu Ay + B'(\bar y)y &= u + B(\bar y) + z_y, \\ y(0) &= y_0 + z_0\end{aligned}$$

and the control constraint $u \in U_{ad}$. The adjoint equations associated with the perturbed problem (P_z) are given by

$$\begin{aligned}(5.10)\quad -\lambda_t + \nu A\lambda + B'(\bar y)^*\lambda &= -B'(y - \bar y)^*\bar\lambda + \alpha_Q(y - y_Q) + \alpha_R \vec{\operatorname{curl}}\operatorname{curl} y + z_Q, \\ \lambda(T) &= \alpha_T(y(T) - y_T) + z_T.\end{aligned}$$

The existence of a unique optimal control of the problem (P_z) cannot be guaranteed by the coercivity assumption (SSC). There, positivity of $\mathcal{L}_{vv}$ was assumed only for the subspace of directions, where the control $\bar u$ is not strong active. Hence, the optimization problem (P_z) is nonconvex in general. At this point, our method of proof differs from that of [13]: There a stronger coercivity assumption was used, which ensures the convexity of (P_z).

We will circumvent this difficulty in the following way: First, we show the existence of a unique solution if we substitute the control constraint by

$$(\widetilde{P_z})\qquad u \in \widetilde{U_{ad}} = \{v \in U_{ad} :\ v_i(x,t) = \bar u_i(x,t) \text{ iff } (x,t) \in Q_{\varepsilon,i}\}.$$

This solution also will be a solution to (P_z) provided the perturbations are small. In what follows, we will denote by $(\widetilde{P_z})$ the linear-quadratic optimization problem (P_z) with a changed set of admissible controls $\widetilde{U_{ad}}$. For the solvability of $(\widetilde{P_z})$, we have the following.

THEOREM 5.2. *Let (SSC) be satisfied for the reference solution $\bar v = (\bar y, \bar u)$ with adjoint state $\bar\lambda$. Moreover, assume that the data satisfy $y_0, y_T \in W_0^{2-2/p,\,p}(\Omega)^2$, $y_Q \in L^p(Q)^2$ with $2 < p < \infty$, and that the bounds u_a, u_b are in $L^\infty(Q)^2$.*

Then problem $(\widetilde{P_z})$ admits a unique solution (y_z, u_z, λ_z). Moreover, the solution mapping $z \to (y_z, u_z, \lambda_z)$ is Lipschitz continuous from Z to $W_p^{2,1} \times L^\infty(Q)^2 \times W_p^{2,1}$.

Proof. Let us denote the Lagrangian associated with (P_z) by $\mathcal{L}^{(z)}$. Then it holds for all y, u, λ that

$$(5.11)\qquad \mathcal{L}_{vv}^{(z)}(y,u,\lambda) = \mathcal{L}_{vv}(\bar y, \bar u, \bar\lambda).$$

Now take two controls $u_1, u_2 \in \widetilde{U_{ad}}$ with associated solutions y_1, y_2 of (5.9). Then the pair $(y_1 - y_2, u_1 - u_2)$ fits into the assumption of (SSC), and we find

$$\mathcal{L}_{vv}^{(z)}(y,u,\lambda)[(y_1-y_2,u_1-u_2)]^2 = \mathcal{L}_{vv}(\bar{y},\bar{u},\bar{\lambda})[(y_1-y_2,u_1-u_2)]^2 \geq \delta\|u_1-u_2\|_2^2.$$

Thus, the problem $(\widetilde{P_z})$ is convex on the space of admissible controls $\widetilde{U_{ad}}$, which yields the existence of a unique optimal control u_z with state y_z and adjoint λ_z.

Let be given two perturbation vectors $z_1,\ z_2 \in Z$. Denote by $u_i := u_{z_i}$ the optimal control of the associated problems $(\widetilde{P_z})$ together with states $y_i := y_{z_i}$ and adjoints $\lambda_i := \lambda_{z_i}$ for $i = 1, 2$.

Applying the convexity of the Lagrangian, the Lipschitz estimate

$$\|u_1-u_2\|_2 + \|y_1-y_2\|_{H^{2,1}} + \|\lambda_1-\lambda_2\|_{H^{2,1}} \leq c\|z_1-z_2\|_Z$$

can be proved following the lines of similar proofs in [13, 29]. It remains to show Lipschitz continuity of the solution mapping in stronger norms.

The space $H^{2,1}$ is continuously imbedded in $L^q(Q)^2$ for all $q < \infty$. Now, we get Lipschitz continuity of the mapping $z \mapsto u_z$ using the projection formula (3.5). Since the pointwise projection is Lipschitz continuous in L^p-spaces, we find

$$\begin{aligned}\|u_1-u_2\|_p &= \left\|\mathrm{Proj}_{\widetilde{U_{ad}}}\left(-\frac{1}{\gamma}\lambda_1\right) - \mathrm{Proj}_{\widetilde{U_{ad}}}\left(-\frac{1}{\gamma}\lambda_2\right)\right\|_p \\ &\leq \frac{1}{\gamma}\|\lambda_1-\lambda_2\|_p \leq c\|\lambda_1-\lambda_2\|_{H^{2,1}} \leq c\|z_1-z_2\|_Z.\end{aligned}$$

The states y_i are solutions of the linearized equation (5.9) with control u_i and perturbed data z_i. This equation is linearized around the state $\bar{y}$. The assumptions of the regularity result of [29] are fulfilled, which gives us $\bar{y} \in W_p^{2,1}$. Thus, we can apply Theorem 2.7 to obtain, for the difference $y_1 - y_2$,

$$\|y_1-y_2\|_{W_p^{2,1}} \leq c\left(\|u_1-u_2\|_p + \|z_1-z_2\|_Z\right) \leq c\|z_1-z_2\|_Z.$$

Finally, we give a Lipschitz estimate for the adjoint states. These are solutions of the adjoint system (5.10). Here, we make use of Theorem 3.3 to conclude

$$\|\lambda_1-\lambda_2\|_{W_p^{2,1}} \leq c\left(\|y_1-y_2\|_{W_p^{2,1}} + \|z_1-z_2\|_Z\right) \leq c\|z_1-z_2\|_Z.$$

After all, we proved the Lipschitz continuity of the solution mapping $z \to (y_z, u_z, \lambda_z)$ associated with the optimal control problem $(\widetilde{P_z})$. □

Note that by using Hilbert space methods it is not possible to derive such a result for the constrained optimal control problem of nonstationary Navier–Stokes equations. This is due to the following: Every regularity result for the nonstationary Navier–Stokes system in Hilbert spaces that gives solutions in $L^\infty(Q)^2$ or $C(\bar{Q})^2$ requires a certain regularity of *derivatives* of the control. But it is impossible to get Lipschitz estimates of the control with respect to $W^{1,p}$-norms, because such an estimation has to be based on the projection formula, which is not Lipschitz continuous between $W^{1,p}$-spaces. This means that one cannot get L^∞-Lipschitz estimates for the state and the adjoint and consequently for the control using Hilbert space theory.

Now, we study the behavior of u_z on the active set Q_ε. To this aim, we have to rely on the L^∞-stability result of the previous theorem.

COROLLARY 5.3. *Let the assumptions of Theorem* 5.2 *be fulfilled. Then there exist* $\rho_z > 0$ *such that for all* $z \in Z$ *with* $\|z\|_Z < \rho_z$ *the optimal control of* $(\widetilde{P_z})$, u_z, *is strongly active a.e. on* $Q_{\varepsilon,i}$; *i.e., it holds that*

$$|\gamma u_{z,i}(x,t) + \lambda_{z,i}(x,t) - z_{u,i}(x,t)| > \frac{\varepsilon}{2},$$

and the signs of $(\gamma\bar{u}_i(x,t) + \bar{\lambda}_i(x,t))$ *and* $(\gamma u_{z,i}(x,t) + \lambda_{z,i}(x,t) - z_{u,i}(x,t))$ *coincide a.e. on* $Q_{\varepsilon,i}$ *for* $i = 1, 2$.

Proof. By Theorem 5.2, the mapping $z \mapsto (y_z, u_z, \lambda_z)$ is Lipschitz continuous from Z to $W_p^{2,1} \times L^\infty(Q)^2 \times W_p^{2,1}$. By imbedding arguments, we find that $z \mapsto \gamma u_z + \lambda_z - z_u$ is Lipschitz as it maps Z to $L^\infty(Q)^2$.

Let $(x,t) \in Q_{\varepsilon,i}$ such that $\gamma\bar{u}_i(x,t) + \bar{\lambda}_i(x,t) > \varepsilon$. Using this, we derive

$$\begin{aligned}
\varepsilon &< \gamma\bar{u}_i(x,t) + \bar{\lambda}_i(x,t) \\
&= \gamma\bar{u}_i(x,t) + \bar{\lambda}_i(x,t) - (\gamma u_{z,i}(x,t) + \lambda_{z,i}(x,t) - z_{u,i}(x,t)) \\
&\qquad + (\gamma u_{z,i}(x,t) + \lambda_{z,i}(x,t) - z_{u,i}(x,t)) \\
&\le c\|z\|_Z + \gamma u_{z,i}(x,t) + \lambda_{z,i}(x,t) - z_{u,i}(x,t).
\end{aligned}$$

Therefore, the choice $\rho_z := c^{-1}\varepsilon/2$ yields $\gamma u_{z,i}(x,t) + \lambda_{z,i}(x,t) - z_{u,i}(x,t) > \varepsilon/2$.

Analogously, if for $(x,t) \in Q_{\varepsilon,i}$ we have $\gamma\bar{u}_i(x,t) + \bar{\lambda}_i(x,t) < -\varepsilon$, then the same value of ρ_z gives $\gamma u_{z,i}(x,t) + \lambda_{z,i}(x,t) - z_{u,i}(x,t) < -\varepsilon/2$. ∎

COROLLARY 5.4. *Let the assumptions of Theorem* 5.2 *be fulfilled. Then the control* u_z *associated with a perturbation* $z \in Z$ *with* $\|z\|_Z < \rho_z$, *where* ρ_z *is given by Corollary* 5.3, *fulfills the variational inequality*

$$(\gamma u_z + \lambda_z - z_u,\, u - u_z) \ge 0 \quad \forall u \in U_{ad}; \tag{5.12}$$

i.e., it satisfies the first-order necessary optimality condition of (P_z).

Proof. Let $u \in U_{ad}$ be given. We begin with

$$\begin{aligned}
\int_Q (\gamma u_{z,i} + \lambda_{z,i} - z_{u,i})(u_i - u_{z,i}) &= \int_{Q\setminus Q_{\varepsilon,i}} (\gamma u_{z,i} + \lambda_{z,i} - z_{u,i})(u_i - u_{z,i}) \\
&\quad + \int_{Q_{\varepsilon,i}} (\gamma u_{z,i} + \lambda_{z,i} - z_{u,i})(u_i - \bar{u}_i),
\end{aligned} \tag{5.13}$$

since $u_z \in \widetilde{U_{ad}}$ means $u_{z,i}(x,t) = \bar{u}_i(x,t)$ a.e. on $Q_{\varepsilon,i}$. The first integral is part of the first-order necessary optimality conditions of $(\widetilde{P_z})$. Therefore, it is nonnegative.

By Corollary 5.3, $(\gamma\bar{u}_i(x,t) + \bar{\lambda}_i(x,t))$ and $(\gamma u_{z,i}(x,t) + \lambda_{z,i}(x,t) - z_{u,i}(x,t))$ have the same sign a.e. on $Q_{\varepsilon,i}$. Furthermore, $\bar{u}_i(x,t)$ is active on this set, so that $u_i(x,t) - \bar{u}_i(x,t)$ always has the same sign regardless of the choice of $u_i(x,t)$. Since $\gamma\bar{u} + \bar{\lambda}$ satisfies $\int_{Q_{\varepsilon,i}} (\gamma\bar{u}_i + \bar{\lambda}_i)(u_i - \bar{u}_i) \ge 0$, the same is true for $\gamma u_z + \lambda_z - z_u$; i.e.,

$$\int_{Q_{\varepsilon,i}} (\gamma u_{z,i} + \lambda_{z,i} - z_{u,i})(u_i - \bar{u}_i) \ge 0$$

is satisfied. So, we proved that both integrals in (5.13) are nonnegative. Adding them, we derived the claim (5.12). ∎

So far, we showed that (y_z, u_z, λ_z) fulfills the optimality system of the perturbed problem (P_z), or equivalently, the linearized and perturbed generalized equation (5.6).

We have to ask whether it might be a local minimizer of (P_z). With the previous corollaries and the identity (5.11) we have all ingredients at hand to prove that (y_z, u_z, λ_z) satisfies a second-order sufficient optimality condition for the problem (P_z); i.e., it is indeed a locally optimal solution.

THEOREM 5.5. *Let the assumptions of Theorem* 5.2 *be fulfilled. Then, there are* $\rho_z, \rho_u > 0$ *such that the control* u_z *associated with a perturbation* $z \in Z$ *with* $\|z\|_Z < \rho_z$ *is a locally optimal solution of* (P_z)*, and it satisfies*

$$J^{(z)}(y_z, u_z) \le J^{(z)}(y, u)$$

for all $u \in U_{ad}$ *with* $\|u - u_z\|_\infty \le \rho_u$*. Here* y_z *and* y *are the solutions of* (5.9) *associated with the controls* u_z *and* u*.*

Proof. Let $u \in U_{ad}$ with $u_i = \bar{u}_i$ a.e. on $Q_{\varepsilon,i}$ be given. Denote by y the associated solution of (5.9). Set $h = u - u_z$ and $w = y - y_z$. This implies $h = 0$ a.e. on $Q_{\varepsilon,i}$. Therefore, h fits into the assumptions of (SSC). The triple $(\bar{y}, \bar{u}, \bar{\lambda})$ satisfies the second-order sufficient optimality condition (SSC), which means

$$\mathcal{L}^{(z)}_{vv}(y_z, u_z, \lambda_z)[(w,h)]^2 = \mathcal{L}_{vv}(\bar{y}, \bar{u}, \bar{\lambda})[(w,h)]^2 \ge \delta \|h\|_2^2. \tag{5.14}$$

Corollaries 5.3 and 5.4 and the coercivity relation (5.14) build up the second-order sufficient optimality condition connected with (P_z). Following the lines of [27], we conclude that u_z is locally optimal: There exists a constant $\rho_u > 0$ such that $J^{(z)}(y_z, u_z) \le J^{(z)}(y, u)$ holds for all $u \in U_{ad}$ with $\|u - u_z\|_\infty \le \rho_u$. ∎

COROLLARY 5.6. *Let the assumptions of Theorem* 5.2 *be fulfilled. Then the generalized equation* (5.3) *is strongly regular at* $(\bar{y}, \bar{u}, \bar{\lambda})$*.*

Proof. At first, the function F is a $C^{1,1}$-mapping in the setting (5.4), because all its components are linear with respect to the variables (y, u, λ) except the nonlinear term in the state equation. The mapping $y \mapsto B(y)$ is $C^{1,1}$ from $W_p^{2,1}$ to $L^p(Q)^2$.

Theorem 5.5 states that the perturbed linearized optimization problem (P_z) has a unique optimal solution in the ball $B_{L^\infty}(\bar{u}, \rho_u)$ for perturbations from $B_Z(0, \rho_z)$. By Theorem 5.2, the associated state y lies in the ball $B_{W_p^{2,1}}(\bar{y},\, c_y \rho_z)$, whereas the adjoint state λ_z is in $B_{W_p^{2,1}}(\bar{\lambda}, c_\lambda\, \rho_z)$. Here, c_y and c_λ are the Lipschitz constants given by Theorem 5.2. This altogether yields the unique solvability of the perturbed linearized generalized equation (5.6) in $B_{W_p^{2,1}}(\bar{y}, c_y\, \rho_z) \times B_{L^\infty}(\bar{u}, \rho_u) \times B_{W_p^{2,1}}(\bar{\lambda}, c_\lambda\, \rho_z)$ for perturbations from $B_Z(0, \rho_z)$. As already mentioned, the solution mapping $z \mapsto (y_z, u_z, \lambda_z)$ is Lipschitz. Therefore, all requirements for strong regularity are fulfilled. ∎

5.3. Local convergence of the SQP-type algorithm. With the help of the previous section, we are in the situation to apply the abstract convergence result of Theorem 5.1. However, we are not allowed to carry it over one-to-one. The SQP-method as stated in the beginning of section 5 requires us to find the *global* minimizer of the linear-quadratic subproblems (P^n). The analysis done so far guarantees only the existence of a *local* solution of those subproblems in the neighborhood of the reference control. Consequently, we have to modify the SQP-method to enforce the solutions of the subproblems to remain near the reference solution in the following way.

Given iterates y_n, u_n, λ_n, compute the next iterates $y_{n+1}, u_{n+1}, \lambda_{n+1}$ as the solution of (P^n) subject to the control constraint

$$u \in U_{ad}^\rho := U_{ad} \cap \{v \in L^\infty(Q)^2 :\ \|v - \bar{u}\|_\infty \le \rho\}. \tag{5.15}$$

See also [26], where those aspects are discussed in more detail.

Then Theorem 5.1 yields quadratic convergence in a neighborhood of the solution.

THEOREM 5.7. *Let the assumptions of Theorem* 5.2 *be satisfied. Then there is a constant* $\rho_s > 0$*, such that for every starting value* (y_1, u_1, λ_1) *with* $u_1 \in U_{ad}^{\rho_s}$ *the SQP-method with control constraint* (5.15) *generates a uniquely determined sequence* (y_n, u_n, λ_n) *with* $u_n \in U_{ad}^{\rho_s}$*, and it holds that*

$$\|y_{n+1} - \bar{y}\|_{W_p^{2,1}} + \|u_{n+1} - \bar{u}\|_\infty + \|\lambda_{n+1} - \bar{\lambda}\|_{W_p^{2,1}} \\ \leq c_s \left(\|y_n - \bar{y}\|_{W_p^{2,1}}^2 + \|u_n - \bar{u}\|_\infty^2 + \|\lambda_n - \bar{\lambda}\|_{W_p^{2,1}}^2 \right)$$

with a constant c_s *independently of* n*. Here,* y_n *and* λ_n *are the states and adjoints associated with the control* u_n*.*

The a priori unknown solution $\bar{u}$ appears in the definition of U_{ad}^{ρ}, which is necessary to establish the convergence theory. To overcome this difficulty, one has to use globalization techniques. For an application of a globalized SQP-method to compute optimal controls of nonstationary Navier–Stokes equations, we refer to [12]. However, in the numerical computations it was not necessary to enforce the method to stay in a neighborhood of the last iterate.

6. Numerical results. Here, we provide a computational example that confirms the convergence analysis of the SQP-method. The following control problem is given: We want to reduce the recirculation bubble after the backward-facing step. We try this by minimization of the objective functional

$$J(y,u) = \frac{1}{2} \int_{Q_c} |y(x,t) - y_Q(x,t)|^2 dxdt + \frac{\gamma}{2} \int_{Q_c} |u(x,t)|^2 dxdt$$

with $T = 1$ and $\gamma = 0.3$. The control has to satisfy $|u_i(x,t)| \leq 0.3$ a.e. on Q, $i = 1, 2$.

The computational domain Ω is the backward-facing step. Here, observation and control take place in the same part of the domain $Q_c = \Omega_c \times (0,T)$; cf. Figure 6.1.

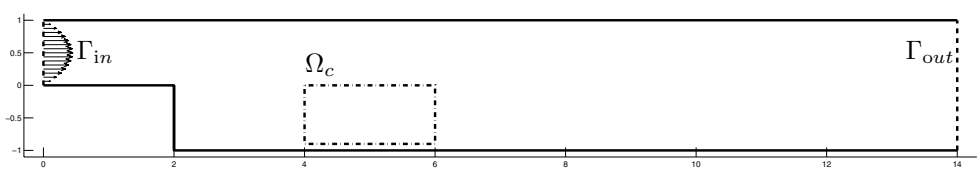


FIG. 6.1. *Flow configuration.*

As desired flow y_Q we chose the Stokes flow (see Figure 6.2), which is the solution of the stationary Stokes equation with the same boundary conditions as used for the nonstationary simulation.

At the inflow boundary Γ_{in} a parabolic velocity profile is prescribed, whereas at the boundary Γ_{out} we use the "do-nothing" boundary condition (cf. [11]): $\nu \frac{\partial y}{\partial n} - pn = 0$.

At the rest of the boundary we use homogeneous Dirichlet conditions. All computations were done with Reynolds number $Re = 400$, which yields a viscosity parameter $\nu = 1/400$. The initial velocity profile was chosen as the stationary limit of the uncontrolled Navier–Stokes equations; cf. Figure 6.3.

The continuous problem was discretized using Taylor–Hood finite elements with different mesh sizes. Further, we use a semi-implicit Euler scheme for time integration

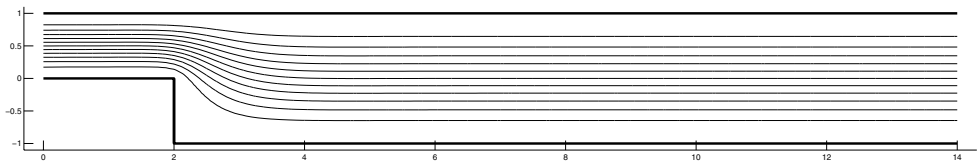


FIG. 6.2. *Desired profile is the Stokes flow.*

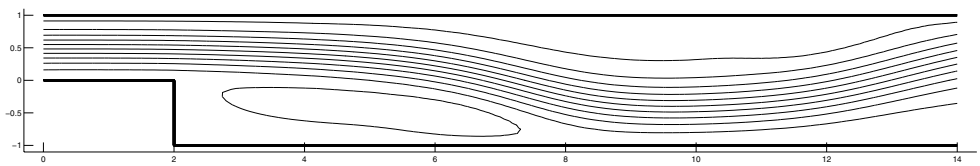


FIG. 6.3. *Initial flow profile* y_0.

with an equidistant time discretization with different step lengths. The computations are based on a finite element code of Hinze; see [16]. We computed solutions of the optimal control problem for different spatial and time discretizations. The discretization parameters can be found in Table 6.1.

TABLE 6.1. *Discretization parameters.*

	Triangles	Velocity nodes	Pressure nodes	Mesh size h	Time step τ
Coarse	416	905	245	0.5	0.01
	1664	3473	905	0.25	0.0025
Fine	6656	13601	3473	0.125	0.000625

On the finest discretization the number of unknown control variables is $1,388,800$, whereas the number of state and adjoint variables is each about 21 million. A further uniform refinement will result in an optimization problem that is very expensive to solve even on today's computers.

The arising discrete control problems are solved by the SQP-method without any globalization. The constrained SQP-subproblems (P^n) were solved by a primal-dual active-set method (see, for instance, [13, 18]) using the method of conjugate gradients (CGs) for the inner loop. Since those subproblems are linear-quadratic optimization problems, this active-set strategy can be interpreted as a semismooth Newton method [14] to solve the nonsmooth equation

$$u = \mathrm{Proj}_{U_{ad}}\left(-\frac{1}{\gamma}\lambda(u)\right);$$

cf. (3.5). Here, $\lambda(u)$ denotes the adjoint state for a given control u of the SQP-subproblem (P^n). This method is known to converge locally with a superlinear convergence rate [14] if the quadratic form $\mathcal{L}''$ is coercive. Under some strong assumptions it converges even globally [18]. Moreover, the SQP-method, as well as the semismooth Newton method, is known to exhibit a mesh-independence convergence; see [4, 15].

In all examples, the stopping criteria of the nested methods are balanced in the following way as proposed in [13].

The outer SQP-loop is terminated if two successive iterates are close enough:

$$\|y^n - y^{n-1}\|_\infty + \|u^n - u^{n-1}\|_\infty + \|\lambda^n - \lambda^{n-1}\|_\infty \le \varepsilon_{SQP}.$$

The primal-dual active-set method is stopped if either the active sets of two successive control iterates coincide or the error in the variational inequality given by

$$\phi(u) = \left\| u - \mathrm{Proj}_{U_{ad}}\left(-\frac{1}{\gamma}\lambda\right)\right\|_2$$

is reduced by a factor of 0.1. The innermost iteration procedure—the CG method—was stopped if the norm of the starting residual was reduced by a factor of 0.01. The initial guesses u^0 and y^0 for control and state were set to zero in all computations.

The convergence behavior of the method for the three different discretizations is listed in Table 6.2. We give an estimation of the convergence speed of the method with respect to the L^2- and L^∞-norms by

$$q_2^n = \frac{\|u^n - u^{n-1}\|_2}{\|u^{n-1} - u^{n-2}\|_2^2}, \quad q_\infty^n = \frac{\|u^n - u^{n-1}\|_\infty}{\|u^{n-1} - u^{n-2}\|_\infty^2}.$$

Now, let us have a look on the convergence history of the SQP-method for the three discretizations. It can be found in Table 6.2. The results for the two finest discretizations are close, such that one can see a mesh-independence behavior. The differences in the iterations on the coarse grid are due to the fact that the coarsest spatial grid was too coarse, and the accuracy of solving the state equation was too low. Mesh-independence results state that for discretizations that are sufficiently fine the iteration rates do not depend on the mesh; see, e.g., [4, 15].

TABLE 6.2. *Convergence history.*

Grid	Iteration	$\|u^n - u^{n-1}\|_\infty$	q_2^n	q_∞^n
Coarse	1	$3.00 \cdot 10^{-1}$		
	2	$1.92 \cdot 10^{-1}$	$5.87 \cdot 10^{-1}$	$2.14 \cdot 10^{0}$
	3	$2.24 \cdot 10^{-2}$	$3.73 \cdot 10^{0}$	$6.06 \cdot 10^{-1}$
	4	$1.24 \cdot 10^{-3}$	$1.32 \cdot 10^{1}$	$2.47 \cdot 10^{0}$
	5	$4.24 \cdot 10^{-5}$	$8.98 \cdot 10^{2}$	$2.76 \cdot 10^{1}$
	1	$3.00 \cdot 10^{-1}$		
	2	$1.97 \cdot 10^{-1}$	$6.24 \cdot 10^{-1}$	$2.19 \cdot 10^{0}$
	3	$3.71 \cdot 10^{-2}$	$4.70 \cdot 10^{0}$	$9.54 \cdot 10^{-1}$
	4	$1.49 \cdot 10^{-3}$	$4.74 \cdot 10^{0}$	$1.08 \cdot 10^{0}$
	5	$5.67 \cdot 10^{-5}$	$2.76 \cdot 10^{2}$	$2.49 \cdot 10^{1}$
Fine	1	$3.00 \cdot 10^{-1}$		
	2	$2.31 \cdot 10^{-1}$	$5.90 \cdot 10^{-1}$	$2.56 \cdot 10^{0}$
	3	$2.54 \cdot 10^{-2}$	$2.61 \cdot 10^{0}$	$4.76 \cdot 10^{-1}$
	4	$1.24 \cdot 10^{-3}$	$1.01 \cdot 10^{1}$	$1.91 \cdot 10^{0}$
	5	$5.89 \cdot 10^{-5}$	$2.12 \cdot 10^{2}$	$3.84 \cdot 10^{1}$

7. Conclusion. In this article, we investigated the SQP-method to solve optimal control problems with control constraints for the nonstationary Navier–Stokes equations. We were able to prove locally quadratic convergence of the SQP-method by using a sufficient condition, which is weaker than required in other articles. The control iterates will converge with respect to the L^∞-norm. The method of proof requires L^p-theory of the respective nonlinear and linearized state and adjoint equations, which were provided in the course of the article.

Acknowledgments. The author wishes to thank the two anonymous referees for their valuable comments leading to an improvement of the entire presentation.

REFERENCES

[1] F. Abergel and R. Temam, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Dynam., 1 (1990), pp. 303–325.
[2] R. A. Adams, *Sobolev Spaces*, Academic Press, San Diego, 1978.
[3] W. Alt, *The Lagrange-Newton method for infinite dimensional optimization problems*, Numer. Funct. Anal. Optim., 11 (1990), pp. 201–224.
[4] W. Alt, *Discretization and mesh-independence of Newton's method for generalized equations*, in Mathematical Programming with Data Perturbations, Marcel Dekker, New York, 1998, pp. 1–30.
[5] E. Casas, *An optimal control problem governed by the evolution Navier–Stokes equations*, in Optimal Control of Viscous Flow, S. S. Sritharan, ed., SIAM, Philadelphia, 1998, pp. 79–95.
[6] K. Deckelnick and M. Hinze, *Error estimates in space and time for tracking-type control of the instationary Stokes system*, in Control and Estimation of Distributed Parameter Systems, Internat. Ser. Numer. Math. 143, Birkhäuser Boston, Boston, MA, 2002, pp. 87–103.
[7] A. L. Dontchev, *Local analysis of a Newton-type method based on partial linearization*, in Proceedings of the AMS-SIAM Summer Seminar in Applied Mathematics, Lect. Appl. Math. 32, AMS, Providence, RI, 1996, pp. 295–306.
[8] H. O. Fattorini and S. Sritharan, *Necessary and sufficient conditions for optimal controls in viscous flow problems*, Proc. Royal Soc. Edinburgh Sect. A, 124 (1994), pp. 211–251.
[9] M. D. Gunzburger and S. Manservisi, *The velocity tracking problem for Navier–Stokes flows with bounded distributed controls*, SIAM J. Control Optim., 37 (1999), pp. 1913–1945.
[10] M. D. Gunzburger and S. Manservisi, *Analysis and approximation of the velocity tracking problem for Navier–Stokes flows with distributed control*, SIAM J. Numer. Anal., 37 (2000), pp. 1481–1512.
[11] J. Heywood, R. Rannacher, and S. Turek, *Artificial boundaries and flux and pressure conditions for the incompressible Navier–Stokes equations*, Int. J. Numer. Methods Fluids, 22 (1996), pp. 325–352.
[12] M. Hintermüller and M. Hinze, *Globalization of SQP-methods in control of the instationary Navier-Stokes equations*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 725–746.
[13] M. Hintermüller and M. Hinze, *A SQP-semismooth Newton-type algorithm applied to control of the instationary Navier–Stokes system subject to control constraints*, SIAM J. Optim., 16 (2006), pp. 1177–1200.
[14] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.
[15] M. Hintermüller and M. Ulbrich, *A mesh-independence result for semismooth Newton methods*, Math. Program. Ser. B, 101 (2004) pp. 151–184.
[16] M. Hinze, *Optimal and Instantaneous Control of the Instationary Navier–Stokes Equations*, Habilitation, Technische Universität Berlin, Berlin, 2002.
[17] M. Hinze and K. Kunisch, *Second-order methods for optimal control of time-dependent fluid flow*, SIAM J. Control Optim., 40 (2001), pp. 925–946.
[18] K. Kunisch and A. Rösch, *Primal-dual active set strategy for a general class of constrained optimal control problems*, SIAM J. Optim., 13 (2002), pp. 321–334.
[19] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.
[20] J. L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications,* Vol. I, Springer, Berlin, 1972.
[21] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[22] V. A. Solonnikov, *Estimates of the solutions of a nonstationary linearized system of Navier–Stokes equations*, Amer. Math. Soc. Transl., 75 (1968), pp. 1–116; translated from Trudy Mat. Inst. Steklov, 70 (1964), pp. 213–317.
[23] V. A. Solonnikov, *Estimates of solutions of nonstationary Navier–Stokes equations*, J. Sov. Math., 8 (1977), pp. 467–529.
[24] R. Temam, *Navier–Stokes Equations*, North–Holland, Amsterdam, 1979.
[25] H. Triebel, *Function spaces in Lipschitz domains and on Lipschitz manifolds. Characteristic functions as pointwise multipliers*, Rev. Mat. Complut., 15 (2002), pp. 475–524.

[26] F. Tröltzsch, *On the Lagrange–Newton–SQP method for the optimal control of semilinearparabolic equations*, SIAM J. Control Optim., 38 (1999), pp. 294–312.

[27] F. Tröltzsch and D. Wachsmuth, *Second-order sufficient optimality conditions for the optimal control of Navier–Stokes equations*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 93–119.

[28] M. Ulbrich, *Constrained optimal control of Navier–Stokes flow by semismooth Newton methods*, Systems Control Lett., 48 (2003), pp. 297–311.

[29] D. Wachsmuth, *Regularity and stability of optimal controls of instationary Navier–Stokes equations*, Control Cybernet., 34 (2005), pp. 387–410.

[30] W. von Wahl, *Instationary Navier–Stokes equations and parabolic systems*, Pacific J. Math., 72 (1977), pp. 557–569.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1155–1179

ASYMPTOTIC PROPERTIES OF HYBRID DIFFUSION SYSTEMS*

C. ZHU† AND G. YIN‡

Abstract. In response to the increasing needs for control and optimization of hybrid systems, this work is concerned with such asymptotic properties as recurrence (also known as weak stochastic stability in the literature) and ergodicity of regime-switching diffusions. Using Liapunov functions, necessary and sufficient conditions for positive recurrence are developed. Then, ergodicity of positive recurrent regime-switching diffusions is obtained by constructing cycles using the associated discrete-time Markov chains.

Key words. switching diffusion, Liapunov function, weak stochastic stability, positive recurrence, ergodicity

AMS subject classifications. 93D30, 93E03, 93E15, 60J60

DOI. 10.1137/060649343

1. Introduction.

1.1. Motivation. Owing to the increasing demands for modeling large-scale and complex systems, designing optimal control, and conducting optimization tasks, hybrid systems are lately receiving growing attention. A distinctive feature of these systems is the coexistence of continuous dynamics and discrete events. The collection of hybrid diffusion systems (also known as regime-switching diffusions) is such a class. Recent research efforts in such systems to capture random evolutions stem from emerging applications in financial engineering, wireless communications, manufacturing systems, and other related fields. For instance, there have been resurgent interests in using regime-switching diffusions to depict the financial market, where the switching or jump processes are used to describe stochastic volatility resulting from market modes, interest rates, as well as other economic factors. Regime-switching diffusions have also been used to enhance the versatility in risk management practice to better understand ruin probability in insurance and to carry out dividend optimization.

Since the underlying systems in applications are often in operation for a relatively long time, it is of foremost importance to understand the systems' asymptotic behavior. Considering average cost per unit time problems, we often wish to "replace" the time-dependent instantaneous measure by a steady state (or ergodic) measure. Thus we face the following questions. Do the systems possess ergodic property? Under what conditions do the systems have the desired ergodicity? In accordance with [22], a deterministic system $\dot{x} = g(t, x)$, which satisfies appropriate conditions, is Lagrange stable if the solutions are ultimately uniformly bounded. When stochastic systems are considered, almost sure boundedness excludes many systems. Thus, in lieu of such boundedness, one seeks stability in a certain weak sense [31]. One question of fundamental importance is, Under what conditions will the systems return to a prescribed

*Received by the editors January 8, 2006; accepted for publication (in revised form) February 13, 2007; published electronically September 5, 2007. This research was supported in part by the National Science Foundation under DMS-0603287 and in part by the National Security Agency under MSPF-068-029.

http://www.siam.org/journals/sicon/46-4/64934.html

†Department of Mathematical Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI 53201 (zhu@uwm.edu).

‡Department of Mathematics, Wayne State University, Detroit, MI 48202 (gyin@math.wayne.edu).

compact region in finite time? In this paper, we focus on asymptotic behaviors and address these issues. More specifically, we deal with such properties as recurrence, positive recurrence, and ergodicity. One of the main features of our approach is the use of appropriate Liapunov functions. We first develop Liapunov function–based general criteria for positive recurrence, followed by a further study on ergodicity in which we construct cycles using discrete-time Markov chains.

1.2. Formulation. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbf{P})$ be a complete probability space with a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ satisfying the usual condition (i.e., it is right continuous and $\mathcal{F}_0$ contains all $\mathbf{P}$-null sets). Let $x \in \mathbb{R}^r$, $\mathcal{M} = \{1, \ldots, m_0\}$, and $Q(x) = (q_{ij}(x))$ be an $m_0 \times m_0$ matrix depending on x and satisfying that for any $x \in \mathbb{R}^r$, $q_{ij}(x) \geq 0$ for $i \neq j$ and $\sum_{j=1}^{m_0} q_{ij}(x) = 0$. For each $i \in \mathcal{M}$, and for any twice continuously differentiable function $g(\cdot, i)$, define $\mathcal{L}$ by

$$\begin{aligned}\mathcal{L}g(x,i) &= \frac{1}{2}\sum_{j,k=1}^{r} a_{jk}(x,i)\frac{\partial^2 g(x,i)}{\partial x_j \partial x_k} + \sum_{j=1}^{r} b_j(x,i)\frac{\partial g(x,i)}{\partial x_j} + Q(x)g(x,\cdot)(i)\\ &= \frac{1}{2}\mathrm{tr}(a(x,i)\nabla^2 g(x,i)) + \big\langle b(x,i), \nabla g(x,i)\big\rangle + Q(x)g(x,\cdot)(i),\end{aligned} \tag{1.1}$$

where $\nabla g(\cdot, i)$ and $\nabla^2 g(\cdot, i)$ denote the gradient and Hessian of $g(\cdot, i)$, respectively, and

$$Q(x)g(x,\cdot)(i) = \sum_{j=1}^{m_0} q_{ij}(x)g(x,j) = \sum_{j\neq i, j\in\mathcal{M}} q_{ij}(x)(g(x,j) - g(x,i)), \quad i \in \mathcal{M}. \tag{1.2}$$

Consider a Markov process $Y(t) = (X(t), \gamma(t))$, whose associated operator is given by $\mathcal{L}$; see [28] for further references. Note that $Y(t)$ has two components, an r-dimensional diffusion component $X(t)$ and a jump component $\gamma(t)$ taking value in $\mathcal{M} = \{1, \ldots, m_0\}$.

The process $Y(t) = (X(t), \gamma(t))$ can be described by the following equations:

$$dX(t) = b(X(t), \gamma(t))dt + \sigma(X(t), \gamma(t))dw(t), \quad X(0) = x, \ \gamma(0) = \gamma, \tag{1.3}$$

and

$$\mathbf{P}\{\gamma(t+\Delta t) = j | \gamma(t) = i, X(s), \gamma(s), s \leq t\} = q_{ij}(X(t))\Delta t + o(\Delta t), \quad i \neq j, \tag{1.4}$$

where $w(t)$ is a d-dimensional standard Brownian motion, $b(\cdot,\cdot) : \mathbb{R}^r \times \mathcal{M} \mapsto \mathbb{R}^r$, and $\sigma(\cdot,\cdot) : \mathbb{R}^r \times \mathcal{M} \mapsto \mathbb{R}^{r\times d}$ satisfying $\sigma(x,i)\sigma'(x,i) = a(x,i)$ (where z' denotes the transpose of z for $z \in \mathbb{R}^{\iota_1 \times \iota_2}$ with $\iota_1, \ \iota_2 \geq 1$). We refer the reader to [28] for related stochastic differential equations involving Poisson measures describing the evolution of the jump processes. In this paper, our study will mainly be concerned with the use of the operator $\mathcal{L}$ given in (1.1). Throughout the paper, we assume that both $b(\cdot, i)$ and $\sigma(\cdot, i)$ satisfy the usual local Lipschitz condition and linear growth condition for each $i \in \mathcal{M}$ and that $Q(\cdot)$ is bounded and continuous. It is well known that under these conditions, the system (1.3)–(1.4) has a unique strong solution; see [12] or [28] for details. In what follows, denote the solution of (1.3)–(1.4) by $(X^{x,\gamma}(t), \gamma(t))$ if the emphasis on the initial data is needed. To study recurrence and ergodicity of the process $Y(t) = (X(t), \gamma(t))$, we further assume that the following condition (A) holds throughout the paper. For convenience, we also collect the boundedness and continuity of $Q(\cdot)$ in (A).

(A) The operator $\mathcal{L}$ satisfies the following conditions:

(i) For each $i \in \mathcal{M}$, $a(x,i) = (a_{jk}(x,i))$ is symmetric and satisfies

$$\kappa_1|\xi|^2 \le \langle a(x,i)\xi,\xi\rangle \le \kappa_1^{-1}|\xi|^2 \quad \text{for all } \xi \in \mathbb{R}^r, \tag{1.5}$$

with some constant $\kappa_1 \in (0,1]$ for all $x \in \mathbb{R}^r$.

(ii) For $i \ne j$, $q_{ij}(x) > 0$. The matrix-valued function $Q(\cdot)$ is bounded and continuous.

As mentioned earlier, the motivation of our study stems from recent interests in regime-switching diffusion processes that include a random process with a finite-state space in addition to the usual diffusion component. The finite-state process depicts a random environment that has right-continuous sample paths and that cannot be described by a diffusion. Consequently, both continuous dynamics (diffusions) and discrete events (jumps) coexist and yield hybrid dynamic systems, which provide a more realistic formulation for many applications.

Regime-switching diffusions lately have received much attention. For instance, optimal controls of switching diffusions were studied in [4] using a martingale problem formulation; jump-linear systems were considered in [13]; stability of semilinear stochastic differential equations with Markovian switching was considered in [2]; ergodic control problems of switching diffusions were studied in [9]; stability of stochastic differential equations with Markovian switching was treated in [23, 25, 33]; asymptotic expansions for solutions of integrodifferential equations for transition densities of singularly perturbed switching-diffusion processes were developed in [11]; and switching diffusions were used for stock liquidation models in [34]. For some recent applications of hybrid systems in communication networks, air traffic management, and control problems, etc., we refer the reader to [14, 15, 24, 26, 29] and references therein.

In [2, 23, 33, 34], $Q(x) = Q$, a constant matrix. In such cases, $\gamma(\cdot)$ is a continuous-time Markov chain. Moreover, it is assumed that the Markov chain $\gamma(\cdot)$ is independent of the Brownian motion. In our formulation, x-dependent $Q(x)$ is considered, and as a result, the transition rates of the discrete event $\gamma(\cdot)$ depend on the continuous dynamic $X(\cdot)$, as depicted in (1.4). Although the pair $(X(\cdot),\gamma(\cdot))$ is a Markov process, for x-dependent $Q(x)$, only for each fixed x, the discrete-event process $\gamma(\cdot)$ is a Markov chain. Such a formulation enables us to describe complex systems and their inherent uncertainty and randomness in the environment. However, it adds much difficulty to our analysis. Our formulation is motivated by the fact that in many applications, the discrete event and continuous dynamic are intertwined, and the independence assumption of the discrete-event process and the Brownian motion appears to be restrictive.

One of the important problems concerning switching models is their longtime behavior. Despite the growing interests in treating regime-switching systems (see the works mentioned in the previous paragraphs and references therein), the results regarding such issues as recurrence and positive recurrence (or weak stochastic stability as termed in [31]) are still scarce. Furthermore, these are not simple extensions of their diffusion counterparts. Due to the coupling and interactions, elliptic systems instead of a single elliptic equation must be treated. Moreover, even though the classical approaches such as Liapunov function methods and Dynkin's formula are still applicable for switching diffusions, the analysis is much more delicate than the diffusion counterparts. It requires careful handling of discrete-event component $\gamma(\cdot)$; see, for example, the proofs of Lemma 3.7, Lemma 3.8, and Theorem 3.12.

In addition to recurrence, many applications in control and optimization require minimizing an expected cost of a certain objective function. The computation is

difficult and complicated. Significant effort has been devoted to approximating such expected values by replacing the measures with stationary measures when the time horizon is long enough. To justify such a replacement, ergodicity is needed. For diffusion processes, much effort has been devoted to ergodicity; see, for example, [3, 20], among others. For regime-switching diffusions, asymptotic stability for the density of the so-called two-state diffusion process $(X(t), \gamma(t))$ was established in [25]; asymptotic stability in distribution for the process $(X(t), \gamma(t))$ was obtained in [33], where the jump component $\gamma(\cdot)$ is generated by some constant matrix Q and is independent of the Brownian motion. In this work, we will address ergodicity for $(X(t), \gamma(t))$ under different conditions than those in [25, 33]. Moreover, our work is applicable to more general settings. The discrete component $\gamma(\cdot)$ has an x-dependent generator $Q(x)$ and takes value in a finite-state space $\mathcal{M} = \{1, 2, \ldots, m_0\}$. Another highlight of this paper is that we obtain the explicit representation of the invariant measure of the process $(X(t), \gamma(t))$ by considering certain cylinder sets and by defining cycles appropriately. As a by-product, we demonstrate a strong law of large numbers type of theorem for positive recurrent regime-switching diffusions.

Compared with the existing work in the literature, the novelty and contribution of this paper are as follows. (a) By considering the x-dependent generator $Q(x)$, our model provides a more realistic formulation which allows the switching component to depend on the continuous states. This, in turn, allows for the coupling and correlation between $X(t)$ and $\gamma(t)$. (b) By appropriately defining cycles, we establish the ergodicity of the underlying process. (c) Moreover, explicit representation of the invariant measure for positive recurrent regime-switching diffusions is given.

The rest of the paper is arranged as follows. In section 2, in addition to introducing certain notation, we also provide definitions of regularity, recurrence, positive recurrence, and null recurrence. Section 3 focuses on positive recurrence. We present results of necessary and sufficient conditions for recurrence using Liapunov functions, along with two examples as applications of the general results. Section 4 develops ergodicity of switching-diffusion processes. Discussions and further remarks are made in section 5. An appendix containing the proofs of several technical lemmas is placed at the end of the paper to facilitate the reading.

2. Regularity, recurrence, positive recurrence, and null recurrence. This section is devoted to the definitions of regularity, recurrence, positive recurrence, and null recurrence. For simplicity, we introduce some notation as follows. For any $U = D \times J \subset \mathbb{R}^r \times \mathcal{M}$, where $D \subset \mathbb{R}^r$ and $J \subset \mathcal{M}$, denote

$$\begin{aligned} \tau_U &:= \inf\{t \geq 0 : (X(t), \gamma(t)) \notin U\}, \\ \sigma_U &:= \inf\{t \geq 0 : (X(t), \gamma(t)) \in U\}. \end{aligned} \tag{2.1}$$

In particular, if $U = D \times \mathcal{M}$ is a "cylinder," we set

$$\begin{aligned} \tau_D &:= \inf\{t \geq 0 : X(t) \notin D\}, \\ \sigma_D &:= \inf\{t \geq 0 : X(t) \in D\}. \end{aligned} \tag{2.2}$$

Definition 2.1 (regularity). *A Markov process $(X^{x,\gamma}(t), \gamma(t))$ is said to be* regular*, if for any $0 < T < \infty$,*

$$\mathbf{P}\left\{\sup_{0 \leq t \leq T} |X^{x,\gamma}(t)| = \infty\right\} = 0. \tag{2.3}$$

Remark 2.2. Let β_n be the first exit time of the process $(X^{x,\gamma}(t), \gamma(t))$ from the bounded set $\{\widetilde{x} : |\widetilde{x}| < n\} \times \mathcal{M}$, that is,

$$\beta_n = \inf\{t : |X^{x,\gamma}(t)| = n\}. \tag{2.4}$$

Then, the sequence $\{\beta_n\}$ is monotonically increasing and hence has a (finite or infinite) limit. It is not difficult to see that the process $(X^{x,\gamma}(t), \gamma(t))$ is regular if and only if

$$\beta_n \to \infty \text{ almost surely (a.s.) as } n \to \infty. \tag{2.5}$$

In what follows, we assume that the process $(X^{x,\gamma}(t), \gamma(t))$ is regular. Subsequently we will use (2.5) often.

DEFINITION 2.3. *Recurrence, positive recurrence, and null recurrence are defined as follows.*

(i) Recurrence. *For $U := D \times J$, where $J \subset \mathcal{M}$ and $D \subset \mathbb{R}^r$ is an open set with compact closure, let $\sigma_U^{x,\gamma} = \inf\{t : (X^{x,\gamma}(t), \gamma(t)) \in U\}$. A regular process $(X^{x,\gamma}(\cdot), \gamma(\cdot))$ is* recurrent *with respect to U if $\mathbf{P}\{\sigma_U^{x,\gamma} < \infty\} = 1$ for any $(x, \gamma) \in D^c \times \mathcal{M}$, where D^c denotes the complement of D.*

(ii) Positive recurrence and null recurrence. *A recurrent process with finite mean recurrence time for some set $U = D \times J$, where $J \subset \mathcal{M}$ and $D \subset \mathbb{R}^r$ is a bounded open set with compact closure, is said to be* positive recurrent *with respect to U; otherwise, the process is* null recurrent *with respect to U.*

3. Positive recurrence. This section takes up the positive recurrence issue. It entails the use of appropriate Liapunov functions. We begin this section with certain preparatory results, which indicate that the process $Y(t) = (X(t), \gamma(t))$ is recurrent (resp., positive recurrent) with respect to some "cylinder" $D \times \mathcal{M}$ if and only if it is recurrent (resp., positive recurrent) with respect to $D \times \{\ell\}$, where $D \subset \mathbb{R}^r$ is a nonempty open set with compact closure and $\ell \in \mathcal{M}$. We will also prove that the properties of recurrence and positive recurrence do not depend on the choice of the open set $D \subset \mathbb{R}^r$ or $\ell \in \mathcal{M}$. After the preparatory results, two subsections follow. The first presents Liapunov function–based criteria on positive recurrence. As applications of the general results, a subsection containing two examples is provided. Note that Example 3.16 is quite interesting because it shows that the combination of a transient diffusion and a positive recurrent diffusion is a positive recurrent switching diffusion.

3.1. Preparatory results. We first prove the following theorem, which asserts that under assumption (A), the process $Y(t) = (X(t), \gamma(t))$ will exit every bounded "cylinder" with a finite mean exit time.

THEOREM 3.1. *Let $D \subset \mathbb{R}^r$ be a nonempty open set with compact closure $\overline{D}$. Let $\tau_D := \inf\{t \geq 0 : X(t) \notin D\}$. Then*

$$\mathbf{E}_{x,i}\tau_D < \infty \quad \textit{for any} \quad (x, i) \in D \times \mathcal{M}. \tag{3.1}$$

Proof. First, note that from the uniform ellipticity condition (1.5), we have

$$\kappa_1 \leq a_{11}(x, i) \leq \kappa^{-1} \quad \text{for any} \quad (x, i) \in D \times \mathcal{M}. \tag{3.2}$$

For each $i \in \mathcal{M}$, consider $W(x, i) = k - (x_1 + \beta)^c$, where the constants k, c (with $c \geq 2$), and β are to be specified, and $x_1 = e_1'x$ is the first component of x, where $e_1 = (1, 0, \ldots, 0)'$. Direct computation leads to

$$\mathcal{L}W(x, i) = -c(x_1 + \beta)^{c-2}\left[b_1(x, i)(x_1 + \beta) + \frac{c-1}{2}a_{11}(x, i)\right].$$

Set

$$c = \frac{2}{\kappa_1}\left(\sup_{(x,i)\in\overline{D}\times\mathcal{M}} |b_1(x,i)(x_1+\beta)| + 1\right) + 1.$$

Then we have from (3.2) that

$$\frac{c-1}{2}a_{11}(x,i) + b_1(x,i)(x_1+\beta) \geq \frac{c-1}{2}\kappa_1 - \sup_{(x,i)\in\overline{D}\times\mathcal{M}} |b_1(x,i)(x_1+\beta)| \geq 1.$$

Meanwhile, since $x \in D \subset \overline{D}$ and $\overline{D}$ is compact, we can choose β such that $1 \leq x_1 + \beta \leq M$ for all $x \in D$, where M is some positive constant. Thus we have $(x_1+\beta)^{c-2} \geq 1^{c-2} = 1$. Finally, we choose k large enough so that $W(x,i) = k - (x_1+\beta)^c > 0$ for all $(x,i) \in D \times \mathcal{M}$. Therefore, $W(x,i), i \in \mathcal{M}$, are Liapunov functions satisfying

$$\mathcal{L}W(x,i) \leq -c \quad \text{for all} \quad (x,i) \in D \times \mathcal{M}. \tag{3.3}$$

Now let $\tau_D(t) := \min\{t, \tau_D\}$. Then we have from Dynkin's formula and (3.3) that

$$\begin{aligned}&\mathbf{E}_{x,i}W(X(\tau_D(t)),\gamma(\tau_D(t))) - W(x,i)\\ &\quad = \mathbf{E}_{x,i}\int_0^{\tau_D(t)} \mathcal{L}W(X(u),\gamma(u))du \leq -c\mathbf{E}_{x,i}\tau_D(t).\end{aligned}$$

Since the function $W(x,i)$ is nonnegative, we have

$$\mathbf{E}_{x,i}\tau_D(t) \leq \frac{1}{c}W(x,i). \tag{3.4}$$

Because

$$\mathbf{E}_{x,i}\tau_D(t) = \mathbf{E}_{x,i}\tau_D\chi_{[\tau_D\leq t]} + \mathbf{E}_{x,i}t\chi_{[\tau_D>t]},$$

we have from (3.4) that $t\mathbf{P}_{x,i}[\tau_D > t] \leq \frac{1}{c}W(x,i)$. Letting $t \to \infty$, we obtain $\mathbf{P}_{x,i}[\tau_D = \infty] = 0$ or $\mathbf{P}_{x,i}[\tau_D < \infty] = 1$. This yields that $\tau_D(t) \to \tau_D$ a.s. $\mathbf{P}_{x,i}$ as $t \to \infty$. Now applying Fatou's lemma, as $t \to \infty$, we obtain

$$\mathbf{E}_{x,i}\tau_D \leq \frac{1}{c}W(x,i) < \infty,$$

as desired. □

Remark 3.2. A closer examination of the proof shows that the conclusion of Theorem 3.1 remains valid if we replace the uniform ellipticity condition (1.5) by a weaker condition as follows: There exist some $\iota = 1, 2, \ldots, r$ and positive constant κ such that

$$a_{\iota\iota}(x,i) \geq \kappa \quad \text{for any} \quad (x,i) \in D \times \mathcal{M}. \tag{3.5}$$

To facilitate subsequent discussions, in what follows we present a twice continuously differentiable (with respect to the variable x) function $u(\cdot,\cdot) : \mathbb{R}^r \times \mathcal{M} \mapsto \mathbb{R}$ that is called $\mathcal{L}$-*harmonic* in a domain $U \subset \mathbb{R}^r \times \mathcal{M}$ if $\mathcal{L}u(x,i) = 0$ for all $(x,i) \in U$. Following the well-known arguments in [8, Vol. II, Chapter 13], we obtain the following two lemmas. (Note that Lemma 3.3 was also proved in [9, Lemma 4.3], and in [6] for the case when the operator $\mathcal{L}$ is in divergence form.)

LEMMA 3.3. *Let $U = D \times \mathcal{M} \subset \mathbb{R}^r \times \mathcal{M}$, where $D \subset \mathbb{R}^r$ is a nonempty open set. Assume that $\mathbf{P}_{x,i}\{\tau_U < \infty\} = 1$ for any $(x,i) \in D \times \mathcal{M}$ and that $f(\cdot,i) \in C^2(D) \cap C(\overline{D})$ for each $i \in \mathcal{M}$. Then*

$$\mathcal{L}f(x,i) = 0 \quad \text{for any} \quad (x,i) \in D \times \mathcal{M} \tag{3.6}$$

if and only if

$$f(x,i) = \mathbf{E}_{x,i} f(X(\tau_U), \gamma(\tau_U)) = \sum_{j=1}^{m_0} \int_{\partial D} \mathbf{P}_{x,i} \left\{ (X(\tau_U), \gamma(\tau_U)) \in (dy \times \{j\}) \right\} f(y,j).$$

Moreover, we further assume that ∂D is sufficiently smooth, $\overline{D}$ is compact, and $\varphi(\cdot,i)$ is an arbitrary continuous function on ∂D for any $i \in \mathcal{M}$. Then

$$f(x,i) = \mathbf{E}_{x,i} \varphi(X(\tau_U), \gamma(\tau_U)) \tag{3.7}$$

is the unique solution of the differential equation (3.6) *with boundary condition*

$$\lim_{x \to x_0} f(x,i) = \varphi(x_0,i) \quad \text{for any} \quad (x_0,i) \in \partial D \times \mathcal{M}. \tag{3.8}$$

LEMMA 3.4. *Let $U = D \times \mathcal{M} \subset \mathbb{R}^r \times \mathcal{M}$, where $D \subset \mathbb{R}^r$ is a nonempty open set with compact closure. Suppose $g(\cdot,i) \in C_b(\overline{D})$ and $f(\cdot,i) \in C^2(D)$ for each $i \in \mathcal{M}$. Then f solves the boundary value problem*

$$\begin{cases} \mathcal{L}f(x,i) = -g, & (x,i) \in D \times \mathcal{M}, \\ f(x,i) = 0, & (x,i) \in \partial D \times \mathcal{M} \end{cases}$$

if and only if

$$f(x,i) = \mathbf{E}_{x,i} \int_0^{\tau_U} g(X(t), \gamma(t)) dt \quad \text{for all} \quad (x,i) \in D \times \mathcal{M}.$$

Using Lemmas 3.3 and 3.4, we proceed to prove that if the process $Y(t) = (X(t), \gamma(t))$ is recurrent (resp., positive recurrent) with respect to some "cylinder" $D \times \mathcal{M} \subset \mathbb{R}^r \times \mathcal{M}$, then it is recurrent (resp., positive recurrent) with respect to any "cylinder" $E \times \mathcal{M} \subset \mathbb{R}^r \times \mathcal{M}$, where D is any nonempty domain in $\mathbb{R}^r$ with compact closure. These results are proved in the following two lemmas. To preserve the flow of presentation, the proofs are postponed to the appendix.

LEMMA 3.5. *Let $D \subset \mathbb{R}^r$ be a nonempty open set with compact closure. Suppose that*

$$\mathbf{P}_{x,i}\{\sigma_D < \infty\} = 1 \quad \text{for any} \quad (x,i) \in D^c \times \mathcal{M}. \tag{3.9}$$

Then for any nonempty open set $E \subset \mathbb{R}^r$, we have

$$\mathbf{P}_{x,i}\{\sigma_E < \infty\} = 1 \quad \text{for any} \quad (x,i) \in E^c \times \mathcal{M}.$$

LEMMA 3.6. *Let $D \subset \mathbb{R}^r$ be a nonempty open set with compact closure. Suppose that*

$$\mathbf{E}_{x,i} \sigma_D < \infty \quad \text{for any} \quad (x,i) \in D^c \times \mathcal{M}. \tag{3.10}$$

Then for any nonempty open set $E \subset \mathbb{R}^r$, we have

$$\mathbf{E}_{x,i}\sigma_E < \infty \quad \text{for any} \quad (x,i) \in E^c \times \mathcal{M}.$$

The following lemma shows that if the process $Y(t) = (X(t), \gamma(t))$ reaches the "cylinder" $D \times \mathcal{M}$ in finite time a.s. $\mathbf{P}_{x,i}$, then it will visit the set $D \times \{\ell\}$ in finite time a.s. $\mathbf{P}_{x,i}$ for any $\ell \in \mathcal{M}$. Its proof, together with the proof of Lemma 3.8, is also placed in the appendix.

LEMMA 3.7. *Let $D \subset \mathbb{R}^r$ be a nonempty open set with compact closure satisfying*

$$\mathbf{P}_{y,j}\{\sigma_D < \infty\} = 1 \quad \text{for any} \quad (y,j) \in D^c \times \mathcal{M}. \tag{3.11}$$

Then for any $(x,i) \in \mathbb{R}^r \times \mathcal{M}$,

$$\mathbf{P}_{x,i}\{\sigma_{D,\ell} < \infty\} = 1 \quad \text{for any} \quad \ell \in \mathcal{M}. \tag{3.12}$$

With Lemma 3.7, we can now prove that if the process $Y(t) = (X(t), \gamma(t))$ is positive recurrent with respect to some "cylinder" $D \times \mathcal{M}$, then it is positive recurrent with respect to the set $D \times \{\ell\} \subset \mathbb{R}^r \times \mathcal{M}$.

LEMMA 3.8. *Let $D \subset \mathbb{R}^r$ be a nonempty open set with compact closure satisfying*

$$\mathbf{E}_{y,j}\sigma_D < \infty \quad \text{for any} \quad (y,j) \in D^c \times \mathcal{M}. \tag{3.13}$$

Then for any $(x,i) \in \mathbb{R}^r \times \mathcal{M}$,

$$\mathbf{E}_{x,i}\sigma_{D,\ell} < \infty \quad \text{for any} \quad \ell \in \mathcal{M}. \tag{3.14}$$

Remark 3.9. By virtue of Lemmas 3.5–3.8, under assumption (A), the process $Y(t) = (X(t), \gamma(t))$ is recurrent (resp., positive recurrent) with respect to some "cylinder" $D \times \mathcal{M}$ if and only if it is recurrent (resp., positive recurrent) with respect to the product set $D \times \{\ell\} \subset \mathbb{R}^r \times \mathcal{M}$ for any $\ell \in \mathcal{M}$. Also we have proved that the properties of recurrence and positive recurrence are independent of the choice of the set D. We summarize these into the following theorem.

THEOREM 3.10. *Suppose that* (A) *holds. Then the following assertions hold:*

(i) *The process $Y(t) = (X(t), \gamma(t))$ is recurrent (resp., positive recurrent) with respect to $D \times \mathcal{M}$ if and only if it is recurrent (resp., positive recurrent) with respect to $D \times \{\ell\}$, where $D \subset \mathbb{R}^r$ is a nonempty open set with compact closure and $\ell \in \mathcal{M}$.*

(ii) *If the process $Y(t) = (X(t), \gamma(t))$ is recurrent (resp., positive recurrent) with respect to some $U = D \times \mathcal{M}$, where $D \subset \mathbb{R}^r$ is a nonempty open set with compact closure, then it is recurrent (resp., positive recurrent) with respect to any $\widetilde{U} = \widetilde{D} \times \mathcal{M}$, where $\widetilde{D} \subset \mathbb{R}^r$ is any nonempty open set.*

Remark 3.11. In view of Theorem 3.10, we make the following remarks.

(i) A regular process $Y(t) = (X(t), \gamma(t))$ with the associated generator $\mathcal{L}$ satisfying (A) is said to be *recurrent* if it is recurrent with respect to some $U = D \times \{\ell\}$, where $D \subset \mathbb{R}^r$ is a nonempty bounded open set and $\ell \in \mathcal{M}$; otherwise it is said to be *transient*.

(ii) Henceforth, we call a recurrent process $Y(t) = (X(t), \gamma(t))$ *positive recurrent* if it is positive recurrent with respect to some bounded domain $U = D \times \{\ell\} \subset \mathbb{R}^r \times \mathcal{M}$; otherwise, we have a *null recurrent* process.

3.2. General criteria.

THEOREM 3.12. *A necessary and sufficient condition for positive recurrence with respect to a domain $U = D \times \{\ell\} \subset \mathbb{R}^r \times \mathcal{M}$ is that for each $i \in \mathcal{M}$, there exists a nonnegative function $V(\cdot, i) : D^c \mapsto \mathbb{R}$ such that $V(\cdot, i)$ is twice continuously differentiable and that*

$$\mathcal{L}V(x,i) = -1, \quad (x,i) \in D^c \times \mathcal{M}. \tag{3.15}$$

Let $u(x,i) = \mathbf{E}_{x,i}\sigma_D$. Then $u(x,i)$ is the smallest positive solution of

$$\begin{cases} \mathcal{L}u(x,i) = -1, & (x,i) \in D^c \times \mathcal{M}, \\ u(x,i) = 0, & (x,i) \in \partial D \times \mathcal{M}, \end{cases} \tag{3.16}$$

where ∂D denotes the boundary of D.

Proof. The proof is organized into three steps.

Step 1. Show that the process $Y(t) = (X(t), \gamma(t))$ is positive recurrent if there exists a nonnegative function $V(\cdot, \cdot)$ satisfying the conditions of the theorem. Choose n_0 to be a positive integer sufficiently large so that $D \subset \{|x| < n_0\}$. Fix any $(x,i) \in D^c \times \mathcal{M}$. For any $t > 0$ and $n \in \mathbb{N}$ with $n > n_0$, we define

$$\sigma_D^{(n)}(t) = \min\{\sigma_D, t, \beta_n\},$$

where β_n is defined as in (2.4) and σ_D is the first entrance time to D. That is, $\sigma_D = \inf\{t : X(t) \in D\}$. Now Dynkin's formula and (3.15) imply that

$$\begin{aligned} &\mathbf{E}_{x,i} V\Big(X\Big(\sigma_D^{(n)}(t)\Big), \gamma\Big(\sigma_D^{(n)}(t)\Big)\Big) - V(x,i) \\ &\quad = \mathbf{E}_{x,i} \int_0^{\sigma_D^{(n)}(t)} \mathcal{L}V(X(s), \gamma(s))ds = -\mathbf{E}_{x,i}\sigma_D^{(n)}(t). \end{aligned}$$

Note that the function V is nonnegative; hence we have $\mathbf{E}_{x,i}\sigma_D^{(n)}(t) \le V(x,i)$. Meanwhile, since the process $Y(t) = (X(t), \gamma(t))$ is regular, it follows from (2.5) that $\sigma_D^{(n)}(t) \to \sigma_D(t)$ a.s. as $n \to \infty$, where $\sigma_D(t) = \min\{\sigma_D, t\}$. By virtue of Fatou's lemma, we obtain

$$\mathbf{E}_{x,i}\sigma_D(t) \le V(x,i). \tag{3.17}$$

Now the argument after (3.4) in the proof of Theorem 3.1 yields that $\mathbf{E}_{x,i}\sigma_D \le V(x,i) < \infty$. Then Lemma 3.8 implies that $\mathbf{E}_{x,i}\sigma_U = \mathbf{E}_{x,i}\sigma_{D,\ell} < \infty$. Since $(x,i) \in D^c \times \mathcal{M}$ is arbitrary, we conclude that $Y(t)$ is positive recurrent with respect to U.

Step 2. Show that $u(x,i) := \mathbf{E}_{x,i}\sigma_D$ is the smallest positive solution of (3.16). To this end, let n_0 be defined as before, that is, a positive integer sufficiently large so that $D \subset \{|x| < n_0\}$. For $n \ge n_0$, set $\sigma_D^{(n)} = \min\{\sigma_D, \beta_n\}$. Clearly, we have $\sigma_D^{(n)} \le \sigma_D^{(n+1)}$ for all $n \ge n_0$. Then the regularity of the process $Y(t)$ implies that $\sigma_D^{(n)} \nearrow \sigma_D$ a.s. as $n \to \infty$. Hence the monotone convergence theorem implies that as $n \to \infty$,

$$\mathbf{E}_{x,i}\sigma_D^{(n)} \nearrow \mathbf{E}_{x,i}\sigma_D. \tag{3.18}$$

Note that $\mathbf{E}_{x,i}\sigma_D < \infty$ from Step 1. Meanwhile, Lemma 3.4 implies that the function $u_n(x,i) = \mathbf{E}_{x,i}\sigma_D^{(n)}$ solves the boundary value problem

$$\mathcal{L}u_n(x,i) = -1, \quad u_n(x,i)|_{x \in \partial D} = 0, \quad u_n(x,i)|_{|x|=n} = 0, \quad i \in \mathcal{M}. \tag{3.19}$$

Thus the function $v_n(x,i) := u_{n+1}(x,i) - u_n(x,i)$ is $\mathcal{L}$-harmonic in the domain $(D^c \cap \{|x| < n\}) \times \mathcal{M}$. Since $\sigma_D^{(n)} \le \sigma_D^{(n+1)}$, it follows that $\mathbf{E}_{x,i}\sigma_D^{(n)} \le \mathbf{E}_{x,i}\sigma_D^{(n+1)}$, and hence $v_n(x,i) \ge 0$. Now (3.18) implies that

$$u(x,i) = u_{n_0}(x,i) + \sum_{k=n_0}^{\infty} v_k(x,i). \tag{3.20}$$

Using Harnack's inequality for $\mathcal{L}$-elliptic systems of equations (see [1, 7], and also [30] for general references on elliptic systems), it can be shown by a slight modification of the well-known arguments (see, for example, [10, pp. 21–22]) that the sum of a convergent series of positive $\mathcal{L}$-harmonic functions is also an $\mathcal{L}$-harmonic function. Hence we conclude that $u(x,i)$ is twice continuously differentiable and satisfies (3.16). To verify that $u(x,i)$ is the smallest positive solution of (3.16), let $w(x,i)$ be any positive solution of (3.16). Note that $u_n(x,i) = \mathbf{E}_{x,i}\sigma_D^{(n)}$ satisfies the boundary conditions

$$u_n(x,i)|_{x\in\partial D} = 0, \quad u_n(x,i)|_{|x|=n} = 0, \quad i \in \mathcal{M}.$$

Then the functions $u_n(x,i) - w(x,i), i \in \mathcal{M}$, are $\mathcal{L}$-harmonic and satisfy $u_n(x,i) - w(x,i) = 0$ for $(x,i) \in \partial D \times \mathcal{M}$ and $u_n(x,i) - w(x,i) < 0$ for $(x,i) \in \{|x| = n\} \times \mathcal{M}$. Hence it follows from the maximum principle for $\mathcal{L}$-elliptic system of equations [27, p. 192] that $u_n(x,i) \le w(x,i)$ in $(D^c \cap \{|x| < n\}) \times \mathcal{M}$ for all $n \ge n_0$. Letting $n \to \infty$, we obtain $u(x,i) \le w(x,i)$, as desired.

Step 3. Show that there exists a nonnegative function V satisfying the conditions of the theorem if the process $Y(t) = (X(t), \gamma(t))$ is positive recurrent with respect to the domain $U = D \times \{\ell\}$. Then $\mathbf{E}_{x,i}\sigma_D < \infty$ for all $(x,i) \in D^c \times \mathcal{M}$, and consequently (3.20) and Harnack's inequality for the $\mathcal{L}$-elliptic system of equations [1, 7] imply that the bounded monotone increasing sequence $u_n(x,i)$ converges uniformly on every compact subset of $D^c \times \mathcal{M}$. Moreover, its limit $u(x,i)$ satisfies the equation $\mathcal{L}u(x,i) = -1$ for every $i \in \mathcal{M}$. Therefore the function $V(x,i) := u(x,i)$ satisfies (3.15). This completes the proof of the theorem. □

THEOREM 3.13. *A necessary and sufficient condition for positive recurrence with respect to a domain $U = D \times \{\ell\} \subset \mathbb{R}^r \times \mathcal{M}$ is that for each $i \in \mathcal{M}$, there exists a nonnegative function $V(\cdot,i) : D^c \mapsto \mathbb{R}$ such that $V(\cdot,i)$ is twice continuously differentiable and that for some $\alpha > 0$,*

$$\mathcal{L}V(x,i) \le -\alpha, \quad (x,i) \in D^c \times \mathcal{M}. \tag{3.21}$$

Proof. Necessity. This part follows immediately from the necessity of Theorem 3.12 with $\alpha = -1$.

Sufficiency. Suppose that there exists a nonnegative function V satisfying the conditions of the theorem. Define the stopping time $\sigma_D^{(n)}(t) = \min\{\sigma_D, t, \beta_n\}$ as in the proof of Theorem 3.12. Now Dynkin's formula and (3.21) imply that for any $(x,i) \in D^c \times \mathcal{M}$,

$$\begin{aligned}
&\mathbf{E}_{x,i}V\left(X\left(\sigma_D^{(n)}(t)\right), \gamma\left(\sigma_D^{(n)}(t)\right)\right) - V(x,i) \\
&\quad = \mathbf{E}_{x,i}\int_0^{\sigma_D^{(n)}(t)} \mathcal{L}V(X(s),\gamma(s))ds \le -\alpha\mathbf{E}_{x,i}\sigma_D^{(n)}(t).
\end{aligned}$$

Hence we have by the nonnegativity of the function V that $\mathbf{E}_{x,i}\sigma_D^{(n)}(t) \le \frac{1}{\alpha}V(x,i)$. Meanwhile, the regularity of the process $Y(t) = (X(t), \gamma(t))$ implies that $\sigma_D^{(n)}(t) \to$

$\sigma_D(t)$ a.s. as $n \to \infty$, where $\sigma_D(t) = \min\{\sigma_D, t\}$. Therefore Fatou's lemma leads to $\mathbf{E}_{x,i}\sigma_D(t) \le \frac{1}{\alpha}V(x,i)$. Moreover, from the proof of Theorem 3.12, $\sigma_D(t) \to \sigma_D$ a.s. as $t \to \infty$. Thus we obtain $\mathbf{E}_{x,i}\sigma_D \le \frac{1}{\alpha}V(x,i)$ by applying Fatou's lemma again. Then Lemma 3.8 implies that $\mathbf{E}_{x,i}\sigma_U = \mathbf{E}_{x,i}\sigma_{D,\ell} < \infty$. Since $(x,i) \in D^c \times \mathcal{M}$ is arbitrary, we conclude that $Y(t)$ is positive recurrent with respect to U. This completes the proof of the theorem. □

3.3. Examples. In this subsection, we provide two examples to illustrate Theorems 3.12 and 3.13.

Example 3.14. Suppose that for each $x \in \mathbb{R}^r$ and each $i \in \mathcal{M}$, there exist positive constants c and α such that for all x with $|x| \ge c$,

$$\left\langle b(x,i), \frac{x}{|x|} \right\rangle < -\alpha, \tag{3.22}$$

where $|x|$ denotes the norm of x. Then the process $Y(t) = (X(t), \gamma(t))$ is positive recurrent.

First note that (1.5) implies that for all x with $|x| \ge \frac{r}{\alpha\kappa_1}$, we have

$$\operatorname{tr}(a(x,i)) = \sum_{j=1}^{r} \langle a(x,i)e_j, e_j \rangle \le \sum_{j=1}^{r} \kappa_1^{-1} \le \alpha|x|.$$

Then by Theorem 3.10, it is enough to prove that the process $Y(t) = (X(t), \gamma(t))$ is positive recurrent with respect to the domain $U := \{|x| < \varrho\} \times \{\ell\}$ for some $\ell \in \mathcal{M}$, where $\varrho := \max\{c, \frac{r}{\alpha\kappa_1}\}$. To this end, consider the function

$$V(x,i) = \frac{1}{2}\langle x, x \rangle \quad \text{for each} \quad i \in \mathcal{M} \quad \text{and for all } |x| \ge \varrho.$$

Then for each $i \in \mathcal{M}$, $\nabla V(\cdot, i) = x$ and $\nabla^2 V(\cdot, i) = I$, where I is the $r \times r$ identity matrix. Thus by the definition of $\mathcal{L}$, we have for all $(x,i) \in \{|x| \ge \varrho\} \times \mathcal{M}$ that

$$\mathcal{L}V(x,i) = \frac{1}{2}\operatorname{tr}(a(x,i)) + \left\langle b(x,i), \frac{x}{|x|} \right\rangle |x| < \frac{1}{2}\alpha|x| - \alpha|x| = -\frac{1}{2}\alpha|x| \le -\frac{1}{2}\alpha\varrho.$$

Then the conclusion immediately follows from Theorem 3.13.

Remark 3.15. Suppose that the diffusion component $X(t)$ of the process $Y(t) = (X(t), \gamma(t))$ is one-dimensional and that there exist constants $c_0 > 0$ and $c_1 > 0$ such that for each $i \in \mathcal{M}$,

$$b(x,i) \begin{cases} < -c_1 & \text{for } x > c_0, \\ > c_1 & \text{for } x < -c_0. \end{cases} \tag{3.23}$$

Then the process $Y(t) = (X(t), \gamma(t))$ is positive recurrent. In fact, the conclusion follows immediately if we observe that (3.23) satisfies (3.22). Alternatively, we can verify this directly by defining the Liapunov function $V(x,i) = |x|$ for each $i \in \mathcal{M}$.

Example 3.16. To illustrate the utility of Theorem 3.13, consider a real-valued process

$$dX(t) = b(X(t), \gamma(t))dt + \sigma(X(t), \gamma(t))dw(t), \tag{3.24}$$

where $\gamma(t)$ is a two-state random jump process, with x-dependent generator

$$Q(x) = \begin{pmatrix} -\frac{1}{3} - \frac{1}{4}\cos x & \frac{1}{3} + \frac{1}{4}\cos x \\ \frac{7}{3} + \frac{1}{2}\sin x & -\frac{7}{3} - \frac{1}{2}\sin x \end{pmatrix},$$

and

$$b(x,1) = -x, \quad \sigma(x,1) = 1, \quad b(x,2) = x, \quad \sigma(x,2) = 1.$$

Thus (3.24) can be regarded as the result of the following two diffusions:

$$dX(t) = -X(t)dt + dw(t), \tag{3.25}$$

$$dX(t) = X(t)dt + dw(t), \tag{3.26}$$

switching back and forth from one to the other according to the movement of $\gamma(t)$.

Note that (3.25) is positive recurrent while (3.26) is a transient diffusion process. But, the switching diffusion (3.24) is positive recurrent. We verify these as follows. Consider the Liapunov function $V(x,1) = |x|$. Let $\mathcal{L}_1$ be the operator associated with (3.25). Then we have for all $|x| \geq 1$, $\mathcal{L}_1 V(x,1) = -x \cdot \operatorname{sign} x = -|x| \leq -1 < 0$. Thus it follows from [16, Theorem 3.7.3] that (3.25) is positive recurrent. Recall that the real-valued diffusion process $dX(t) = b(X(t))dt + \sigma(X(t))dw(t)$ with $\sigma(x) \neq 0$ for all $x \in \mathbb{R}$ is recurrent if and only if $\int_0^x \exp\{-2\int_0^u \frac{b(z)}{\sigma^2(z)}dz\}du \to \pm\infty$ as $x \to \pm\infty$; see [16, p. 105]. Direct computation shows that (3.26) fails to satisfy this condition and hence is transient.

Next, we use Theorem 3.13 to demonstrate that the switching diffusion (3.24) is positive recurrent for appropriate Q. Consider Liapunov functions

$$V(x,1) = |x|, \quad V(x,2) = \frac{7}{3}|x|.$$

Then we have

$$\mathcal{L}V(x,1) = -x \cdot \operatorname{sign} x + \left(\frac{1}{3} + \frac{1}{4}\cos x\right)\left(\frac{7}{3} - 1\right)|x| \leq -\frac{2}{9}|x| \leq -\frac{2}{9},$$

$$\mathcal{L}V(x,2) = x \cdot \frac{7}{3}\operatorname{sign} x + \left(\frac{7}{3} + \frac{1}{2}\sin x\right)\left(1 - \frac{7}{3}\right)|x| \leq -\frac{1}{9}|x| \leq -\frac{1}{9}$$

for all $|x| \geq 1$. Then the switching diffusion (3.24) is positive recurrent by Theorem 3.13.

4. Ergodicity. In this section, we study the ergodic properties of the process $Y(t) = (X(t), \gamma(t))$ under the assumption that the process is positive recurrent with respect to some bounded domain $U = E \times \{\ell\}$, where $E \subset \mathbb{R}^r$ and $\ell \in \mathcal{M}$ are fixed throughout this section. We also assume that the boundary ∂E of E is sufficiently smooth. Let the operator $\mathcal{L}$ satisfy (A). Then it follows from Theorem 3.10 that the process is positive recurrent with respect to any nonempty open set.

Let $D \subset \mathbb{R}^r$ be a bounded ball with sufficiently smooth boundary ∂D such that $E \cup \partial E \subset D$. Let $\varsigma_0 = 0$ and define the stopping times $\varsigma_1, \varsigma_2, \ldots$ inductively as follows: ς_{2n+1} is the first time after ς_{2n} at which the process $Y(t) = (X(t), \gamma(t))$ reaches the set $\partial E \times \{\ell\}$, and ς_{2n+2} is the first time after ς_{2n+1} at which the path reaches the set $\partial D \times \{\ell\}$. Now we can divide an arbitrary sample path of the process $Y(t) = (X(t), \gamma(t))$ into cycles:

$$[\varsigma_0, \varsigma_2), [\varsigma_2, \varsigma_4), \ldots, [\varsigma_{2n}, \varsigma_{2n+2}), \ldots. \tag{4.1}$$

Figure 1 presents a demonstration of such cycles when the discrete component $\gamma(\cdot)$ has three states.

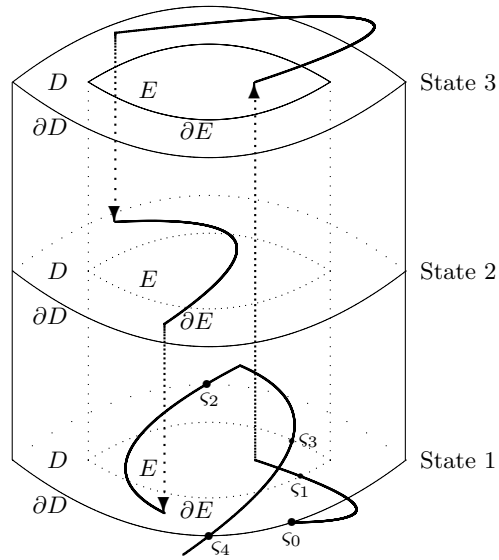


FIG. 1. *A sample path of the process* $Y(t) = (X(t), \gamma(t))$ *when* $m_0 = 3$.

The process $Y(t) = (X(t), \gamma(t))$ is positive recurrent with respect to $E \times \{\ell\}$ and hence positive recurrent with respect to $D \times \{\ell\}$ by Theorem 3.10. It follows that all the stopping times $\varsigma_0 < \varsigma_1 < \varsigma_2 < \varsigma_3 < \varsigma_4 < \cdots$ are finite a.s. Since the process $Y(t) = (X(t), \gamma(t))$ is positive recurrent, we may assume without loss of generality that $Y(0) = (X(0), \gamma(0)) = (x, \ell) \in \partial D \times \{\ell\}$. It follows from the strong Markov property of the process $Y(t) = (X(t), \gamma(t))$ that the sequence $\{Y_n\}$ is a Markov chain on $\partial D \times \{\ell\}$, where $Y_n = Y(\varsigma_{2n}) = (X_n, \ell)$, $n = 0, 1, 2, \ldots$. Let $\widetilde{P}(x, A)$ denote the one-step transition probabilities of this Markov chain, that is,

$$\widetilde{P}(x, A) = \mathbf{P}\left(Y_1 \in (A \times \{\ell\}) \mid Y_0 = (x, \ell)\right),$$

for any $x \in \partial D$ and $A \in \mathcal{B}(\partial D)$, where $\mathcal{B}(\partial D)$ denotes the collection of Borel measurable sets on ∂D. Note that the process $Y(t) = (X(t), \gamma(t))$, starting from (x, ℓ), may jump many times before it reaches the set (A, ℓ); see [28] for more details. Denote by $\widetilde{P}^{(n)}(x, A)$ the n-step transition probability of the Markov chain for any $n \geq 1$. For any Borel measurable function $f : \mathbb{R}^r \mapsto \mathbb{R}$, set

$$\mathbf{E}_x f(X_1) := \mathbf{E}_{x,\ell} f(X_1) = \int_{\partial D} f(y) \widetilde{P}(x, dy). \tag{4.2}$$

Throughout this section, for simplicity we write $\mathbf{E}_x$ for $\mathbf{E}_{x,\ell}$. We will show that the process $Y(t) = (X(t), \gamma(t))$ possesses a unique stationary distribution. To this end, we need the following lemma.

LEMMA 4.1. *The Markov chain* $Y_i = (X_i, \ell)$ *has a unique stationary distribution* $m(\cdot)$ *such that*

$$\left|\widetilde{P}^{(n)}(x, A) - m(A)\right| < \lambda^n \quad \textit{for any} \quad A \in \mathcal{B}(\partial D), \tag{4.3}$$

for some constant $0 < \lambda < 1$.

Proof. Note that

$$\begin{aligned}\widetilde{P}(x,A) &= \mathbf{P}\left\{Y_1 \in (A\times\{\ell\})|Y_0=(x,\ell)\right\}\\ &= \int_{\partial E} \mathbf{P}_{x,\ell}\left\{(X(\varsigma_1),\gamma(\varsigma_1))\in(dy\times\{\ell\})\right\}\cdot \mathbf{P}_{y,\ell}\left\{(X(\varsigma_2),\gamma(\varsigma_2))\in(A\times\{\ell\})\right\}.\end{aligned}$$

Using the harmonic measure defined in and Lemmas 2.2 and 2.3 of [7], relating the kernel and surface area (similar to the solution for the diffusion process without switching in the form of double layer potential given in the first displayed equation in [16, p. 97]) and the harmonic measure, we can finish the proof of this lemma analogously to that of [16, Lemma 4.4.1]. The details are omitted here. □

Remark 4.2. Note that

$$(X^{s,X(s),\gamma(s)}(t),\gamma(t)) = (X^{0,X(0),\gamma(0)}(t+s),\gamma(t+s)), \tag{4.4}$$

where $(X^{0,X(0),\gamma(0)}(u),\gamma(u))$ denotes the sample path of the process $(X(\cdot),\gamma(\cdot))$ with initial point $(X(0),\gamma(0))$ at time $t=0$, and note the similar definition for $(X^{s,X(s),\gamma(s)}(t),\gamma(t))$. When no confusion arises, we simply write $(X(u),\gamma(u)) = (X^{0,X(0),\gamma(0)}(u),\gamma(u))$.

Let τ be an $\mathcal{F}_t$ stopping time with $\mathbf{E}_{x,i}\tau<\infty$ and let $f:\mathbb{R}^r\times\mathcal{M}\mapsto\mathbb{R}$ be a Borel measurable function. Then

$$\mathbf{E}_{x,i}\int_0^\tau f(X(s+t),\gamma(s+t))ds = \mathbf{E}_{x,i}\int_0^\tau \mathbf{E}_{X(s),\gamma(s)}f(X(s+t),\gamma(s+t))ds. \tag{4.5}$$

Now we can explicitly construct the stationary distribution of the process $Y(t)=(X(t),\gamma(t))$.

THEOREM 4.3. *The positive recurrent process $Y(t)=(X(t),\gamma(t))$ has a unique stationary distribution $\widehat{\nu}(\cdot,\cdot)=(\widehat{\nu}(\cdot,i):i\in\mathcal{M})$.*

Proof. Let $A\in\mathcal{B}(\mathbb{R}^r)$ and $i\in\mathcal{M}$. Denote by $\tau^{A\times\{i\}}$ the time spent by the path of $Y(t)=(X(t),\gamma(t))$ in the set $(A\times\{i\})$ during the first cycle (the cycles were defined in (4.1)). Set

$$\nu(A,i) := \int_{\partial D} m(dx)\mathbf{E}_x\tau^{A\times\{i\}}, \tag{4.6}$$

where $m(\cdot)$ is the stationary distribution of $Y_i=(X_i,\ell)$, whose existence is guaranteed by Lemma 4.1. It is easy to verify that $\nu(\cdot,\cdot)$ is a positive measure defined on $\mathcal{B}(\mathbb{R}^r)\times\mathcal{M}$. Thus for any bounded Borel measurable function $g(\cdot):\mathbb{R}^r\mapsto\mathbb{R}$, it follows from (4.2) and Fubini's theorem that

$$\int_{\partial D}\mathbf{E}_x g(X_1)m(dx) = \int_{\partial D} m(dx)\int_{\partial D} g(y)\widetilde{P}(x,dy) = \int_{\partial D} g(y)m(dy). \tag{4.7}$$

Now we claim that for any bounded and continuous function $f(\cdot,\cdot)$,

$$\sum_{j=1}^{m_0}\int_{\mathbb{R}^r} f(y,j)\nu(dy,j) = \int_{\partial D} m(dx)\mathbf{E}_x\int_0^{\varsigma_2} f(X(t),\gamma(t))dt \tag{4.8}$$

holds. In fact, if $f(y,j)=\chi_{[A\times\{i\}]}(y,j)$ for some $A\in\mathcal{B}(\mathbb{R}^r)$ and $i\in\mathcal{M}$, then from (4.6),

$$\begin{aligned}\sum_{j=1}^{m_0}\int_{\mathbb{R}^r}\chi_{[A\times\{i\}]}(y,j)\nu(dy,j) = \nu(A,i) &= \int_{\partial D} m(dx)\mathbf{E}_x\tau^{A\times\{i\}}\\ &= \int_{\partial D} m(dx)\mathbf{E}_x\int_0^{\varsigma_2}\chi_{[A\times\{i\}]}(X(t),\gamma(t))dt.\end{aligned}$$

Similarly, we obtain that (4.8) holds for f being a simple function:

$$f(y,j)=\sum_{p=1}^{n} c_p \chi_{U_p}(y,j), \text{ where } U_p \subset \mathbb{R}^r \times \mathcal{M}.$$

Finally, if f is a bounded and continuous function, (4.8) follows by approximating f by simple functions. It follows from (4.8), (4.4), and (4.5) that

$$\begin{aligned}
&\sum_{i=1}^{m_0} \int_{\mathbb{R}^r} \mathbf{E}_{x,i} f(X(t),\gamma(t))\nu(dx,i)\\
&\quad = \int_{\partial D} m(dx)\mathbf{E}_x \int_0^{\varsigma_2} \mathbf{E}_{X(s),\gamma(s)} f(X(t+s),\gamma(t+s))ds\\
&\quad = \int_{\partial D} m(dx)\mathbf{E}_x \int_0^{\varsigma_2} f(X(t+s),\gamma(t+s))ds\\
&\quad = \int_{\partial D} m(dx)\mathbf{E}_x \int_0^{\varsigma_2} f(X(u),\gamma(u))du\\
&\qquad + \int_{\partial D} m(dx)\mathbf{E}_x \int_{\varsigma_2}^{t+\varsigma_2} f(X(u),\gamma(u))du - \int_{\partial D} m(dx)\mathbf{E}_x \int_0^{t} f(X(u),\gamma(u))du.
\end{aligned}$$

Now applying (4.7) with $g(x)=\mathbf{E}_x \int_{\varsigma_2}^{\varsigma_2+t} f(X(u),\gamma(u))du$, we obtain

$$\begin{aligned}
&\int_{\partial D} m(dx)\mathbf{E}_x \int_{\varsigma_2}^{\varsigma_2+t} f(X(u),\gamma(u))du\\
&\qquad = \int_{\partial D} m(dx)\mathbf{E}_x \mathbf{E}_{X_1,\ell} \int_{\varsigma_2}^{\varsigma_2+t} f(X(u+\varsigma_2),\gamma(u+\varsigma_2))du\\
&\qquad = \int_{\partial D} m(dx)\mathbf{E}_x \int_0^{t} f(X(u),\gamma(u))du.
\end{aligned}$$

Note that in the above deduction, we used (4.4) again. Therefore, the above two equations and (4.8) yield that

$$\sum_{i=1}^{m_0} \int_{\mathbb{R}^r} \mathbf{E}_{x,i} f(X(t),\gamma(t))\nu(dx,i) = \sum_{i=1}^{m_0} \int_{\mathbb{R}^r} f(x,i)\nu(dx,i).$$

Thus, the normalized measure

$$\widehat{\nu}(A,i)=\frac{\nu(A,i)}{\sum_{j=1}^{m_0}\nu(\mathbb{R}^r,j)},\ i\in\mathcal{M}, \tag{4.9}$$

defines the desired stationary distribution. The theorem thus follows. □

THEOREM 4.4. *Denote by $\mu(\cdot,\cdot)$ the stationary density associated with the stationary distribution $\widehat{\nu}(\cdot,\cdot)$ constructed in Theorem* 4.3*, and let $f(\cdot,\cdot):\mathbb{R}^r\times\mathcal{M}\mapsto\mathbb{R}$ be a Borel measurable function such that*

$$\sum_{i=1}^{m_0}\int_{\mathbb{R}^r} |f(x,i)|\mu(x,i)dx<\infty. \tag{4.10}$$

Then

$$\mathbf{P}_{x,i}\left(\frac{1}{T}\int_0^T f(X(t),\gamma(t))dt \to \overline{f}\right)=1 \tag{4.11}$$

for any $(x,i)\in\mathbb{R}^r\times\mathcal{M}$, *where*

$$\overline{f}=\sum_{i=1}^{m_0}\int_{\mathbb{R}^r} f(x,i)\mu(x,i)dx. \tag{4.12}$$

Proof. We first prove (4.11) if the initial distribution is the stationary distribution of the Markov chain $Y_i=(X_i,\gamma_i)$, that is,

$$\mathbf{P}\{(X(0),\gamma(0))\in(A\times\{\ell\})\}=m(A) \tag{4.13}$$

for any $A\in\mathcal{B}(\partial D)$. Consider the sequence of random variables

$$\eta_n=\int_{\varsigma_{2n}}^{\varsigma_{2n+2}} f(X(t),\gamma(t))dt. \tag{4.14}$$

Then it follows from (4.13) that $\{\eta_n\}$ is a strictly stationary sequence. Also from (4.6) and (4.8), we have

$$\mathbf{E}\eta_n=\sum_{i=1}^{m_0}\int_{\mathbb{R}^r} f(x,i)\nu(dx,i) \tag{4.15}$$

for all $n=0,1,2,\ldots$. Meanwhile, (4.3) implies that the sequence η_n is metrically transitive. Let $\upsilon(T)$ denote the number of cycles completed up to time T. That is,

$$\upsilon(T):=\max\left\{n\in\mathbb{N}:\sum_{k=1}^{n}(\varsigma_{2k}-\varsigma_{2k-2})\le T\right\}.$$

Then we can decompose $\int_0^T f(X(t),\gamma(t))dt$ into

$$\int_0^T f(X(t),\gamma(t))dt=\sum_{n=0}^{\upsilon(T)}\eta_n+\int_{\varsigma_{2\upsilon(T)}}^{T} f(X(t),\gamma(t))dt, \tag{4.16}$$

with η_n as given in (4.14). We may assume without loss of generality that $f(x,i)\ge 0$; for the general case, we can write $f(x,i)$ as a difference of two nonnegative functions. Then it follows from (4.16) that

$$\sum_{n=0}^{\upsilon(T)}\eta_n\le\int_0^T f(X(t),\gamma(t))dt\le\sum_{n=0}^{\upsilon(T)+1}\eta_n.$$

Since the sequence $\{\eta_n\}$ is stationary and metrically transitive, the law of large numbers for such sequences implies that

$$\mathbf{P}\left\{\frac{1}{n}\sum_{k=0}^{n}\eta_k\overset{n\to\infty}{\longrightarrow}\sum_{i=1}^{m_0}\int_{\mathbb{R}^r} f(x,i)\nu(dx,i)\right\}=1. \tag{4.17}$$

In particular, if $f(x,i)\equiv 1$, then the above equation reduces to

$$\mathbf{P}\left\{\frac{\varsigma_{2n+2}}{n}\overset{n\to\infty}{\longrightarrow}\sum_{i=1}^{m_0}\nu(\mathbb{R}^r,i)\right\}=1. \tag{4.18}$$

Note that the positive recurrence of the process $Y(t) = (X(t), \gamma(t))$ implies that $\upsilon(T) \to \infty$ as $T \to \infty$. Clearly, $\upsilon(T)/(\upsilon(T)+1) \to 1$ a.s. as $T \to \infty$. Thus, it follows from (4.18) that as $T \to \infty$,

$$\frac{\varsigma_{2\upsilon(T)}}{\varsigma_{2\upsilon(T)+2}} = \frac{\frac{\varsigma_{2\upsilon(T)}}{\upsilon(T)}}{\frac{\varsigma_{2\upsilon(T)+2}}{\upsilon(T)+1}} \frac{\upsilon(T)}{\upsilon(T)+1} \to 1 \quad \text{a.s.} \tag{4.19}$$

Meanwhile, since $\varsigma_{2\upsilon(T)} \le T \le \varsigma_{2\upsilon(T)+2}$, we have

$$\frac{\varsigma_{2\upsilon(T)}}{\varsigma_{2\upsilon(T)+2}} \le \frac{\varsigma_{2\upsilon(T)}}{T} \le \frac{\varsigma_{2\upsilon(T)}}{\varsigma_{2\upsilon(T)}} = 1.$$

Therefore, we have from (4.19) that

$$\frac{\varsigma_{2\upsilon(T)}}{T} \to 1 \quad \text{a.s. as} \quad T \to \infty. \tag{4.20}$$

Moreover, (4.18) implies that

$$\frac{\upsilon(T)}{\varsigma_{2\upsilon(T)}} \to \frac{1}{\sum_{i=1}^{m_0} \nu(\mathbb{R}^r, i)} \quad \text{a.s. as} \quad T \to \infty. \tag{4.21}$$

Now using (4.17), (4.20), and (4.21), we obtain

$$\mathbf{P}\left\{ \frac{1}{T}\int_0^T f(X(t),\gamma(t))dt = \frac{\int_0^T f(X(t),\gamma(t))dt}{\upsilon(T)} \cdot \frac{\upsilon(T)}{\varsigma_{2\upsilon(T)}} \cdot \frac{\varsigma_{2\upsilon(T)}}{T} \right.$$
$$\left. \overset{T\to\infty}{\longrightarrow} \sum_{i=1}^{m_0} \int_{\mathbb{R}^r} f(x,i)\widehat{\nu}(dx,i) \right\} = 1.$$

Finally, we note that

$$\int_{\mathbb{R}^r} f(x,i)\widehat{\nu}(dx,i) = \int_{\mathbb{R}^r} f(x,i)\mu(x,i)dx$$

by the definition of $\mu(\cdot,\cdot)$. Thus, (4.11) holds. This proves (4.11) if the initial distribution is (4.13).

Now let $(x,i) \in \mathbb{R}^r \times \mathcal{M}$. Since the process $Y(t) = (X(t), \gamma(t))$ is positive recurrent with respect to the domain $D \times \{\ell\}$, we have

$$\begin{aligned}
&\mathbf{P}_{x,i}\left\{ \lim_{T\to\infty} \frac{1}{T}\int_0^T f(X(t),\gamma(t))dt = a \right\} \\
&\quad = \mathbf{P}_{x,i}\left\{ \lim_{T\to\infty} \frac{1}{T}\int_{\varsigma_2}^T f(X(t),\gamma(t))dt = a \right\} \\
&\quad = \int_{\partial D} \mathbf{P}_{x,i}\{(X(\varsigma_2),\gamma(\varsigma_2)) \in (dy,\ell)\} \cdot \mathbf{P}_{y,\ell}\left\{ \lim_{T\to\infty} \frac{1}{T}\int_0^T f(X(t),\gamma(t))dt = a \right\} \\
&\quad = \mathbf{P}_{y,\ell}\left\{ \lim_{T\to\infty} \frac{1}{T}\int_0^T f(X(t),\gamma(t))dt = a \right\}.
\end{aligned}$$

Therefore, (4.11) holds for all $(x,i) \in \mathbb{R}^r \times \mathcal{M}$. This completes the proof of the theorem. □

As a consequence of Theorem 4.4, we obtain the following corollary.

COROLLARY 4.5. *Let the assumptions of Theorem* 4.4 *be satisfied and let* $u(t,x,i)$ *be the solution of the Cauchy problem*

$$\begin{cases} \dfrac{\partial u(t,x,i)}{\partial t} &= \mathcal{L}u(x,i),\ i\in\mathcal{M},\\ u(0,x,i) &= f(x,i). \end{cases} \tag{4.22}$$

Then as $T\to\infty$,

$$\frac{1}{T}\int_0^T u(t,x,i)dt \to \sum_{i=1}^{m_0}\int_{\mathbb{R}^r} f(x,i)\mu(x,i)dx. \tag{4.23}$$

Proof. The generalized Itô's formula (see, for example, [5] or [28, Lemma 3, p. 104]) implies that $u(t,x,i)=\mathbf{E}_{x,i}f(X(t),\gamma(t))$. Thus we have

$$\frac{1}{T}\int_0^T u(t,x,i)dt = \mathbf{E}_{x,i}\left(\frac{1}{T}\int_0^T f(X(t),\gamma(t))dt\right). \tag{4.24}$$

Meanwhile (4.11) implies that

$$\frac{1}{T}\int_0^T f(X(t),\gamma(t))dt \overset{T\to\infty}{\longrightarrow} \sum_{i=1}^{m_0}\int_{\mathbb{R}^r} f(x,i)\mu(x,i)dx \text{ a.s.}$$

with respect to the probability $\mathbf{P}_{x,i}$. Then (4.23) follows from the dominated convergence theorem. □

5. Discussions and remarks.

5.1. Discussions. The recurrence and ergodicity obtained enable us to further study asymptotic properties of hybrid diffusion systems and to carry out control and optimization tasks. We outline several directions in what follows.

Easily verifiable conditions. In many applications, it is often more convenient to analyze weak stability through conditions on the coefficients of the corresponding stochastic differential equations. Assume for simplicity that $X(\cdot)$ is a real-valued process; assume also that condition (A) holds. Motivated by Examples 3.14 and 3.16, next we present easily verifiable conditions for positive recurrence when the coefficients of the switching diffusions (1.3)–(1.4) are linearizable in an x-neighborhood of ∞. Suppose that for each $i\in\mathcal{M}$, there exists $b_i\in\mathbb{R}$ such that

$$\frac{b(x,i)}{x} = b_i + o(1), \quad \text{and} \quad Q(x)\to\widetilde{Q}, \quad \text{as} \quad |x|\to\infty,$$

where $\widetilde{Q}=(\widetilde{q}_{ij})$ is the generator of a continuous-time ergodic Markov chain $\widetilde{\gamma}(t)$ whose stationary distribution is $\mu=(\mu_1,\mu_2,\ldots,\mu_m)\in\mathbb{R}^{1\times m}$. Then using Theorems 3.10 and 3.13, we can prove that the process is positive recurrent if $\sum_{i=1}^m \mu_i b_i<0$. The result can be strengthened if, in addition,

$$\frac{\sigma(x,i)}{x} = \sigma_i + o(1) \quad \text{as} \quad |x|\to\infty,$$

where $\sigma_i^2 > 0$. Then in this case, the process is positive recurrent if

$$\sum_{i=1}^{m} \mu_i \Big(b_i - \frac{\sigma_i^2}{2} \Big) < 0.$$

The details are omitted for brevity.

Path excursions. Applications of the positive recurrence criteria enable us to establish path excursions of the underlying processes. Suppose that $Y(t) = (X(t), \gamma(t))$ is positive recurrent. Suppose that the Liapunov functions $V(x,i)$ (with $i \in \mathcal{M}$) are given as in Theorem 3.13, as is the set D. Let D_0 be a bounded open set with a compact closure satisfying $D \subset D_0$, and τ be a random time such that $(X(\tau), \gamma(\tau)) \in D_0^c \times \mathcal{M}$, and $\tau_1 = \min\{t > \tau : (X(t), \gamma(t)) \in D_0 \times \mathcal{M}\}$. We can obtain

$$\begin{aligned} &\mathbf{P}\left(\sup_{\tau \le t \le \tau_1} V(X(t), \gamma(t)) \ge \kappa \right) \le \frac{\mathbf{E}V(X(\tau), \gamma(\tau))}{\kappa} \quad \text{for} \quad \kappa > 0, \\ &\mathbf{E}(\tau_1 - \tau) \le \frac{\mathbf{E}V(X(\tau), \gamma(\tau))}{\alpha}, \end{aligned}$$

where α is as given in Theorem 3.13.

Tightness. Under positive recurrence, we may obtain tightness (or boundedness in the sense of probability) of the underlying process. Suppose that $(X(t), \gamma(t))$ is positive recurrent. It is then possible to prove that for any compact set $\overline{D}$, the set $\cup_{x \in \overline{D}} \{(X(t), \gamma(t)) : t \ge 0,\ X(0) = x,\ \gamma(0) = \gamma\}$ is tight (or bounded in probability). For a study on the diffusion counterpart, we refer the reader to [19, p. 146].

Occupation measures. To illustrate the utility of Theorem 4.4, take $f(x,i) = \chi_{[B \times J]}(x,i)$, the indicator function of the set $B \times J$, where $B \subset \mathbb{R}^r$ and $J \subset \mathcal{M}$. Then Theorem 4.4 becomes a result regarding an occupation measure. In fact, we have

$$\frac{1}{T} \int_0^T \chi_{[B \times J]}(X(t), \gamma(t)) dt \to \sum_{i \in J} \int_B \mu(x,i) dx \quad \text{a.s. as} \quad T \to \infty.$$

Stochastic approximation. Consider a parameter optimization problem. We wish to find θ_*, a vector-valued parameter, so that the cost function

$$J(\theta) = \lim_{T \to \infty} E \frac{1}{T} \int_0^T \widehat{J}(\theta, Y(t)) dt$$

is minimized, where $Y(t)$ is a positive recurrent switching diffusion as considered in this paper and where, for each θ, $\widehat{J}(\theta, \cdot, \cdot)$ satisfies the conditions of Theorem 4.4. For simplicity, we assume that the gradient of $\widehat{J}(\cdot, x, i)$ with respect to θ is available for each x and each $i \in \mathcal{M}$. Then we consider a constant stepsize recursive algorithm

$$\theta_{n+1} = \theta_n - \varepsilon \frac{1}{T} \int_{nT}^{nT+T} \nabla \widehat{J}(\theta_n, Y(t)) dt,$$

or a decreasing stepsize algorithm

$$\theta_{n+1} = \theta_n - \varepsilon_n \frac{1}{T} \int_{nT}^{nT+T} \nabla \widehat{J}(\theta_n, Y(t)) dt,$$

where $\varepsilon > 0$, and $\varepsilon_n \to 0$ as $n \to \infty$ and $\sum_n \varepsilon_n = \infty$. Modifications and variants are possible. For example, we may include additional measurement noise, and the gradient of $\widehat{J}(\cdot)$ may be changed to its gradient estimates. The motivation for such algorithms stems from optimization of average cost per unit time problems arising from parameter estimations in switching systems of stochastic differential equations, manufacturing systems, and queueing networks; see related work in [21, Chapter 9] and [32]. The ergodicity of the switching diffusion is crucial in the study of the asymptotic behavior of the algorithms.

5.2. Further remarks. This work developed asymptotic properties of positive recurrent switching diffusions. Under general conditions, necessary and sufficient conditions for positive recurrence were developed. Then ergodicity was established for positive recurrent Markov processes with switching. Also provided were explicit representations of the invariant measures. For new results on related problems of stability for regime-switching diffusions, we refer the reader to our recent work [18].

A number of problems remains open. Obtaining large deviations type of bounds is a worthwhile undertaking, which will have an important impact on studying the associated control and optimization problems. Next, concerning null recurrent switching diffusions (see [16, 17]), can we obtain necessary and sufficient conditions? It appears that the desired criteria will be more difficult to obtain compared to a single diffusion process since one needs to solve systems of boundary value problems.

Appendix. Proofs of the lemmas.

Proof of Lemma 3.5. It suffices to prove the lemma for the case when $E \cup \partial E \subset D$ and ∂E is sufficiently smooth. Fix any $(x,i) \in E^c \times \mathcal{M}$. Let $G \subset \mathbb{R}^r$ be an open and bounded set with a sufficiently smooth boundary such that $D \cup \partial D \subset G$. Without loss of generality, we may further assume that $(x,i) \in G \times \mathcal{M}$. Define a sequence of stopping times by

$$\varsigma_1 := \inf\{t \geq 0 : X(t) \in \partial G\}, \tag{A.1}$$

and for $n = 1, 2, \ldots$,

$$\begin{aligned} \varsigma_{2n} &:= \inf\{t \geq \varsigma_{2n-1} : X(t) \in \partial D\}, \\ \varsigma_{2n+1} &:= \inf\{t \geq \varsigma_{2n} : X(t) \in \partial G\}. \end{aligned} \tag{A.2}$$

It follows from (3.9) and Theorem 3.1 that $\varsigma_n < \infty$ a.s. $\mathbf{P}_{x,i}$ for $n = 1, 2, \ldots$. Let $H := G - \overline{E}$ and define $u(x,i) := \mathbf{P}_{x,i}\{X(\tau_H) \in \partial E\}$. Note that $u(x,j)|_{x \in \partial E} = 1$ and $u(x,j)|_{x \in \partial G} = 0$ for all $j \in \mathcal{M}$. Therefore, it follows that

$$\begin{aligned} u(x,i) &= \sum_{j=1}^{m_0} \int_{\partial E} \mathbf{P}_{x,i}\{(X(\tau_H), \gamma(\tau_H)) \in (dy \times \{j\})\}\, u(y,j) \\ &\quad + \sum_{j=1}^{m_0} \int_{\partial G} \mathbf{P}_{x,i}\{(X(\tau_H), \gamma(\tau_H)) \in (dy \times \{j\})\}\, u(y,j) \\ &= \mathbf{E}_{x,i} u(X(\tau_H), \gamma(\tau_H)). \end{aligned}$$

Thus $u(x,i) \geq 0$ is $\mathcal{L}$-harmonic in $H \times \mathcal{M}$ by Lemma 3.3. Moreover, u is not identically zero since $u(x,i) = 1$ for $(x,i) \in \partial E \times \mathcal{M}$. Therefore the maximum principle for $\mathcal{L}$-harmonic functions [9] implies that

$$\inf_{(x,i) \in K \times \mathcal{M}} u(x,i) \geq \delta_1 > 0, \tag{A.3}$$

where K is some compact subset of H containing x and ∂D. Define

$$A_0 := \{X(t) \in \partial E \ \text{ for some } \ t \in [0, \varsigma_1)\}, \tag{A.4}$$

and for $n = 1, 2, \ldots$,

$$A_n := \{X(t) \in \partial E, \ \text{for some } t \in [\varsigma_{2n}, \varsigma_{2n+1})\}. \tag{A.5}$$

Note that the event A_0^c implies that $X(\tau_H) = X(\varsigma_1) \in \partial G$. Hence we have from (A.3) that

$$\mathbf{P}_{x,i}(A_0^c) \le \mathbf{P}_{x,i}(X(\tau_H) \in \partial G) = 1 - u(x,i) \le 1 - \delta_1.$$

Then it follows from the strong Markov property and (A.3) that

$$\mathbf{P}_{x,i}\left\{\bigcap_{k=0}^{n} A_k^c\right\} \le (1-\delta_1)^{n+1}. \tag{A.6}$$

Thus, we have

$$\begin{aligned}
\mathbf{P}_{x,i}\{\sigma_E = \infty\} &= \mathbf{P}_{x,i}\{X(t) \notin \partial E \text{ for any } t \ge 0\} \\
&\le \lim_{n\to\infty} \mathbf{P}_{x,i}\left\{\bigcap_{k=0}^{n} A_k^c\right\} \\
&\le \lim_{n\to\infty} (1-\delta_1)^{n+1} = 0.
\end{aligned}$$

Hence it follows that $\mathbf{P}_{x,i}\{\sigma_E < \infty\} = 1$ as desired. □

Proof of Lemma 3.6. As in Lemma 3.5, it is enough to prove the lemma for the case when $E \cup \partial E \subset D$ and ∂E is sufficiently smooth. Fix any $(x,i) \in E^c \times \mathcal{M}$. Let $G \subset \mathbb{R}^r$ be an open and bounded set with a sufficiently smooth boundary such that $D \cup \partial D \subset G$. As in the proof of Lemma 3.5, we may further assume that $(x,i) \in G \times \mathcal{M}$. Define stopping times $\varsigma_1, \varsigma_2, \ldots$ and events $A_0, A_1, A_2, \ldots$ as in (A.1), (A.2), (A.4), and (A.5) in the proof of Lemma 3.5. It follows from (3.10) and Lemma 3.5 that $\mathbf{P}_{x,i}\{\sigma_E < \infty\} = 1$. Note that if $\varsigma_{2n} < \sigma_E < \varsigma_{2n+1}$, then the event $\bigcap_{k=0}^{n-1} A_k^c$ happens a.s. Hence, it follows from (A.6) that

$$\mathbf{P}_{x,i}\{\varsigma_{2n} < \sigma_E < \varsigma_{2n+1}\} \le \mathbf{P}_{x,i}\left\{\bigcap_{k=0}^{n-1} A_k^c\right\} \le (1-\delta_1)^n.$$

Therefore, we have

$$\begin{aligned}
\mathbf{E}_{x,i}\tau_{E^c} &= \mathbf{E}_{x,i}\sigma_E \chi_{[0<\sigma_E<\varsigma_1]} + \sum_{n=1}^{\infty} \mathbf{E}_{x,i}\sigma_E \chi_{[\varsigma_{2n}<\sigma_E<\varsigma_{2n+1}]} \\
&\le \mathbf{P}_{x,i}[0 < \sigma_E < \varsigma_1]\mathbf{E}_{x,i}\varsigma_1 + \sum_{n=1}^{\infty} \mathbf{P}_{x,i}[\varsigma_{2n} < \sigma_E < \varsigma_{2n+1}]\mathbf{E}_{x,i}\varsigma_{2n+1} \\
&\le \sum_{n=0}^{\infty} (1-\delta_1)^n \mathbf{E}_{x,i}\varsigma_{2n+1}.
\end{aligned}$$

In what follows, denote by M_i $(i = 1, 2, 3)$ positive real numbers. Since $(x, i) \in G \times \mathcal{M}$, it follows from Theorem 3.1 that $\mathbf{E}_{x,i}\varsigma_1 = \mathbf{E}_{x,i}\tau_G \le M_1 < \infty$. Consequently, using σ_D and τ_G defined in (2.1),

$$\begin{aligned}\mathbf{E}_{x,i}\varsigma_3 &= \mathbf{E}_{x,i}\varsigma_1 + \mathbf{E}_{x,i}\mathbf{E}_{X(\varsigma_1),\gamma(\varsigma_1)}(\varsigma_3 - \varsigma_1)\\ &\le M_1 + \sup_{(y,j)\in\partial G\times\mathcal{M}} \mathbf{E}_{y,j}\sigma_D + \sup_{(z,k)\in\partial D\times\mathcal{M}} \mathbf{E}_{z,k}\tau_G\\ &\le M_1 + M_2 + M_3 \le 2M,\end{aligned}$$

where $M = \max\{M_1, M_2 + M_3\} < \infty$. Note that in the above deductions, we used (3.10) and Theorem 3.1. Likewise, in general, we have $\mathbf{E}_{x,i}\varsigma_{2n+1} \le (n+1)M$ for any $n = 1, 2, \ldots$ Therefore, it follows that

$$\mathbf{E}_{x,i}\sigma_E \le \sum_{n=0}^{\infty}(1-\delta_1)^n(n+1)M < \infty.$$

This completes the proof of the lemma. □

Proof of Lemma 3.7. Fix any $\ell \in \mathcal{M}$. It suffices to prove (3.12) when $(x, i) \in D \times (\mathcal{M} - \{\ell\})$ since the process $Y(t) = (X(t), \gamma(t))$, starting from $(y, j) \in D^c \times \mathcal{M}$, will reach $D \times \mathcal{M}$ in finite time a.s. $\mathbf{P}_{y,j}$ by (3.11). Choose $\varepsilon > 0$ sufficiently small such that $B \subset \overline{B} \subset B_1 \subset \overline{B_1} \subset D$, where

$$B = B(x, \varepsilon) = \{y \in \mathbb{R}^r : |y - x| < \varepsilon\} \quad \text{and} \quad B_1 = B(x, 2\varepsilon). \tag{A.7}$$

Redefine

$$\varsigma_1 := \inf\{t \ge 0 : X(t) \in \partial B\}, \tag{A.8}$$

and for $n = 1, 2, \ldots$,

$$\begin{aligned}\varsigma_{2n} &:= \inf\{t \ge \varsigma_{2n-1} : X(t) \in \partial B_1\},\\ \varsigma_{2n+1} &:= \inf\{t \ge \varsigma_{2n} : X(t) \in \partial B\}.\end{aligned} \tag{A.9}$$

Note that (3.11), Theorem 3.1, and Lemma 3.5 imply that $\varsigma_n < \infty$ a.s. $\mathbf{P}_{x,i}$. Set

$$u(x, i) := \mathbf{P}_{x,i}\left\{\sigma_{\overline{B}\times\{\ell\}} < \tau_{B_1}\right\}.$$

As in the proof of Lemma 3.5, we can verify that $u(x, i)$ is $\mathcal{L}$-harmonic in $B_1 \times \mathcal{M}$. Moreover, u is not identically zero, since $u(x, \ell)|_{x\in\partial B} = 1$. Therefore, the maximum principle [9] implies that

$$\inf_{(x,i)\in\overline{B}\times\mathcal{M}} u(x, i) \ge \delta_2 > 0. \tag{A.10}$$

Redefine

$$A_0 := \{\gamma(t) = \ell \text{ for some } t \in [0, \varsigma_2)\}, \tag{A.11}$$

and for $n = 1, 2, \ldots$,

$$A_n := \{\gamma(t) = \ell \text{ for some } t \in [\varsigma_{2n+1}, \varsigma_{2n+2})\}. \tag{A.12}$$

Using almost the same argument as in the proof of Lemma 3.5, we obtain that

$$\mathbf{P}_{x,i}(A_0^c) \le 1-\delta_2 \quad \text{and} \quad \mathbf{P}_{x,i}\left\{\bigcap_{k=0}^{n} A_k^c\right\} \le (1-\delta_2)^{n+1}. \tag{A.13}$$

Thus, we have

$$\begin{aligned}
&\mathbf{P}_{x,i}\left\{(X(t),\gamma(t)) \notin D\times\{\ell\} \text{ for any } t\ge 0\right\}\\
&\quad\le \mathbf{P}_{x,i}\left\{(X(t),\gamma(t)) \notin \overline{B_1}\times\{\ell\} \text{ for any } t\ge 0\right\}\\
&\quad\le \lim_{n\to\infty} \mathbf{P}_{x,i}\left\{\bigcap_{k=0}^{n} A_k^c\right\}\\
&\quad\le \lim_{n\to\infty} (1-\delta_2)^{n+1} = 0.
\end{aligned}$$

As a result, $\mathbf{P}_{x,i}\{\sigma_{D,\ell}=\infty\} = \mathbf{P}_{x,i}\left\{(X(t),\gamma(t)) \notin D\times\{\ell\} \text{ for any } t\ge 0\right\} = 0$, or $\mathbf{P}_{x,i}\{\sigma_{D,\ell}<\infty\}=1$. This completes the proof of the lemma. ∎

Proof of Lemma 3.8. Fix any $\ell\in\mathcal{M}$. As in Lemma 3.7, it is enough to prove (3.14) when $(x,i)\in D\times(\mathcal{M}-\{\ell\})$. Let the balls B and B_1, stopping times $\varsigma_1,\varsigma_2,\ldots$, and events $A_0,A_1,\ldots$ as in (A.7)–(A.9), (A.11), and (A.12) in the proof of Lemma 3.7. It follows from (3.13) and Lemma 3.7 that $\mathbf{P}_{x,i}\{\sigma_{D,\ell}<\infty\}=1$. Observe that if $\varsigma_{2n}\le\sigma_{D,\ell}<\varsigma_{2n+2}$, then the event $\bigcap_{k=0}^{n-1}A_k^c$ happens a.s. Hence we have from (A.13) that

$$\mathbf{P}_{x,i}\{\varsigma_{2n}\le\sigma_{D,\ell}<\varsigma_{2n+2}\} \le \mathbf{P}_{x,i}\left\{\bigcap_{k=0}^{n-1}A_k^c\right\} \le (1-\delta_2)^n.$$

It follows that

$$\begin{aligned}
\mathbf{E}_{x,i}\sigma_{D,\ell} &= \mathbf{E}_{x,i}\sigma_{D,\ell}\chi_{[0\le\sigma_{D,\ell}<\varsigma_2]} + \sum_{n=1}^{\infty}\mathbf{E}_{x,i}\sigma_{D,\ell}\chi_{[\varsigma_{2n}\le\sigma_{D,\ell}<\varsigma_{2n+2}]}\\
&\le \mathbf{P}_{x,i}[0\le\sigma_{D.\ell}<\varsigma_2]\mathbf{E}_{x,i}\varsigma_2 + \sum_{n=1}^{\infty}\mathbf{P}_{x,i}[\varsigma_{2n}\le\sigma_{D,\ell}<\varsigma_{2n+2}]\mathbf{E}_{x,i}\varsigma_{2n+2}\\
&\le \sum_{n=0}^{\infty}(1-\delta_2)^n\mathbf{E}_{x,i}\varsigma_{2n+2}.
\end{aligned}$$

Following almost the same argument as that for the proof of Lemma 3.6, we can show that $\mathbf{E}_{x,i}\varsigma_{2n}\le nM$ for some positive constant M. Consequently, $\mathbf{E}_{x,i}\sigma_{D,\ell}\le\sum_{n=0}^{\infty}(1-\delta_2)^n(n+1)M<\infty$. The proof of the lemma is thus completed. ∎

Acknowledgments. We thank Professor Rafail Z. Khasminskii for discussions on regime-switching diffusions, for reading an earlier version of the manuscript, and for his comments and suggestions leading to much improvement.

REFERENCES

[1] A. Arapostathis, M.K. Ghosh, and S.I. Marcus, *Harnack's inequality for cooperative weakly coupled elliptic systems*, Comm. Partial Differential Equations, 24 (1999), pp. 1555–1571.

[2] G.K. Basak, A. Bisi, and M.K. Ghosh, *Stability of a random diffusion with linear drift*, J. Math. Anal. Appl., 202 (1996), pp. 604–622.

[3] A. Bensoussan, *Perturbation Methods in Optimal Control*, John Wiley, Chichester, UK, 1988.

[4] A. Bensoussan and P.L. Lions, *Optimal control of random evolutions*, Stochastics, 5 (1981), pp. 169–190.

[5] T. Björk, *Finite dimensional optimal filters for a class of Ito processes with jumping parameters*, Stochastics, 4 (1980), pp. 167–183.

[6] Z.Q. Chen and Z. Zhao, *Potential theory for elliptic systems*, Ann. Probab., 24 (1996), pp. 293–319.

[7] Z.Q. Chen and Z. Zhao, *Harnack inequality for weakly coupled elliptic systems*, J. Differential Equations, 139 (1997), pp. 261–282.

[8] E.B. Dynkin, *Markov Processes*, Vols. I and II, Academic Press, New York, Springer-Verlag, Berlin, 1965.

[9] M.K. Ghosh, A. Arapostathis, and S.I. Marcus, *Ergodic control of switching diffusions*, SIAM J. Control Optim., 35 (1997), pp. 1952–1988.

[10] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 2001.

[11] A.M. Il'in, R.Z. Khasminskii, and G. Yin, *Asymptotic expansions of solutions of integro-differential equations for transition densities of singularly perturbed switching diffusions*, J. Math. Anal. Appl., 238 (1999), pp. 516–539.

[12] J. Jacod and A.N. Shiryayev, *Limit Theorems for Stochastic Processes*, Springer-Verlag, New York, 1980.

[13] Y. Ji and H.J. Chizeck, *Controllability, stabilizability, and continuous-time Markovian jump linear quadratic control*, IEEE Trans. Automat. Control, 35 (1990), pp. 777–788.

[14] J.P. Hespanha, *Stochastic Hybrid Systems: Application to Communication Networks*, Springer, Berlin, 2004.

[15] J.P. Hespanha, *A model for stochastic hybrid systems with application to communication networks*, Nonlinear Anal., 62 (2005), pp. 1353–1383.

[16] R.Z. Khasminskii, *Stochastic Stability of Differential Equations*, Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, 1980.

[17] R.Z. Khasminskii and G. Yin, *Asymptotic behavior of parabolic equations arising from one-dimensional null-recurrent diffusions*, J. Differential Equations, 161 (2000), pp. 154–173.

[18] R.Z. Khasminskii, C. Zhu, and G. Yin, *Stability of regime-switching diffusions*, Stochastic Process. Appl., to appear.

[19] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[20] H.J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser Boston, Boston, MA, 1990.

[21] H.J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Springer-Verlag, New York, 2003.

[22] J.P. LaSalle and S. Lefschitz, *The Stability by Liapunov Direct Method*, Academic Press, New York, 1961.

[23] X. Mao, *Stability of stochastic differential equations with Markovian switching*, Stochastic Process. Appl., 79 (1999), pp. 45–67.

[24] M. Mariton, *Jump Linear Systems in Automatic Control*, Marcel Dekker, New York, 1990.

[25] K. Pichór and R. Rudnicki, *Stability of Markov semigroups and applications to parabolic systems*, J. Math. Anal. Appl., 215 (1997), pp. 56–74.

[26] M. Prandini, J. Hu, J. Lygeros, and S. Sastry, *A probabilistic approach to aircraft conflict detection*, IEEE Trans. Intell. Transportation Systems, 1 (2000), pp. 199–220.

[27] M.H. Protter and H.F. Weinberger, *Maximum Principles in Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1967.

[28] A.V. Skorohod, *Asymptotic Methods in the Theory of Stochastic Differential Equations*, Amer. Math. Soc., Providence, RI, 1989.

[29] D.D. Sworder and V.G. Robinson, *Feedback regulators for jump parameters systems with state and control dependent transition rates*, IEEE Tran. Automat. Control, AC-18 (1973), pp. 355–360.

[30] J.T. Wloka, B. Rowley, and B. Lawruk, *Boundary Value Problems for Elliptic Systems*, Cambridge University Press, Cambridge, UK, 1995.

[31] W.M. Wonham, *Liapunov criteria for weak stochastic stability*, J. Differential Equations, 2 (1966), pp. 195–207.

[32] G. Yin, H.M. Yan, and X.C. Lou, *On a class of stochastic optimization algorithms with applications to manufacturing models*, in Model-Oriented Data Analysis, W.G. Müller, H.P. Wynn, and A.A. Zhigljavsky, eds., Physica-Verlag, Heidelberg, 1993, pp. 213–226.

[33] C. Yuan and X. Mao, *Asymptotic stability in distribution of stochastic differential equations with Markovian switching*, Stochastic Process. Appl., 103 (2003), pp. 277–291.

[34] Q. Zhang, *Stock trading: An optimal selling rule*, SIAM J. Control Optim., 40 (2001), pp. 64–87.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1180–1210

FINITE FUEL PROBLEM IN NONLINEAR SINGULAR STOCHASTIC CONTROL*

MONICA MOTTA† AND CATERINA SARTORI‡

Abstract. We investigate, via the dynamic programming approach, a finite fuel nonlinear singular stochastic control problem of Bolza type. We prove that the associated value function is continuous and that its continuous extension to the closure of the domain coincides with the value function of a nonsingular control problem, for which we prove the existence of an optimal control. Moreover, such a continuous extension is characterized as the unique viscosity solution of a quasi-variational inequality with suitable boundary conditions of mixed type.

Key words. singular stochastic control problems, degenerate parabolic HJB equations, viscosity solutions, representation formulas

AMS subject classification. 49J20, 93E20, 49L25

DOI. 10.1137/050637236

1. Introduction.

We study a finite fuel stochastic control problem with finite horizon via the dynamic programming approach. For any initial condition $(\bar{t}, \bar{k}, \bar{x}) \in [0, T[\times[0, K] \times \mathbb{R}^n$ we consider the nonlinear stochastic differential equation

$$x_t = \bar{x} + \int_{\bar{t}}^{t} A(r, x_r)\, dr + \int_{\bar{t}}^{t} B(r, x_r) u_r\, dr + \int_{\bar{t}}^{t} D(r, x_r)\, d\mathcal{W}_r, \tag{1}$$

where the functions A, B, and D are deterministic, $\{\mathcal{W}_t\}$ is a Brownian motion, and $\{u_t\}$ is a control. All the processes are assumed to be defined on a probability space $(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\})$. Given a closed convex cone $\mathcal{K} \subset \mathbb{R}^m$, the class of admissible controls, denoted by $\mathcal{C}(\bar{t}, \bar{k}, \bar{x})$, is given by the set of $\mathcal{K}$-valued, $\{\mathcal{G}_t\}$-predictable processes verifying the constraint

$$\int_{\bar{t}}^{T} |u_t|\, dt \le K - \bar{k}. \tag{2}$$

For any admissible control u we consider a cost of the form

$$\mathcal{J}(\bar{t}, \bar{k}, \bar{x}, u) = E_Q \left[\int_{\bar{t}}^{T} \left(l_0(r, x_r) + \langle l_1(r, x_r), u_r \rangle \right) dr + g(x_T) \right], \tag{3}$$

where l_0, l_1, and g are deterministic functions. The value function is defined as

$$\mathcal{V}(\bar{t}, \bar{k}, \bar{x}) = \inf_{u \in \mathcal{C}(\bar{t}, \bar{k}, \bar{x})} \mathcal{J}(\bar{t}, \bar{k}, \bar{x}, u). \tag{4}$$

*Received by the editors August 1, 2005; accepted for publication (in revised form) February 21, 2007; published electronically September 5, 2007. This research was partially supported by the M.U.R.S.T. project "Viscosity, Metric, and Control Theoretic Methods for Nonlinear Partial Differential Equations."

http://www.siam.org/journals/sicon/46-4/63723.html

†Dipartimento di Matematica Pura e Applicata, Università di Padova, Via Trieste, 7–35131 Padova, Italy (monica.motta@unipd.it).

‡Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, Università di Padova, Via Trieste, 7–35131 Padova, Italy (caterina.sartori@unipd.it).

In this paper we prove the continuity of the value function, and, via a dynamic programming principle, we show that the function V, which is the continuous extension of $\mathcal{V}$ to $[0,T]\times[0,K]\times\mathbb{R}^n$, is a viscosity solution of the following generalized Cauchy problem:

$$\max\left\{-\tfrac{\partial v}{\partial t}+\mathcal{F}(t,x,Dv,D^2v),-\tfrac{\partial v}{\partial k}+\mathcal{H}(t,x,Dv)\right\}=0 \quad \text{in }]0,T[\times]0,K[\times\mathbb{R}^n, \tag{5}$$

$$\max\left\{-\tfrac{\partial v}{\partial t}+\mathcal{F}(t,x,Dv,D^2v),-\tfrac{\partial v}{\partial k}+\mathcal{H}(t,x,Dv)\right\}\geq 0 \quad \text{on }]0,T[\times\{K\}\times\mathbb{R}^n, \tag{6}$$

$$v\leq g \ \text{ and } \ \max\left\{-\tfrac{\partial \mathrm{v}}{\partial \mathrm{t}}+\mathcal{F}(\mathrm{t},\mathrm{x},\mathrm{Dv},\mathrm{D}^2\mathrm{v}),-\tfrac{\partial \mathrm{v}}{\partial \mathrm{k}}+\mathcal{H}(\mathrm{t},\mathrm{x},\mathrm{Dv})\right\}\geq 0 \text{ if } \mathrm{v}<\mathrm{g} \quad \text{on } \{T\}\times]0,K]\times\mathbb{R}^n, \tag{7}$$

where Dv and D^2v denote the gradient and the matrix of the second derivatives of the function $v=v(t,k,x)$ with respect to the x variable,

$$\mathcal{F}(t,x,p,S)\doteq-\langle A(t,x),p\rangle-l_0(t,x)-\frac{1}{2}\mathrm{Tr}\{\tilde{D}(t,x)S\},$$

where $\tilde{D}(t,x)\doteq D(t,x)D(t,x)^T$, and

$$\mathcal{H}(t,x,p)\doteq\max_{w\in\mathcal{K},\ |w|=1}\{-\langle B(t,x)w,p\rangle-\langle l_1(t,x),w\rangle\}$$

for any $(t,x,p,S)\in\mathbb{R}_+\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbf{M}(n,n)$, where $\mathbf{M}(n,n)$ denotes the set of $n\times n$ real matrices. A uniqueness theorem proven in [MS2] allows us to characterize V as the unique viscosity solution to the above boundary value problem. V is in fact the value function of a more regular problem for which we can also prove the existence of an optimal control.

This paper presents some new results on the dynamic programming approach of the theory of singular stochastic control problems as well as on its probabilistic aspects.

From the probabilistic point of view, we consider a singular control problem with a dynamic and a cost function in which the terms B and l_1 depend explicitly on the state variable x, an important difference with respect to previous works on singular stochastic controls (see, e.g., [HS1], [BC], [CMR], [SS], and the many references in [FS]). In order to deal with such general dynamics and cost, we introduce an extension of our problem by considering a new set of controls, called auxiliary controls, justified by the observation that optimal controls for the above problem may not exist and in fact quasi-optimal controls may be as close as desired to a control of impulsive type (see, e.g., [FS]), so that discontinuous trajectories should be allowed as solutions. It is known that a measure approach works if the terms B in the dynamics and l_1 in the cost do not depend on the state variable x. Such a special class of *state-independent* problems has been widely investigated in recent years and there are several results on the existence of (generalized) optimal controls, on the regularity, and on the characterization of $\mathcal{V}$ as the unique solution of a Cauchy problem for a suitable nonlinear PDE (see [HS1], [HS2], [SS], [CMR], and the references therein). If, instead, B and l_1 depend on the state variable x, a good definition of the solution to (1) requires a completely different approach in order to guarantee its robustness. In the deterministic context Bressan and Rampazzo [BR] introduced a definition of the generalized

solution to (1) based on a time change of the completion of the graphs (t, x_t) (see also [Be], [Se]). Only recently has such a definition been extended to the study of some stochastic control problems by Miller and Runggaldier [MiRu] in 1997, by Dorroh, Ferreyra, and Sundar [DFS] in 1999, and by Miller and Dufour [DM] in 2002. In particular, in [DM] the authors prove the existence of an auxiliary optimal control for a Mayer problem assuming (2). Using their stochastic framework we study the boundary value problem (5)–(7) associated to the minimization problem (1)–(4). In addition, we prove the existence of an optimal control, consistently improving their work since we avoid the convexity assumption, under which their main existence result was proven, but which does not hold in many cases.

This paper is a starting point in the program of extending to stochastic control problems, with nonlinear dynamics and nonlinear costs of the form considered here, several results obtained in singular or coercive linear control problems (that is, with B and l_1 not depending on x) concerning, in particular, the existence of optimal controls and the properties of the associated value functions. Besides the obvious goal of considering nonlinear versions of classical applications such as the finite fuel problem, introduced in [BC] to model an aircraft motion, the study of nonlinear problems is also motivated by some applications to economics for which we refer the interested reader to the recent works [A1], [A2].

From the PDE point of view, we are able to show that the value function is continuous and solves the quasi-variational inequality (5) which can be derived from the dynamic programming principle, either heuristically (as usually done in the literature on singular control), or from an equivalent formulation of the minimum problem that uses compact valued controls (as we do). Here a key tool is the concept of control rules together with the compactification method introduced by El Karoui, Ngoyen, and Jeanblanc-Picqué in [EKNP]. In fact we use an abstract version of the dynamic programming principle (DPP) introduced by Haussmann and Lepeltier [HL] and formulated in terms of control rules, in which, among other things, the terminal time is allowed to be an exit time or even a stopping time chosen by the controller, hard constraints (i.e., state constraints that must be met almost surely) as well as soft constraints (i.e., constraints that must be met in the mean) are considered, and very mild regularity of the data is required.

Thus, the notion of auxiliary controls allows us to reduce the minimization problem to an equivalent one where the controls take values in a compact set. The dynamic programming principle, given in terms of control rules, is the key point for proving that the value function $\mathcal{V}$ defined in (4) is continuous. Both of the concepts are essential in order to write a Hamilton–Jacobi equation like (5) which, a priori, is *not* the formal equation associated to the unbounded control problem, and to show that $\mathcal{V}$ solves (5)–(7) in the viscosity sense. Indeed, the formal equation associated to (1)–(4) involves a different Hamiltonian studied in section 6, which is obtained through a maximization over the unbounded control set $\mathcal{K}$.

A comment about the boundary conditions is in order since conditions (6) and (7) seem original in the setting of singular stochastic control problems. First of all, since we deal with problems of impulsive type, even when considering a finite horizon problem, the limit $\lim_{\bar{t}\to T^-} \mathcal{V}(\bar{t}, \bar{k}, \bar{x})$ does not coincide in general with the final cost $g(\bar{x})$ and, therefore, at time $\bar{t} = T$, we impose (7), which is an alternative between the quasi-variational inequality (5) and $v = g$. Such a generalized boundary condition was introduced in order to characterize continuous value functions of Dirichlet problems in [I], but it also perfectly fits our Cauchy problem. At the boundary $\bar{k} = K$,

instead, we introduce the supersolution condition (6) which replaces the Dirichlet condition $v(\bar{t}, K, \bar{x}) = J(\bar{t}, K, \bar{x}, 0)$, usually assumed in finite fuel control problems (see, e.g., [BJM], [FS]). It has the advantage that it does not require the computation of $J(\bar{t}, K, \bar{x}, 0)$. Supersolution type conditions have been considered first by [So] for problems with state constraints, and in fact by considering the fuel consumed at time t as a new variable, in view of (2) such a variable turns out to be constrained in $[0, K]$. Boundary value problems similar to (5)–(7) for first order Hamiltonians were already investigated by the authors in the context of impulsive deterministic control problems when either a constraint on the L^1 norm of the controls or a weak coercivity condition on the Lagrangian is imposed (see, e.g., [MoRa] and [MS1]). In such a context, it is worth mentioning that our approach leads to approximation schemes for the numerical evaluation of the value functions for first order Hamilton–Jacobi equations (see, e.g., [CF]), which for the second order case has not yet been done.

The paper is organized as follows. In section 2 we state the problem precisely and, following [DM], we introduce an auxiliary control problem whose value function V turns out to coincide with $\mathcal{V}$, but with the essential property that the auxiliary controls are *compact-valued*. In section 3 we introduce relaxed controls and control rules and prove, thanks to some technical results contained in the appendix, that there exists an auxiliary optimal control for our problem and that V is in fact the minimum value over relaxed controls. In section 4, using the DPP, we obtain the continuity of $\mathcal{V}$ (see Theorem 4.1). Section 5 is devoted to deducing the boundary value problem (5)–(7) and to showing that $\mathcal{V}$ is a viscosity solution to it. Then we apply a uniqueness theorem proven in [MS2] and prove in Theorem 5.3 that $\mathcal{V}$ is in fact the only solution to (5)–(7) in the class of the bounded functions which are continuous on $\partial(]0, T[\times]0, K[\times\mathbb{R}^n)$. Moreover, in section 6 we show that $\mathcal{V}$ also turns out to represent a solution of a generalized Cauchy problem for a second order semilinear degenerate parabolic PDE involving a noncoercive Hamiltonian defined via maximization over an unbounded set.

Notation. Throughout this paper we shall adopt the following notation. The symbol $|\cdot|$ denotes the norm of vectors and matrices and $\langle\cdot,\cdot\rangle$ denotes the scalar product for vectors. For any positive integer N and any $r > 0$, $B_N(r) = \{v \in \mathbb{R}^N : |v| < r\}$ and $\overline{B}_N(r) = \{v \in \mathbb{R}^N : |v| \le r\}$. $\mathbb{R}_+ = [0, +\infty[$. For arbitrary positive integers N, M, $\mathbf{M}(N, M)$ denotes the set of the $N \times M$ real matrices. $(^T)$ denotes the transposed operator. $\mathcal{C}_b^2(\mathbb{R}^N)$ is the set of the bounded real maps which are continuous on $\mathbb{R}^N$ with their first and second partial derivatives. Given a function $v : E \to \mathbb{R}$, $E \subset \mathbb{R}^N$, the upper and lower semicontinuous envelopes of v are defined by $v^*(x) \doteq \lim_{s\to 0^+} \sup\{v(y) : y \in E, |y - x| \le s\}$, $v_*(x) \doteq \lim_{s\to 0^+} \inf\{v(y) : y \in E, \quad |y - x| \le s\}$ for any $x \in \overline{E}$. Of course, v^* is upper semicontinuous and v_* is lower semicontinuous. Let $(\Omega, \mathcal{F}, P)$ be a probability space. We will use $E_P[\cdot]$ to denote the mathematical expectation on such a space. Given two random variables X, Y, the notation $X = Y$, $X \le Y$ means $P(X = Y) = 1$, $P(X \le Y) = 1$, respectively, $\delta_{\{w\}}$ denotes the Dirac measure at a fixed $w \in \mathcal{K}$, and T and K are fixed positive real numbers.

2. Statement of the problem. In this section we give the precise formulation of the nonlinear singular stochastic control problem described in the introduction, introduce the auxiliary control problem, and prove their equivalence.

2.1. The control problem. Throughout the paper we will use the following hypotheses.

(A0): There are some constants L_1, L_2 such that the deterministic functions $A : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^n$, $B : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbf{M}(n,m)$, and $D : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbf{M}(n,p)$ verify for all $t, s \in \mathbb{R}_+$ and $x, y \in \mathbb{R}^n$,

$$|A(t,x)| + |B(t,x)| + |D(t,x)| \le L_1(1+|x|),$$
$$|A(t,x) - A(s,y)| + |B(t,x) - B(s,y)| + |D(t,x) - D(s,y)| \le L_2(|t-s| + |x-y|).$$

(A1): There are some constants L, L_3 such that the functions $l_0 : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}$, $l_1 : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^m$, and $g : \mathbb{R}^n \to \mathbb{R}$ verify for all $t, s \in \mathbb{R}_+$ and $x, y \in \mathbb{R}^n$,

$$|l_0(t,x) - l_0(s,y)| + |l_1(t,x) - l_1(s,y)| \le L(|t-s| + |x-y|),$$
$$|g(x) - g(y)| \le L|x-y|,$$

and

$$|l_0(t,x)| + |l_1(t,x)| + |g(x)| \le L_3. \tag{8}$$

DEFINITION 2.1. *Given an initial condition* $(\bar t, \bar k, \bar x) \in [0,T[\times[0,K] \times \mathbb{R}^n$, *a* control *is a term*

$$c = (\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\}, \{u_t\}, \{\mathcal{W}_t\}, \{x_t\}),$$

where

- $(\Omega, \mathcal{G}, Q)$ *is a complete probability space with a right continuous complete filtration* $\{\mathcal{G}_t\}$,
- $\{u_t\}$ *is a* $\mathcal{K}$*-valued process (*$\mathcal{K}$ *a closed, convex cone of* $\mathbb{R}^m$*) defined on* $[\bar t, T] \times \Omega$, *which is* $\{\mathcal{G}_t\}$*-predictable,*
- $\{\mathcal{W}_t\}$ *is a standard* p*-dimensional* $\{\mathcal{G}_t\}$*-Brownian motion,*
- $\{x_t\}$ *is an* $\mathbb{R}^n$*-valued process which is* $\{\mathcal{G}_t\}$*-progressively measurable, with continuous paths, such that*

$$x_t = \bar x + \int_{\bar t}^t A(r, x_r)\, dr + \int_{\bar t}^t B(r, x_r) u_r\, dr + \int_{\bar t}^t D(r, x_r)\, d\mathcal{W}_r \qquad \forall t \in [\bar t, T].$$

A control c *is* admissible *if*

$$\int_{\bar t}^T |u_t|\, dt \le K - \bar k. \tag{9}$$

The set of admissible controls *will be denoted by* $\mathcal{C}(\bar t, \bar k, \bar x)$.
For any admissible control c *we consider a cost of the form*

$$\mathcal{J}(\bar t, \bar k, \bar x, c) \doteq E_Q\left[\int_{\bar t}^T \left(l_0(r, x_r) + \langle l_1(r, x_r), u_r\rangle\right) dr + g(x_T)\right]. \tag{10}$$

The value function is defined for $(\bar t, \bar k, \bar x) \in [0,T[\times[0,K] \times \mathbb{R}^n$ *by*

$$\mathcal{V}(\bar t, \bar k, \bar x) \doteq \inf_{c \in \mathcal{C}(\bar t, \bar k, \bar x)} \mathcal{J}(\bar t, \bar k, \bar x, c). \tag{11}$$

Remark 2.1. If we replace the boundedness hypothesis (8) with

$$|l_0(t,x)| + |l_1(t,x)| + |g(x)| \le L_3(1+|x|) \qquad \forall t \in \mathbb{R}_+,\ x \in \mathbb{R}^n, \tag{12}$$

the main results of the paper remain true, except that, of course, the value function $\mathcal{V}$ is bounded no more but turns out to verify $|\mathcal{V}(t,k,x)| \le \bar C(1+|x|)$ for some $\bar C$ and for all $(t,k,x) \in [0,T[\times[0,K] \times \mathbb{R}^n$, as one can deduce from the proof of Theorem 4.1 (see also Corollary 4.2).

2.2. The auxiliary control problem. In this section, following [DM], we introduce an auxiliary control problem, equivalent to the original one, but with the key property that the controls take values in a compact set.

DEFINITION 2.2. *For any* $(\bar{t}, \bar{k}, \bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n$ *an* auxiliary control *is a term*

$$\beta = (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{w_s\}, \{(t_s, k_s, \xi_s)\}, \theta),$$

where the following **(B1)** *and* **(B2)** *are assumed.*

(B1) • $(\Omega, \mathcal{F}, P)$ *is a complete probability space, with a right continuous complete filtration* $\{\mathcal{F}_s\}$,
- $\{w_s\}$ *is a* $\overline{B}_m(1) \cap \mathcal{K}$*-valued control defined on* $[0, T+K] \times \Omega$ *which is* $\{\mathcal{F}_s\}$*-predictable,*
- θ *is an* $\{\mathcal{F}_s\}$*-stopping time such that* $\theta \le T + K$,

and

(B2) $\{(t_s, k_s, \xi_s)\}$ *is an* $\mathbb{R}^{2+n}$*-valued* $\{\mathcal{F}_s\}$*-progressively measurable process with continuous paths, such that, for* $0 \le s \le T + K$,

$$\begin{cases} t_s = \bar{t} + \int_0^s w_{0_\sigma}\, d\sigma, \\ k_s = \bar{k} + \int_0^s |w_\sigma|\, d\sigma, \\ \xi_s = \bar{x} + \int_0^s (A(t_\sigma, \xi_\sigma) w_{0_\sigma} + B(t_\sigma, \xi_\sigma) w_\sigma)\, d\sigma + \int_0^s D(t_\sigma, \xi_\sigma) \sqrt{w_{0_\sigma}}\, dW_\sigma, \end{cases}$$

where $\{W_s\}$ *is a standard p-dimensional* $\{\mathcal{F}_s\}$*-Brownian motion defined on* $[0, T+K] \times \Omega$ *and where we set* $w_{0_s}(\omega) \doteq 1 - |w_s(\omega)| \forall (s, \omega)$ *just for the sake of notation.*

The cost corresponding to an auxiliary control β *is of the form*

$$J(\bar{t}, \bar{k}, \bar{x}, \beta) \doteq E_P \left[\int_0^\theta (l_0(t_\sigma, \xi_\sigma) w_{0_\sigma} + \langle l_1(t_\sigma, \xi_\sigma), w_\sigma \rangle)\, d\sigma + g(\xi_\theta) + G(t_\theta, k_\theta) \right],$$

where $G(T, k) = 0$ *for all* $k \le K$ *and* $G(t, k) = +\infty$ *otherwise. We use* $\Gamma(\bar{t}, \bar{k}, \bar{x})$ *to denote the set of auxiliary controls, while*

$$\Gamma^a(\bar{t}, \bar{k}, \bar{x}) \doteq \left\{ \beta \in \Gamma(\bar{t}, \bar{k}, \bar{x}) : \quad J(\bar{t}, \bar{k}, \bar{x}, \beta) < +\infty \right\} \tag{13}$$

denotes the subset of admissible *auxiliary controls. We define for every* $(\bar{t}, \bar{k}, \bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n$ *the* auxiliary *value function as*

$$V(\bar{t}, \bar{k}, \bar{x}) \doteq \inf_{\beta \in \Gamma^a(\bar{t}, \bar{k}, \bar{x})} J(\bar{t}, \bar{k}, \bar{x}, \beta). \tag{14}$$

Remark 2.2. The definition of auxiliary controls given in [DM] is slightly different from Definition 2.2. More precisely, fixing an initial condition $(\bar{t}, \bar{k}, \bar{x})$, the natural extension of [DM], to our setting yields controls $\beta = (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{w_s\}, \{(t_s, k_s, \xi_s)\}, \theta)$ with stopping times θ verifying the constraint

$$\theta \le (T - \bar{t}) + (K - \bar{k}), \tag{15}$$

with the cost functional defined by

$$\hat{J}(\bar{t}, \bar{k}, \bar{x}, \beta) = E_P \left[\int_0^\theta (l_0(t_\sigma, \xi_\sigma) w_{0_\sigma} + \langle l_1(t_\sigma, \xi_\sigma), w_\sigma \rangle)\, d\sigma + g(\xi_\theta) + \hat{G}(t_\theta) \right], \tag{16}$$

where $\hat{G}(T) = 0$ and $\hat{G}(t) = +\infty$ for all $t \neq T$. Moreover, a control β is admissible if $\hat{J}(\bar{t}, \bar{k}, \bar{x}, \beta) < +\infty$. In fact, we will show, in the proof of Theorem 2.3 below, that the two sets of admissible auxiliary controls coincide and therefore that the two definitions are equivalent. The reason we choose a different formulation of our problem is that Definition 2.2 is better suited to state a dynamic programming principle. It is well known, indeed, that if (t_s, k_s, ξ_s) is a process starting from $(\bar{t}, \bar{k}, \bar{x})$ at $s = 0$, in order to apply the dynamic programming technique, one needs to consider any state (t_s, k_s, ξ_s) for $s > 0$ as the initial condition, that is, to restate the problem with a random variable in place of a deterministic point as initial datum. Following [DM], therefore, one should deal with the hard constraint $\theta \leq (T - t_s) + (K - k_s)$ coming from condition (15).

The problems in Definitions 2.1 and 2.2 are equivalent in the following sense.

THEOREM 2.3. *Assume* **(A0)**, **(A1)**. *Then for any initial condition* $(\bar{t}, \bar{k}, \bar{x}) \in [0, T[\times[0, K] \times \mathbb{R}^n$ *one has*

(i) $\mathcal{C}(\bar{t}, \bar{k}, \bar{x}) \hookrightarrow \Gamma^a(\bar{t}, \bar{k}, \bar{x})$, *that is, for every control* $c \in \mathcal{C}(\bar{t}, \bar{k}, \bar{x})$ *there exists an admissible auxiliary control* $\beta \in \Gamma^a(\bar{t}, \bar{k}, \bar{x})$ *such that* $J(\bar{t}, \bar{k}, \bar{x}, \beta) = \mathcal{J}(\bar{t}, \bar{k}, \bar{x}, c)$;

(ii) *for any admissible auxiliary control* $\beta \in \Gamma^a(\bar{t}, \bar{k}, \bar{x})$ *there is a sequence of controls* $c^n \in \mathcal{C}(\bar{t}, \bar{k}, \bar{x})$ *such that* $\lim_n \mathcal{J}(\bar{t}, \bar{k}, \bar{x}, c^n) = J(\bar{t}, \bar{k}, \bar{x}, \beta)$;

(iii)

$$V(\bar{t}, \bar{k}, \bar{x}) = \mathcal{V}(\bar{t}, \bar{k}, \bar{x}). \tag{17}$$

Proof. Since the proof is based on that given in [DM], we begin by proving that the set $\Gamma^a(\bar{t}, \bar{k}, \bar{x})$ coincides with the set of admissible auxiliary controls introduced in [DM], that is, controls with stopping time θ verifying (15) and with a cost defined by (16), which must be bounded. To prove this claim, let us fix an admissible auxiliary control β in the sense considered in [DM]. First of all, let us notice that condition (15) in fact plays the role of the integral constraint (9) of Definition 2.1, in that it implies

$$\int_0^\theta |w_\sigma| \, d\sigma \leq K - \bar{k}.$$

Indeed, from $\hat{J}(\bar{t}, \bar{k}, \bar{x}, \beta) < +\infty$ it follows that $E_P[\hat{G}(t_\theta)] = 0$, that is, $t_\theta = T$. Moreover, since the stopping time θ verifies (15), one has

$$k_\theta = \bar{k} + \int_0^\theta |w_\sigma| \, d\sigma = \bar{k} + \theta - (t_\theta - \bar{t}) \leq \bar{k} + (K - \bar{k}) + (T - \bar{t}) - (T - \bar{t}) = K.$$

Thus $E_P[G(t_\theta, k_\theta)] = 0$, $\hat{J}(\bar{t}, \bar{k}, \bar{x}, \beta) = J(\bar{t}, \bar{k}, \bar{x}, \beta)$, and β turns out to belong to the set $\Gamma^a(\bar{t}, \bar{k}, \bar{x})$ defined in (13). On the contrary, given a control $\beta \in \Gamma^a(\bar{t}, \bar{k}, \bar{x})$, from $J(\bar{t}, \bar{k}, \bar{x}, \beta) < +\infty$ it follows that $E_P[G(t_\theta, k_\theta)] = 0$, that is, $t_\theta = T$ and $k_\theta \leq K$. Hence $E_P[\hat{G}(t_\theta)] = 0$, $\hat{J}(\bar{t}, \bar{k}, \bar{x}, \beta) = J(\bar{t}, \bar{k}, \bar{x}, \beta)$, and in order to show that β is an admissible auxiliary control in the sense considered in [DM] it remains to prove that θ verifies condition (15). Since by definition

$$k_s + t_s = \bar{k} + \bar{t} + s \qquad s \geq 0,$$

one has that

$$\theta = (k_\theta - \bar{k}) + (t_\theta - \bar{t}) \leq (K - \bar{k}) + (T - \bar{t}),$$

which concludes the proof of the claim.

If the Lagrangian function $l = l_0 + \langle l_1, u\rangle$ is identically zero, statements (i) and (ii) have been proved by Dufour and Miller in Proposition 4.12, and in Proposition 4.8 and Theorem 4.15 of [DM], respectively. This yields the equality $V = \mathcal{V}$ in (iii) for a problem of Mayer type. The extension of these results to a Bolza problem is standard; therefore the proof is concluded. □

In the deterministic case, following the method of the graphs completion, the equivalence between an original singular control problem and a corresponding auxiliary control problem has been proven for several types of problems (see [MoRa], [MS1], and the references therein).

The following simple example taken from [MoRa] shows that at the points of the form $(T, \bar{k}, \bar{x})$, the value function V associated to the auxiliary control problem does not coincide in general with the terminal cost g.

Example 2.1 (see [MoRa]). Let us consider the deterministic control problem

$$x(t) = \bar{x} + \int_{\bar{t}}^{t} (c + u_1(r) + x(r)u_2(r))\, dr \quad \forall t \in [\bar{t}, T],$$

where $\bar{x} \in \mathbb{R}$, c is a positive constant, the control (u_1, u_2) defined on $[\bar{t}, T]$ assumes values on the closed cone

$$\mathcal{K} \doteq \{(w_1, w_2) \in \mathbb{R}^2 :\ w_1 \le 0,\ w_2 \ge 0\},$$

and it verifies the constraint

$$\int_{\bar{t}}^{T} |(u_1(r), u_2(r))|\, dr \le K - \bar{k},$$

where $0 \le \bar{k} \le K$. Let us minimize the following payoff in Mayer form:

$$\mathcal{J}(\bar{t}, \bar{k}, \bar{x}, u) = \arctan(x(T)).$$

For any $(\bar{t}, \bar{k}, \bar{x}) \in [0, T[\times[0, K] \times \mathbb{R}$, the maximum principle yields the existence of an optimal auxiliary control whose corresponding trajectory has a terminal position given by

$$\xi(T) = \begin{cases} \sinh(\operatorname{arcsinh}(\bar{x}) - (K - \bar{k})) + c(T - \bar{t}), & \bar{x} \le 0, \\ \sinh(\bar{x} - (K - \bar{k})) + c(T - \bar{t}), & 0 < \bar{x} < K - \bar{k}, \\ \bar{x} - (K - \bar{k}) + c(T - \bar{t}), & \bar{x} \ge K - \bar{k}, \end{cases} \tag{18}$$

so that $\mathcal{V}(\bar{t}, \bar{k}, \bar{x}) = V(\bar{t}, \bar{k}, \bar{x}) = \arctan(\xi(T))$. At the points $(T, \bar{k}, \bar{x}) \in \{T\} \times [0, K] \times \mathbb{R}^n$, V is given again by $\arctan(\xi(T))$ once we put $\bar{t} = T$ and obviously it does not coincide with $g(\bar{x}) = \arctan(\bar{x})$ unless $\bar{k} = K$.

This is a general result: V coincides with the continuous extension to $[0, T] \times [0, K] \times \mathbb{R}^n$ of the original value function $\mathcal{V}$ defined on the set $[0, T[\times[0, K] \times \mathbb{R}^n$ (see Corollary 4.2 in section 4) and in general it does not coincide with g at $\bar{t} = T$. For deterministic control problems, there are well known sufficient conditions under which $V(T, \bar{k}, \bar{x}) = g(\bar{x})\ \ \forall (T, \bar{k}, \bar{x}) \in \{T\} \times [0, K] \times \mathbb{R}^n$ (see [RS] and also Notes on [CIL, section 7]).

3. Relaxed controls and control rules. We devote this section to the definition of relaxed controls which are needed in order to introduce the concept of control rules and the compactification method, key tools to prove a dynamic programming principle. We follow here the presentation given by Haussman and Lepeltier in [HL], where an earlier work by El Karoui, Ngoyen, and Jeanblanc-Picqué [EKNP] is generalized to the case of unbounded data and controls and no fixed terminal time.

3.1. The martingale model. We introduce the equivalent formulation of the above auxiliary control problem as a martingale problem, where the ambiguous term represented by the Brownian motion, unknown in advance, is removed (see, e.g., Ikeda and Watanabe in[IW]). To this end, we introduce for all $\varphi \in \mathcal{C}_b^2(\mathbf{R}^{2+n})$, $(t,k,x) \in \mathbf{R}^{2+n}$, and $w \in \overline{B}_m(1) \cap \mathcal{K}$ the operator $\mathcal{L}$ defined by

$$\begin{aligned}&\mathcal{L}\varphi(t,k,x,w) \doteq \\ &\left[\tfrac{1}{2}\sum_{ij}\tilde{D}_{ij}(t,x)\tfrac{\partial^2\varphi}{\partial x_i \partial x_j}(t,k,x) + \sum_i A_i(t,x)\tfrac{\partial\varphi}{\partial x_i}(t,k,x) + \tfrac{\partial\varphi}{\partial t}(t,k,x)\right] w_0 \\ &\qquad + \sum_i \langle B_i(t,x), w\rangle \tfrac{\partial\varphi}{\partial x_i}(t,k,x) + \tfrac{\partial\varphi}{\partial k}(t,k,x)|w|,\end{aligned} \tag{19}$$

where $w_0 \doteq 1 - |w|$, $\tilde{D}_{ij}$ are the entries of $\tilde{D} = DD^T$, A_i are the components of A, and B_i are the rows of B. Notice that in this formulation the diffusion coefficient D disappears and is replaced by $\tilde{D}$, which, differently from D, is something intrinsic to a process ξ_s as defined in **(B2)**.

The following proposition establishes the correspondence between the martingale model and the control problem with the Brownian motion.

PROPOSITION 3.1 (see [HL, Proposition 3.1]). *Let us assume* **(A0)**, **(A1)**. *Let us fix* $(\bar{t},\bar{k},\bar{x}) \in [0,T]\times[0,K]\times\mathbb{R}^n$. *A control* $\beta = (\Omega,\mathcal{F},P,\{\mathcal{F}_s\},\{w_s\},\{(t_s,k_s,\xi_s)\},\theta)$ *such that*

(B3)
- $(\Omega,\mathcal{F},P)$ *is a probability space, with a filtration* $\{\mathcal{F}_s\}$,
- $\{w_s\}$ *is a* $\overline{B}_m(1)\cap\mathcal{K}$*-valued control, defined on* $[0,T+K]\times\Omega$, $\{\mathcal{F}_s\}$*-progressively measurable,*
- θ *is an* $\{\mathcal{F}_s\}$*-stopping time such that* $\theta \le T+K$

verifies **(B2)** *if and only if it verifies*

(B4)
- $\{(t_s,k_s,\xi_s)\}$ *is a* $\mathbb{R}^{2+n}$*-valued,* $\{\mathcal{F}_s\}$*-progressively measurable process for* $s \in [0,T+K]$, *with continuous paths, such that* $(t_s,k_s,\xi_s) = (\bar{t},\bar{k},\bar{x})$ *for* $s=0$, *for any* $\varphi \in \mathcal{C}_b^2(\mathbf{R}^{2+n})$, $\mathcal{M}_s(\varphi,\beta)$ *is a* $(P,\{\mathcal{F}_s\})$ *square integrable martingale for* $s \in [0,T+K]$, *where*

$$\mathcal{M}_s(\varphi,\beta) \doteq \varphi(t_s,k_s,\xi_s) - \int_0^s \mathcal{L}\varphi(t_\sigma,k_\sigma,\xi_\sigma,w_\sigma)\,d\sigma.$$

3.2. Relaxed controls. In a relaxed control, the $\overline{B}_m(1)\cap\mathcal{K}$-valued process $\{w_s\}$ is replaced by an $\mathbf{M}_1(\overline{B}_m(1)\cap\mathcal{K})$-valued process $\{\mu_s\}$, where $\mathbf{M}_1(\overline{B}_m(1)\cap\mathcal{K})$ is the space of probability measures on $\overline{B}_m(1)\cap\mathcal{K}$. We will extend any bounded measurable map $\psi : \overline{B}_m(1)\cap\mathcal{K} \to \mathbb{R}$ to $\mathbf{M}_1(\overline{B}_m(1)\cap\mathcal{K})$ by setting

$$\psi(\mu) = \int_{\overline{B}_m(1)\cap\mathcal{K}} \psi(w)\mu(dw).$$

DEFINITION 3.2. *Given* $(\bar{t},\bar{k},\bar{x}) \in [0,T]\times[0,K]\times\mathbb{R}^n$ *we say that* $\tilde{\alpha}$ *is a* relaxed control *and write* $\tilde{\alpha} \in \tilde{\Gamma}(\bar{t},\bar{k},\bar{x})$ *if*

$$\tilde{\alpha} = (\Omega,\mathcal{F},P,\{\mathcal{F}_s\},\{\mu_s\},\{(t_s,k_s,\xi_s)\},\theta),$$

where the following **(B3′)**, **(B4′)** *are assumed.*

(B3′)
- $(\Omega,\mathcal{F},P)$ *is a probability space with a filtration* $\{\mathcal{F}_s\}$,
- $\{\mu_s\}$ *is a* $\mathbf{M}_1(\overline{B}_m(1)\cap\mathcal{K})$*-valued process defined on* $[0,T+K]\times\Omega$ *which is* $\{\mathcal{F}_s\}$*-progressively measurable,*
- θ *is an* $\{\mathcal{F}_s\}$*-stopping time such that* $\theta \le T+K$,

(B4′) • *$\{(t_s, k_s, \xi_s)\}$ is an $\mathbb{R}^{2+n}$-valued $\{\mathcal{F}_s\}$-progressively measurable process for $s \in [0, T+K]$, with continuous paths, such that $(t_s, k_s, \xi_s) = (\bar{t}, \bar{k}, \bar{x})$ for $s = 0$, for any $\varphi \in \mathcal{C}_b^2(\mathbf{R}^{2+n})$, $\mathcal{M}_s(\varphi, \tilde{\alpha})$ is a $(P, \{\mathcal{F}_s\})$ square integrable martingale for $s \in [0, T+K]$, where*

$$M_s(\varphi, \tilde{\alpha}) \doteq \varphi(t_s, k_s, \xi_s) - \int_0^s \mathcal{L}\varphi(t_\sigma, k_\sigma, \xi_\sigma, \mu_\sigma)\, d\sigma.$$

For any $\tilde{\alpha} \in \tilde{\Gamma}(\bar{t}, \bar{k}, \bar{x})$ we define the cost

$$J(\bar{t}, \bar{k}, \bar{x}, \tilde{\alpha}) = E_P \left[\int_0^\theta \left(l_0(t_\sigma, \xi_\sigma)(1 - |\mu_\sigma|) + \langle l_1(t_\sigma, \xi_\sigma), \mu_\sigma \rangle \right) d\sigma + g(\xi_\theta) + G(t_\theta, k_\theta) \right]. \tag{20}$$

We use $\tilde{\Gamma}^a(\bar{t}, \bar{k}, \bar{x})$ to denote the subset of admissible *relaxed controls, that is,*

$$\tilde{\Gamma}^a(\bar{t}, \bar{k}, \bar{x}) \doteq \left\{ \tilde{\alpha} \in \tilde{\Gamma}(\bar{t}, \bar{k}, \bar{x}) : \quad J(\bar{t}, \bar{k}, \bar{\xi}, \tilde{\alpha}) < +\infty \right\}.$$

Remark 3.1. Following [HL], the processes that appear in Definition 3.2 are progressively measurable and the probability space is arbitrary. The processes that appear in the auxiliary controls of Definition 2.2, instead, are predictable processes and the probability space is complete and right continuous. Thus, it is not obvious a priori that the control problem in Definition 3.2 is the relaxed version of our auxiliary control problem. From Lemmatas A1–A3 in [DM], however, it follows that, given an initial condition $(\bar{t}, \bar{k}, \bar{x})$, for any control $\alpha = (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{w_s\}, \{(t_s, k_s, \xi_s)\}, \theta)$ verifying **(B3)** and **(B2)** (or, equivalently, **(B3)** and **(B4)**, in view of Proposition 3.1), there exists a new control $\hat{\alpha} = (\hat{\Omega}, \hat{\mathcal{F}}, \hat{P}, \{\hat{\mathcal{F}}_s\}, \{\hat{w}_s\}, \{(\hat{t}_s, \hat{k}_s, \hat{\xi}_s)\}, \hat{\theta})$, where $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$ is a suitable modification of $(\Omega, \mathcal{F}, P)$, $\hat{\theta} = \theta$, the process $\{(\hat{t}_s, \hat{k}_s, \hat{\xi}_s)\}$ is indistinguishable from $\{(t_s, k_s, \xi_s)\}$, $\hat{\alpha}$ verifies **(B1)** and **(B2)**, and, moreover, $J(\bar{t}, \bar{k}, \bar{x}, \hat{\alpha}) = J(\bar{t}, \bar{k}, \bar{x}, \alpha)$. Therefore, if $J(\bar{t}, \bar{k}, \bar{x}, \alpha) < +\infty$, then $\hat{\alpha} \in \Gamma^a(\bar{t}, \bar{k}, \bar{x})$.

The set $\Gamma^a(\bar{t}, \bar{k}, \bar{x})$ can be naturally embedded in $\tilde{\Gamma}^a(\bar{t}, \bar{k}, \bar{x})$; therefore, the inequality

$$\inf_{\tilde{\alpha} \in \tilde{\Gamma}^a(\bar{t}, \bar{k}, \bar{x})} J(\bar{t}, \bar{k}, \bar{x}, \tilde{\alpha}) \leq \inf_{\alpha \in \Gamma^a(\bar{t}, \bar{k}, \bar{x})} J(\bar{t}, \bar{k}, \bar{x}, \alpha)$$

is trivially verified. In fact, the converse inequality also holds true.

THEOREM 3.3. *Assume* **(A0)**, **(A1)**. *Then for any $(\bar{t}, \bar{k}, \bar{x}) \in [0, T] \times [0, K] \times \mathbb{R}^n$,*

$$V(\bar{t}, \bar{k}, \bar{x}) = \inf_{\alpha \in \Gamma^a(\bar{t}, \bar{k}, \bar{x})} J(\bar{t}, \bar{k}, \bar{x}, \alpha) = \inf_{\tilde{\alpha} \in \tilde{\Gamma}^a(\bar{t}, \bar{k}, \bar{x})} J(\bar{t}, \bar{k}, \bar{x}, \tilde{\alpha}).$$

Moreover, the infimum over relaxed controls is attained and so is the infimum over auxiliary controls.

Remark 3.2. Dufour and Miller proved in [DM] the existence of an optimal control for the auxiliary problem in the case $l = l_0 + \langle l_1, u \rangle \equiv 0$ and while under the assumption that the set

$$\begin{aligned} &\tilde{M}(t, x) \doteq \big\{ (A(t, x)(1 - |w|) + B(t, x)w, (1 - |w|)D(t, x)D^T(t, x), |w|) : \\ &w \in \overline{B}_m(1) \cap \mathcal{K} \big\} \qquad \text{is convex} \quad \forall (t, x). \end{aligned} \tag{21}$$

It is important to observe that the presence of the terms depending on $|w|$ in (21) implies that such a condition does not hold in most cases. Let us point out that in

Theorem 3.3 the method of the graphs completion yields instead the existence of an optimal control for the auxiliary control problem just under assumptions **(A0)**, **(A1)**, without assumption (21).

Proof of Theorem 3.3. From [HL, Theorem 3.6] it follows straightforwardly that for any initial condition $(\bar t, \bar k, \bar x) \in [0,T]\times[0,K]\times\mathbb{R}^n$ there exists an optimal relaxed control

$$\tilde\alpha = (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{\mu_s\}, \{(t_s, k_s, \xi_s)\}, \theta) \in \tilde\Gamma^a(\bar t, \bar k, \bar x).$$

Let us set $\mu_{0_s}(\omega) = 1 - |\mu_s(\omega)|$ (pointwise) and let us define γ by

$$\begin{aligned}\gamma(s,\omega) &= \Big(A\mu_{0_s}(\omega) + B\mu_s(\omega), \mu_{0_s}(\omega)\tilde D, l_0\mu_{0_s}(\omega) + \langle l_1, \mu_s(\omega)\rangle, \mu_{0_s}(\omega)\Big)\,(t_s(\omega), \xi_s(\omega))\\ &= \textstyle\int_{\overline{B}_m(1)\cap\mathcal{K}} \Big(Aw_0 + Bw, w_0\tilde D, l_0 w_0 + \langle l_1, w\rangle, w_0\Big)\,(t_s(\omega), \xi_s(\omega))\,\mu_s(\omega, dw),\end{aligned}$$

where $w_0 = 1 - |w|$ for $w \in \overline{B}_m(1)\cap\mathcal{K}$. Since for every (t,x),

$$\begin{aligned}&\Big\{(A(t,x)w_0 + B(t,x)w, w_0\tilde D(t,x), z, w_0) :\\ &z \ge l_0(t,x)w_0 + \langle l_1(t,x), w\rangle, \quad (w_0, w) \in \mathbb{R}_+\times\mathcal{K},\ w_0 + |w| = 1\Big\},\\ &\subset \Big\{(A(t,x)w_0 + B(t,x)w, w_0\tilde D(t,x), z, w_0) :\\ &z \ge l_0(t,x)w_0 + \langle l_1(t,x), w\rangle, \quad (w_0, w) \in \mathbb{R}_+\times\mathcal{K},\ w_0 + |w| \le 1\Big\},\end{aligned}$$

where the last set is a compact, convex subset of $\mathbb{R}^n\times\mathbf{M}(n,n)\times\mathbb{R}^2$, arguing as in the proof of Theorem 3.6 in [HL] (see also Theorem A9 in [HL]) one can show that there exist

$$\begin{aligned}&\textit{two } \mathcal{F}_s\text{-progressively measurable processes } \{v_s\},\ \{(w_{0_s}, w_s)\},\\ &R_+ \text{ and } (\mathbb{R}_+\times\mathcal{K})\cap\{(w_0,w) :\ w_0 + |w| \le 1\}\text{-valued, respectively,}\end{aligned} \tag{22}$$

such that one has

$$\begin{aligned}\gamma(s,\omega) = &\Big(Aw_{0_s}(\omega) + Bw_s(\omega), w_{0_s}(\omega)\tilde D, l_0 w_{0_s}(\omega) + \langle l_1, w_s(\omega)\rangle, w_{0_s}(\omega)\Big)\,(t_s(\omega), \xi_s(\omega))\\ &+(0,0,v_s(\omega),0) \qquad\qquad \text{for almost all } (s,\omega).\end{aligned} \tag{23}$$

Let us define the *noncanonical control* α by

$$\alpha \doteq (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{(w_{0_s}, w_s)\}, \{(t_s, k_s, \xi_s)\}, \theta), \tag{24}$$

where $\{(w_{0_s}, w_s)\}$ is given in (22). In the rest of the proof, with a slight abuse of notation, let us use again the symbols J and $\mathcal{L}$ to denote the cost and the operator in (19) once the scalar process $\{w_{0_s}\}$ in their definitions is not subjected to the constraint $w_{0_s} = 1 - |w_s|$, but is an independent process with values in $[0,1]$. Then by (23) it follows that for any $\varphi \in \mathcal{C}_b^2(\mathbf{R}^{2+n})$, $\mathcal{L}\varphi(t_s, k_s, \xi_s, \mu_s) = \mathcal{L}\varphi(t_s, k_s, \xi_s, (w_{0_s}, w_s))$ except on a (s,ω) null set. Hence for all φ and all $s \in [0, T+K]$,

$$\mathcal{M}_s(\varphi, \tilde\alpha) = \mathcal{M}_s(\varphi, \alpha).$$

Moreover, since

$$\begin{aligned}l_0(t_s,\xi_s)\mu_{0_s} + \langle l_1(t_s,\xi_s), \mu_s\rangle &= l_0(t_s,\xi_s)w_{0_s} + \langle l_1(t_s,\xi_s), w_s\rangle + v_s\\ &\ge l_0(t_s,\xi_s)w_{0_s} + \langle l_1(t_s,\xi_s), w_s\rangle,\end{aligned}$$

and $J(\bar{t},\bar{k},\bar{x},\tilde{\alpha}) < +\infty$, then $J(\bar{t},\bar{k},\bar{x},\alpha) \le J(\bar{t},\bar{k},\bar{x},\tilde{\alpha}) < +\infty$. The noncanonical control α does not belong, however, to the set $\Gamma^a(\bar{t},\bar{k},\bar{x})$ since it does not verify **(B1)** and $w_{0_s} \neq 1 - |w_s|$. Therefore, in order to conclude the proof, it remains to prove the following.

Claim. There exists a control $\check{\alpha} = (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{\check{w}_s\}, \{(\check{t}_s, \check{k}_s, \check{\xi}_s)\}, \check{\theta})$ verifying **(B3)** and **(B4)** and such that $J(\bar{t},\bar{k},\bar{x},\check{\alpha}) = J(\bar{t},\bar{k},\bar{x},\alpha)$.

In view of Remark 3.1, indeed, this is sufficient for the existence of a control $\hat{\alpha} \in \Gamma^a(\bar{t},\bar{k},\bar{x})$ such that $J(\bar{t},\bar{k},\bar{x},\hat{\alpha}) = J(\bar{t},\bar{k},\bar{x},\alpha)$ and $\hat{\alpha}$ is the required optimal auxiliary control. The claim will be proved in appendix (Remark 7.1 and Lemma 7.3). □

Remark 3.3. In general, the original control problem described in Definition 2.1 does not have an optimal control while, by Theorem 3.3, the auxiliary control problem does. Thanks to (ii) of Theorem 2.3, this yields a sequence of suboptimal controls $c^n \in \mathcal{C}(\bar{t},\bar{k},\bar{x})$ for the original problem for any $(\bar{t},\bar{k},\bar{x}) \in [0,T[\times[0,K] \times \mathbb{R}^n$.

3.3. Control rules. We are now going to recall very briefly the definition of control rules (for a detailed description, see [HL]). In order to introduce a canonical space for the problem, let us define the following spaces:

$$\mathcal{C}^{2+n} = \{f : [0, T+K] \to \mathbb{R}^{2+n},\ f \text{ continuous}\},$$

endowed with the topology of uniform convergence;

$$\mathcal{U} \doteq \{\nu : [0, T+K] \to \mathbf{M}_1(\overline{B}_m(1) \cap \mathcal{K}),\ \nu \text{ Borel measurable}\},$$

endowed with the stable topology;

$$\mathcal{Z} = \{\zeta : [0, T+K] \to \mathbb{R},\ \zeta = \chi_{s \ge \Delta},\ \Delta \in [0, +\infty]\}, \tag{25}$$

endowed with the topology of weak convergence of the corresponding (point) probability measures. We denote the map $\zeta \to \Delta$ by $\Delta(\cdot)$. Let $\tilde{\mathcal{C}}, \tilde{\mathcal{U}}, \tilde{\mathcal{Z}}$ denote their Borel σ-fields, let $\tilde{\mathcal{C}}_s, \tilde{\mathcal{U}}_s, \tilde{\mathcal{Z}}_s$ denote the σ-fields up to time s (e.g., $\tilde{\mathcal{Z}}_s = \sigma\{\zeta(s') :\ 0 \le s' \le s\}$), and let us introduce the canonical setting

$$\Omega = \mathcal{C}^{2+n} \times \mathcal{U} \times \mathcal{Z}, \qquad \mathcal{F} \doteq \tilde{\mathcal{C}} \times \tilde{\mathcal{U}} \times \tilde{\mathcal{Z}}, \qquad \mathcal{F}_s \doteq \tilde{\mathcal{C}}_s \times \tilde{\mathcal{U}}_s \times \tilde{\mathcal{Z}}_s. \tag{26}$$

Notice that Ω is metrizable and separable under the product topology.

DEFINITION 3.4. *Fix $(\bar{t},\bar{k},\bar{x}) \in [0,T] \times [0,K] \times \mathbf{R}^n$, and let Ω, $\mathcal{F}$ and $\{\mathcal{F}_s\}$ be defined by* (26). *We say that R is a* control rule *and we write that $R \in \mathcal{R}(\bar{t},\bar{k},\bar{x})$ if R is a probability measure on the canonical space $(\Omega,\mathcal{F})$, such that*

$$\tilde{\alpha} = (\Omega, \mathcal{F}, R, \{\mathcal{F}_s\}, \{\mu_s\}, \{(t_s, k_s, \xi_s)\}, \theta)$$

is a relaxed control (i.e., $\tilde{\alpha} \in \tilde{\Gamma}(\bar{t},\bar{k},\bar{x})$), where

$$(t_s, k_s, \xi_s)(\omega) = f_s, \quad \mu_s(\omega) = \nu_s, \quad \theta(\omega) = \Delta(\zeta)$$

for $\omega = (f, \nu, \zeta) \in \Omega$. Finally, we define the cost associated to R as $J(\bar{t},\bar{k},\bar{x},R) \doteq J(\bar{t},\bar{k},\bar{x},\tilde{\alpha})$, where $J(\bar{t},\bar{k},\bar{x},\tilde{\alpha})$ is given in (20). *The subset $\mathcal{R}^a(\bar{t},\bar{k},\bar{x})$ of the* admissible control rules *can be now defined as follows:*

$$\mathcal{R}^a(\bar{t},\bar{k},\bar{x}) \doteq \left\{R \in \mathcal{R}(\bar{t},\bar{k},\bar{x}) :\quad J(\bar{t},\bar{k},\bar{\xi},R) < +\infty\right\}.$$

Remark 3.4. For the sake of notation, in what follows a given element ω of the canonical space $\mathcal{C}^{2+n} \times \mathcal{U} \times \mathcal{Z}$ will be denoted by $\omega = ((t_\cdot, k_\cdot, \xi_\cdot), \mu_\cdot, \theta)$.

By definition, $\mathcal{R}(\bar{t}, \bar{k}, \bar{x}) \hookrightarrow \tilde{\Gamma}(\bar{t}, \bar{k}, \bar{x})$. In fact, the inverse embedding is also valid. In particular, one has the following proposition.

PROPOSITION 3.5. *Fix $(\bar{t}, \bar{k}, \bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n$ and assume* **(A0)**, **(A1)**. *Then*

$$V(\bar{t}, \bar{k}, \bar{x}) = \inf_{\tilde{\alpha} \in \tilde{\Gamma}^a(\bar{t}, \bar{k}, \bar{x})} J(\bar{t}, \bar{k}, \bar{x}, \tilde{\alpha}) = \inf_{R \in \mathcal{R}^a(\bar{t}, \bar{k}, \bar{x})} J(\bar{t}, \bar{k}, \bar{x}, R). \tag{27}$$

Moreover, the infimum is attained in any one of $\mathcal{R}^a(\bar{t}, \bar{k}, \bar{x})$, $\tilde{\Gamma}^a(\bar{t}, \bar{k}, \bar{x})$, and $\Gamma^a(\bar{t}, \bar{k}, \bar{x})$.

Proof. The first equality in (27) has been obtained in Theorem 3.3, while the second one follows from Theorem 3.13 in [HL]. The minimum for the auxiliary (and hence, for any other) control problem exists in view of Theorem 3.3. □

Let us conclude this subsection by recalling a dynamic programming principle established in [HL]. To this end, let us notice that the auxiliary control problem is in fact an *unconstrained* stopping time control problem. Indeed, from Definition 2.2 it follows that for all $(\bar{t}, \bar{k}, \bar{x})$ such that either $\bar{t} > T$ or $\bar{k} > K$, the set of admissible auxiliary controls $\Gamma^a(\bar{t}, \bar{k}, \bar{x})$ is empty. Hence the auxiliary value function V might be extended to the whole set $[0, +\infty[\times [0, +\infty[\times \mathbb{R}^n$ in a natural way by setting $V = +\infty$ outside $[0,T] \times [0,K] \times \mathbb{R}^n$.

PROPOSITION 3.6. *Assume* **(A0)**, **(A1)**. *For any $(\bar{t}, \bar{k}, \bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n$, one has*

$$V(\bar{t}, \bar{k}, \bar{x}) = \inf \left\{ E_R \left[\int_0^{\rho'} (l_0(t_\sigma, \xi_\sigma)(1 - |\mu_\sigma|) + \langle l_1(t_\sigma, \xi_\sigma), \mu_\sigma \rangle)\, d\sigma + V(t_{\rho'}, k_{\rho'}, \xi_{\rho'}) \right] \right\}, \tag{28}$$

where the infimum is taken over the set $\mathcal{R}^a(\bar{t}, \bar{k}, \bar{x})$ and $\rho' = \rho \wedge \theta$, ρ being any finite stopping time such that $0 \le \rho \le \theta$.

4. Continuity of the value function.

THEOREM 4.1. *Let* **(A0)**, **(A1)** *hold. Then the value function V is bounded and continuous. More precisely, there exists some $\bar{C} > 0$ such that V satisfies the following:*

$$|V(\bar{t}, \bar{k}, \bar{x})| \le \bar{C} \qquad \forall (\bar{t}, \bar{k}, \bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n;$$

$$|V(\bar{t}_1, \bar{k}_1, \bar{x}_1) - V(\bar{t}_2, \bar{k}_2, \bar{x}_2)| \le \bar{C} \left[|\bar{x}_1 - \bar{x}_2| + (1 + |\bar{x}_1| \vee |\bar{x}_2|) \left(|\bar{t}_1 - \bar{t}_2|^{1/2} + |\bar{k}_1 - \bar{k}_2| \right) \right]$$

for all $(\bar{t}_1, \bar{k}_1, \bar{x}_1)$, $(\bar{t}_2, \bar{k}_2, \bar{x}_2) \in [0,T] \times [0,K] \times \mathbb{R}^n$.

Proof (*Boundedness*). It is very easy to see that for any initial condition $(\bar{t}, \bar{k}, \bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n$ the set of admissible control rules is nonempty. Since the stopping time θ is bounded from above by $T + K$, the boundedness of V follows, therefore, straightforwardly from the boundedness of both the process $\{w_s\}$ and the data l_0, l_1, and g.

Lipschitz continuity in x. Fix $(\bar{t}, \bar{k}, \bar{x}_1), (\bar{t}, \bar{k}, \bar{x}_2) \in [0,T] \times [0,K] \times \mathbb{R}^n$ and assume that $V(\bar{t}, \bar{k}, \bar{x}_1) \ge V(\bar{t}, \bar{k}, \bar{x}_2)$. One has

$$0 \le V(\bar{t}, \bar{k}, \bar{x}_1) - V(\bar{t}, \bar{k}, \bar{x}_2) \le \sup_{P \in \mathcal{R}^a(\bar{t}, \bar{k}, \bar{x}_2)} \big(J(\bar{t}, \bar{k}, \bar{x}_1, Q) - J(\bar{t}, \bar{k}, \bar{x}_2, P) \big)$$

for every $Q \in \mathcal{R}^a(\bar{t}, \bar{k}, \bar{x}_1)$. Take $P \in \mathcal{R}^a(\bar{t}, \bar{k}, \bar{x}_2)$ arbitrary and let

$$(\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{\mu_s\}, \{(t_s, k_s, \xi_{2_s})\}, \theta)$$

be the associated relaxed control. By the definition of control rules, there exists an extension $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}_s\})$ of $(\Omega, \mathcal{F}, P, \{\mathcal{F}_s\})$, i.e., there exists another probability space $(\Omega', \mathcal{F}', \mathcal{F}'_s, P')$ such that $\tilde{\Omega} = \Omega \times \Omega'$, $\tilde{\mathcal{F}} = \mathcal{F} \times \mathcal{F}'$, $\tilde{\mathcal{F}}_s = \mathcal{F}_s \times \mathcal{F}'_s$, and $\tilde{P} = P \times P'$. We can extend the process $\{((t_\cdot, k_\cdot, \xi_\cdot), \mu_\cdot, \theta)\}$ to $\tilde{\Omega}$ by the following: for $\tilde{\omega} = (\omega, \omega') \in \tilde{\Omega}$,

$$(t_\cdot, k_\cdot, \xi_\cdot)(\tilde{\omega}) = (t_\cdot, k_\cdot, \xi_\cdot)(\omega), \quad \tilde{\mu}_\cdot(\tilde{\omega}) = \mu_\cdot(\omega), \quad \theta(\tilde{\omega}) = \theta(\omega).$$

On $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \tilde{\mathcal{F}}_s)$ there exists a standard p-dimensional Brownian motion $\{W_s\}$ such that for $s \in [0, T + K]$,

$$\begin{aligned}
t_s &= \bar{t} + \textstyle\int_0^s (1 - |\mu|_\sigma)\, d\sigma, \\
k_s &= \bar{k} + \textstyle\int_0^s |\mu_\sigma|\, d\sigma, \\
\xi_{2s} &= \bar{x}_2 + \textstyle\int_0^s \left(A(t_\sigma, \xi_{2\sigma})(1 - |\mu|_\sigma) + B(t_\sigma, \xi_{2\sigma})\mu_\sigma\right) d\sigma + \int_0^s D(t_\sigma, \xi_{2\sigma})\sqrt{1 - |\mu|_\sigma}\, dW_\sigma,
\end{aligned}$$

the control $\tilde{\beta} = (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}_s\}, \{\mu_s\}, \{(t_s, k_s, \xi_{2s})\}, \theta) \in \tilde{\Gamma}^a(\bar{t}, \bar{k}, \bar{x}_2)$, where, by the definition of the set $\mathcal{Z}$ in (25), θ is the first time in which $\chi_{s \geq \theta}$ jumps from 0 to 1 and $J(\bar{t}, \bar{k}, \bar{x}_2, \tilde{\beta}) = J(\bar{t}, \bar{k}, \bar{x}_2, \tilde{P}) = J(\bar{t}, \bar{k}, \bar{x}_2, P)$.

Consider the equations with the initial condition $(\bar{t}, \bar{k}, \bar{x}_1)$, for $s \in [0, T + K]$,

$$\begin{aligned}
t_s &= \bar{t} + \textstyle\int_0^s (1 - |\mu|_\sigma)\, d\sigma, \\
k_s &= \bar{k} + \textstyle\int_0^s |\mu_\sigma|\, d\sigma, \\
\xi_{1s} &= \bar{x}_1 + \textstyle\int_0^s \left(A(t_\sigma, \xi_{1\sigma})(1 - |\mu|_\sigma) + B(t_\sigma, \xi_{1\sigma})\mu_\sigma\right) d\sigma + \int_0^s D(t_\sigma, \xi_{1\sigma})\sqrt{1 - |\mu|_\sigma}\, dW_\sigma
\end{aligned} \tag{29}$$

on the stochastic basis $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}_s\})$. Under assumptions **(A0)**, **(A1)**, the strong solution to (29) exists and one can see that $\tilde{\alpha} = (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}_s\}, \{\mu_s\}, \{(t_s, k_s, \xi_{1s})\}, \theta) \in \tilde{\Gamma}^a(\bar{t}, \bar{k}, \bar{x}_1)$. Therefore, there exists a control rule $Q \in \mathcal{R}^a(\bar{t}, \bar{k}, \bar{x}_1)$ such that

$$J(\bar{t}, \bar{k}, \bar{x}_1, \tilde{\alpha}) = J(\bar{t}, \bar{k}, \bar{x}_1, Q).$$

We have

$$\begin{aligned}
&J(\bar{t}, \bar{k}, \bar{x}_1, Q) - J(\bar{t}, \bar{k}, \bar{x}_2, P) = J(\bar{t}, \bar{k}, \bar{x}_1, \tilde{\alpha}) - J(\bar{t}, \bar{k}, \bar{x}_2, \tilde{\beta}) \\
&\leq E_{\tilde{P}}\left[\int_0^\theta |l_0(t_\sigma, \xi_{1\sigma}) - l_0(t_\sigma, \xi_{2\sigma})|\, |1 - |\mu_\sigma||\, d\sigma + \int_0^\theta |l_1(t_\sigma, \xi_{1\sigma}) - l_1(t_\sigma, \xi_{2\sigma})||\mu_\sigma|\, d\sigma \right. \\
&\qquad \left. + |g(\xi_{1\theta}) - g(\xi_{2\theta})|\right] \leq L\, E_{\tilde{P}}\left[\int_0^\theta |\xi_{1\sigma} - \xi_{2\sigma}|\, d\sigma\right] + L\, E_{\tilde{P}}[|\xi_{1\theta} - \xi_{2\theta}|],
\end{aligned}$$

where we have used the Lipschitz continuity of l_0, l_1, and g and L is the same as in **(A1)**. Let us define $\hat{\xi}_{i_s} \doteq \xi_{i_{s \wedge \theta}}$ for all $s \geq 0$ and $i = 1, 2$. By the Burkholder–Gundy and Gronwall inequalities, we obtain that there exists a constant C, depending on the Lipschitz constant L_2 in **(A0)** and on $T + K$, such that, for all $0 \leq \sigma \leq T + K$,

$$E_{\tilde{P}}\left[\sup_{s \leq \sigma}(|\hat{\xi}_{1_s} - \hat{\xi}_{2s}|^2)\right] \leq C|\bar{x}_1 - \bar{x}_2|^2.$$

Since from the definitions of $\{\hat{\xi}_{1_s}\}$ and $\{\hat{\xi}_{2_s}\}$ it follows that

$$\begin{aligned}
E_{\tilde{P}}\left[\int_0^\theta |\xi_{1\sigma} - \xi_{2\sigma}|\, d\sigma\right] &\leq E_{\tilde{P}}\left[\int_0^{T+K} |\hat{\xi}_{1\sigma} - \hat{\xi}_{2\sigma}|\, d\sigma\right] \\
&\leq \left(\int_0^{T+K} E_{\tilde{P}}\left[\sup_{s \leq \sigma}(|\hat{\xi}_{1_s} - \hat{\xi}_{2s}|^2)\right] d\sigma\right)^{1/2},
\end{aligned} \tag{30}$$

in view of the arbitrariness of $P \in \mathcal{R}^a(\bar t, \bar k, \bar x_2)$, the previous estimates yield that

$$0 \le V(\bar t, \bar k, \bar x_1) - V(\bar t, \bar k, \bar x_2) \le \bar C |\bar x_1 - \bar x_2|$$

for a suitable constant $\bar C$, depending just on L, L_2, and $T + K$.

Hölder continuity in t. Fix $(\bar t_1, \bar k, \bar x)$, $(\bar t_2, \bar k, \bar x) \in [0, T] \times [0, K] \times \mathbb{R}^n$ and assume that $V(\bar t_1, \bar k, \bar x) \ge V(\bar t_2, \bar k, \bar x)$.

Case 1. $\bar t_1 < \bar t_2$. One has

$$0 \le V(\bar t_1, \bar k, \bar x) - V(\bar t_2, \bar k, \bar x) \le \sup_{P \in \mathcal{R}^a(\bar t_2, \bar k, \bar x)} \big(J(\bar t_1, \bar k, \bar x, Q) - J(\bar t_2, \bar k, \bar x, P) \big)$$

for every $Q \in \mathcal{R}^a(\bar t_1, \bar k, \bar x)$. Take $P \in \mathcal{R}^a(\bar t_2, \bar k, \bar x)$ arbitrary and let

$$(\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{\mu_s\}, \{(t_{2_s}, k_s, \xi_{2_s})\}, \theta_2)$$

be the associated relaxed control. Now, as in the previous step, there exist an extension $(\tilde\Omega, \tilde{\mathcal{F}}, \tilde P, \{\tilde{\mathcal{F}}_s\})$ of $(\Omega, \mathcal{F}, P, \{\mathcal{F}_s\})$ and a standard Brownian motion $\{W_s\}$ on $(\tilde\Omega, \tilde{\mathcal{F}}, \tilde P, \{\tilde{\mathcal{F}}_s\})$ such that, for $s \in [0, T + K]$,

$$\begin{aligned}
&t_{2s} = \bar t_2 + \textstyle\int_0^s (1 - |\mu_\sigma|)\, d\sigma,\\
&k_s = \bar k + \textstyle\int_0^s |\mu_\sigma|\, d\sigma,\\
&\xi_{2s} = \bar x + \textstyle\int_0^s \big(A(t_{2_\sigma}, \xi_{2_\sigma})(1 - |\mu_\sigma|) + B(t_{2_\sigma}, \xi_{2_\sigma})\mu_\sigma\big)\, d\sigma + \int_0^s D(t_{2_\sigma}, \xi_{2_\sigma})\sqrt{1 - |\mu_\sigma|}\, dW_\sigma,
\end{aligned}$$

the control $\tilde\beta = (\tilde\Omega, \tilde{\mathcal{F}}, \tilde P, \{\tilde{\mathcal{F}}_s\}, \{\mu_s\}, \{(t_{2_s}, k_s, \xi_{2_s})\}, \theta_2) \in \tilde\Gamma^a(\bar t_2, \bar k, \bar x)$, and $J(\bar t_2, \bar k, \bar x, \tilde\beta) = J(\bar t_2, \bar k, \bar x, P)$. Let us now consider the relaxed control that one obtains from the definition of $\tilde\beta$ when μ_s is replaced by $\mu_s \chi_{\{s \le \theta_2\}}$ for $s \ge 0$. It is easy to see that this control is admissible, that is, it belongs to $\tilde\Gamma^a(\bar t_2, \bar k, \bar x)$ and the corresponding cost coincides with $J(\bar t_2, \bar k, \bar x, P)$. With a small abuse of notation, from now on let us use $\tilde\beta$ to denote such control.

Let us introduce the stopping time $\theta_1 \doteq \theta_2 + (\bar t_2 - \bar t_1)$ and let $\{(t_{1_s}, k_s, \xi_{1_s})\}$ be the strong solution to

$$\begin{aligned}
&t_{1_s} = \bar t_1 + \textstyle\int_0^s (1 - |\mu_\sigma|)\, d\sigma,\\
&k_s = \bar k + \textstyle\int_0^s |\mu_\sigma|\, d\sigma,\\
&\xi_{1_s} = \bar x + \textstyle\int_0^s \big(A(t_{1_\sigma}, \xi_{1_\sigma})(1 - |\mu_\sigma|) + B(t_{1_\sigma}, \xi_{1_\sigma})\mu_\sigma\big)\, d\sigma + \int_0^s D(t_{1_\sigma}, \xi_{1_\sigma})\sqrt{1 - |\mu_\sigma|}\, dW_\sigma
\end{aligned}$$

on the stochastic basis $(\tilde\Omega, \tilde{\mathcal{F}}, \tilde P, \{\tilde{\mathcal{F}}_s\})$ for $s \in [0, T + K]$. As claimed in Remark 2.2, $\tilde\beta$ admissible implies that $\theta_2 \le (T - \bar t_2) + (K - \bar k)$, $t_{2_{\theta_2}} = T$, and $k_{\theta_2} \le K$. Hence one deduces that

$$\theta_1 \le (T - \bar t_2) + (K - \bar k) + (\bar t_2 - \bar t_1) = (T - \bar t_1) + (K - \bar k) \le T + K.$$

Moreover, since we identified μ_s with $\mu_s \chi_{\{s \le \theta_2\}}$ one has

$$t_{1_{\theta_1}} = t_{2_{\theta_2}} + (\theta_1 - \theta_2) - (\bar t_2 - \bar t_1), \qquad k_{\theta_1} = k_{\theta_2} \le K.$$

Therefore, $t_{1_{\theta_1}} = T$, $k_{\theta_1} \le K$, and the control $\tilde\alpha = (\tilde\Omega, \tilde{\mathcal{F}}, \tilde P, \{\tilde{\mathcal{F}}_s\}, \{\mu_s\}, \{(t_{1_s}, k_s, \xi_{1_s})\}, \theta_1)$ is in $\tilde\Gamma^a(\bar t_1, \bar k, \bar x)$. Thus there exists a control rule $Q \in \mathcal{R}^a(\bar t_1, \bar k, \bar x)$ such that

$J(\bar{t}_1, \bar{k}, \bar{x}, \tilde{\alpha}) = J(\bar{t}_1, \bar{k}, \bar{x}, Q)$. We have

$$\begin{aligned}
&J(\bar{t}_1, \bar{k}, \bar{x}, Q) - J(\bar{t}_2, \bar{k}, \bar{x}, P) = J(\bar{t}_1, \bar{k}, \bar{x}, \tilde{\alpha}) - J(\bar{t}_2, \bar{k}, \bar{x}, \tilde{\beta}) \\
&\le E_{\tilde{P}}\left[|g(\xi_{1_{\theta_1}}) - g(\xi_{2_{\theta_2}})|\right] + E_{\tilde{P}}\left[\int_0^{\theta_2} |l_0(t_{1_\sigma}, \xi_{1_\sigma}) - l_0(t_{2_\sigma}, \xi_{2_\sigma})|\ |1 - |\mu_\sigma||\, d\sigma\right] \\
&+E_{\tilde{P}}\left[\int_0^{\theta_2} |l_1(t_{1_\sigma}, \xi_{1_\sigma}) - l_1(t_{2_\sigma}, \xi_{2_\sigma})|\ |\mu_\sigma|\, d\sigma\right] \\
&+E_{\tilde{P}}\left[\int_{\theta_2}^{\theta_2 + (\bar{t}_2 - \bar{t}_1)} |l_0(t_{1_\sigma}, \xi_{1_\sigma})|\ |1 - |\mu_\sigma||\, d\sigma\right] \\
&\le L\left[E_{\tilde{P}}[|\xi_{1_{\theta_1}} - \xi_{2_{\theta_2}}|^2]\right]^{\frac{1}{2}} + LE_{\tilde{P}}\left[\int_0^{\theta_2} (|t_{1_\sigma} - t_{2_\sigma}| + |\xi_{1_\sigma} - \xi_{2_\sigma}|)\, d\sigma\right] + L_3(\bar{t}_2 - \bar{t}_1),
\end{aligned}$$

where the constants L and L_3 are the same as in **(A1)**. In order to conclude the proof, let us introduce for $s \ge 0$ the processes $\hat{\xi}_{i_s} \doteq \xi_{i_{s \wedge \theta_i}}$, for $i = 1, 2$. Since

$$t_{1_{\theta_2}} = t_{2_{\theta_2}} - (\bar{t}_2 - \bar{t}_1),$$

by standard calculations (see, e.g., [F]) one can prove that

$$E_{\tilde{P}}\left[\sup_{s \le \sigma} |\hat{\xi}_{2_s} - \hat{\xi}_{1_s}|^2\right] \le C^2(1 + |\bar{x}|)^2 |\bar{t}_2 - \bar{t}_1|,$$

for every $0 \le \sigma \le T + K$, with C a suitable constant depending on L_1, L_2 in **(A0)** and $T + K$. From the definitions of $\{\hat{\xi}_{2_s}\}$ and $\{\hat{\xi}_{1_s}\}$, this yields

$$E_{\tilde{P}}\left[\left|\xi_{2_{\theta_2}} - \xi_{1_{\theta_1}}\right|^2\right] = E_{\tilde{P}}\left[\left|\hat{\xi}_{2_{T+K}} - \hat{\xi}_{1_{T+K}}\right|^2\right] \le C^2(1 + |\bar{x}|)^2 |\bar{t}_2 - \bar{t}_1|.$$

Therefore, by (30) we obtain

$$J(\bar{t}_1, \bar{k}, \bar{x}, Q) - J(\bar{t}_2, \bar{k}, \bar{x}, P) \le \bar{C}\left[(1 + |\bar{x}|)|\bar{t}_1 - \bar{t}_2|^{\frac{1}{2}} + |\bar{t}_1 - \bar{t}_2|\right],$$

which, by the arbitrariness of P, yields

$$0 \le V(\bar{t}_2, \bar{k}, \bar{x}) - V(\bar{t}_1, \bar{k}, \bar{x}) \le \bar{C}(1 + |\bar{x}|)|\bar{t}_1 - \bar{t}_2|^{\frac{1}{2}}$$

for some constant $\bar{C}$ depending on the constants L, L_2, L_3, and $T + K$ in **(A0)**, **(A1)**.

Case 2. $\bar{t}_1 > \bar{t}_2$. Consider the dynamic programming principle (28) for $V(\bar{t}_2, \bar{k}, \bar{x})$,

$$\begin{aligned}
V(\bar{t}_2, \bar{k}, \bar{x}) = \inf_{R \in \mathcal{R}^a(\bar{t}_2, \bar{k}, \bar{x})} \Bigg\{ E_R\Bigg[\int_0^{r \wedge \theta} &(l_0(t_\sigma, \xi_\sigma)(1 - |\mu_\sigma|) \\
&+ \langle l_1(t_\sigma, \xi_\sigma), \mu_\sigma \rangle)\, d\sigma + V(t_{r \wedge \theta}, k_{r \wedge \theta}, \xi_{r \wedge \theta}) \Bigg] \Bigg\},
\end{aligned} \tag{31}$$

where we choose the (deterministic) time $r = \bar{t}_1 - \bar{t}_2$. It is easy to see that there exists an admissible control rule $P \in \mathcal{R}^a(\bar{t}_2, \bar{k}, \bar{x})$ associated to a relaxed control

$$(\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{\mu_s\}, \{(t_s, k_s, \xi_s)\}, \theta)$$

and such that

$$P(\mu_s = \delta_{\{0\}} \quad 0 \le s \le \theta, \quad \theta = T - \bar{t}_2) = 1.$$

Then $P(\theta \geq r) = 1$; by the boundedness of l_0 one has

$$V(\bar{t}_2, \bar{k}, \bar{x}) \leq E_P \left[\int_0^r |l_0(t_\sigma, \xi_\sigma)| \, d\sigma + V(t_r, k_r, \xi_r) \right] \leq L_3 r + E_P[V(t_r, k_r, \xi_r)],$$

and by the Lipschitz continuity of the value function in x,

$$V(t_r, k_r, \xi_r) \leq V(t_r, k_r, \bar{x}) + C|\xi_r - \bar{x}|.$$

Hence

$$V(\bar{t}_2, \bar{k}, \bar{x}) - E_P[V(t_r, k_r, \bar{x})] \leq L_3 r + C E_P[|\xi_r - \bar{x}|] \leq L_3 r + C(E_P[|\xi_r - \bar{x}|^2])^{\frac{1}{2}}. \tag{32}$$

From the definition of control rules, we know that under P,

$$\begin{aligned} &t_r = \bar{t}_2 + r = \bar{t}_1, \\ &k_r = \bar{k}, \\ &\xi_r = \bar{x} + \textstyle\int_0^r A(t_\sigma, \xi_\sigma) \, d\sigma + M_r, \end{aligned} \tag{33}$$

where $\{M_r\}$ is a continuous square integrable martingale with

$$\langle M \rangle_r = \int_0^r \mathrm{Tr}\{\tilde{D}(t_\sigma, \xi_\sigma)\} \, d\sigma.$$

Therefore, by the Burkholder–Davis–Gundy inequality there exists a constant C, depending on L_1 in **(A0)**, such that

$$E_P[|\xi_r - \bar{x}|^2] \leq C^2(1 + |\bar{x}|)^2(r^2 + r). \tag{34}$$

Therefore, (32), (33), and (34) yield

$$0 \leq V(\bar{t}_2, \bar{k}, \bar{x}) - V(\bar{t}_1, \bar{k}, \bar{x}) \leq \bar{C}(1 + |\bar{x}|)|\bar{t}_2 - \bar{t}_1|^{\frac{1}{2}}$$

for some constant $\bar{C}$ depending on the constants introduced in **(A0)**, **(A1)**.

Lipschitz continuity in k. Fix $(\bar{t}, \bar{k}_1, \bar{x}), (\bar{t}, \bar{k}_2, \bar{x}) \in [0, T] \times [0, K] \times \mathbb{R}^n$, and assume that $V(\bar{t}, \bar{k}_1, \bar{x}) \geq V(\bar{t}, \bar{k}_2, \bar{x})$.

Case 1. $\bar{k}_1 < \bar{k}_2$. One has

$$0 \leq V(\bar{t}, \bar{k}_1, \bar{x}) - V(\bar{t}, \bar{k}_2, \bar{x}) \leq \sup_{P \in \mathcal{R}^a(\bar{t}, \bar{k}_2, \bar{x})} \big(J(\bar{t}, \bar{k}_1, \bar{x}, Q) - J(\bar{t}, \bar{k}_2, \bar{x}, P) \big)$$

for every $Q \in \mathcal{R}^a(\bar{t}, \bar{k}_1, \bar{x})$. As in the previous step, take $P \in \mathcal{R}^a(\bar{t}, \bar{k}_2, \bar{x})$ arbitrary and let $\{W_s\}$ be a standard Brownian motion on a suitable $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}_s\})$ such that

$$\tilde{\beta} = (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}_s\}, \{\mu_s\}, \{(t_s, k_{2_s}, \xi_s)\}, \theta_2) \in \tilde{\Gamma}^a(\bar{t}, \bar{k}_2, \bar{x})$$

is a relaxed control, where, for $s \in [0, T + K]$,

$$\begin{aligned} &t_s = \bar{t} + \textstyle\int_0^s (1 - |\mu_\sigma|) \, d\sigma, \\ &k_{2_s} = \bar{k}_2 + \textstyle\int_0^s |\mu_\sigma| \, d\sigma, \\ &\xi_s = \bar{x} + \textstyle\int_0^s \left(A(t_\sigma, \xi_\sigma)(1 - |\mu_\sigma|) + B(t_\sigma, \xi_\sigma)\mu_\sigma \right) d\sigma + \int_0^s D(t_\sigma, \xi_\sigma)\sqrt{1 - |\mu_\sigma|} \, dW_\sigma, \end{aligned}$$

and $J(\bar{t}, \bar{k}_2, \bar{x}, \tilde{\beta}) = J(\bar{t}, \bar{k}_2, \bar{x}, P)$. Moreover, setting for $s \geq 0$,

$$k_{1_s} \doteq \bar{k}_1 + \int_0^s |\mu_\sigma| \, d\sigma = k_{2_s} - (\bar{k}_2 - \bar{k}_1),$$

one easily sees that the control

$$\tilde{\alpha} = (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}, \{\tilde{\mathcal{F}}_s\}, \{\mu_s\}, \{(t_s, k_{1_s}, \xi_s)\}, \theta_2)$$

is in $\tilde{\Gamma}^a(\bar{t}, \bar{k}_1, \bar{x})$. As before, there exists a control rule $Q \in \mathcal{R}^a(\bar{t}, \bar{k}_1, \bar{x})$ such that $J(\bar{t}, \bar{k}_1, \bar{x}, \tilde{\alpha}) = J(\bar{t}, \bar{k}_1, \bar{x}, Q)$. Since the cost functional J and the state process $\{\xi_s\}$ do not depend explicitly on the k variable, one has that

$$J(\bar{t}, \bar{k}_2, \bar{x}, P) = J(\bar{t}, \bar{k}_2, \bar{x}, \tilde{\beta}) = J(\bar{t}, \bar{k}_1, \bar{x}, \tilde{\alpha}) = J(\bar{t}, \bar{k}_1, \bar{x}, Q).$$

As a consequence, in this case,

$$V(\bar{t}, \bar{k}_1, \bar{x}) = V(\bar{t}, \bar{k}_2, \bar{x}).$$

Case 2. $\bar{k}_1 > \bar{k}_2$. Consider the dynamic programming principle (28) for $V(\bar{t}, \bar{k}_2, \bar{x})$,

$$V(\bar{t}, \bar{k}_2, \bar{x}) = \inf_{R \in \mathcal{R}^a(\bar{t}, \bar{k}_2, \bar{x})} \left\{ E_R \left[\int_0^{r \wedge \theta} (l_0(t_\sigma, \xi_\sigma)(1 - |\mu_\sigma|) + \langle l_1(t_\sigma, \xi_\sigma), \mu_\sigma \rangle) \, d\sigma + V(t_{r\wedge\theta}, k_{r\wedge\theta}, \xi_{r\wedge\theta}) \right] \right\},$$

where we choose the (deterministic) time $r = \bar{k}_1 - \bar{k}_2$. Let us fix an arbitrary $w \in \mathcal{K}$ with $|w| = 1$. Then there exists a control rule $P \in \mathcal{R}^a(\bar{t}, \bar{k}_2, \bar{x})$ associated to a relaxed control

$$\tilde{\beta} = (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{\mu_s\}, \{(t_s, k_s, \xi_s)\}, \theta)$$

such that

$$P(\mu_s = \delta_{\{w\}} \quad 0 \le s \le \theta, \quad \theta = K - \bar{k}_2) = 1,$$

and $J(\bar{t}, \bar{k}_2, \bar{x}, \tilde{\beta}) = J(\bar{t}, \bar{k}_2, \bar{x}, P)$. Then, arguing as in the case "$\bar{t}_1 > \bar{t}_2$" of the proof of the continuity in t, we can deduce that an estimate analogous to (32) is still verified, that is

$$V(\bar{t}, \bar{k}_2, \bar{x}) - E_P[V(t_r, k_r, \bar{x})] \le L_3 r + C E_P[|\xi_r - \bar{x}|] \le L_3 r + C(E_P[|\xi_r - \bar{x}|^2])^{\frac{1}{2}}. \tag{35}$$

Now, under P we have

$$\begin{aligned} &t_r = \bar{t}, \\ &k_r = \bar{k}_2 + r = \bar{k}_1, \\ &\xi_r = \bar{x} + \textstyle\int_0^r B(t_\sigma, \xi_\sigma) \, d\sigma. \end{aligned} \tag{36}$$

Therefore, since $E_P[|B(t_s, \xi_s)|^2] \le E_P[[L_1(1 + |\xi_s|)]^2] \le C^2(1 + |\bar{x}|)^2$, we deduce that for $0 \le r \le \theta$,

$$E_P[|\xi_r - \bar{x}|^2] \le C^2(1 + |\bar{x}|)^2 r^2. \tag{37}$$

Then (35), (36), and (37) yield

$$0 \le V(\bar{t}, \bar{k}_2, \bar{x}) - V(\bar{t}, \bar{k}_1, \bar{x}) \le \bar{C}(1 + |\bar{x}|)|\bar{k}_2 - \bar{k}_1|.$$

The proof of Theorem 4.1 is thus concluded. □

By Theorem 2.3, as a straightforward consequence of Theorem 4.1 one has the following corollary.

COROLLARY 4.2. *Assume* **(A0)**, **(A1)**. *Then there exists a unique, bounded, continuous extension of the value function* $\mathcal{V} : [0,T[\times[0,K] \times \mathbb{R}^n \to \mathbb{R}$, *still denoted by* $\mathcal{V}$, *to the closed set* $[0,T] \times [0,K] \times \mathbb{R}^n$ *which coincides with the auxiliary value function* V. *Hence there exists some* $\bar{C} > 0$ *such that*

$$|\mathcal{V}(\bar{t},\bar{k},\bar{x})| \le \bar{C} \qquad \forall (\bar{t},\bar{k},\bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n,$$

$$|\mathcal{V}(\bar{t}_1,\bar{k}_1,\bar{x}_1) - \mathcal{V}(\bar{t}_2,\bar{k}_2,\bar{x}_2)| \le \bar{C}[|\bar{x}_1 - \bar{x}_2| + (1 + |\bar{x}_1| \vee |\bar{x}_2|)(|\bar{t}_1 - \bar{t}_2|^{1/2} + |\bar{k}_1 - \bar{k}_2|)]$$

for all $(\bar{t}_1,\bar{k}_1,\bar{x}_1)$, $(\bar{t}_2,\bar{k}_2,\bar{x}_2) \in [0,T] \times [0,K] \times \mathbb{R}^n$.

Therefore, from now on $\mathcal{V}$ will denote the extension of $\mathcal{V}$ to $[0,T] \times [0,K] \times \mathbb{R}^n$, which exists being equal to V.

5. Dynamic programming equation and boundary conditions. This section is devoted to showing that the value function $\mathcal{V}$ is a viscosity solution of (5)–(7). To this end, we will recall below the definition of viscosity sub- and supersolutions with generalized boundary conditions (see, e.g., [CIL]). A formal derivation of the boundary value problem described in the introduction is given in the following subsection.

5.1. Heuristic derivation of the quasi-variational inequality and of the boundary conditions. It is quite easy to deduce heuristically the boundary value problem (5)–(7) once we consider the value function V of the auxiliary optimization control problem defined in Definition 2.2 to which our original control problem, introduced in Definition 2.1, is equivalent. The auxiliary control problem is indeed formulated as an *unconstrained stopping time problem*, with bounded controls and discontinuous final cost given by

$$\tilde{G}(t,k,x) \doteq g(x) - G(t,k) \qquad \forall (t,k,x) \in \mathbb{R}^{2+n}.$$

Therefore, assuming V of class $C^{1,2}$, using Ito's formula and arguing as usual (see, e.g., [FS]), we can deduce from the dynamic programming principle (28) that V verifies the following equation:

$$\tilde{\mathcal{F}}\left(x, DV, \frac{\partial V}{\partial t}, \frac{\partial V}{\partial k}, D^2V\right) = 0 \qquad \text{in }]0,T[\times]0,K[\times\mathbb{R}^n,$$

where

$$\tilde{\mathcal{F}}(x,p_x,p_t,p_k,S) \doteq \max_{\{(w_0,w):\ w_0\ge 0,\ \ w\in\mathcal{K},\ w_0+|w|=1\}}\{-\tfrac{1}{2}w_0\mathrm{Tr}\{\tilde{D}(t,x)S\}$$
$$-\langle A(t,x)w_0 + B(t,x)w, p_x\rangle - l_0(t,x)w_0 - \langle l_1(t,x), w\rangle - p_t\, w_0 - p_k|w|\},$$

which is, in turn, equivalent to the quasi-variational inequality (5), as shown in [MS2].

More precisely, one can show that the value function of an optimal stopping time problem verifies

$$\max\left\{\tilde{\mathcal{F}}\left(x, DV, \frac{\partial V}{\partial t}, \frac{\partial V}{\partial k}, D^2V\right); V - \tilde{G}\right\} = 0 \qquad \text{in } \mathbb{R}^{2+n}, \tag{38}$$

due to the fact that the controller can decide to stop as soon as it is convenient (for the derivation of (38) in a viscosity framework we refer to [BP] and [BCD]). Since

$V(t,k,x) = +\infty$ outside $[0,T] \times [0,K] \times \mathbb{R}^n$ and the lower semicontinuous exit cost $\tilde{G}(t,k,x)$ is equal to $g(x)$ for $(t,x,k) \in \{T\} \times [0,K] \times \mathbb{R}^n$ and to $+\infty$ otherwise, by (38) it follows easily that (7) holds for every $(t,x,k) \in \{T\} \times [0,K] \times \mathbb{R}^n$ and that (6) holds for $(t,x,k) \in [0,T[\times[0,K] \times \mathbb{R}^n$.

We underline that, as far as we know, there is not in the literature a dynamic programming principle for the problem of Definition 2.1; hence, even assuming the value function $\mathcal{V}$ to be regular enough, there is no way to deduce the equation and the boundary conditions (5)–(7) directly for the original control problem.

5.2. Viscosity solution.

DEFINITION 5.1. *A locally bounded function v defined on $[0,T] \times [0,K] \times \mathbb{R}^n$ is a* viscosity subsolution *of* (5)–(7) *if for every point $\bar{z} = (\bar{t}, \bar{k}, \bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n$ and for every map $\phi \in C_b^2([0,T] \times [0,K] \times \mathbb{R}^n)$ such that $v^* - \phi$ has a local maximum at $\bar{z}$ one has*

$$\max\left\{-\frac{\partial \phi}{\partial t}(\bar{z}) + \mathcal{F}(\bar{z}, D\phi(\bar{z}), D^2\phi(\bar{z})), -\frac{\partial \phi}{\partial k}(\bar{z}) + \mathcal{H}(\bar{z}, D\phi(\bar{z}))\right\} \le 0$$

if $\bar{z} \in]0,T[\times]0,K[\times\mathbb{R}^n$, and

$$v^*(\bar{z}) \le g(\bar{x})$$

if $\bar{z} \in \{T\}\times]0,K] \times \mathbb{R}^n$.

A locally bounded function v defined on $[0,T] \times [0,K] \times \mathbb{R}^n$ is a viscosity supersolution *of* (5)–(7) *if for every point $\bar{z} = (\bar{t}, \bar{k}, \bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n$ and for every map $\phi \in C_b^2([0,T] \times [0,K] \times \mathbb{R}^n)$ such that $v_* - \phi$ has a local minimum at $\bar{z}$ one has*

$$\max\left\{-\frac{\partial \phi}{\partial t}(\bar{z}) + \mathcal{F}(\bar{z}, D\phi(\bar{z}), D^2\phi(\bar{z})), -\frac{\partial \phi}{\partial k}(\bar{z}) + \mathcal{H}(\bar{z}, D\phi(\bar{z}))\right\} \ge 0$$

if $\bar{z} \in]0,T[\times]0,K] \times \mathbb{R}^n$, and

$$\max\left\{-\frac{\partial \phi}{\partial t}(\bar{z}) + \mathcal{F}(\bar{z}, D\phi(\bar{z}), D^2\phi(\bar{z})), -\frac{\partial \phi}{\partial k}(\bar{z}) + \mathcal{H}(\bar{z}, D\phi(\bar{z}))\right\} \ge 0 \quad \text{or} \quad v_*(\bar{z}) \ge g(\bar{x})$$

if $\bar{z} \in \{T\}\times]0,K] \times \mathbb{R}^n$.

A locally bounded function v defined on $[0,T] \times [0,K] \times \mathbb{R}^n$ is called a viscosity solution *of* (5)–(7) *if it is both a viscosity sub- and supersolution of* (5)–(7).

Example 5.1. Consider the control problem introduced in Example 2.1 and the following boundary value problem:

$$\max\left\{-\frac{\partial v}{\partial t} - cDv, -\frac{\partial v}{\partial k} + \mathcal{H}(x, Dv)\right\} = 0 \quad \text{in }]0,T[\times]0,K[\times\mathbb{R}, \tag{39}$$

$$\max\left\{-\frac{\partial v}{\partial t} - cDv, -\frac{\partial v}{\partial k} + \mathcal{H}(x, Dv)\right\} \ge 0 \quad \text{on }]0,T[\times\{K\} \times \mathbb{R}, \tag{40}$$

$$\begin{aligned} v(T,k,x) \le \arctan(x) \ \text{ and } \ & \max\left\{-\tfrac{\partial v}{\partial t} - cDv, -\tfrac{\partial v}{\partial k} + \mathcal{H}(x, Dv)\right\} \ge 0 \\ \text{if } \ v(T,k,x) < \arctan(x) \ & \text{ on } \{T\}\times]0,K] \times \mathbb{R}, \end{aligned} \tag{41}$$

where

$$\mathcal{H}(x,p) = \max_{(w_1,w_2) \in \mathcal{K},\, |(w_1,w_2)| = 1} \{-(w_1 + xw_2)p\}.$$

It is not difficult to prove that the value function $\mathcal{V}$ is a classical solution of (39, while we refer to [MoRa] to show that it satisfies (40)–(41) in the viscosity sense.

THEOREM 5.2. *Assume* **(A0)**, **(A1)**. *Then the value function* $\mathcal{V} : [0,T]\times[0,K]\times\mathbb{R}^n \to \mathbb{R}$ *solves the boundary value problem* (5)–(7) *in the viscosity sense.*

Proof. We postpone the proof of this theorem to the end of this section. □

THEOREM 5.3. *Assume* **(A0)**, **(A1)**. *Then the value function* $\mathcal{V} : [0,T]\times[0,K]\times\mathbb{R}^n \to \mathbb{R}$ *is the unique viscosity solution of* (5)–(7) *among the bounded functions defined on* $[0,T]\times[0,K]\times\mathbb{R}^n$ *which are continuous on* $\partial(]0,T[\times]0,K[\times\mathbb{R}^n)$.

Proof. This result follows straightforwardly from Corollary 6.1 of Theorem 3.1 in [MS2] in view of Theorem 5.2 and Corollary 4.2. □

Remark 5.1. When the boundedness assumption (8) in **(A1)** is weakened in the linear growth condition (12) introduced in Remark 2.1, in order to apply Corollary 6.1 in [MS2] we have to introduce the following stronger growth hypothesis on A, B, and $\tilde{D}$:

(A2) for any $\varepsilon > 0$ there is a constant $C_\varepsilon > 0$ for which

$$|A(t,x)| + |B(t,x)| \le C_\varepsilon + \varepsilon|x|, \quad |\tilde{D}(t,x)| \le C_\varepsilon + \varepsilon|x|^2 \quad \forall (t,x) \in [0,T]\times\mathbb{R}^n.$$

Then under hypotheses **(A0)**, **(A1)** with (8) replaced by (12) and **(A2)**, we obtain the uniqueness for the viscosity solution to (5)–(7) among the functions v which are continuous on $\partial(]0,T[\times]0,K]\times\mathbb{R}^n)$ and such that

$$\sup_{x\in\mathbb{R}^n} \frac{|v(t,k,x)|}{1+|x|} < +\infty \qquad \text{uniformly for } \ (t,k) \in [0,T]\times[0,K].$$

Proof of Theorem 5.2. Since $\mathcal{V} = V$, let us prove the theorem for V. Owing to Theorem 4.1, V is continuous, so that $V^* = V_* = V$. In what follows, for any $\bar{z} = (\bar{t},\bar{k},\bar{x}) \in [0,T]\times[0,K]\times\mathbb{R}^n$ and any $r' > 0$ let us set $\Theta_{r'} \doteq [0 \vee (\bar{t}-r'), T \wedge (\bar{t}+r')[\times[0 \vee (\bar{k}-r'), K \wedge (\bar{k}+r')[\times B_n(r')$.

Step 1. Let us start by showing that V is a viscosity subsolution of (5)–(7). Since at any point of the form $(T,\bar{k},\bar{x})$ (with $\bar{k} < K$ and $\bar{x} \in \mathbb{R}^n$) it is clear that there exists a control rule $P \in \mathcal{R}^a(T,\bar{k},\bar{x})$ such that

$$P(\theta = 0) = 1,$$

from the very definition of V it follows that $V(T,\bar{k},\bar{x}) \le g(\bar{x})$. Hence it remains to prove that V is a viscosity subsolution of (5). We argue by contradiction. If this fails to hold, then there is a point $\bar{z} = (\bar{t},\bar{k},\bar{x}) \in]0,T[\times]0,K[\times\mathbb{R}^n$, a test function $\phi \in C_b^2([0,T]\times[0,K]\times\mathbb{R}^n)$, and a constant $r_1 > 0$ such that $\bar{z}$ is a local maximum point for $V - \phi$, that is,

$$V(z) - \phi(z) \le V(\bar{z}) - \phi(\bar{z}) \quad \forall z = (t,k,x) \in \overline{\Theta}_{r_1},$$

and

$$\max\left\{ -\frac{\partial\phi}{\partial t}(\bar{z}) + \mathcal{F}(\bar{z}, D\phi(\bar{z}), D^2\phi(\bar{z})), -\frac{\partial\phi}{\partial k}(\bar{z}) + \mathcal{H}(\bar{z}, D\phi(\bar{z}))\right\} > 0.$$

Here, either

$$-\frac{\partial\phi}{\partial t}(\bar{z}) + \mathcal{F}(\bar{z}, D\phi(\bar{z}), D^2\phi(\bar{z})) > 0 \tag{42}$$

or

$$-\frac{\partial \phi}{\partial k}(\bar{z}) + \mathcal{H}(\bar{z}, D\phi(\bar{z})) > 0 \tag{43}$$

is verified. Since $\bar{t} < T$ and $\bar{k} < K$, in the definition of Θ_{r_1}, reducing r_1 if necessary, we can always assume that $\bar{t} + r_1 < T$ and $\bar{k} + r_1 < K$. If (42) is true, from the definition of $\mathcal{F}$, from the regularity hypotheses in **(A0)**, **(A1)** and from the fact that $\phi \in C^2$, it then follows that there exists some positive constant $r_2 \leq r_1$ such that

$$-\frac{\partial \phi}{\partial t}(t,k,x) - \langle A(t,x), D\phi(t,k,x)\rangle - l_0(t,x) - \frac{1}{2}\mathrm{Tr}\{\tilde{D}(t,x)D^2\phi(t,k,x)\} > 0$$

for all $z = (t,k,x) \in \overline{\Theta}_{r_2}$, and $\bar{t} + r_2 < T$, $\bar{k} + r_2 < K$. Take a control rule $P \in \mathcal{R}^a(\bar{t},\bar{k},\bar{x})$ such that

$$P\left(\mu_s = \delta_{\{0\}} \quad 0 \leq s \leq \theta, \quad \theta = T - \bar{t}\right) = 1.$$

It is easy to see that such a control rule exists; that, setting

$$\rho \doteq \inf\{s \in]0, T+K] : \ (t_s, k_s, \xi_s) \notin \Theta_{r_2}\},$$

by the continuity of the state process (t_s, k_s, ξ_s) one gets

$$P(T - \bar{t} > \rho) = 1, \qquad P(\rho > 0) = 1;$$

and that, for $0 \leq s < \rho$,

$$-\frac{\partial \phi}{\partial t}(t_s,k_s,\xi_s) - \langle A(t_s,\xi_s), D\phi(t_s,k_s,\xi_s)\rangle - \frac{1}{2}\mathrm{Tr}\{\tilde{D}(t_s,\xi_s)D^2\phi(t_s,k_s,\xi_s)\} - l_0(t_s,\xi_s) > 0.$$

Since $\mu_s = \delta_{\{0\}}$, this yields that

$$E_P\left[\int_0^\rho \left(-\mathcal{L}\phi(t_s,k_s,\xi_s,\mu_s) - l_0(t_s,\xi_s)(1-|\mu_s|) - \langle l_1(t_s,\xi_s),\mu_s\rangle\right) ds\right] > 0. \tag{44}$$

By the definition of control rule one has

$$\phi(t_\rho,k_\rho,\xi_\rho) = \phi(\bar{t},\bar{k},\bar{x}) + \int_0^\rho \mathcal{L}\phi(t_s,k_s,\xi_s,\mu_s)\, ds + \mathcal{M}_\rho\phi,$$

where $\mathcal{M}_\rho\phi$ is a continuous square-integrable martingale with respect to P. Hence,

$$E_P\left[\phi(t_\rho,k_\rho,\xi_\rho) - \phi(\bar{t},\bar{k},\bar{x})\right] = E_P\left[\int_0^\rho \mathcal{L}\phi(t_s,k_s,\xi_s,\mu_s)\, ds\right].$$

Since $(t_\rho,k_\rho,\xi_\rho) \in \overline{\Theta}_{r_2}$, one has

$$\begin{aligned} &E_P\left[V(t_\rho,k_\rho,\xi_\rho) - V(\bar{t},\bar{k},\bar{x})\right] \leq E_P\left[\phi(t_\rho,k_\rho,\xi_\rho) - \phi(\bar{t},\bar{k},\bar{x})\right] \\ &= E_P\left[\int_0^\rho \mathcal{L}\phi(t_s,k_s,\xi_s,\mu_s)\, ds\right] < -E_P\left[\int_0^\rho \left(l_0(t_s,\xi_s)(1-|\mu_s|) + \langle l_1(t_s,\xi_s),\mu_s\rangle\right) ds\right], \end{aligned}$$

where the last inequality follows from (44). This can be rewritten as

$$V(\bar{t},\bar{k},\bar{x}) > E_P\left[\int_0^\rho \left(l_0(t_s,\xi_s)(1-|\mu_s|) + \langle l_1(t_s,\xi_s),\mu_s\rangle\right) + V(t_\rho,k_\rho,\xi_\rho)\right],$$

in contradiction with the dynamic programming principle (28).

If (43) is true, reasoning as before one can deduce that there exist some positive constant $r_2 \le r_1$ and a vector $\bar{w} \in \mathcal{K}$ with $|\bar{w}| = 1$ such that

$$-\frac{\partial \phi}{\partial k}(t,k,x) - \langle B(t,x)\bar{w}, p\rangle - \langle l_1(t,x), \bar{w}\rangle > 0 \quad \forall z = (t,k,x) \in \overline{\Theta}_{r_2}.$$

Then, let us introduce a control rule $P \in \mathcal{R}^a(\bar{t},\bar{k},\bar{x})$ such that

$$P\left(\mu_s = \delta_{\{\bar{w}\}} \quad 0 \le s \le \theta, \quad \theta = K - \bar{k}\right) = 1.$$

From now on, the proof proceeds, with obvious changes, as in the previous case. The proof that $\mathcal{V}$ is a viscosity subsolution of (5)–(7) is therefore concluded.

Step 2. Let us assume by contradiction that V fails to be a viscosity supersolution of (5)–(7). Thus there is a point $\bar{z} = (\bar{t},\bar{k},\bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n$, a test function $\phi \in C_b^2([0,T] \times [0,K] \times \mathbb{R}^n)$, and a constant $r_1 > 0$ such that $\bar{z}$ is a local minimum point for $V - \phi$, that is,

$$V(z) - \phi(z) \ge V(\bar{z}) - \phi(\bar{z}) \quad \forall z = (t,k,x) \in \overline{\Theta}_{r_1},$$

and either of the following cases hold.

Case 1. $(\bar{t},\bar{k},\bar{x})$ is such that $\bar{t} < T$, $\bar{k} \le K$, and

$$-\frac{\partial \phi}{\partial t}(\bar{z}) + \mathcal{F}(\bar{z}, D\phi(\bar{z}), D^2\phi(\bar{z})) < 0 \ \text{ and } \ -\frac{\partial \phi}{\partial k}(\bar{z}) + \mathcal{H}(\bar{z}, D\phi(\bar{z})) < 0.$$

Case 2. $(\bar{t},\bar{k},\bar{x})$ is such that $\bar{t} = T$, $\bar{k} < K$, and

$$\max\left\{-\frac{\partial \phi}{\partial t}(\bar{z}) + \mathcal{F}(\bar{z}, D\phi(\bar{z}), D^2\phi(\bar{z})), -\frac{\partial \phi}{\partial k}(\bar{z}) + \mathcal{H}(\bar{z}, D\phi(\bar{z}))\right\} < 0 \ \text{ and } \ V(\bar{z}) < g(\bar{x}).$$

Let us first consider Case 1. Since $\bar{t} < T$, in the definition of Θ_{r_1}, reducing r_1 if necessary, we can always assume that $\bar{t} + r_1 < T$. From the regularity hypotheses in **(A0)**, **(A1)** and from the fact that $\phi \in C^2$, it follows that there exist some positive constants $r_2 \le r_1$ and $\varepsilon > 0$ such that

$$(45) \quad -\frac{\partial \phi}{\partial t}(z) + \mathcal{F}(z, D\phi(z), D^2\phi(z)) < -\varepsilon \ \text{ and } \ -\frac{\partial \phi}{\partial k}(z) + \mathcal{H}(z, D\phi(z)) < -\varepsilon$$

for all $z = (t,k,x) \in \overline{\Theta}_{r_2}$, and $\bar{t} + r_2 \le \bar{t} + r_1 < T$. Take an optimal control rule $P \in \mathcal{R}^a(\bar{t},\bar{k},\bar{x})$. Such a control rule exists in view of Proposition 3.5. Let us notice that at any point $(\bar{t},\bar{k},\bar{x})$ with $\bar{t} < T$, every control rule $Q \in \mathcal{R}^a(\bar{t},\bar{k},\bar{x})$ verifies

$$(46) \qquad Q\left(\theta \ge T - \bar{t} > 0\right) = 1.$$

Moreover (see Remark 2.2), in both cases, $\bar{k} < K$ and $\bar{k} = K$, one has

$$(47) \qquad Q\left(k_s \le K \quad 0 \le s \le \theta\right) = 1.$$

Equations (46) and (47) hold, in particular, for $Q = P$. Let us define the exit time

$$\rho \doteq \inf\{s \in]0, T+K] : \\ (t_s, k_s, \xi_s) \notin [0 \vee (\bar{t} - r_2), \bar{t} + r_2[\times [0 \vee (\bar{k} - r_2), \bar{k} + r_2[\times B_n(r_2)\}.$$

Observe that, if $\bar{k} < K$, then it is not restrictive to assume that $\bar{k} + r_2 < K$, so that ρ coincides with the first exit time from the set Θ_{r_2}. In case $\bar{k} = K$, instead, $\bar{k} + r_2 > K$ and ρ may be greater than the first exit time from Θ_{r_2}. However, taking into account (47), it is not difficult to see that in both cases $\bar{k} < K$ and $\bar{k} = K$, one has

$$(t_s, k_s, \xi_s) \in \overline{\Theta}_{r_2} \qquad 0 \le s \le \rho \wedge \theta.$$

Now, from the continuity of the process (t_s, k_s, ξ_s) it follows that

$$P(\rho > 0) = 1,$$

which together with (46) yields that the stopping time $\rho' \doteq \rho \wedge \theta$ verifies

$$P(\rho' > 0) = 1.$$

Therefore by (45) it follows that

$$E_P\left[\int_0^{\rho'} (\mathcal{L}\phi(t_s, k_s, \xi_s, \mu_s) + l_0(t_s, \xi_s)(1 - |\mu_s|) + \langle l_1(t_s, \xi_s), \mu_s\rangle)\, ds\right] \ge \varepsilon E_P[\rho'].$$

Applying Ito's formula, we have

$$E_P[\phi(t_{\rho'}, k_{\rho'}, \xi_{\rho'})] = \phi(\bar{t}, \bar{k}, \bar{x}) + E_P\left[\int_0^{\rho'} \mathcal{L}\phi(t_s, k_s, \xi_s, \mu_s)\, ds\right]$$

which yields

$$\begin{aligned} &E_P\left[\phi(t_{\rho'}, k_{\rho'}, \xi_{\rho'}) - \phi(\bar{t}, \bar{k}, \bar{x})\right] = E_P\left[\textstyle\int_0^{\rho'} \mathcal{L}\phi(t_s, k_s, \xi_s, \mu_s)\, ds\right] \\ &\ge E_P\left[\textstyle\int_0^{\rho'} (-l_0(t_s, \xi_s)(1 - |\mu_s|) - \langle l_1(t_s, \xi_s), \mu_s\rangle)\, ds\right] + \varepsilon E_P[\rho']. \end{aligned}$$

Since $(t_{\rho'}, k_{\rho'}, \xi_{\rho'}) \in \overline{\Theta}_{r_2}$ and $E_P[\rho'] > 0$, setting $\varepsilon' \doteq \varepsilon E_P[\rho']$, $\varepsilon' > 0$ one has

$$\begin{aligned} &E_P\left[V(t_{\rho'}, k_{\rho'}, \xi_{\rho'}) - V(\bar{t}, \bar{k}, \bar{x})\right] \ge E_P\left[\phi(t_{\rho'}, k_{\rho'}, \xi_{\rho'}) - \phi(\bar{t}, \bar{k}, \bar{x})\right] \\ &\ge E_P\left[\textstyle\int_0^{\rho'} (-l_0(t_s, \xi_s)(1 - |\mu_s|) - \langle l_1(t_s, \xi_s), \mu_s\rangle)\, ds\right] + \varepsilon', \end{aligned}$$

which, rewritten as

$$V(\bar{t}, \bar{k}, \bar{x}) \le E_P\left[\int_0^{\rho'} (l_0(t_s, \xi_s)(1 - |\mu_s|) + \langle l_1(t_s, \xi_s), \mu_s\rangle)\, ds + V(t_{\rho'}, k_{\rho'}, \xi_{\rho'})\right] - \varepsilon',$$

contradicts the dynamic programming principle (28).

Let $(T, \bar{k}, \bar{x})$ be some point satisfying the assumptions of Case 2. Since by definition $V(T, K, \bar{x}) = g(\bar{x}) \ \ \forall \bar{x} \in \mathbb{R}^n$, it must be that $\bar{k} < K$. Hence in the definition of Θ_{r_1}, reducing r_1 if necessary, we can always assume that $\bar{k} + r_1 < K$, $r_1 < T$ (while $(T + r_1) \wedge T = T$), and from the regularity hypotheses in **(A0)**, **(A1)** and from $\phi \in C^2$, it follows that there exist some positive constants $r_2 \le r_1$ and $\varepsilon > 0$ such that

$$-\frac{\partial \phi}{\partial t}(z) + \mathcal{F}(z, D\phi(z), D^2\phi(z)) < -\varepsilon, \quad -\frac{\partial \phi}{\partial k}(z) + \mathcal{H}(z, D\phi(z)) < -\varepsilon, \quad V(z) < g(x) - \varepsilon$$

for all $z = (t, k, x) \in \overline{\Theta}_{r_2}$, where now $\Theta_{r_2} = [T - r_2, T[\times[0 \vee (\bar{k} - r_2), \bar{k} + r_2[\times B_n(r_2)$. Let $P \in \mathcal{R}^a(T, \bar{k}, \bar{x})$ be an optimal control rule, which exists in view of Proposition 3.5. It is not difficult to see that every control rule $Q \in \mathcal{R}^a(T, \bar{k}, \bar{x})$ is such that

$$Q\,(t_s = T, \quad 0 \le s \le \theta) = 1, \tag{48}$$

and, if in addition $J(T, \bar{k}, \bar{x}, Q) < g(\bar{x})$, then ($\bar{k} < K$ and)

$$Q\,(\theta > 0) = 1. \tag{49}$$

Since $V(T, \bar{k}, \bar{x}) < g(\bar{x})$ by hypothesis, (48) and (49) hold, in particular, for $Q = P$. Let us define

$$\rho \doteq \inf\left\{s \in]0, T + K] : \ (t_s, k_s, \xi_s) \notin [T - r_2, T + r_2[\times[0 \vee (\bar{k} - r_2), \bar{k} + r_2[\times B_n(r_2)\right\}.$$

The exit time ρ may be greater than the first exit time from Θ_{r_2}, but from (48) it follows that

$$(t_s, k_s, \xi_s) \in \overline{\Theta}_{r_2}, \qquad 0 \le s \le \rho \wedge \theta.$$

Now the continuity of the process (t_s, k_s, ξ_s) implies that

$$P\,(\rho > 0) = 1.$$

Owing to (49), this yields that the stopping time $\rho' \doteq \rho \wedge \theta$ verifies

$$P\,(\rho' > 0) = 1.$$

From now on, the proof is analogous to the proof of Case 1, so we omit it. This proves that $\mathcal{V}$ is a viscosity supersolution of (5)–(7).

The proof that $\mathcal{V}$ is a viscosity solution of (5)–(7) is therefore concluded. □

6. Existence of solutions for generalized Cauchy problems with discontinuous Hamiltonians. The results of the previous sections allow us to prove the existence of a viscosity solution to a boundary value problem involving a second order semilinear Hamilton–Jacobi–Bellman equation such as

$$-\frac{\partial v}{\partial t} - \frac{1}{2}\mathrm{Tr}\{\tilde{D}(t,x)D^2 v\} + H\left(t, x, \frac{\partial v}{\partial k}, Dv\right) = 0 \qquad \text{on} \quad]0, T[\times]0, K[\times\mathbb{R}^n, \tag{50}$$

and mixed boundary conditions such as

$$-\frac{\partial v}{\partial t} - \frac{1}{2}\mathrm{Tr}\{\tilde{D}(t,x)D^2 v\} + H\left(t, x, \frac{\partial v}{\partial k}, Dv\right) \ge 0 \quad \text{on }]0, T[\times\{K\}\times\mathbb{R}^n, \tag{51}$$

$$v \le g \ \text{ and } \ -\tfrac{\partial v}{\partial t} - \tfrac{1}{2}\mathrm{Tr}\{\tilde{D}(t,x)D^2 v\} + H(t, x, \tfrac{\partial v}{\partial k}, Dv) \ge 0 \text{ if } v < g \text{ on } \{T\}\times]0, K]\times\mathbb{R}^n, \tag{52}$$

for a (possibly discontinuous) Hamiltonian of the form

$$H(t, x, p_k, p) \doteq \sup_{w\in\mathcal{K}}\{-\langle A(t,x) + B(t,x)w, p\rangle - l_0(t,x) - \langle l_1(t,x), w\rangle - p_k|w|\} \tag{53}$$

$\forall (t, x, p_k, p) \in [0,T] \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$. We will refer to such a problem as a *generalized Cauchy problem with discontinuous Hamiltonian.*

We point out that the Hamiltonian H in (53), defined via a maximization over the unbounded control set $\mathcal{K}$, is the natural Hamiltonian related to the original minimization problem (4). In other words, at least formally, one expects that the value function $\mathcal{V}$ is a viscosity solution to the generalized Cauchy problem (50)–(52) rather than to (5)–(7). Such an observation motivates the study of such a boundary value problem (also in more general form, as in [MS2]), mainly dealing with existence and uniqueness of solutions.

Since H in (53) is in general discontinuous and equal to $+\infty$ in many points, we interpret solutions to (50)–(52) in the sense of the definition, due to Ishii [I], of discontinuous viscosity solutions for discontinuous Hamiltonians which we recall in the definition below.

DEFINITION 6.1. *A locally bounded function v defined on $[0,T[\times[0,K] \times \mathbb{R}^n$ is a* viscosity subsolution *of* (50)–(52) *if for every point $\bar{z} = (\bar{t}, \bar{k}, \bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n$ and for every map $\phi \in C_b^2([0,T] \times [0,K] \times \mathbb{R}^n)$ such that $v^* - \phi$ has a local maximum at $\bar{z}$ one has*

$$-\frac{\partial \phi}{\partial t}(\bar{z}) - \mathrm{Tr}\{\tilde{D}(\bar{z}) D^2 \phi(\bar{z})\} + H_* \left(\bar{z}, \frac{\partial \phi}{\partial k}(\bar{z}), D\phi(\bar{z}) \right) \leq 0$$

if $\bar{z} \in]0,T[\times]0,K[\times\mathbb{R}^n$, and

$$v^*(\bar{z}) \leq g(\bar{x})$$

if $\bar{z} \in \{T\}\times]0,K] \times \mathbb{R}^n$.

A locally bounded function v defined on $[0,T[\times[0,K] \times \mathbb{R}^n$ is a viscosity supersolution *of* (50)–(52) *if for every point $\bar{z} = (\bar{t}, \bar{k}, \bar{x}) \in [0,T] \times [0,K] \times \mathbb{R}^n$ and for every map $\phi \in C_b^2([0,T] \times [0,K] \times \mathbb{R}^n)$ such that $v_* - \phi$ has a local minimum at $\bar{z}$ one has*

$$-\frac{\partial \phi}{\partial t}(\bar{z}) - \mathrm{Tr}\{\tilde{D}(\bar{z}) D^2 \phi(\bar{z})\} + H^* \left(\bar{z}, \frac{\partial \phi}{\partial k}(\bar{z}), D\phi(\bar{z}) \right) \geq 0$$

if $\bar{z} \in]0,T[\times]0,K] \times \mathbb{R}^n$, and

$$-\frac{\partial \phi}{\partial t}(\bar{z}) - \mathrm{Tr}\{\tilde{D}(\bar{z}) D^2 \phi(\bar{z})\} + H^* \left(\bar{z}, \frac{\partial \phi}{\partial k}(\bar{z}), D\phi(\bar{z}) \right) \geq 0 \quad \text{or} \quad v_*(\bar{z}) \geq g(\bar{x})$$

if $\bar{z} \in \{T\}\times]0,K] \times \mathbb{R}^n$.

A locally bounded function v defined on $[0,T[\times[0,K] \times \mathbb{R}^n$ is a called a viscosity solution *of* (50)–(52) *if it is both a viscosity sub- and supersolution of* (50)–(52).

We can show that $\mathcal{V}$ is a viscosity solution to the generalized Cauchy problem (50)–(52) since there exists a one-to-one correspondence among solutions to (50)–(52) and solutions to (5)–(7), as specified by the following theorem.

THEOREM 6.2 (see [MS2, Theorem 3.4]). *Assume* **(A0)**, **(A1)**. *Let* v *(resp., v)*: $[0,T] \times [0,K] \times \mathbb{R}^n \to \mathbb{R}$ *be a upper (resp., lower) semicontinuous locally bounded function. Then*

(a) v *is a viscosity subsolution to* (50)–(52) *if and only if it is a viscosity subsolution to* (5)–(7),

(b) *v is a viscosity supersolution to* (50)–(52) *if and only if it is a viscosity supersolution to* (5)–(7).

Therefore we can state the following existence (and uniqueness) theorem whose proof is a consequence of Theorems 6.2 and 5.3.

THEOREM 6.3. *Assume* **(A0)**, **(A1)**. *Then the value function* $\mathcal{V}$ *solves the boundary value problem* (50)–(52) *in the viscosity sense. Moreover, its continuous extension to* $[0,T]\times[0,K]\times\mathbb{R}^n$ *is the unique viscosity solution of* (50)–(52) *among the bounded functions defined on* $[0,T]\times[0,K]\times\mathbb{R}^n$ *which are continuous on* $\partial(]0,T[\times]0,K[\times\mathbb{R}^n)$.

7. Appendix. Before proving Lemma 7.3 on the claim stated in Theorem 3.3, we need to introduce some definitions and prove some technical results in Lemmas 7.1 and 7.2.

Let us consider the noncanonical control $\alpha = (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{(w_{0_s}, w_s)\}, \{(t_s, k_s, \xi_s)\}, \theta)$ given by (24), whose existence is proved in Theorem 3.3. Let us understand that whenever the operators J and $\mathcal{L}$ have to be evaluated on such α, the constraint $w_{0_s} = 1 - |w_s|$ in their definition must be dropped. Let us notice that since we are only interested with the (random) time interval $0 \le s \le \theta$, with a small abuse of notation, we will denote still by α the control in which (w_{0_s}, w_s) is replaced by $(w_{0_s}, w_s)\chi_{\{s\le\theta\}} + (1,0)\chi_{\{s>\theta\}}$.

Remark 7.1. Given the control $\alpha = (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{(w_{0_s}, w_s)\}, \{(t_s, k_s, \xi_s)\}, \theta)$, by (24), in Case 1 has

$$\int_0^\theta (w_{0_s} + |w_s|)\, ds = 0$$

(such an eventuality can happen only if $\bar{t} = T$), it is easy to check that any control $\check{\alpha}$, $\check{\alpha} = (\Omega, \{\mathcal{F}\}, P, \{\mathcal{F}_s\}, \check{w}_s, (\check{t}_s, \check{k}_s, \check{\xi}_s), \check{\theta}) \in \Gamma^a(\bar{t}, \bar{k}, \bar{x})$, such that

$$P\left(\check{\theta} = 0\right) = 1,$$

verifies the claim in Theorem 3.3 and $J(\bar{t}, \bar{k}, \bar{x}, \check{\alpha}) = J(\bar{t}, \bar{k}, \bar{x}, \alpha)$.

LEMMA 7.1. *Assume* **(A0)**, **(A1)**. *Let us consider the noncanonical control* $\alpha = (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{(w_{0_s}, w_s)\}, \{(t_s, k_s, \xi_s)\}, \theta)$, *given by* (24). *Assume that* $\int_0^\theta (w_{0_s} + |w_s|)\, ds > 0$. *Let us define*

$$\Phi_s \doteq \int_0^s (w_{0_r} + |w_r|)\, dr,$$

for $0 \le s \le T + K$. *Let us denote by* $\{\Psi_\sigma\}$ *the right inverse of* Φ:

$$\Psi_\sigma \doteq \inf\{s \ge 0 :\ \Phi_s > \sigma\},$$

for $0 \le \sigma \le \Phi_{T+K}$. *Then* $\Phi_{T+K} > 0$ *and* $\{\Psi_\sigma\}$ *is a right continuous time change satisfying the following properties:*

(i) $\Psi_{\Phi_s} \ge s\ \forall s \ge 0$, $\Phi_{\Psi_\sigma} = \sigma\ \forall \sigma \ge 0$;

(ii) *let*

$$\check{\mathcal{F}}_\sigma \doteq \mathcal{F}_{\Psi_\sigma} \forall \sigma > 0; \tag{54}$$

then $\check{\mathcal{F}}_\sigma$ *is a filtration on the probability space* $(\Omega, \mathcal{F}, P)$;

(iii) Φ_θ *is a* $\check{\mathcal{F}}_\sigma$*-stopping time such that* $\Phi_\theta \le T + K$.

Proof. The proof follows from the definition of time change and right inverse and from Proposition 1.1, Chapter V, in [RY]. $\Phi_\theta \le T + K$ since $w_{0_s} + |w_s| \le 1$ for $s \ge 0$ by definition. □

LEMMA 7.2. *Assume* **(A0)**, **(A1)**. *Let us consider the noncanonical control* $\alpha = (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{(w_{0_s}, w_s)\}, \{(t_s, k_s, \xi_s)\}, \theta)$, *given by* (24). *Assume that* $\int_0^\theta (w_{0_s} + |w_s|)\, ds > 0$. *On the probability space* $(\Omega, \mathcal{F}, P)$ *let us consider the filtration* $\check{\mathcal{F}}_\sigma$ *given by* (54). *Let us define the processes*

$$\check{t}_\sigma \doteq t_{\Psi_\sigma}, \quad \check{u}_\sigma \doteq \int_0^{\Psi_\sigma} w_r\, dr, \quad \check{k}_\sigma \doteq k_{\Psi_\sigma},$$

for $0 \le \sigma \le \Phi_{T+K}$. *Then there exists a* $\check{\mathcal{F}}_\sigma$*-progressively measurable process* $\{\check{w}_\sigma\}$, $\overline{B}_m(1) \cap \mathcal{K}$ *valued, such that, for* $0 \le \sigma \le \Phi_{T+K}$,

$$\begin{aligned} \check{t}_\sigma &= \bar{t} + \textstyle\int_0^\sigma (1 - |\check{w}_r|)\, dr \left(= \bar{t} + \int_0^{\Psi_\sigma} w_{0_r}\, dr \right), \\ \check{u}_\sigma &= \textstyle\int_0^\sigma \check{w}_r\, dr \left(= \int_0^{\Psi_\sigma} w_r\, dr \right), \\ \check{k}_\sigma &= \bar{k} + \textstyle\int_0^\sigma |\check{w}_r|\, dr \left(= \bar{k} + \int_0^{\Psi_\sigma} |w_r|\, dr \right). \end{aligned} \tag{55}$$

Proof. Let $\{u_s\}$ denote the (strong) solution to

$$u_s = \int_0^s w_r\, dr. \tag{56}$$

Since $\sigma = \Phi_s$ is the arclength parameter of both the processes (t_s, u_s) and (t_s, k_s), we know that $(\check{t}_\sigma, \check{u}_\sigma, \check{k}_\sigma)$ is absolutely continuous for $0 \le \sigma \le \Phi_{T+K}$. Consequently, from Proposition 3.13, Chapter I, in [JS], it follows that there exists a $\check{\mathcal{F}}_\sigma$-progressively measurable process $(\tilde{w}_{0_\sigma}, \tilde{w}_\sigma, z_\sigma)$, $\mathbb{R}_+ \times \mathcal{K} \times \mathbb{R}_+$-valued, such that, for $0 \le \sigma \le \Phi_{T+K}$,

$$\check{t}_\sigma = \bar{t} + \int_0^\sigma \tilde{w}_{0_r}\, dr, \quad \check{u}_\sigma = \int_0^\sigma \tilde{w}_r\, dr, \quad \check{k}_\sigma = \bar{k} + \int_0^\sigma z_r\, dr.$$

Moreover, by the properties of the arclength parameter for almost all $\omega \in \Omega$ there exists a set of measure zero $\mathcal{N}_\omega$ such that $\check{t}'_\sigma(\omega) + |\check{u}'_\sigma(\omega)| = 1$ and $\check{t}'_\sigma(\omega) + \check{k}'_\sigma(\omega) = 1$ for every $\sigma \notin \mathcal{N}_\omega$. This implies that $\tilde{w}_{0_\sigma}(\omega) + |\tilde{w}_\sigma(\omega)| = 1$ and $z_\sigma(\omega) = |\tilde{w}_\sigma(\omega)|$ for $\sigma \notin \mathcal{N}_\omega$. Let us define for every $\sigma \ge 0$ and for $i = 1, \dots, m$ the process

$$\check{w}^i_\sigma \doteq (-1 \vee \tilde{w}^i_\sigma) \wedge 1.$$

$\check{w}$ is $\check{\mathcal{F}}_\sigma$-progressively measurable, $\overline{B}_m(1) \cap \mathcal{K}$-valued. Moreover, $(\check{t}_\sigma, \check{u}_\sigma, \check{k}_\sigma)$ is indistinguishable from $(\bar{t} + \int_0^\sigma (1 - |\check{w}_r|)\, dr, \int_0^\sigma \check{w}_r\, dr, \bar{k} + \int_0^\sigma |\check{w}_r|\, dr)$. Indeed for almost all $\omega \in \Omega$ we have, for $0 \le \sigma \le \Phi_{T+K}$,

$$\check{t}_\sigma(\omega) = \bar{t} + \textstyle\int_{[0,\sigma]} \tilde{w}_{0_r}(\omega)\, dr = \bar{t} + \int_{[0,\sigma]\setminus\mathcal{N}_\omega} (1 - |\tilde{w}_r(\omega)|)\, dr = \bar{t} + \int_0^\sigma (1 - |\check{w}_r(\omega)|)\, dr,$$

$$\check{u}_\sigma(\omega) = \textstyle\int_{[0,\sigma]} \tilde{w}_r(\omega)\, dr = \int_{[0,\sigma]\setminus\mathcal{N}_\omega} \tilde{w}_r(\omega)\, dr = \int_0^\sigma \check{w}_r(\omega)\, dr,$$

$$\check{k}_\sigma(\omega) = \bar{k} + \textstyle\int_{[0,\sigma]} z_r(\omega)\, dr = \bar{k} + \int_{[0,\sigma]\setminus\mathcal{N}_\omega} z_r(\omega)\, dr = \bar{k} + \int_0^\sigma |\check{w}_r(\omega)|\, dr. \qquad \square$$

LEMMA 7.3. *Assume* **(A0)**, **(A1)**. *Let us consider the noncanonical control* $\alpha = (\Omega, \mathcal{F}, P, \{\mathcal{F}_s\}, \{(w_{0_s}, w_s)\}, \{(t_s, k_s, \xi_s)\}, \theta)$, *given by* (24). *Assume that* $\int_0^\theta (w_{0_s} + |w_s|)\, ds > 0$ *and* $J(\bar{t}, \bar{k}, \bar{x}, \alpha) < +\infty$. *Then the control*

$$\check{\alpha} \doteq (\Omega, \mathcal{F}, P, \check{\mathcal{F}}_\sigma, \{\check{w}_\sigma\}, \{(\check{t}_\sigma, \check{k}_\sigma, \check{\xi}_\sigma)\}, \check{\theta}),$$

where $\check\theta \doteq \Phi_\theta$, $\{\check w_\sigma\}$ *is the process whose existence is proved in* Lemma 7.2, $\{\check{\mathcal{F}}_\sigma\}$ *is given by* (54), $\{(\check t_\sigma, \check k_\sigma)\}$ *are given by* (55), $\check\xi_\sigma \doteq \xi_{\Psi_\sigma}$, *for* $0 \le \sigma \le \Phi_{T+K}$, *verifies* **(B3)**, **(B4)**, *and is such that*

$$J(\bar t, \bar k, \bar x, \check\alpha) = J(\bar t, \bar k, \bar x, \alpha).$$

Proof. We already proved that $\check\alpha$ verifies **(B3)** in Lemmas 7.1 and 7.2; therefore, in order to prove that condition **(B4)** holds, one has to show that for every $\varphi \in \mathcal{C}_b^2(\mathbb{R}^{2+n})$ $\check{\mathcal{M}}_\sigma(\varphi, \check\alpha)$ is a $(P, \{\check{\mathcal{F}}_\sigma\})$ square integrable martingale for $\sigma \in [0, \Phi_{T+K}]$, where

$$\check{\mathcal{M}}_\sigma(\varphi, \check\alpha) \doteq \varphi(\check t_\sigma, \check k_\sigma, \check\xi_\sigma) - \int_0^\sigma \mathcal{L}\varphi(\check t_r, \check k_r, \check\xi_r, \check w_r)\, dr.$$

To this end let us consider the square integrable martingale $\mathcal{M}_s(\varphi, \alpha)$ associated to the control α and let us notice that it is Ψ-continuous (that is, it is constant on each stochastic interval where $(w_{0_s}, w_s) = (0,0)$; see Definition 1.3 Chapter V, [RY]) and that by (56) and (55) one has

$$\begin{aligned} &dt_s = w_{0_s} d_s, \quad dt_{\Psi_\sigma} = d\check t_\sigma = (1 - |\check w_\sigma|) d\sigma, \\ &du_s = w_s ds, \quad du_{\psi_\sigma} = \check w_\sigma d\sigma, \\ &dk_s = |w_s| ds, \quad dk_{\Psi_\sigma} = d\check k_\sigma = |\check w_\sigma| d\sigma. \end{aligned}$$

By Proposition 1.4, Chapter V of [RY], for every process H which is $\mathcal{F}_s$-progressively measurable, if we denote by $\check H_\sigma \doteq H_{\psi_\sigma}$, since the process (t_s, u_s, k_s) is Ψ-continuous, one has that

$$\begin{aligned} &\textstyle\int_0^{\Psi_\sigma} H_s w_{0_s} ds = \int_0^{\Psi_\sigma} H_s\, dt_s = \int_0^\sigma \check H_\sigma d\check t_\sigma = \int_0^\sigma \check H_\sigma (1 - |\check w_\sigma|) d\sigma, \\ &\textstyle\int_0^{\Psi_\sigma} H_s w_s ds = \int_0^{\Psi_\sigma} H_s\, du_s = \int_0^\sigma \check H_\sigma d\check u_\sigma = \int_0^\sigma \check H_\sigma \check w_\sigma\, d\sigma, \\ &\textstyle\int_0^{\Psi_\sigma} H_s |w_s| ds = \int_0^{\Psi_\sigma} H_s\, dk_s = \int_0^\sigma \check H_\sigma d\check k_\sigma = \int_0^\sigma \check H_\sigma |\check w_\sigma| d\sigma. \end{aligned} \tag{57}$$

Therefore, for any $\varphi \in \mathcal{C}_b^2(\mathbb{R}^{2+n})$, by applying the first equation of (57) with $H_s = \frac{1}{2}\sum_{ij} \tilde D_{ij}(t_s, \xi_s)\, \frac{\partial^2 \varphi}{\partial x_i \partial x_j}(t_s, k_s, \xi_s) + \sum_i A_i(t_s, \xi_s) \frac{\partial \varphi}{\partial \xi_i}(t_s, k_s, \xi_s) + \frac{\partial \varphi}{\partial t}(t_s, k_s, \xi_s)$, the second one with $H_s = (B_1(t_s, \xi_s) \frac{\partial \varphi}{\partial x_1}(t_s, k_s, \xi_s), \dots, B_n(t_s, \xi_s) \frac{\partial \varphi}{\partial x_n}(t_s, k_s, \xi_s))^T$, the third one with $H_s = \frac{\partial \varphi}{\partial k}(t_s, k_s, \xi_s)$, by Proposition 1.5, Chapter V of [RY] one concludes that $\check{\mathcal{M}}_\sigma(\varphi, \check\alpha)$ is a martingale since it coincides with $\mathcal{M}_{\psi_\sigma}(\varphi, \alpha)$. Finally, it is also easy to see that $\check\alpha$ is admissible. Indeed since $J(\bar t, \bar k, \bar x, \alpha) < +\infty$, then $t_\theta = T$ and $k_\theta \le K$, which implies by Proposition 1.4, Chapter V of [RY] that

$$\check t_{\Phi_\theta} = \bar t + \int_0^{\Phi_\theta} (1 - |\check w_r|)\, dr = t_\theta = T \quad \text{and} \quad \check k_{\Phi_\theta} = \bar k + \int_0^{\Phi_\theta} |\check w_r|\, dr = k_\theta \le K$$

and

$$\begin{aligned} J(\bar t, \bar k, \bar k, \check\alpha) = E_P\Big[&\int_0^{\Phi_\theta} (l_0(\check t_\sigma, \check\xi_\sigma)(1 - |\check w_\sigma|) + \langle l_1(\check t_\sigma, \check\xi_\sigma), \check w_\sigma\rangle)\, d\sigma \\ &+ g(\check\xi_{\Phi_\theta}) + G(\check t_{\Phi_\theta}, \check k_{\Phi_\theta})\Big]. \end{aligned} \tag{58}$$

By the first and second equations in (57) with $H_s = l_0(t_s, \xi_s)$ and $H_s = l_1(t_s, \xi_s)$, respectively, and by the fact that, by Proposition 3.1, $\check\xi_{\Phi_\theta} = \xi_\theta$, one has $J(\bar t, \bar k, \bar k, \check\alpha) = J(\bar t, \bar k, \bar x, \alpha)$. ∎

Acknowledgment. The authors are grateful to Professor Paolo Dai Prà for some very helpful discussions and comments on the topics of this paper.

REFERENCES

[A1] L. Alvarez, *Singular stochastic control, linear diffusions, and optimal stopping: A class of solvable problems*, SIAM J. Control Optim., 39 (2001), pp. 1697–1710.

[A2] L. Alvarez, *Singular stochastic control in the presence of a state-dependent yield structure*, Stochastic Process. Appl., 86 (2000), pp. 323–343.

[BCD] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Solutions of Hamilton-Jacobi-Bellman Equations*, in Systems & Control: Foundations & Applications, Birkhäuser, Boston, 1997.

[B] G. Barles, *An approach of deterministic control problems with unbounded data*, Ann. Inst. H. Anal. Non Linéaire Poincaré 7 (1990), pp. 235–258.

[BP] G. E. Barles and B. Perthame, *Discontinuous solutions of deterministic optimal stopping time problems*, RAIRO Modél. Math. Anal. Numer., 21 (1987), pp. 557–579.

[BJM] E. N. Barron, R. Jensen, and J.-L. Menaldi, *Optimal control and differential games with measures*, Nonlinear Anal., 21 (1993), pp. 241–268.

[BC] J. A. Bather and H. Chernoff, *Sequential decisions in the control of a spaceship*, Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California, Press Berkeley, CA, 3 (1967), pp. 181–207.

[Be] M. S. Berger, *Nonlinearity and functional analysis*, in Lectures on nonlinear problems in mathematical Analysis, Pure and Applied Mathematics, Academic Press, New York-London, 1977

[BR] A. Bressan and F. Rampazzo, *On differential systems with vector-valued impulsive controls*, Boll. Un. Mat. Ital. B (7), 2 (1988), pp. 641–656.

[CF] F. Camilli and M. Falcone, *Approximation of control problems involving ordinary and impulsive controls*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 159–176.

[CMR] P. L. Chow, J.-L. Menaldi, and M. Robin, *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23 (1985), pp. 858–899.

[CIL] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[DFS] J. R. Dorroh, G. Ferreyra, and P. Sundar, *A technique for stochastic control problems with unbounded control set*, J. Theoret. Probab., 12 (1999), pp. 255–270.

[DM] F. Dufour and B. M. Miller, *Generalized solutions in nonlinear stochastic control problems*, SIAM J. Control Optim., 40 (2002), pp. 1724–1745.

[EKNP] N. El Karoui, D. H. Nguyen, and M. Jeanblanc-Picqué, *Compactification methods in the control of degenerate diffusions: Existence of an optimal control*, Stochastics, 20 (1987), pp. 169–219.

[FS] W. Fleming and H. M. Soner, *Controlled Markov processes and viscosity solutions*, Appl. Math. 25, Springer-Verlag, New York, 1993.

[F] A. Friedman, *Stochastic Differential Equations and Application*, Probab. Math. Statist., 28, Academic Press, New York, 1975.

[HL] U. G. Haussmann and J. P. Lepeltier, *On the existence of optimal controls*, SIAM J. Control Optim., 28 (1990), pp. 851–902.

[HS1] U. G. Haussmann and W. Suo, *Singular optimal stochastic controls.* I. *Existence*, SIAM J. Control Optim., 33 (1995), pp. 916–936.

[HS2] U. G. Haussmann and W. Suo, *Singular optimal stochastic controls.* II. *Dynamic programming*, SIAM J. Control Optim., 33 (1995), pp. 937–959.

[IW] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, in North-Holland Mathematical Library 24, North-Holland, Amsterdam, 1981.

[I] H. Ishii, *A boundary value problem of the Dirichlet type for Hamilton-Jacobi equations*, Ann. Sc. Norm. Sup. Pisa IV, 16 (1989), pp. 105–135.

[JS] J. Jacod and A. N. Shiryaev, *Limit Theorems for Stochastic Processes*, 2nd edition. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 288, Springer-Verlag, Berlin, 2003.

[MiRu] B. M. Miller and W. J. Runggaldier, *Optimization of observations: A stochastic control approach*, SIAM J. Control Optim., 35 (1997), pp. 1030–1052.

[MoRa] M. Motta and F. Rampazzo, *Dynamic programming for nonlinear systems driven by ordinary and impulsive controls*, SIAM J. Control Optim., 34 (1996), pp. 199–225.

[MS1] M. Motta and C. Sartori, *Semicontinuous viscosity solutions to mixed boundary value*

problems with degenerate convex Hamiltonians, Nonlinear Anal. 49, 2002, pp. 905–927.

[MS2] M. Motta and C. Sartori, *Second Order Bellman-Isaacs Equations with Mixed Boundary Conditions. Part* II. *Finite Fuel and Other Singular and Unbounded Stochastic Control Problems*, submitted.

[RS] F. Rampazzo and C. Sartori, *Hamilton-Jacobi-Bellman equations with fast gradient-dependence*, Indiana Univ. Math. J., 49 (2000), pp. 1043–1077.

[RY] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, 2nd edition. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 293, Springer-Verlag, Berlin, 1994.

[Se] S. P. Sethi, *Dynamic optimal control models in advertising: A survey*, SIAM Rev., 19 (1977), pp. 685–725.

[So] H. M. Soner, *Optimal control with state-space constraint.* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.

[SS] H. M. Soner and S. E. Shreve, *Regularity of the value function for a two-dimensional singular stochastic control problem*, SIAM J. Control Optim., 27 (1989), pp. 876–907.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1211–1238

ON THE UNIFORM CONTROLLABILITY OF THE BURGERS EQUATION*

O. GLASS† AND S. GUERRERO†

Abstract. In this paper, we deal with the viscous Burgers equation with a small dissipation coefficient ν. We prove the (global) exact controllability property to nonzero constant states, that is to say, the possibility of finding boundary values such that the solution of the associated Burgers equation is driven to a constant state. The main objective of this paper is to do so with control functions whose norms in an appropriate space are bounded independently of ν, which belongs to a suitably small interval. This result is obtained for a sufficiently large time.

Key words. controllability, Burgers equation, vanishing viscosity limit

AMS subject classifications. 93B05, 35Q20, 35B25

DOI. 10.1137/060664677

1. Introduction.

1.1. Statement of the result and background. We are interested in the controllability of the Burgers equation in a bounded interval:

$$u_t + uu_x - \nu u_{xx} = 0 \text{ in } (0,T) \times (0,1), \tag{1}$$

where T is a positive real number. We complete this equation with the following: We give an initial condition,

$$u_{|t=0} = u_0 \text{ in } (0,1), \tag{2}$$

and controlled boundary values,

$$u_{|x=0} = v_1(t) \text{ in } (0,T), \quad u_{|x=1} = v_2(t) \text{ in } (0,T). \tag{3}$$

Here, v_1 and v_2 stand for control functions, which translates into the possibility of acting over the system through both endpoints of the boundary $x = 0$ and $x = 1$. Let $\bar{u}$ be a solution of (1) with Dirichlet boundary conditions (also called a trajectory). The *exact controllability to trajectories* holds if we can find controls v_1 and v_2 such that the associated solution coincides with $\bar{u}$ at time $t = T$. In this paper, we are interested in proving an exact controllability result uniformly with respect to the viscosity coefficient ν in a sufficiently small range.

Let us be more specific on the problem under review. We consider the system constituted by (1), (2), and (3). Let us fix the initial condition u_0 in some Banach space X and a constant $M \neq 0$; then, our goal is to find two controls v_1 and v_2 such that the associated solution u satisfies

$$u|_{t=T} = M \text{ in } (0,1). \tag{4}$$

Moreover, as we said above we are interested in finding controls whose norms in some Banach space Y are uniformly bounded with respect to ν, whenever ν is sufficiently

*Received by the editors July 10, 2006; accepted for publication (in revised form) February 8, 2007; published electronically September 12, 2007.
http://www.siam.org/journals/sicon/46-4/66467.html

†Université Pierre et Marie Curie-Paris6, UMR 7598 Laboratoire Jacques-Louis Lions, Paris, F-75005, France (glass@ann.jussieu.fr, guerrero@ann.jussieu.fr).

small and T is sufficiently large. The main result of this paper is given in the following theorem, where we prove the previous results for $X = L^\infty(0,1)$ and $Y = L^\infty(0,T)$.

THEOREM 1. *There is a constant $\alpha_0 \geq 1$ such that for any $M \in \mathbb{R} \setminus \{0\}$ there exists $\nu_0 > 0$ such that for any $u_0 \in L^\infty([0,1])$, any time $T > \alpha_0/|M|$, and any $\nu \in (0,\nu_0)$ there exist controls v_1^ν and v_2^ν satisfying the following properties:*

- *$\|v_1^\nu\|_\infty$ and $\|v_2^\nu\|_\infty$ are uniformly bounded for $\nu \in (0,\nu_0)$; that is to say, there exists a constant $C(\alpha_0) > 0$ such that*

$$\|v_1^\nu\|_\infty + \|v_2^\nu\|_\infty \leq C(\|u_0\|_\infty + |M|).$$

- *The solution u of (1), (2), (3) associated with $v_1 = v_1^\nu$ and $v_2 = v_2^\nu$ satisfies (4).*

The controllability of the Burgers equation for fixed ν has been studied by several authors. In particular, two kinds of controllability properties have been studied:

- On the one hand, the *local exact controllability to trajectories*, which stands for the concept of exact controllability with the additional assumption that the initial state u_0 is close to the initial state of the targeted trajectory $\bar{u}_{|t=0}$; this has been established for the Burgers equation in [8]. It is also proved in [8] that the exact controllability does not hold when the control acts in a subinterval (a,b) of $(0,1)$, which is equivalent to control at one endpoint. In the more recent work [9], the authors prove that the global exact controllability for (1) does not hold even if the control is acting on both sides of the domain.
- On the other hand, in [4] the author establishes a global result between 0 and constant states; more precisely, the author proves that for $u_0 = 0$ and for any $T > 0$ one can drive the solution of (1) to any constant M satisfying that $|M|$ is sufficiently large with respect to T.

Here, as we are interested in the properties of uniform controllability as $\nu \to 0^+$, it seems natural to regard the inviscid framework ($\nu = 0$). In this case, and in the context of entropy solutions, the controllability of the equation

$$u_t + (u^2/2)_x = 0 \tag{5}$$

was studied in [12], where some conditions are given on the final state in order to ensure this property. More general convex scalar conservation laws

$$u_t + (f(u))_x = 0 \tag{6}$$

were considered in [1], for which the controllability problem is posed in the half line with a null initial condition. The set of attainable states is completely described.

We recall that for conservation laws such as (5), solutions generally develop singularities in finite time, regardless of the regularity of the initial condition. This leads to considering distributional solutions, but in this setting, uniqueness is lost. From both physical and mathematical standpoints, it is then natural to consider solutions that fulfill entropy conditions in order to extract the physically relevant solution. These conditions are the following: For any regular couple (η, q) defined on $\mathbb{R}$ and such that $\eta' f' = q'$ and η is convex, the following stands in the sense of measures:

$$\eta(u)_t + q(u)_x \leq 0.$$

We emphasize that entropy solutions are the ones which can be obtained by vanishing viscosity. One can summarize the situation by saying that the viscosity has

disappeared from the equation and is effective only for the selection of admissible discontinuities. Concerning the Cauchy problem, (1) was first approached by Hopf in [11], where an explicit formula is given and the limit as $\nu \to 0^+$ is considered. The convergence of vanishing viscosity approximations to the entropy solutions of a general scalar conservation law was studied in the celebrated work of Kružkov [13]. For a general reference to conservation laws, we refer to [6].

It is therefore very natural, when considering control problems for conservation laws, to consider the cost of the viscosity, that is, to determine if known controllability properties for the hyperbolic equation are still valid for the model with small viscosity, and how the size of the control evolves as the viscosity approaches 0. Note that some problems, such as the global approximate controllability of the Navier–Stokes equation with Navier slip boundary conditions [3], are obtained through controllability results for the inviscid equation (in this case, the Euler equation).

1.2. Some remarks. Let us make some remarks on the above theorem and state a corollary which concerns the system with $\nu = 1$.

Remark 1. In general, entropy solutions of (5) cannot reach a state M (starting, for instance, from $u_0 = 0$) in a time less than $1/|M|$. In particular, the state 0 cannot be reached unless one has $u_0 = 0$. This is easily seen by considering generalized backward characteristics (see [1]). Hence the time of control $O(1/M)$ is not surprising. Note that even in the case of a linear transport equation, the uniform controllability results [5, 10] consider a time of control of the form $C/|M|$, $C > 1$.

Remark 2. Following the proof of Theorem 1, one can check that Theorem 1 holds for the choices $\alpha_0 = 9$ (or in fact, as obtained numerically, approximately 6.3) and

$$\nu_0 = \nu_1 \min\{1, |M|/|\log(|M|)|\}, \tag{7}$$

where ν_1 is a small enough constant independent of M but depending on α_0.

Now we can present the following result as a consequence of Theorem 1.

COROLLARY 1. *Consider for $w^0 \in L^\infty(0,1)$ and $\tilde{T} > 0$ the following control problem:*

$$\begin{cases} w_t + ww_x - w_{xx} = 0 & \text{in } (0,\tilde{T}) \times (0,1), \\ w_{|x=0} = \tilde{v}_1(t), \quad w_{|x=1} = \tilde{v}_2(t) & \text{in } (0,\tilde{T}), \\ w_{|t=0} = w_0 & \text{in } (0,1). \end{cases} \tag{8}$$

Assume that $|M_0|$ is large enough in order that

$$|M_0|\nu_1 \geq |\log(|M_0|\nu_1)|,$$

where ν_1 is defined as in (7)*; then, for every $\tilde{T} > \alpha_0/|M_0|$ (where α_0 can be chosen as* 9*), there exist controls $\tilde{v}_1(t)$ and $\tilde{v}_2(t)$ in $L^\infty((0,T),\mathbb{R})$ such that the solution of* (8) *satisfies*

$$w_{|t=\tilde{T}} = M_0. \tag{9}$$

Note in particular that Corollary 1 implies that the result of [4] is valid for any $u_0 \in L^\infty(0,1)$. Let us also emphasize that the time we use to control the system depends only on the final state and is independent from the initial state.

Proof. This is a simple scaling argument. Indeed, let us set

$$\begin{cases} u(t,x) = \nu_1 w(\nu_1 t, x),\ v_1(t) = \nu_1 \tilde{v}_1(\nu_1 t),\ v_2(t) = \nu_1 \tilde{v}_2(\nu_1 t), \text{ and } u_0(x) = \nu_1 w_0(x), \\ t \in (0,T),\ x \in (0,1), \end{cases}$$

where we have denoted $T = \tilde{T}/\nu_1$. Then we have

$$\begin{cases} u_t + uu_x - \nu_1 u_{xx} = 0 & \text{in } (0,T)\times(0,1), \\ u_{|x=0} = v_1(t),\ u_{|x=1} = v_2(t) & \text{in } (0,T), \\ u_{|t=0} = w_0 & \text{in } (0,1). \end{cases} \tag{10}$$

Let us set $M := \nu_1 M_0$. From Theorem 1 and since $T > \alpha_0/|M|$ (thanks to the choice of $\tilde{T}$) and $\nu_1 \le \nu_1 \min\{1, |M|/|\log|M||\}$ (thanks to the choice of M_0), we know of the existence of v_1 and v_2 such that the solution of (10) satisfies $u|_{t=T} = M$ in $(0,1)$. Going back to w, this shows the existence of two controls $\tilde{v}_1$ and v_2 such that the associated solution of (8) satisfies $w|_{t=\tilde{T}} = M_0$ in $(0,1)$, as we wanted to prove. ∎

1.3. Structure of the paper. One of the main ingredients of the proof is the use of the *return method* by J. M. Coron, which consists of finding a particular trajectory of the system which moves far away from the initial state to get back to the final state afterward. In the present situation we steer the system toward a large constant state N, and then we get back to the constant state M.

Consequently, the proof of Theorem 1 is divided into two parts, which we summarize in the following propositions.

PROPOSITION 1. *There are some constants $\alpha_1 \ge 1$ and $\nu_1 > 0$ such that for any $u_0 \in L^\infty([0,1])$, for any $N \in \mathbb{R}$ with $|N|$ large enough (depending on $\|u_0\|_\infty$), and for any $\nu \in (0,\nu_1)$ there are controls w_1^ν and w_2^ν in $L^\infty(0,T_1)$, where $T_1 = \alpha_1/|N|$, satisfying the following properties:*

- *$\|w_1^\nu\|_\infty$ and $\|w_2^\nu\|_\infty$ are uniformly bounded for $\nu \in [0,\nu_1]$.*
- *The associated solution u satisfies $u_{|t=T_1} = N$ in $(0,1)$.*

PROPOSITION 2. *The conclusion of Theorem 1 is true when $M > 0$ and u_0 is a positive constant large enough with respect to M.*

The plan of the paper is the following. Proposition 1 is established in section 2. Section 3 is devoted to proving Proposition 2 and, finally, in section 4 we prove some technical results we need for the previous propositions.

2. Proof of Proposition 1. Due to the invariance of the solutions of (1) by the transformation $u(t,x) \leftrightarrow -u(t,1-x)$, we can assume that $N > 0$. Now the proof of Proposition 1 is divided into two parts. First, we prove that we can reach a state close to N in a time $O(1/N)$ (which is a kind of global approximate controllability, but where the target is a constant that depends on the initial state) and then we prove that we can steer the latter state exactly toward N in a time $O(1/N)$ (local exact controllability).

2.1. Reaching N approximately. In the following proposition we prove that, starting from an L^∞ initial condition, we can construct a solution of (1)–(2) which is close in the sense of the $W^{1,\infty}$ norm to some large constant.

PROPOSITION 3. *Given $u_0 \in L^\infty([0,1])$, one can find $N > 0$ large enough such that for any $\nu > 0$, one can find controls v_1 and v_2 such that the solution of* (1), (2), (3) *satisfies*

$$\|u(t,\cdot) - N\|_{L^\infty([0,1])} \le \left(\|u_0\|_\infty + \frac{N}{2}\right) \exp\left\{-\frac{3N^2}{16\nu}\left(t - \frac{8}{N}\right)\right\} \tag{11}$$

for any $t > 0$ *and*

$$\|u_x(t,\cdot)\|_{L^\infty([0,1])} \le C\frac{N^2}{\nu}\exp\left\{-\frac{3N^2}{16\nu}\left(t-\frac{8}{N}\right)\right\} \tag{12}$$

for any $t > (8/N)$ *and some* $C > 0$. *Moreover, the controls satisfy, independently from* ν,

$$\max(\|v_1\|_{L^\infty(0,T)}, \|v_2\|_{L^\infty(0,T)}) \le N. \tag{13}$$

Remark 3. All the above constants (such as 8 or 16) are not optimal (see the proof below) but are sufficient for our purpose (because N is arbitrarily large).

Proof of Proposition 3. The proof of this proposition relies on the comparison principle and on traveling waves for (1). Let us state precisely the comparison principle for the reader's convenience.

LEMMA 1 (comparison principle). *Consider* $\overline{u}_1$ *and* $\overline{u}_2$ *in* $L^\infty(\mathbb{R})$, *and the corresponding solutions* u_1 *and* u_2 *of the Burgers equation on the whole real line with initial conditions* $\overline{u}_1$ *and* $\overline{u}_2$, *respectively. Then, if*

$$\overline{u}_1 \le \overline{u}_2 \text{ in } \mathbb{R}, \tag{14}$$

we have

$$u_1(t,x) \le u_2(t,x) \text{ in } \mathbb{R}^+ \times \mathbb{R}. \tag{15}$$

Proof. A simple way to prove Lemma 1 (although it could be proven in a far more general setting) is to use Hopf's formula for solutions of the viscous Burgers equation [11]:

$$u_i(t,x) = \frac{\int_{-\infty}^{\infty} \frac{x-y}{t}\exp\left\{-\frac{1}{2\nu}\left(\frac{(x-y)^2}{2t} + \int_0^y \overline{u}_i(\eta)d\eta\right)\right\}dy}{\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\nu}\left(\frac{(x-y)^2}{2t} + \int_0^y \overline{u}_i(\eta)d\eta\right)\right\}dy}, \quad i = 1,2. \tag{16}$$

We consider the function

$$\rho_i(x,y) := \exp\left\{-\frac{1}{2\nu}\left(\frac{(x-y)^2}{2t} + \int_0^y \overline{u}_i(\eta)d\eta\right)\right\},$$

and $d\mu_i$ the probability measure (depending on x) given by

$$d\mu_i := \frac{\rho_i(x,y)}{\int_{-\infty}^{\infty}\rho_i(x,\cdot)}dy.$$

Now we have

$$
\begin{aligned}
u_2(t,x) &= \int_{-\infty}^{+\infty} \frac{x-y}{t} d\mu_2 \\
&= \int_{-\infty}^{+\infty} \frac{x-y}{t} \exp\Big(-\frac{1}{2\nu} \int_0^y (\overline{u}_2(\eta) - \overline{u}_1(\eta)) d\eta \Big) d\mu_1 \cdot \frac{\int_{-\infty}^{+\infty} \rho_1(x,y)dy}{\int_{-\infty}^{+\infty} \rho_2(x,y)dy} \\
&\geq \int_{-\infty}^{+\infty} \frac{x-y}{t} d\mu_1 . \int_{-\infty}^{+\infty} \exp\Big(-\frac{1}{2\nu} \int_0^y (\overline{u}_2(\eta) - \overline{u}_1(\eta)) d\eta \Big) d\mu_1 \\
&\qquad \cdot \frac{\int_{-\infty}^{+\infty} \rho_1(x,y)dy}{\int_{-\infty}^{+\infty} \rho_2(x,y)dy} \\
&= \int_{-\infty}^{+\infty} \frac{x-y}{t} d\mu_1 = u_1(t,x).
\end{aligned}
$$

The inequality used above is Chebyshev's inequality: for μ a probability measure on $\mathbb{R}$ and two nondecreasing functions f and g, one has

$$\int_{\mathbb{R}} f g d\mu \geq \int_{\mathbb{R}} f d\mu \times \int_{\mathbb{R}} g d\mu.$$

(This follows easily by considering $\int\int (f(x)-f(y))(g(x)-g(y)) d\mu(x) d\mu(y)$.) □

Back to the proof of Proposition 3. We introduce traveling wave profiles for (1). These are solutions of the viscous Burgers equations on the whole real line, of the form

$$u(t,x) = U(x-ct) \quad t \in \mathbb{R}^+,\ x \in \mathbb{R},$$

with c a fixed real number, which will be chosen later on. Furthermore, u satisfies the following asymptotic properties:

$$u(t,x) \to U^- \text{ as } x \to -\infty, \quad u(t,x) \to U^+ \text{ as } x \to +\infty, \quad u_x(t,x) \to 0 \text{ as } x \to \pm\infty,$$

where U^- and U^+ are constant states. Straightforward computations show that these traveling waves are given by the following:

$$U^- \geq U^+, \tag{17}$$

$$c = \frac{U^- + U^+}{2}, \tag{18}$$

$$U(y) = \frac{U^- + U^+}{2} - \frac{U^- - U^+}{2} \tanh\Big(\frac{U^- - U^+}{2\nu} (y - y_0) \Big), \tag{19}$$

where of course y_0 is arbitrary (we will refer to y_0 as the center of the wave).

Now let us go back to the proof of Proposition 3. Given u_0, we choose $N > 2\|u_0\|_\infty$; later we will also take $N > L\|u_0\|_\infty$ for some $L \geq 2$, and in section 3 we will additionally require $N > |M|$.

We introduce the following:

- u as the solution of the Burgers equation on $\mathbb{R}$ with initial value

$$u(0,x) = \begin{cases} N \text{ for } x < 0, \\ u_0(x) \text{ for } 0 \le x \le 1, \\ 0 \text{ for } x > 1. \end{cases} \tag{20}$$

- $\check{u}$ as the traveling wave solution of the Burgers equation with $U^- = N$ and $U^+ = -2\|u_0\|_\infty$, initially centered at y_0, if $u_0 \not\equiv 0$. If $u_0 \equiv 0$, then we take, for instance, $U_+ := -N/4$.

The goal is to prove that the restriction of u to $[0,1]$ is a suitable solution for Proposition 3 (of course, v_1 and v_2 are defined as the traces of u at $x = 0$ and $x = 1$, respectively). Now if y_0 is such that

$$\check{u}(0,\cdot) \le u_0(\cdot) \text{ in } (0,1), \tag{21}$$

it follows from the comparison principle that

$$\check{u}(t,x) \le u(t,x) \le N \text{ in } \mathbb{R}^+ \times \mathbb{R}. \tag{22}$$

Let us make a choice of y_0 so that (21) is satisfied. Indeed, if we take, in the case $u_0 \not\equiv 0$,

$$y_0 = -\frac{2\nu}{N + 2\|u_0\|_\infty} \operatorname{arctanh}\left(\frac{N}{N + 2\|u_0\|_\infty}\right),$$

one can easily check that

$$\check{u}(0,x) \le -\|u_0\|_\infty \quad \forall x \in [0,1],$$

just taking into account that the maximum value of the function

$$x \in [0,1] \longmapsto -\tanh\left(\frac{U^- - U^+}{2\nu}(x - y_0)\right)$$

is reached at $x = 0$. The case $u_0 \equiv 0$ is similar.

Clearly, for N large enough, one has

$$y_0 \ge -1 \tag{23}$$

and

$$N - 2\|u_0\|_\infty \ge \frac{N}{2}. \tag{24}$$

Let us now prove that estimate (11) holds. We first recall the expression of $\check{u}$: for $t \in \mathbb{R}^+$ and $x \in \mathbb{R}$,

$$\check{u}(t,x) = \frac{N}{2} - \|u_0\|_\infty - \frac{N + 2\|u_0\|_\infty}{2} \tanh\left(\frac{N + 2\|u_0\|_\infty}{2\nu}\left(x - \frac{N - 2\|u_0\|_\infty}{2}t - y_0\right)\right). \tag{25}$$

From (23) and (24), we get for $t \in \mathbb{R}^+$ and $x \in [0,1]$,

$$\begin{aligned}
&\tanh\left(\frac{N+2\|u_0\|_\infty}{2\nu}\left(x-\frac{N-2\|u_0\|_\infty}{2}t-y_0\right)\right)\\
&\quad\le \tanh\left(\frac{N+2\|u_0\|_\infty}{2\nu}\left(-\frac{N}{4}t+2\right)\right)\\
&\quad\le \tanh\left(\frac{-3N^2}{16\nu}\left(t-\frac{8}{N}\right)\right).
\end{aligned}$$

On the other hand, from the definition of the function tanh, we readily deduce that

$$\tanh\left(\frac{-3N^2}{16\nu}\left(t-\frac{8}{N}\right)\right) \le -1+\exp\left\{\frac{-3N^2}{16\nu}\left(t-\frac{8}{N}\right)\right\}.$$

Going back to (25), we obtain

$$\check{u}(t,x)-N \ge -\left(\frac{N}{2}+\|u_0\|_\infty\right)\exp\left\{\frac{-3N^2}{16\nu}\left(t-\frac{8}{N}\right)\right\}.$$

Since $\check{u}(t,x)-N\le 0$ already holds (see (22)), we deduce (11).

In order to prove (12) we begin by giving an explicit representation of the spatial derivative of u in the following lemma.

LEMMA 2. *Let u be a solution of*

$$\begin{cases} u_t-\nu u_{xx}+uu_x=0, & (t,x)\in\mathbb{R}_+\times\mathbb{R},\\ u_{|t=0}=u_0, & x\in\mathbb{R}. \end{cases}$$

Then one has

$$\partial_x u(t,x)=\frac{\int_{-\infty}^{+\infty}\frac{y-x}{t}(u_0(y)-u(t,x))\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+\int_0^y u_0(\eta)d\eta\right]\right\}dy}{2\nu\int_{-\infty}^{+\infty}\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+\int_0^y u_0(\eta)d\eta\right]\right\}dy}. \tag{26}$$

Proof. Let us define

$$f(t,x):=\int_{-\infty}^{+\infty}\frac{x-y}{t}\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+\int_0^y u_0(\eta)d\eta\right]\right\}dy$$

and

$$g(t,x):=\int_{-\infty}^{+\infty}\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+\int_0^y u_0(\eta)d\eta\right]\right\}dy.$$

From Hopf's formula (see (16)), we find

$$u(t,x)=\frac{f(t,x)}{g(t,x)},\quad (t,x)\in\mathbb{R}_+\times\mathbb{R}.$$

We notice that $\partial_x g=-\frac{1}{2\nu}f$ and

$$\partial_x f(t,x)=\frac{g}{t}-\frac{1}{2\nu t}\int_{-\infty}^{+\infty}\frac{(y-x)^2}{t}\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+\int_0^y u_0(\eta)d\eta\right]\right\}dy,$$

so that

$$\partial_x u(t,x) = \frac{1}{t} - \frac{1}{2\nu}\left[\frac{\int_{-\infty}^{+\infty} \frac{(y-x)^2}{t^2} \exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} dy}{g} - \frac{f^2}{g^2}\right]. \tag{27}$$

Now let us consider the second term in the above right-hand side:

$$\begin{aligned} I &:= \int_{-\infty}^{+\infty} \frac{(y-x)^2}{t^2} \exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} dy \\ &= \int_{-\infty}^{+\infty} \frac{y-x}{t}\left(\frac{y-x}{t} + u_0(y)\right) \exp\left\{-\frac{1}{2\nu}\left[\frac{(y-x)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} dy \\ &\quad - \int_{-\infty}^{+\infty} \frac{y-x}{t} u_0(y) \exp\left\{-\frac{1}{2\nu}\left[\frac{(y-x)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} dy \\ &= \frac{2\nu g}{t} - \int_{-\infty}^{+\infty} \frac{y-x}{t} u_0(y) \exp\left\{-\frac{1}{2\nu}\left[\frac{(y-x)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} dy. \end{aligned}$$

In the last identity we have used the fact that

$$\begin{aligned} &-2\nu \frac{d}{dy} \exp\left\{-\frac{1}{2\nu}\left[\frac{(y-x)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} \\ &\qquad = \left(\frac{y-x}{t} + u_0(y)\right) \exp\left\{-\frac{1}{2\nu}\left[\frac{(y-x)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} \end{aligned} \tag{28}$$

and we have integrated by parts. Injecting I into (27) yields

$$\partial_x u(t,x) = -\frac{1}{2\nu g}\Bigg[-\int_{-\infty}^{+\infty} \frac{y-x}{t} u_0(y) \exp\left\{-\frac{1}{2\nu}\left[\frac{(y-x)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} dy - u(t,x)f(t,x)\Bigg],$$

which yields (26). □

Back to the proof of (12). We consider $x \in [0,1]$ and use (26) to estimate $\partial_x u(t,x)$:

$$\begin{aligned} \partial_x u(t,x) &= \frac{1}{2\nu} \frac{\int_{-\infty}^{+\infty} \frac{y-x}{t}(u_0(y) - N) \exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} dy}{\int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} dy} \\ &\quad + \frac{1}{2\nu} \frac{\int_{-\infty}^{+\infty} \frac{y-x}{t}(N - u(t,x)) \exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} dy}{\int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t} + \int_0^y u_0(\eta)d\eta\right]\right\} dy}. \end{aligned}$$

Let us denote by A and B, respectively, the first and second term of the above right-hand side. Clearly,

$$B = -\frac{1}{2\nu}(N - u(t,x))u(t,x),$$

and this term is easily estimated using the L^∞ estimate on $N-u$. Concerning A, we first notice that, due to the initial condition of u (see (20)), one has

$$A = \frac{1}{2\nu}\frac{\int_0^{+\infty} \frac{y-x}{t}(u_0(y)-N)\exp\Big\{-\frac{1}{2\nu}\big[\frac{(x-y)^2}{2t}+\int_0^y u_0(\eta)d\eta\big]\Big\}dy}{\int_{-\infty}^{+\infty}\exp\Big\{-\frac{1}{2\nu}\big[\frac{(x-y)^2}{2t}+\int_0^y u_0(\eta)d\eta\big]\Big\}dy}.$$

Note that simple computations yield

$$\int_a^b \exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+\alpha\, y\right]\right\}dy = 2\sqrt{\nu t}\exp\left\{\frac{\alpha(-2x+\alpha t)}{4\nu}\right\}\int_{\xi_-}^{\xi_+} e^{-t^2}dt \tag{29}$$

with

$$\xi_- := \frac{a-x+\alpha\, t}{2\sqrt{\nu t}} \quad \text{and} \quad \xi_+ := \frac{b-x+\alpha\, t}{2\sqrt{\nu t}},$$

and yield

$$\begin{aligned}
&\int_a^b y\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+\alpha\, y\right]\right\}dy \\
&= 2\sqrt{\nu t}(x-\alpha t)\exp\left\{\frac{\alpha(-2x+\alpha t)}{4\nu}\right\}\int_{\xi_-}^{\xi_+} e^{-t^2}dt \qquad (30)\\
&-2\nu t\left\{\exp\left[-\frac{b^2-2bx+2b\alpha t+x^2}{4\nu t}\right]-\exp\left[-\frac{a^2-2ax+2a\alpha t+x^2}{4\nu t}\right]\right\}.
\end{aligned}$$

Also, we note that for $y>0$, one has

$$\int_y^{+\infty} e^{-s^2}ds \le \frac{e^{-y^2}}{2y}. \tag{31}$$

We estimate from below the denominator of A in the following way:

$$\begin{aligned}
\int_{-\infty}^{+\infty}\exp\Big\{-\frac{1}{2\nu}\Big[\frac{(x-y)^2}{2t}+\int_0^y u_0(\eta)d\eta\Big]\Big\}dy &\\
&\ge \int_{-\infty}^{0}\exp\Big\{-\frac{1}{2\nu}\Big[\frac{(x-y)^2}{2t}+\int_0^y u_0(\eta)d\eta\Big]\Big\}dy\\
&= \int_{-\infty}^{0}\exp\Big\{-\frac{1}{2\nu}\Big[\frac{(x-y)^2}{2t}+Ny\Big]\Big\}dy\\
&\ge \sqrt{\pi\nu t}\exp\left\{\frac{N}{2\nu}\left(\frac{Nt}{2}-x\right)\right\}.
\end{aligned}$$

The numerator is bounded by

$$\begin{aligned}
&\left|\int_0^{+\infty}\frac{y-x}{t}(u_0(y)-N)\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+\int_0^y u_0(\eta)d\eta\right]\right\}dy\right|\\
&\le \int_0^{+\infty}\frac{y+1}{t}(\|u_0\|_\infty+N)\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}-\|u_0\|_\infty y\right]\right\}dy\\
&\le 12\nu N\exp\left\{\frac{-x^2}{4\nu t}\right\},
\end{aligned}$$

where we used $N > \|u_0\|_\infty$ and $Nt - 1 > 1$ for the times under review.

Then (12) follows by taking $N \gg \|u_0\|_\infty$. Note that the estimate on the term A is better than the estimate on the term B, which comes directly from the L^∞ estimate.

Finally, estimate (13) comes directly from the maximum principle. □

Remark 4. Taking N large enough (as we may), it is easy to see that we can replace the properties (23)–(24) with the following ones:

$$y_0 \geq -\delta,$$

$$N - 2\|u_0\|_\infty \geq (1-\delta)N,$$

for some arbitrarily small parameter $\delta > 0$. Then following the above computations we obtain estimates of the form

$$\|u(t,\cdot) - N\|_{W^{1,\infty}([0,1])} \leq \frac{CN^2}{\nu} \exp\left\{-\frac{K}{\nu}\left(t - \frac{2(1+\delta)}{N(1-\delta)}\right)\right\}.$$

Hence the time of approximate controllability can be made close to $2/N$ for N large enough (as $\nu \to 0^+$). This is not surprising, since for the hyperbolic equation, the solution of the Riemann problem (5) with initial condition

$$u|_{t=0} = \begin{cases} N \text{ on } \mathbb{R}^-, \\ M \text{ on } \mathbb{R}^+, \end{cases}$$

with $N, M \in \mathbb{R}$, $N > M$, is given by a simple shock with speed $(N+M)/2$.

2.2. Reaching N exactly. Observe that from subsection 2.1, and after a time $t = \alpha_1/N$ has gone by (with, for instance, $\alpha_1 = 8$ or even $\alpha_1 > 2$, taking Remark 4 into account), we can assume that our new initial condition (which we also denote u_0) satisfies

$$\|u_0 - N\|_{W^{1,\infty}(0,1)} \leq e^{-CN/\nu} \tag{32}$$

for some $C > 0$, provided that $\nu \in (0, \nu_0)$.

Remark 5. As stated above, condition (32) trivially follows from Proposition 3, since ν_0 is small enough and N is large enough. In what follows we will prove that condition (32) suffices to reach N exactly with uniform bounds on the controls, which concludes the proof of Proposition 1. □

In what concerns the proof of Proposition 2, at this stage of the analysis, we will also have some initial state u_0 satisfying

$$\|u_0 - M\|_{W^{1,\infty}(0,1)} \leq e^{-CM/\nu},$$

provided that $\nu < \nu_0$, where ν_0 is small enough (see Proposition 5 below). As a consequence, arguing as in this subsection, we will be able to drive the solution exactly to M as well (with uniform bounds on the controls).

In this subsection we prove that we have a local exact (uniform in ν) controllability result for a time $T = O(1/N)$. Precisely, we establish the next proposition.

PROPOSITION 4. *Assume that* $u_0 \in W^{1,\infty}(0,1)$ *and there exists* $K_0 > 0$ *such that*

$$\|u_0 - N\|_{W^{1,\infty}(0,1)} \leq e^{-K_0 N/\nu}. \tag{33}$$

Then, one can find controls v_1 and v_2 such that the solution of (1), (2), (3) *satisfies, for* $T = \frac{\alpha_0 - 2}{N}$,

$$u|_{t=T} = N \quad \text{in } (0,1). \tag{34}$$

Moreover, the controls satisfy the following estimate, which is independent of $\nu \in (0, \nu_0)$:

$$\max(\|v_1\|_{W^{1,\infty}(0,T)}, \|v_2\|_{W^{1,\infty}(0,T)}) \leq 2N. \tag{35}$$

Proof. First, we set $y(t,x) = u(t,x) - N$ so that y fulfills

$$\begin{cases} y_t + yy_x - \nu y_{xx} + Ny_x = 0 & \text{in } (0,T) \times (0,1), \\ y|_{t=0} = y^0 := u^0 - N & \text{in } (0,1). \end{cases} \tag{36}$$

Now, our objective is to find boundary controls $y|_{x=0}(t) = v_1(t) - N$ and $y|_{x=1}(t) = v_2(t) - N$ such that

$$y|_{t=T} = 0 \quad \text{in } (0,1) \tag{37}$$

and

$$\|v_1(t) - N\|_{W^{1,\infty}(0,T)} + \|v_2(t) - N\|_{W^{1,\infty}(0,T)} \leq N. \tag{38}$$

We will prove this by means of a fixed point argument posed in a suitable Hilbert space Z.

2.2.1. Uniform null controllability of the linearized Burgers equation. In this subsection, we consider the following linearized Burgers equation:

$$\begin{cases} y_t - \nu y_{xx} + ((N + z(t,x)/2)y)_x = 0 & \text{in } (0,T) \times (0,1), \\ y|_{x=0} = \widetilde{v}_1, \quad y|_{x=1} = \widetilde{v}_2 & \text{in } (0,T), \\ y|_{t=0} = y^0 & \text{in } (0,1). \end{cases} \tag{39}$$

For this system, we prove the null controllability with controls bounded independently of ν. Specifically, we have the next lemma.

LEMMA 3. *Let z be in $L^1(0,T; W^{1,\infty}(0,1)) \cap L^\infty((0,T) \times (0,1))$ with*

$$\|z_x\|_{L^1_t L^\infty_x} + \|z\|_{L^\infty_t L^\infty_x} \leq \exp\left(-\frac{K_0 N}{5\nu}\right), \tag{40}$$

where K_0 is as introduced in Proposition 4, *and let us introduce the quantity*

$$D(T,N,z) := \frac{3e^{-2\|z_x\|_{L^1_t(L^\infty_x)}}}{4}\left(\frac{2T(N - \|z\|_\infty/2)}{3} - 1\right)^2 - 6\chi, \tag{41}$$

where χ is some positive constant defined in (84). *Assume that the initial condition $y^0 = u^0 - N$ satisfies* (33) *and that the final time T satisfies*

$$(N - \|z\|_\infty/2)2T/3 > 1. \tag{42}$$

Then, for any $\nu \in (0, \nu_0)$, there exist two controls $\widetilde{v}_1, \widetilde{v}_2 \in W^{1,\infty}(0,T)$ such that the associated solution to (39) *satisfies*

$$y|_{t=T} = 0 \quad \text{in } (0,1)$$

and

$$\|\widetilde{v}_1\|_{W^{1,\infty}(0,T)} + \|\widetilde{v}_2\|_{W^{1,\infty}(0,T)} \le e^{-K_0 N/(3\nu)} \left(e^{-D(T^*,N,z)/(\nu T^*)} + 1 \right), \tag{43}$$

where $T^* := \min\{T, 3/(N - \|z\|_\infty/2)\}$.

Remark 6. It will follow from Lemma 4 that one can take $6\chi = 4$ in (41). When we follow Remark 7 below (as found numerically), we see that one can estimate 6χ by 2.61.

Proof. First, choosing the controls $\widetilde{v}_1$ and $\widetilde{v}_2$ to be zero close to $t = 0$, and on account of the regularizing effect of the heat equation, one can always suppose that our initial condition y^0 belongs to $W^{2,\infty}(0,1)$, and thanks to (33), we can also assume that

$$\|y^0\|_{W^{2,\infty}(0,1)} \le e^{-K_0 N/(2\nu)}. \tag{44}$$

Now, we introduce a function $\widetilde{y}^0 \in W^{2,\infty}(-1,2)$ such that $\widetilde{y}^0 = y^0$ in $(0,1)$ and

$$\|\widetilde{y}^0\|_{W^{2,\infty}(-1,2)} \le C\|y^0\|_{W^{2,\infty}(0,1)} \tag{45}$$

for some $C > 0$. Let us first suppose that we can find two controls $\widetilde{v}_3$, $\widetilde{v}_4 \in L^2(0,T)$ satisfying

$$\begin{aligned} \|\widetilde{v}_3\|_{L^2(0,T^*)} + \|\widetilde{v}_4\|_{L^2(0,T^*)} &\le Ce^{-D(T^*,N,z)/(\nu T^*)}\|\widetilde{y}^0\|_{L^2(-1,2)} \\ &\le Ce^{-D(T^*,N,z)/(\nu T^*)}e^{-K_0 N/(2\nu)} \end{aligned} \tag{46}$$

for some $C > 0$, such that the solution $\widetilde{y} \in L^2((0,T)\times(-1,2))$ of

$$\begin{cases} \widetilde{y}_t - \nu\widetilde{y}_{xx} + ((N + z(t,x)/2)\widetilde{y})_x = 0 & \text{in } (0,T^*)\times(-1,2), \\ \widetilde{y}|_{x=-1} = \widetilde{v}_3, \quad \widetilde{y}|_{x=2} = \widetilde{v}_4 & \text{in } (0,T^*), \\ \widetilde{y}|_{t=0} = \widetilde{y}^0 & \text{in } (-1,2) \end{cases} \tag{47}$$

satisfies

$$\widetilde{y}|_{t=T^*} = 0 \quad \text{in } (-1,2).$$

Then, the function $y := \widetilde{y}|_{[0,1]}1_{(0,T^*)}$ fulfills system (39) with

$$\widetilde{v}_1(t) = \widetilde{y}|_{x=0}(t)1_{(0,T^*)} \quad \text{and} \quad \widetilde{v}_2(t) = \widetilde{y}|_{x=1}(t)1_{(0,T^*)} \tag{48}$$

and satisfies

$$y|_{t=T} = 0 \quad \text{in } (0,1)$$

(indeed, $y \equiv 0$ in (T^*,T)).

In order to prove estimate (43), it suffices to use classical localization arguments together with regularity estimates for the solution of (47). Precisely, for any $\delta > 0$, we can prove that there exists a positive constant $C > 0$ such that

$$\|\widetilde{y}\|_{W^{1,\infty}(0,T;H^1(-1+\delta,2-\delta))} \le (C/\nu)(\|\widetilde{y}^0\|_{W^{2,\infty}(-1,2)} + \|\widetilde{v}_3\|_{L^2(0,T^*)} + \|\widetilde{v}_4\|_{L^2(0,T^*)}).$$

In particular, this implies that $\widetilde{y}|_{x=0}$ and $\widetilde{y}|_{x=1}$ belong to $W^{1,\infty}(0,T^*)$, and thanks to (46), (33), and $\nu \in (0,\nu_0)$, they satisfy estimate (43).

Consequently, our task now will be to find $\widetilde{v}_3$ and $\widetilde{v}_4$ satisfying the above properties. We use the classical approach, consisting of obtaining a suitable observability inequality for the adjoint system of (47). For simplicity, we will suppose that we are working in the space interval $(0,1)$ instead of $(-1,2)$ (so, in particular, we will refer to system (39) instead of (47)). Thus, let us introduce the adjoint problem associated with (39):

$$\begin{cases} -\varphi_t - \nu\varphi_{xx} - (N + z(t,x)/2)\varphi_x = 0 & \text{in } (0,T^*)\times(0,1), \\ \varphi|_{x=0} = 0, \quad \varphi|_{x=1} = 0 & \text{in } (0,T^*), \\ \varphi|_{t=T^*} = \varphi^0 & \text{in } (0,1), \end{cases} \tag{49}$$

where $\varphi^0 \in H_0^1(0,1)$ is the initial condition.

We prove the following observability inequality for the solutions of (49):

$$\|\varphi|_{t=0}\|^2_{L^2(0,1)} \leq K(T^*,\nu)\int_0^{T^*} (|\varphi_x|_{x=0}|^2 + |\varphi_x|_{x=1}|^2)\,dt \tag{50}$$

for some positive constant $K(T^*,\nu)$. Then, it is not difficult to prove that the null controllability of system (39) holds with controls $\widetilde{v}_1$ and $\widetilde{v}_2$, whose L^2 norms are bounded by $K(T^*,\nu)/\nu$. We omit the proof of this fact for the sake of simplicity.

In order to prove estimate (50), we will follow the same ideas as in [5]. That is to say, we will combine a suitable Carleman inequality with a dissipation result for system (49).

Dissipation result. Let $t_0 \in (0,T^*)$. Then, following the steps of the proof in [7], one can prove

$$\begin{cases} \|\varphi|_{t=t_0}\|^2_{L^2(0,1)} \leq \exp\left\{ \dfrac{\|z_x\|_{L^1_t(L^\infty_x)}}{8} - \dfrac{((N-\|z\|_\infty/2)t^*-1)^2}{2\nu t^*} e^{-2\|z_x\|_{L^1_t(L^\infty_x)}} \right\} \\ \qquad\qquad \times \|\varphi|_{t=t_0+t^*}\|^2_{L^2(0,1)} \end{cases} \tag{51}$$

for any $t^* \in (0, T^* - t_0)$ such that

$$(N - \|z\|_\infty/2)t^* > 1. \tag{52}$$

Carleman inequality. Let $0 < \gamma < 1/3$ (one can take, for example, $\gamma := 1/6$). We will prove that, if (40) and (42) hold, then we have the following inequality:

$$\begin{aligned} &\int_0^1 \int_{2T^*/3}^{(2+3\gamma)T^*/3} |\varphi|^2\,dt\,dx \\ &\leqslant Ce^{6\chi/(\nu T^*)} \left(\frac{\nu^2 T^*}{N}\int_0^{T^*} |\varphi_x(t,0)|^2\,dt + \frac{1}{N}\int_0^1 |\varphi(0,x)|^2\,dx \right). \end{aligned} \tag{53}$$

The proofs of these results are postponed to the last section of the paper.

Now we are in position to establish our central observability inequality (50). We apply the dissipativity result (51) to φ (which is a solution of (49)) for each $t^* \in (2T^*/3, (2+3\gamma)T^*/3)$. This is possible by (42). We obtain

$$\int_0^1 \int_{2T^*/3}^{(2+3\gamma)T^*/3} |\varphi|^2\,dt\,dx \geq C(\nu,N,T^*,z)\int_0^1 |\varphi(0,x)|^2\,dx,$$

with

$$C(\nu, N, T^*, z) = \exp\left\{ -\frac{\|z_x\|_{L^1_t(L^\infty_x)}}{8} + 3e^{-2\|z_x\|_{L^1_t(L^\infty_x)}} \frac{(\frac{2T^*}{3}(N - \frac{\|z\|_\infty}{2}) - 1)^2}{4\nu T^*} \right\}.$$

Here, we have used the fact that

$$t^* \longmapsto \frac{((N - \|z\|_\infty/2)t^* - 1)^2}{4\nu t^*}$$

is an increasing function as long as (52) is satisfied.

Finally, we obtain the observability inequality (50) with

$$K(T^*, \nu) = C \exp\left(-\frac{D(T^*, N, z)}{\nu T^*} \right),$$

for some $C = C(T^*, N)$; recall that $D(T^*, N, z)$ was introduced in (41). In particular, estimate (46) holds.

This concludes the proof of Lemma 3. □

2.2.2. Fixed point argument. In this subsection, we end the proof of the null controllability of system (36) by performing a fixed point argument applied to the following application: With each $z \in W^{1,\infty}((0,T) \times (0,1))$ such that (41) holds, we associate a y solution of (39) given by Lemma 3. More precisely, let us first define the set of controls:

$$A(z) = \{(\widetilde{v}_1, \widetilde{v}_2) \in W^{1,\infty}(0,T)^2 : \tag{54}$$
$$y \text{ solution of (39) satisfies } y|_{t=T} = 0 \text{ and } \widetilde{v}_1,\ \widetilde{v}_2 \text{ satisfy (43)}\}.$$

Then, $\Lambda(z) = y$, where y fulfills system (39) for some controls $(\widetilde{v}_1, \widetilde{v}_2) \in A(z)$.

Let us recall Kakutani's fixed point theorem (see, for instance, [2]).

THEOREM 2. *Let Z be a Hilbert space and let $\Lambda : Z \mapsto Z$ be a set-valued mapping satisfying the following assumptions:*

1. *$\Lambda(z)$ is a nonempty closed convex set of Z for every $z \in Z$.*
2. *There exists a nonempty convex compact set $E \subset Z$ such that $\Lambda(E) \subset E$.*
3. *Λ is upper-hemicontinuous in Z; i.e., for each $\sigma \in Z'$ the single-valued mapping*

$$z \mapsto \sup_{y \in \Lambda(z)} \langle \sigma, y \rangle_{Z',Z} \tag{55}$$

is upper-semicontinuous.

Then Λ possesses a fixed point in the set E; i.e., there exists $z \in E$ such that $z \in \Lambda(z)$.

Let us check that Kakutani's theorem can be applied to Λ and to

$$Z = H^{3/4}(0,T; L^2(0,1)) \cap L^2(0,T; H^{7/4}(0,1)).$$

Observe that $Z \subset L^\infty(Q) \cap L^2(0,T; W^{1,\infty}(0,1))$.

Let us check the three assumptions of Theorem 2 separately:

• The fact that $\Lambda(z)$ is a nonempty closed convex set of Z for every $z \in Z$ is very easy to verify, so we leave it to the reader.

• Let us prove that Λ maps a compact set into itself. For this, we consider the Hilbert space

$$H = H^1(0,T;L^2(0,1)) \cap L^2(0,T;H^2(0,1)).$$

Then we introduce the space

$$E = \{w \in H : \|w\|_H \leq e^{-K_0 N/(5\nu)}\},$$

where K_0 is the constant in (33). Observe in particular that if $w \in E$, then $\|w\|_\infty \leq N$ as long as $\nu \in (0, \nu_0)$.

It is very easy to check that E is a compact set of Z. Moreover, since $E \subset L^2(0,T;W^{1,\infty}(0,1))$, one can prove that for each $z \in E$ the solution y of (39) belongs to H and there exists a constant $C > 0$ such that

$$\|y\|_H \leq C(\|\widetilde{v}_1\|_{W^{1,\infty}(0,T)} + \|\widetilde{v}_2\|_{W^{1,\infty}(0,T)} + \|y^0\|_{W^{1,p}(0,1)}). \tag{56}$$

Indeed, let us consider the following lifting of the boundary conditions:

$$V(t,x) = (1-x)\widetilde{v}_1(t) + x\widetilde{v}_2(t), \quad t \in (0,T),\ x \in (0,1).$$

By introducing $w := y - V$, our problem (39) is transformed into

$$\begin{cases} w_t - \nu w_{xx} + ((N + z(t,x)/2)w)_x = f(t,x) & \text{in } (0,T)\times(0,1), \\ w|_{x=0} = 0, \quad w|_{x=1} = 0 & \text{in } (0,T), \\ w|_{t=0} = y^0 - (1-x)\widetilde{v}_1(0) + x\widetilde{v}_2(0) & \text{in } (0,1), \end{cases} \tag{57}$$

where

$$f(t,x) = (1-x)\widetilde{v}_{1,t}(t) + x\widetilde{v}_{2,t}(t) + (N/2 + z(t,x))(-\widetilde{v}_1(t) + \widetilde{v}_2(t)) + z_x(t,x)V.$$

This is a linear parabolic equation with an $L^2_t(L^\infty_x)$ coefficient for the zero order term, an $L^\infty(Q)$ coefficient for the first order (in space) term, an $L^\infty(Q)$ right-hand side, and a $W^{1,\infty}(0,1)$ initial condition. In this situation, it is not difficult to prove that the solution of (57) belongs to H and (56) holds. (Observe that, thanks to (48), the initial data in (57) satisfies the required compatibility condition.)

Now, looking at the definition of D given in (41), we see that as long as $T > (\alpha_0 - 2)/N$ (for some $\alpha_0 < 9$), we have

$$D(T,N,z) \geq 0 \qquad \forall z \in E,\ \nu \in (0,\nu_0).$$

Then, from (56) and taking into account estimates (33) and (43), we obtain for some $C > 0$

$$\|y\|_H \leq C(e^{-K_0/\nu} + e^{-K_0/(3\nu)}) \leq e^{-K_0 N/(5\nu)}, \qquad \nu \in (0,\nu_0),$$

and so $y \in E$.

• It remains to check that Λ is upper-hemicontinuous. Thus, assume that $\sigma \in Z'$ and let a sequence $\{z_n\}$ be given, with $z_n \to z$ strongly in Z. We must prove that

$$\overline{\lim}_{n\to+\infty} \sup_{y\in\Lambda(z_n)} \langle \sigma, y\rangle_{Z',Z} \leq \sup_{y\in\Lambda(z)} \langle \sigma, y\rangle_{Z',Z}\ .$$

Let $\{z_{n'}\}$ be a subsequence of $\{z_n\}$ such that

$$\overline{\lim}_{n\to+\infty} \sup_{y\in\Lambda(z_n)} \langle \sigma, y\rangle_{Z',Z} = \lim_{n'\to+\infty} \sup_{y\in\Lambda(z_{n'})} \langle \sigma, y\rangle_{Z',Z} .$$

Since each $\Lambda(z_{n'})$ is a compact set of Z, for every n' we have

$$\sup_{y\in\Lambda(z_{n'})} \langle \sigma, y\rangle_{Z',Z} = \langle \sigma, y_{n'}\rangle_{Z',Z}$$

for some $y_{n'} \in \Lambda(z_{n'})$. On the other hand, since all the states $y_{n'}$ belong to the same compact set E, at least for a new subsequence (again indexed by n'), we must have $y_{n'} \to y$ strongly in Z. We will now prove that $y \in \Lambda(z)$. This will achieve the proof of the upper-hemicontinuity of Λ.

Indeed, it can be assumed that the controls $\widetilde{v}_{1,n'}$ and $\widetilde{v}_{2,n'}$ converge to some functions $\widetilde{v}_1$ and $\widetilde{v}_2$ weakly-$*$ in $W^{1,\infty}(0,T)$. Then y solves (39) and $y|_{t=T} = 0$. Moreover, since inequality (43) is independent of n, $\widetilde{v}_1$ and $\widetilde{v}_2$ also satisfy (43). Therefore, $(\widetilde{v}_1, \widetilde{v}_2) \in A(z)$. Consequently, it is immediate that y is the solution to (39) associated with the controls $\widetilde{v}_1$ and $\widetilde{v}_2$.

This shows that $y \in \Lambda(z)$ and, therefore, Λ is upper-hemicontinuous.

Consequently, Kakutani's theorem applies and this implies that there exists $y \in \Lambda(y)$; that is to say, we have found a function y solution of (36) such that (37) and (38) (thanks to (43)) are satisfied. The proof of Proposition 4 is finished. □

3. Proof of Proposition 2. Again relying on the invariance of the solutions of (1) by the transformation $u(t,x) \leftrightarrow -u(t,1-x)$, we can always assume that $M > 0$. Now Proposition 2 is proved approximately as Proposition 1, but here a (viscous) rarefaction wave is used in place of a traveling wave. More precisely, we start from N, which can be chosen larger than M. First, we reach M approximately and then reach M exactly by using the same argument as above.

3.1. Reaching M approximately. Let us prove the next proposition.

PROPOSITION 5. *One can find controls v_1 and v_2 such that the solution of* (1)–(3) *with initial condition $u_{|t=0} = N$ satisfies, for some constant $C > 0$ independent from M, N, and ν,*

$$\|u(t,\cdot) - M\|_{W^{1,\infty}([0,1])} \le CM\sqrt{\nu t} \exp\left\{-\frac{M^2}{4\nu}\left(t - \frac{2}{M}\right)\right\} \tag{58}$$

for any $t > \frac{2}{M}$ and, moreover, the controls satisfy, independently from ν,

$$\|v_1\|_{L^\infty(0,T)} + \|v_2\|_{L^\infty(0,T)} \le N. \tag{59}$$

Proof of Proposition 5. In this situation, the solution u is obtained by taking the restriction to $[0,T]\times[0,1]$ of the solution defined on the whole space domain $\mathbb{R}$ as the unique solution with initial condition:

$$u(0,x) := \hat{u}_0 = \begin{cases} M \text{ if } x \le 0, \\ N \text{ if } x > 0. \end{cases} \tag{60}$$

Then v_1 and v_2 are obtained by taking the traces of u along the lines $(0,T)\times\{0\}$ and $(0,T)\times\{1\}$. As before, (59) follows directly from the maximum principle, so we have only to check (58).

In a first step we consider only the L^∞ norm. The solution $u(t,x)$ has the following explicit form:

$$u(t,x) = \frac{\int_{-\infty}^{0} \frac{x-y}{t} \exp\Big\{ -\frac{1}{2\nu}\Big(\frac{(x-y)^2}{2t} + My\Big)\Big\}dy + \int_{0}^{+\infty} \frac{x-y}{t} \exp\Big\{ -\frac{1}{2\nu}\Big(\frac{(x-y)^2}{2t} + Ny\Big)\Big\}dy}{\int_{-\infty}^{0} \exp\Big\{ -\frac{1}{2\nu}\Big(\frac{(x-y)^2}{2t} + My\Big)\Big\}dy + \int_{0}^{+\infty} \exp\Big\{ -\frac{1}{2\nu}\Big(\frac{(x-y)^2}{2t} + Ny\Big)\Big\}dy}.$$

We note that

$$\int_{-\infty}^{0} \frac{x-y}{t} \exp\Big\{ -\frac{1}{2\nu}\Big(\frac{(x-y)^2}{2t} + My\Big)\Big\}dy = 2\nu \exp\Big(-\frac{x^2}{4\nu t}\Big) + M\int_{-\infty}^{0} \exp\Big\{ -\frac{1}{2\nu}\Big(\frac{(x-y)^2}{2t} + My\Big)\Big\}dy$$

(as seen by adding and subtracting M inside the integral). In the same way, we have

$$\int_{0}^{+\infty} \frac{x-y}{t} \exp\Big\{ -\frac{1}{2\nu}\Big(\frac{(x-y)^2}{2t} + Ny\Big)\Big\}dy = -2\nu \exp\Big(-\frac{x^2}{4\nu t}\Big) + N\int_{0}^{+\infty} \exp\Big\{ -\frac{1}{2\nu}\Big(\frac{(x-y)^2}{2t} + Ny\Big)\Big\}dy.$$

Hence we get that

$$u(t,x) - M = \frac{(N-M)\int_{0}^{+\infty} \exp\Big\{ -\frac{1}{2\nu}\Big(\frac{(x-y)^2}{2t} + Ny\Big)\Big\}dy}{\int_{-\infty}^{0} \exp\Big\{ -\frac{1}{2\nu}\Big(\frac{(x-y)^2}{2t} + My\Big)\Big\}dy + \int_{0}^{+\infty} \exp\Big\{ -\frac{1}{2\nu}\Big(\frac{(x-y)^2}{2t} + Ny\Big)\Big\}dy}.$$

Note that $\hat{u}_0$ is nondecreasing; hence $u(t,\cdot)$ is also nondecreasing (as seen from Lemma 1 and comparing the solutions corresponding to $\hat{u}_0$ and $\hat{u}_0(\cdot + h)$). Using the fact that $\hat{u}_0 \geq M$, we deduce, together with the maximum principle, that $u \geq M$. Consequently, it is sufficient to have an upper estimate for $u(t,1) - M$. Now from (29) we have

$$\int_{-\infty}^{0} \exp\left\{-\frac{1}{2\nu}\left(\frac{(1-y)^2}{2t} + My\right)\right\} dy = \sqrt{4\nu t} \exp\left(\frac{M(Mt-2)}{4\nu}\right) \int_{-\infty}^{\xi_M} e^{-s^2} ds,$$

$$\int_{0}^{+\infty} \exp\left\{-\frac{1}{2\nu}\left(\frac{(1-y)^2}{2t} + Ny\right)\right\} dy = \sqrt{4\nu t} \exp\left(\frac{N(Nt-2)}{4\nu}\right) \int_{\xi_N}^{+\infty} e^{-s^2} ds,$$

with

$$\xi_M := \frac{Mt-1}{2\sqrt{\nu t}} \text{ and } \xi_N := \frac{Nt-1}{2\sqrt{\nu t}}.$$

We deduce

$$u(t,1)-M \le \frac{(N-M)\int_0^{+\infty}\exp\Big\{-\frac{1}{2\nu}\Big(\frac{(1-y)^2}{2t}+Ny\Big)\Big\}dy}{\int_{-\infty}^{0}\exp\Big\{-\frac{1}{2\nu}\Big(\frac{(1-y)^2}{2t}+My\Big)\Big\}dy}$$

$$\le \frac{(N-M)\exp\big(\frac{M(Mt-2)}{4\nu}\big)\int_{-\infty}^{\xi_M} e^{-s^2}ds}{\exp\big(\frac{N(Nt-2)}{4\nu}\big)\int_{\xi_N}^{+\infty} e^{-s^2}ds}$$

$$\le \frac{(N-M)\exp\big(\frac{M(Mt-2)}{4\nu}\big)\int_{-\infty}^{\xi_M} e^{-s^2}ds}{\sqrt{\pi}/2}.$$

With (31), we get

$$u(t,1)-M \le \frac{2\sqrt{\nu t}}{\sqrt{\pi}}\frac{N-M}{Nt-1}\exp\left\{\frac{-1}{4\nu t}\right\}\exp\left\{-\frac{M^2}{4\nu}\left(t-\frac{2}{M}\right)\right\},$$

and the result in the L^∞ norm follows using $t>(2/M)$.

The L^∞ estimate on $\partial_x u$ is done approximately as in section 2.1 by using Lemma 2. We fix $x\in[0,1]$; we get

$$\partial_x u(t,x) = \frac{1}{2\nu}\frac{\int_{-\infty}^{+\infty}\frac{y-x}{t}(\hat{u}_0(y)-M)\exp\Big\{-\frac{1}{2\nu}\big[\frac{(x-y)^2}{2t}+\int_0^y\hat{u}_0(\eta)d\eta\big]\Big\}dy}{\int_{-\infty}^{+\infty}\exp\Big\{-\frac{1}{2\nu}\big[\frac{(x-y)^2}{2t}+\int_0^y\hat{u}_0(\eta)d\eta\big]\Big\}dy}$$

$$+\frac{1}{2\nu}\frac{\int_{-\infty}^{+\infty}\frac{y-x}{t}(M-u(t,x))\exp\Big\{-\frac{1}{2\nu}\big[\frac{(x-y)^2}{2t}+\int_0^y\hat{u}_0(\eta)d\eta\big]\Big\}dy}{\int_{-\infty}^{+\infty}\exp\Big\{-\frac{1}{2\nu}\big[\frac{(x-y)^2}{2t}+\int_0^y\hat{u}_0(\eta)d\eta\big]\Big\}dy}$$

$$=: A+B.$$

Using again the fact that $u(t,\cdot)$ is nondecreasing, we see that we have only to give an upper bound for $\partial_x u$. The second term B satisfies

$$B=\frac{1}{2\nu}(M-u(t,x))u(t,x)$$

and thus is clearly nonpositive, as follows from the maximum principle. Therefore, it remains to estimate the first term A. To this aim, we estimate the denominator from below as follows:

$$\int_{-\infty}^{+\infty}\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+\int_0^y\hat{u}_0(\eta)d\eta\right]\right\}dy$$

$$\ge \int_{-\infty}^{0}\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+\int_0^y\hat{u}_0(\eta)d\eta\right]\right\}dy$$

$$= \int_{-\infty}^{0}\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+My\right]\right\}dy$$

$$\ge \sqrt{\pi\nu t}\exp\left\{\frac{M}{2\nu}\left(\frac{Mt}{2}-x\right)\right\}.$$

For the numerator, thanks to (30), we have

$$\begin{aligned}
\mathcal{N} &:= \int_0^{+\infty} \frac{y-x}{t}(\hat{u}_0(y)-M)\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+\int_0^y \hat{u}_0(\eta)d\eta\right]\right\}dy \\
&\le \int_0^{+\infty} \frac{y}{t}(N-M)\exp\left\{-\frac{1}{2\nu}\left[\frac{(x-y)^2}{2t}+Ny\right]\right\}dy \\
&\le \frac{N-M}{t}\left[2\sqrt{\nu t}(x-Nt)\exp\left\{\frac{N(-2x+Nt)}{4\nu}\right\}\right. \\
&\qquad\qquad \left.\times\int_{\frac{-x+Nt}{2\sqrt{\nu t}}}^{+\infty} e^{-s^2}ds + 2\nu t\exp\left\{-\frac{x^2}{4\nu t}\right\}\right] \\
&= \frac{N-M}{t}\exp\left\{\frac{N(-2x+Nt)}{4\nu}\right\} \\
&\qquad\qquad \times\left[2\sqrt{\nu t}(x-Nt)\int_{\frac{-x+Nt}{2\sqrt{\nu t}}}^{+\infty} e^{-s^2}ds + 2\nu t\exp\left\{-\frac{(x-Nt)^2}{4\nu t}\right\}\right].
\end{aligned}$$

Simple integrations by parts prove that

$$-\int_y^{+\infty} e^{-s^2}ds + \frac{e^{-y^2}}{2y} \le \frac{e^{-y^2}}{4y^3}. \tag{61}$$

Plugging (61) with $y=\frac{-x+Nt}{2\sqrt{\nu t}}$ into the previous estimate of $\mathcal{N}$ and using $Nt \ge 1 \ge x$, we deduce

$$\mathcal{N} \le 4\frac{N-M}{(Nt-1)^2}\nu^2 t\exp\left\{\frac{-x^2}{4\nu t}\right\}.$$

Using $N > M$, we finally obtain

$$\mathcal{N} \le 4M\nu^2 t,$$

which yields the result.

3.2. Reaching M exactly. Reaching M exactly is done identically to reaching N exactly (see section 2.2). This is due to the fact that we did not use the size of N in subsection 2.2. This ends the proof of section 2 and hence of Theorem 1.

4. Technical results.

4.1. Proof of the dissipation result. In this first part, we will prove the estimate presented in (51):

$$\begin{aligned}
\|\varphi|_{t=t_0}\|_{L^2(0,1)} \le \exp\left\{\frac{\|z_x\|_{L^1_t(L^\infty_x)}}{8} - \frac{((N-\|z\|_\infty/2)t^*-1)^2}{4\nu t^*}e^{-2\|z_x\|_{L^1_t(L^\infty_x)}}\right\} \\
\times \|\varphi|_{t=t_0+t^*}\|_{L^2(0,1)},
\end{aligned} \tag{62}$$

for any $t_0 \in (0, T^*)$ and any $t^* \in (0, T^* - t_0)$ such that (52) is satisfied. Here, φ designs the solution of the system

$$\begin{cases} -\varphi_t - \nu\varphi_{xx} - (N + z(t,x)/2)\varphi_x = 0 & \text{in } (0,T^*)\times(0,1), \\ \varphi|_{x=0} = 0, \quad \varphi|_{x=1} = 0 & \text{in } (0,T^*), \\ \varphi|_{t=T^*} = \varphi^0 & \text{in } (0,1). \end{cases} \tag{63}$$

Let us first define a function $\Theta(t,x) = e^{\theta(t,x)}$, where $\theta \in L^\infty(0,T^*; W^{1,\infty}(0,1))$ is chosen as follows: We set $\theta(t,x) := r_0|\psi^{-1}(t,x)|$, with $r_0 > 0$ a constant which will be chosen later on, and where ψ is the backward flow associated with $N + z(t,x)/2$. More precisely, ψ is given by

$$\begin{cases} \dfrac{d\psi}{dt}(t,x) = N + \dfrac{z(t,\psi(t,x))}{2}, \\ \psi_{|t=t_0+t^*} = x. \end{cases}$$

Here, we have extended $z(t,\cdot)$ by $z(t,0)$ on the left of 0 and by $z(t,1)$ on the right of 1. In particular, Θ is defined for $x \in \mathbb{R}$ and satisfies

$$\Theta_t + (N + z(t,x)/2)\Theta_x = 0. \tag{64}$$

We regard the equation satisfied by $\Theta\,\varphi$ as the following (in fact, to be complete, we should regularize Θ and z, establish estimates for regularized Θ and z, and then pass to the limit):

$$-(\Theta\varphi)_t - \nu(\Theta\varphi)_{xx} - (N + z(t,x)/2)(\Theta\varphi)_x = -\nu\varphi\Theta_{xx} - 2\nu\varphi_x\Theta_x. \tag{65}$$

We multiply by $\Theta\varphi$ and integrate on $(0,1)$. After integration by parts we obtain

$$\begin{aligned} &-\frac{1}{2}\frac{d}{dt}\int_0^1 |e^\theta\varphi|^2\,dx + \frac{1}{4}\int_0^1 z_x|e^\theta\varphi|^2\,dx + \nu\int_0^1 |(e^\theta\varphi)_x|^2\,dx \\ &= \nu\int_0^1 |\Theta\theta_x\varphi|^2\,dx \le \nu\|\theta_x\|_\infty^2\int_0^1 |e^\theta\varphi|^2\,dx. \end{aligned} \tag{66}$$

After an application of Gronwall's lemma in the time interval $(t_0, t_0 + t^*)$, we find the following from (66):

$$\begin{aligned} &\int_0^1 |e^\theta\varphi|^2(t_0)\,dx \\ &\le \exp\left\{\frac{\|z_x\|_{L^1_t(L^\infty_x)}}{4} + 2\nu r_0^2 t^* \exp\left(2\|z_x\|_{L^1_t(L^\infty_x)}\right)\right\}\int_0^1 |e^\theta\varphi|^2(t_0+t^*)\,dx. \end{aligned} \tag{67}$$

Here, we have used the expression of θ together with the estimate

$$|\psi_x^{-1}(t,x)|^2 \le \exp\left\{2\int_0^{T^*} \|z_x(s)\|_\infty\,ds\right\}, \quad t \in (0,T),\ x \in (0,1).$$

Now, from the expression of $\psi^{-1}(t,x)$ we observe that

$$\psi^{-1}(t_0,x) \le 1 - (N - \|z\|_\infty/2)t^* \text{ for } x \in (0,1).$$

Then from (67) we find that

$$\int_0^1 |\varphi|^2(t_0)\,dx \leq C(r_0,t^*)\int_0^1 |\varphi|^2(t_0+t^*)\,dx, \tag{68}$$

with

$$C=\exp\left\{\frac{\|z_x\|_{L^1_t(L^\infty_x)}}{4}+\ 2\nu r_0^2 t^*\exp\left(2\|z_x\|_{L^1_t(L^\infty_x)}\right)+2r_0(1-(N-\|z\|_\infty/2)t^*)\right\}.$$

Finally, we choose

$$r_0=\frac{e^{-2\|z_x\|_{L^1_t(L^\infty_x)}}((N-\|z\|_\infty/2)t^*-1)}{2\nu t^*}$$

and we find the desired inequality (62) squared.

4.2. Proof of the Carleman inequality. In this, the last section of the paper, we will provide the proof of the Carleman inequality which was presented in (53). In order to prove this estimate, we follow the steps of the proof in [5].

Hence, let us first perform a change of variables in order to restrict ourselves to the case where $\nu=1$:

$$\begin{cases}\widetilde{t}=\nu t,\\ \widetilde{x}=x.\end{cases} \tag{69}$$

In the new variables, we have, with $\widetilde{\varphi}(\widetilde{t},\widetilde{x}):=\varphi(t,x)$ and $\widetilde{z}(\widetilde{t},\widetilde{x})=z(t,x)$,

$$\begin{cases}\widetilde{\varphi}_{\widetilde{t}}+\widetilde{\varphi}_{\widetilde{x}\widetilde{x}}+\nu^{-1}N\widetilde{\varphi}_{\widetilde{x}}=-\nu^{-1}(\widetilde{z}(\widetilde{t},\widetilde{x})/2)\widetilde{\varphi}_{\widetilde{x}}, & (\widetilde{t},\widetilde{x})\in(0,\nu T^*)\times(0,1),\\ \widetilde{\varphi}(\widetilde{t},0)=\widetilde{\varphi}(\widetilde{t},1)=0, & \widetilde{t}\in(0,\nu T^*),\\ \widetilde{\varphi}(\nu T^*,\widetilde{x})=\widetilde{\varphi}^0(\widetilde{x}), & \widetilde{x}\in(0,1).\end{cases} \tag{70}$$

Let

$$\widetilde{N}:=\frac{N}{\nu}, \tag{71}$$

$$\widetilde{T}:=\nu T^*. \tag{72}$$

Then, condition (42) implies

$$\widetilde{N}\widetilde{T}\geqslant(3/2). \tag{73}$$

Let us define a weight function, similar to the one introduced by Fursikov and Imanuvilov in [8],

$$\alpha(\widetilde{t},\widetilde{x}):=\frac{\beta(\widetilde{x})}{\widetilde{T}-\widetilde{t}}\quad(\widetilde{t},\widetilde{x})\in(0,\widetilde{T})\times(0,1), \tag{74}$$

where $0\leqslant\beta\in C^2([0,1])$ will be chosen below. We also introduce the function

$$\psi:=e^{-\alpha}\widetilde{\varphi},$$

which verifies

$$P_1\psi + P_2\psi = P_3\psi, \tag{75}$$

with

$$P_1\psi := \psi_{\widetilde{x}\widetilde{x}} + \alpha_{\widetilde{x}}^2\psi + \widetilde{N}\alpha_{\widetilde{x}}\psi + \alpha_{\widetilde{t}}\psi,$$

$$P_2\psi := \psi_{\widetilde{t}} + 2\alpha_{\widetilde{x}}\psi_{\widetilde{x}} + \widetilde{N}\psi_{\widetilde{x}},$$

$$P_3\psi := -\alpha_{\widetilde{x}\widetilde{x}}\psi - \nu^{-1}(\widetilde{z}(\widetilde{t},\widetilde{x})/2)(\alpha_{\widetilde{x}}\psi + \psi_{\widetilde{x}}).$$

We develop here the classical proof, consisting of taking the L^2 norm in identity (75), and then develop all the double products:

$$\|P_1\psi\|^2_{L^2(Q)} + \|P_2\psi\|^2_{L^2(Q)} + 2(P_1\psi, P_2\psi)_{L^2(Q)} = \|P_3\psi\|^2_{L^2(Q)}, \tag{76}$$

where Q stands for the open set $(0,\widetilde{T})\times(0,1)$.

Let us compute $2(P_1\psi, P_2\psi)_{L^2(Q)}$. Let us first compute the terms concerning $\psi_{\widetilde{x}\widetilde{x}}$. We have

$$(\psi_{\widetilde{x}\widetilde{x}}, \psi_{\widetilde{t}})_{L^2(Q)} = \frac{1}{2}\int_0^1 |\psi_{\widetilde{x}}(0,\widetilde{x})|^2\, d\widetilde{x}.$$

Moreover,

$$\begin{aligned} &2(\psi_{\widetilde{x}\widetilde{x}}, \alpha_{\widetilde{x}}\psi_{\widetilde{x}})_{L^2(Q)} \\ &\quad = \int_0^{\widetilde{T}} (\alpha_{\widetilde{x}}(\widetilde{t},1)|\psi_{\widetilde{x}}(\widetilde{t},1)|^2 - \alpha_{\widetilde{x}}(\widetilde{t},0)|\psi_{\widetilde{x}}(\widetilde{t},0)|^2)\, d\widetilde{t} - \iint_Q \alpha_{\widetilde{x}\widetilde{x}}|\psi_{\widetilde{x}}|^2\, d\widetilde{x}\, d\widetilde{t}. \end{aligned} \tag{77}$$

Finally,

$$\widetilde{N}(\psi_{\widetilde{x}\widetilde{x}}, \psi_{\widetilde{x}})_{L^2(Q)} = (\widetilde{N}/2)\int_0^{\widetilde{T}} (|\psi_{\widetilde{x}}(\widetilde{t},1)|^2 - |\psi_{\widetilde{x}}(\widetilde{t},0)|^2)\, d\widetilde{t}.$$

As far as the term $\alpha_{\widetilde{x}}^2\psi$ is concerned, we first have

$$(\alpha_{\widetilde{x}}^2\psi, \psi_{\widetilde{t}})_{L^2(Q)} = -\iint_Q \alpha_{\widetilde{x}}\alpha_{\widetilde{x}\widetilde{t}}|\psi|^2\, d\widetilde{x}\, d\widetilde{t} - \frac{1}{2}\int_0^1 \alpha_{\widetilde{x}}^2(0,\widetilde{x})|\psi(0,\widetilde{x})|^2\, d\widetilde{x}.$$

Next,

$$2(\alpha_{\widetilde{x}}^2\psi, \alpha_{\widetilde{x}}\psi_{\widetilde{x}})_{L^2(Q)} = -3\iint_Q \alpha_{\widetilde{x}\widetilde{x}}\alpha_{\widetilde{x}}^2|\psi|^2\, d\widetilde{x}\, d\widetilde{t}.$$

Finally,

$$\widetilde{N}(\alpha_{\widetilde{x}}^2\psi, \psi_{\widetilde{x}})_{L^2(Q)} = -\widetilde{N}\iint_Q \alpha_{\widetilde{x}\widetilde{x}}\alpha_{\widetilde{x}}|\psi|^2\, d\widetilde{x}\, d\widetilde{t}.$$

Let us next perform the terms concerning $\widetilde{N}\alpha_{\widetilde{x}}\psi$. First, we have

$$\widetilde{N}(\alpha_{\widetilde{x}}\psi, \psi_{\widetilde{t}})_{L^2(Q)} = -(\widetilde{N}/2)\iint_Q \alpha_{\widetilde{x}\widetilde{t}}|\psi|^2\, d\widetilde{x}\, d\widetilde{t} - (\widetilde{N}/2)\int_0^1 \alpha_{\widetilde{x}}(0,\widetilde{x})|\psi(0,\widetilde{x})|^2\, d\widetilde{x}.$$

Then, we find

$$2\widetilde{N}(\alpha_{\widetilde{x}}\psi, \alpha_{\widetilde{x}}\psi_{\widetilde{x}})_{L^2(Q)} = -2\widetilde{N}\iint_Q \alpha_{\widetilde{x}}\alpha_{\widetilde{x}\widetilde{x}}|\psi|^2\, d\widetilde{x}\, d\widetilde{t}.$$

The last term provides

$$\widetilde{N}^2(\alpha_{\widetilde{x}}\psi, \psi_{\widetilde{x}})_{L^2(Q)} = -(\widetilde{N}^2/2)\iint_Q \alpha_{\widetilde{x}\widetilde{x}}|\psi|^2\, d\widetilde{x}\, d\widetilde{t}.$$

Lastly, we deal with the computations of the term $\alpha_{\widetilde{t}}\psi$. First, we obtain

$$(\alpha_{\widetilde{t}}\psi, \psi_{\widetilde{t}})_{L^2(Q)} = -(1/2)\iint_Q \alpha_{\widetilde{t}\widetilde{t}}|\psi|^2\, d\widetilde{x}\, d\widetilde{t} - (1/2)\int_0^1 \alpha_{\widetilde{t}}(0,\widetilde{x})|\psi(0,\widetilde{x})|^2\, d\widetilde{x}.$$

Additionally, we find

$$(\alpha_{\widetilde{t}}\psi, 2\alpha_{\widetilde{x}}\psi_{\widetilde{x}})_{L^2(Q)} = -\iint_Q (\alpha_{\widetilde{t}}\alpha_{\widetilde{x}\widetilde{x}} + \alpha_{\widetilde{t}\widetilde{x}}\alpha_{\widetilde{x}})|\psi|^2\, d\widetilde{x}\, d\widetilde{t}.$$

Finally,

$$(\alpha_{\widetilde{t}}\psi, \widetilde{N}\psi_{\widetilde{x}})_{L^2(Q)} = -(\widetilde{N}/2)\iint_Q \alpha_{\widetilde{t}\widetilde{x}}|\psi|^2\, d\widetilde{x}\, d\widetilde{t}.$$

Putting all these computations together, we conclude that the double product term is

$$\begin{aligned}
2(P_1\psi, P_2\psi)_{L^2(Q)} &= \int_0^1 |\psi_{\widetilde{x}}(0,\widetilde{x})|^2\, d\widetilde{x} \\
&+ \int_0^{\widetilde{T}} ((2\alpha_{\widetilde{x}}(\widetilde{t},1) + \widetilde{N})|\psi_{\widetilde{x}}(\widetilde{t},1)|^2 - (2\alpha_{\widetilde{x}}(\widetilde{t},0) + \widetilde{N})|\psi_{\widetilde{x}}(\widetilde{t},0)|^2)\, d\widetilde{t} \\
&- 2\iint_Q \alpha_{\widetilde{x}\widetilde{x}}|\psi_{\widetilde{x}}|^2\, d\widetilde{x}\, d\widetilde{t} - 4\iint_Q \alpha_{\widetilde{x}}\alpha_{\widetilde{x}\widetilde{t}}|\psi|^2\, d\widetilde{x}\, d\widetilde{t} \\
&- \int_0^1 \alpha_{\widetilde{x}}^2(0,\widetilde{x})|\psi(0,\widetilde{x})|^2\, d\widetilde{x} - 2\iint_Q (3\alpha_{\widetilde{x}\widetilde{x}}\alpha_{\widetilde{x}}^2 + \widetilde{N}\alpha_{\widetilde{x}\widetilde{t}})|\psi|^2\, d\widetilde{x}\, d\widetilde{t} \\
&- \widetilde{N}\int_0^1 \alpha_{\widetilde{x}}(0,\widetilde{x})|\psi(0,\widetilde{x})|^2\, d\widetilde{x} - \widetilde{N}\iint_Q (6\alpha_{\widetilde{x}}\alpha_{\widetilde{x}\widetilde{x}} + \widetilde{N}\alpha_{\widetilde{x}\widetilde{x}})|\psi|^2\, d\widetilde{x}\, d\widetilde{t} \\
&- \iint_Q \alpha_{\widetilde{t}\widetilde{t}}|\psi|^2\, d\widetilde{x}\, d\widetilde{t} - \int_0^1 \alpha_{\widetilde{t}}(0,\widetilde{x})|\psi(0,\widetilde{x})|^2\, d\widetilde{x} - 2\iint_Q \alpha_{\widetilde{t}}\alpha_{\widetilde{x}\widetilde{x}}|\psi|^2\, d\widetilde{x}\, d\widetilde{t}.
\end{aligned} \tag{78}$$

On the other hand, we have the following for the right-hand side term:

$$\|P_3\psi\|^2_{L^2(Q)} \le \iint_Q (2\alpha_{\widetilde{x}\widetilde{x}}^2|\psi|^2 + \nu^{-2}|\widetilde{z}(\widetilde{t},\widetilde{x})|^2(\alpha_{\widetilde{x}}^2|\psi|^2 + |\psi_{\widetilde{x}}|^2))\, d\widetilde{x}\, d\widetilde{t}. \tag{79}$$

Combining (78)–(79) with (76), we obtain

$$\begin{aligned}
&\int_0^{\widetilde{T}} (2\alpha_{\widetilde{x}}(\widetilde{t},1)+\widetilde{N})|\psi_{\widetilde{x}}(\widetilde{t},1)|^2\,d\widetilde{t} - 2\iint_Q \alpha_{\widetilde{x}\widetilde{x}}|\psi_{\widetilde{x}}|^2\,d\widetilde{x}\,d\widetilde{t} - 6\iint_Q \alpha_{\widetilde{x}\widetilde{x}}\alpha_{\widetilde{x}}^2|\psi|^2\,d\widetilde{x}\,d\widetilde{t}\\
&-2\iint_Q \alpha_{\widetilde{t}}\alpha_{\widetilde{x}\widetilde{x}}|\psi|^2\,d\widetilde{x}\,d\widetilde{t} - 6\widetilde{N}\iint_Q \alpha_{\widetilde{x}}\alpha_{\widetilde{x}\widetilde{x}}|\psi|^2\,d\widetilde{x}\,d\widetilde{t} - \widetilde{N}^2\iint_Q \alpha_{\widetilde{x}\widetilde{x}}|\psi|^2\,d\widetilde{x}\,d\widetilde{t}\\
&\qquad\leqslant \iint_Q (2\alpha_{\widetilde{x}\widetilde{x}}^2|\psi|^2+\nu^{-2}|\widetilde{z}(\widetilde{t},\widetilde{x})|^2(\alpha_{\widetilde{x}}^2|\psi|^2+|\psi_{\widetilde{x}}|^2))\,d\widetilde{x}\,d\widetilde{t}\\
&\qquad+\int_0^{\widetilde{T}} (2\alpha_{\widetilde{x}}(\widetilde{t},0)+\widetilde{N})|\psi_{\widetilde{x}}(\widetilde{t},0)|^2\,d\widetilde{t} + 4\iint_Q \alpha_{\widetilde{x}}\alpha_{\widetilde{x}\widetilde{t}}|\psi|^2\,d\widetilde{x}\,d\widetilde{t}\\
&\qquad+2\widetilde{N}\iint_Q \alpha_{\widetilde{x}\widetilde{t}}|\psi|^2\,d\widetilde{x}\,d\widetilde{t} + \iint_Q \alpha_{\widetilde{t}\widetilde{t}}|\psi|^2\,d\widetilde{x}\,d\widetilde{t} + \int_0^1 \alpha_{\widetilde{x}}^2(0,\widetilde{x})|\psi(0,\widetilde{x})|^2\,d\widetilde{x}\\
&\qquad+\widetilde{N}\int_0^1 \alpha_{\widetilde{x}}(0,\widetilde{x})|\psi(0,\widetilde{x})|^2\,d\widetilde{x} + \int_0^1 \alpha_{\widetilde{t}}(0,\widetilde{x})|\psi(0,\widetilde{x})|^2\,d\widetilde{x}.
\end{aligned}$$

From the definition of α (given in (74)), we find

$$\begin{aligned}
&\int_0^{\widetilde{T}} \left(2\frac{\beta'(1)}{\widetilde{T}-\widetilde{t}}+\widetilde{N}\right)|\psi_{\widetilde{x}}(\widetilde{t},1)|^2\,d\widetilde{t} - 2\iint_Q \frac{\beta''(\widetilde{x})}{\widetilde{T}-\widetilde{t}}|\psi_{\widetilde{x}}|^2\,d\widetilde{x}\,d\widetilde{t}\\
&-6\iint_Q \frac{\beta''(\widetilde{x})(\beta'(\widetilde{x}))^2}{(\widetilde{T}-\widetilde{t})^3}|\psi|^2\,d\widetilde{x}\,d\widetilde{t} - 2\iint_Q \frac{\beta(\widetilde{x})\beta''(\widetilde{x})}{(\widetilde{T}-\widetilde{t})^3}|\psi|^2\,d\widetilde{x}\,d\widetilde{t}\\
&-6\widetilde{N}\iint_Q \frac{\beta'(\widetilde{x})\beta''(\widetilde{x})}{(\widetilde{T}-\widetilde{t})^2}|\psi|^2\,d\widetilde{x}\,d\widetilde{t} - \widetilde{N}^2\iint_Q \frac{\beta''(\widetilde{x})}{\widetilde{T}-\widetilde{t}}|\psi|^2\,d\widetilde{x}\,d\widetilde{t}\\
&\qquad\leqslant \frac{\|z\|_\infty^2}{\nu^2}\iint_Q |\psi_{\widetilde{x}}|^2\,d\widetilde{x}\,d\widetilde{t} + \iint_Q \frac{2(\beta''(\widetilde{x}))^2+\nu^{-2}\|z\|^2(\beta'(\widetilde{x}))^2}{(\widetilde{T}-\widetilde{t})^2}|\psi|^2\,d\widetilde{x}\,d\widetilde{t}\\
&\qquad+\int_0^{\widetilde{T}} \left(2\frac{\beta'(0)}{\widetilde{T}-\widetilde{t}}+\widetilde{N}\right)|\psi_{\widetilde{x}}(\widetilde{t},0)|^2\,d\widetilde{t} + 4\iint_Q \frac{(\beta'(\widetilde{x}))^2}{(\widetilde{T}-\widetilde{t})^3}|\psi|^2\,d\widetilde{x}\,d\widetilde{t}\\
&\qquad+2\iint_Q \left(\widetilde{N}\beta'(\widetilde{x})+\frac{\beta(\widetilde{x})}{\widetilde{T}-\widetilde{t}}\right)\frac{|\psi|^2}{(\widetilde{T}-\widetilde{t})^2}\,d\widetilde{x}\,d\widetilde{t} + \int_0^1 \frac{\beta'(\widetilde{x})^2}{\widetilde{T}^2}|\psi(0,\widetilde{x})|^2\,d\widetilde{x}\\
&\qquad+\frac{\widetilde{N}}{\widetilde{T}}\int_0^1 \beta'(\widetilde{x})|\psi(0,\widetilde{x})|^2\,d\widetilde{x} + \frac{1}{\widetilde{T}^2}\int_0^1 \beta(\widetilde{x})|\psi(0,\widetilde{x})|^2\,d\widetilde{x}.
\end{aligned} \tag{80}$$

Let us now define the function $\beta : [0,1] \mapsto \mathbb{R}$. We will take a function satisfying

$$\beta''(\widetilde{x}) = -\frac{1}{1-\delta}\frac{2(\beta'(\widetilde{x}))^2+\beta(\widetilde{x})}{3(\beta'(\widetilde{x}))^2+\beta(\widetilde{x})}, \quad \widetilde{x}\in[0,1], \tag{81}$$

together with the initial conditions

$$\beta(0) = \delta \text{ and } \beta'(0) = \lambda, \tag{82}$$

where $\lambda > 0$ and $\delta \in (0,1)$ are parameters to be determined.

Now we claim that for proper λ, the function β is well defined and satisfies

$$\beta > 0,\ \beta' > 0 \text{ and } \beta'' < 0 \text{ on } [0,1]. \tag{83}$$

Once such a function β is obtained, we consider χ a constant satisfying

$$\beta(1) < \chi. \tag{84}$$

In what follows, the smaller χ is, the better the estimates will be.

Remark 7. One can check, for instance with MATLAB, that, if $\delta > 0$ is small enough and if one fixes $\lambda := 0.807$, then the corresponding β is defined on $[0,1]$ and satisfies (84) with $\chi = 0.435$.

At the end of the paper, we will establish elementarily the following lemma.

LEMMA 4. *There are some values of $\lambda > 0$ and $\delta \in (0,1)$ such that the unique solution β of* (81)–(82) *is well defined in* $[0,1]$ *and satisfies*

$$\beta(1) < (2/3). \tag{85}$$

From now on, we suppose that we have such a β satisfying (81), (82), (83), and (84). We remark that the first term in the left-hand side of (80) is nonnegative, and we regroup the third and fourth terms of the left-hand side, together with the fourth and sixth terms of the right-hand side. We deduce

$$\begin{aligned}
&-2\iint_Q \frac{\beta''(\widetilde{x})}{\widetilde{T}-\widetilde{t}}|\psi_{\widetilde{x}}|^2\,d\widetilde{x}\,d\widetilde{t} - 6\delta\iint_Q \beta''(\widetilde{x})\frac{\beta(\widetilde{x})+3(\beta'(\widetilde{x}))^2}{(\widetilde{T}-\widetilde{t})^3}|\psi|^2\,d\widetilde{x}\,d\widetilde{t}\\
&-6\widetilde{N}\iint_Q \frac{\beta'(\widetilde{x})\beta''(\widetilde{x})}{(\widetilde{T}-\widetilde{t})^2}|\psi|^2\,d\widetilde{x}\,d\widetilde{t} - \widetilde{N}^2\iint_Q \frac{\beta''(\widetilde{x})}{\widetilde{T}-\widetilde{t}}|\psi|^2\,d\widetilde{x}\,d\widetilde{t}\\
&\quad\leqslant \frac{\|z\|_\infty^2}{\nu^2}\iint_Q |\psi_{\widetilde{x}}|^2\,d\widetilde{x}\,d\widetilde{t} + \iint_Q \frac{2(\beta''(\widetilde{x}))^2+\nu^{-2}\|z\|^2(\beta'(\widetilde{x}))^2}{(\widetilde{T}-\widetilde{t})^2}|\psi|^2\,d\widetilde{x}\,d\widetilde{t}\\
&\qquad+\int_0^{\widetilde{T}}\left(2\frac{\beta'(0)}{\widetilde{T}-\widetilde{t}}+\widetilde{N}\right)|\psi_{\widetilde{x}}(\widetilde{t},0)|^2\,d\widetilde{t} + 2\widetilde{N}\iint_Q \frac{\beta'(\widetilde{x})}{(\widetilde{T}-\widetilde{t})^2}|\psi|^2\,d\widetilde{x}\,d\widetilde{t}\\
&\qquad+\frac{1}{\widetilde{T}}\int_0^1\left(\frac{\beta'(\widetilde{x})^2}{\widetilde{T}}+\widetilde{N}\beta'(\widetilde{x})\right)|\psi(0,\widetilde{x})|^2\,d\widetilde{x} + \frac{1}{\widetilde{T}^2}\int_0^1 \beta(\widetilde{x})|\psi(0,\widetilde{x})|^2\,d\widetilde{x}.
\end{aligned} \tag{86}$$

Additionally, using (40), the definition of $\widetilde{T} = \nu T^*$ and the fact that $\beta''(\widetilde{x}) \leqslant -2/((1-\delta)3)$, we can absorb the first term in the right-hand side of (86) with

$$-2\iint_Q \frac{\beta''(\widetilde{x})}{\widetilde{T}-\widetilde{t}}|\psi_{\widetilde{x}}|^2\,d\widetilde{x}\,d\widetilde{t}$$

as long as ν is small enough. Furthermore, using $\beta \geq \delta$ and $-\beta'' \geq (2/3)$, the second term in the left-hand side of (86) can be estimated in the following way:

$$-6\delta\iint_Q \beta''(\widetilde{x})\frac{\beta(\widetilde{x})+3(\beta'(\widetilde{x}))^2}{(\widetilde{T}-\widetilde{t})^3}|\psi|^2\,d\widetilde{x}\,d\widetilde{t} \geq \frac{4\delta^2}{\nu T^*}\iint_Q \frac{|\psi|^2}{(\widetilde{T}-\widetilde{t})^2}\,d\widetilde{x}\,d\widetilde{t}.$$

Then, thanks to (40) and taking $\nu \in (0,\nu_0)$, we have that

$$\frac{4\delta^2}{\nu T^*}\iint_Q \frac{|\psi|^2}{(\widetilde{T}-\widetilde{t})^2}\,d\widetilde{x}\,d\widetilde{t} \geqslant \iint_Q \frac{2(\beta''(\widetilde{x}))^2+\nu^{-2}\|z\|^2(\beta'(\widetilde{x}))^2+2\widetilde{N}\beta'(\widetilde{x})}{(\widetilde{T}-\widetilde{t})^2}|\psi|^2\,d\widetilde{x}\,d\widetilde{t}.$$

From (83) and (86), we have

$$-\widetilde{N}^2 \iint_Q \frac{\beta''(\widetilde{x})}{\widetilde{T}-\widetilde{t}}|\psi|^2\,d\widetilde{x}\,d\widetilde{t} \leqslant \int_0^{\widetilde{T}} (2\frac{\beta'(0)}{\widetilde{T}-\widetilde{t}}+\widetilde{N})|\psi_{\widetilde{x}}(\widetilde{t},0)|^2\,d\widetilde{t} + \frac{1}{\widetilde{T}}\int_0^1 \left(\frac{\beta'(\widetilde{x})^2}{\widetilde{T}}+\widetilde{N}\beta'(\widetilde{x})\right)|\psi(0,\widetilde{x})|^2\,d\widetilde{x} + \frac{1}{\widetilde{T}^2}\int_0^1 \beta(\widetilde{x})|\psi(0,\widetilde{x})|^2\,d\widetilde{x}. \tag{87}$$

Let us recall that $\psi := e^{-\alpha}\widetilde{\varphi}$. Then, from (73), (74), (83), and (87), we deduce that

$$\widetilde{N}\iint_Q \frac{1}{\widetilde{T}-\widetilde{t}}e^{-2\alpha}|\widetilde{\varphi}|^2\,d\widetilde{x}\,d\widetilde{t} \leqslant C\left(\int_0^{\widetilde{T}}|\widetilde{\varphi}_{\widetilde{x}}(\widetilde{t},0)|^2\,d\widetilde{t}+\frac{1}{\widetilde{T}}\int_0^1|\widetilde{\varphi}(0,\widetilde{x})|^2\,d\widetilde{x}\right). \tag{88}$$

In (88) and what follows, C will stand for generic positive constants independent of ν, N, T^*, and φ^0.

By (74) and (83), $e^{-2\alpha}$ reaches its minimum in the region $[2\widetilde{T}/3,(2+3\gamma)\widetilde{T}/3]\times[0,1]$ at $(\widetilde{t},\widetilde{x})=(2\widetilde{T}/3,1)$ (recall that $0<\gamma<(1/3)$ was introduced right before (53)). Hence

$$\begin{aligned}&\frac{\widetilde{N}}{\widetilde{T}}e^{-2\alpha(2\widetilde{T}/3,1)}\int_0^1\int_{2\widetilde{T}/3}^{(2+3\gamma)\widetilde{T}/3}|\widetilde{\varphi}|^2\,d\widetilde{t}\,d\widetilde{x}\\ &\leqslant C\left(\int_0^{\widetilde{T}}|\widetilde{\varphi}_{\widetilde{x}}(\widetilde{t},0)|^2\,d\widetilde{t}+\frac{1}{\widetilde{T}}\int_0^1|\widetilde{\varphi}(0,\widetilde{x})|^2\,d\widetilde{x}\right).\end{aligned} \tag{89}$$

From (74), we deduce, with (84) that

$$\exp\{-2\alpha(2\widetilde{T}/3,1)\}=\exp\{-6\beta(1)/\widetilde{T}\}>\exp\{-6\chi/\widetilde{T}\}. \tag{90}$$

Coming back to our original variables (see (69), (71), and (72)), we get from (89) and (90) the desired inequality (53).

Proof of Lemma 4. First, we introduce the unique maximal solution $\overline{\beta}$ of

$$\overline{\beta}''=-\frac{2\overline{\beta}'^2+\overline{\beta}}{3\overline{\beta}'^2+\overline{\beta}}, \tag{91}$$

with initial conditions

$$\overline{\beta}(0)=0 \text{ and } \overline{\beta}'(0)=1. \tag{92}$$

Clearly, for $a\in\mathbb{R}$ and $b\geq 0$, we have

$$\frac{2}{3}\leq\frac{2a^2+b}{3a^2+b}\leq 1. \tag{93}$$

Hence a solution of (91) can be locally extended as long as, for instance, $\overline{\beta}(x)\geq 0$. It straightforwardly follows that $\overline{\beta}$ is well defined on $[0,1]$ and, moreover, satisfies

$$\overline{\beta}'>0 \text{ on } [0,1), \text{ and hence } \overline{\beta}>0 \text{ on } (0,1].$$

Now it follows that

$$\overline{\beta}'<1-\frac{2x}{3} \text{ on } (0,1],$$

which yields

$$\overline{\beta}(1) < \frac{2}{3}.$$

Now, if we consider, instead of $\overline{\beta}$, the solution β_δ of (81)–(82) for $\lambda = 1$ and δ small enough, it follows easily (for instance, using Gronwall's lemma) that

$$\beta_\delta \longrightarrow \overline{\beta} \text{ uniformly on } [0,1], \text{ as } \delta \to 0^+.$$

Hence for δ small enough, $\beta := \beta_\delta$ satisfies (85).

REFERENCES

[1] F. Ancona and A. Marson, *On the attainable set for scalar nonlinear conservation laws with boundary control*, SIAM J. Control Optim., 36 (1998), pp. 290–312.

[2] J.-P. Aubin, *L'analyse non linéaire et ses motivations économiques*, Masson, Paris, 1984.

[3] J.-M. Coron, *On the controllability of the* 2*-D incompressible Navier-Stokes equations with the Navier slip boundary conditions*, ESAIM Contrôle Optim. Calc. Var., 1 (1995/96), pp. 35–75.

[4] J.-M. Coron, *Some open problems on the control of nonlinear partial differential equations*, in Perspectives in Nonlinear Partial Differential Equations: In Honor of Haïm Brezis, H. Berestycki, M. Bertsch, B. Peletier, and L. Véron, eds., Contemp. Math., AMS, Providence, RI, to appear.

[5] J.-M. Coron and S. Guerrero, *Singular optimal control: A linear* 1*-D parabolic-hyperbolic example*, Asympot. Anal., 44 (2005), pp. 237–257.

[6] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. 325, Springer-Verlag, Berlin, 2000.

[7] R. Danchin, *Poches de tourbillon visqueuses*, J. Math. Pures Appl., 76 (1997), pp. 609–647.

[8] A. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes 34, Seoul National University, Seoul, Korea, 1996.

[9] S. Guerrero and O. Yu. Imanuvilov, *Remarks on global controllability for the Burgers equation with two control forces*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[10] S. Guerrero and G. Lebeau, *Singular Optimal Control for a Transport-Diffusion Equation*, preprint, 2006.

[11] E. Hopf, *The partial differential equation* $u_t + uu_x = \mu u_{xx}$, Comm. Pure Appl. Math., 3 (1950), pp. 201–230.

[12] T. Horsin, *On the controllability of the Burgers equation*, ESAIM: Control Opt. Calc. Var., 3 (1998), pp. 83–95.

[13] S. N. Kružkov, *First order quasilinear equations with several independent variables*, Mat. Sb. (N.S.), 81 (1970), pp. 228–255 (in Russian); Math. USSR Sb., 10 (1970), pp. 217–243 (in English).

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1239–1276

A CURSE-OF-DIMENSIONALITY-FREE NUMERICAL METHOD FOR SOLUTION OF CERTAIN HJB PDES*

WILLIAM M. McENEANEY†

Abstract. In previous works of the author and others, max-plus methods have been explored for the solution of first-order, nonlinear Hamilton–Jacobi–Bellman partial differential equations (HJB PDEs) and corresponding nonlinear control problems. These methods exploit the max-plus linearity of the associated semigroups. In particular, although the problems are nonlinear, the semigroups are linear in the max-plus sense. These methods have been used successfully to compute solutions. Although they provide certain computational-speed advantages, they still generally suffer from the curse of dimensionality. Here we consider HJB PDEs in which the Hamiltonian takes the form of a (pointwise) maximum of linear/quadratic forms. The approach to the solution will be rather general, but in order to ground the work, we consider only constituent Hamiltonians corresponding to long-run average-cost-per-unit-time optimal control problems for the development. We obtain a numerical method not subject to the curse of dimensionality. The method is based on construction of the dual-space semigroup corresponding to the HJB PDE. This dual-space semigroup is constructed from the dual-space semigroups corresponding to the constituent linear/quadratic Hamiltonians. The dual-space semigroup is particularly useful due to its form as a max-plus integral operator with a kernel obtained from the originating semigroup. One considers repeated application of the dual-space semigroup to obtain the solution.

Key words. partial differential equations, curse of dimensionality, dynamic programming, max-plus algebra, Legendre transform, Fenchel transform, semiconvexity, Hamilton–Jacobi–Bellman equations, idempotent analysis

AMS subject classifications. 49LXX, 93C10, 35B37, 35F20, 65N99, 47D99

DOI. 10.1137/040610830

1. Introduction.

One approach to nonlinear control is through dynamic programming (DP). With DP, the solution of the control problem "reduces" to the solution of the corresponding partial differential equation (PDE). In the case of deterministic optimal control or deterministic games, where one player's feedback is prespecified, the PDE is a Hamilton–Jacobi–Bellman (HJB) PDE. If one can solve the HJB PDE, then this approach is ideal in that one obtains the optimal control for the given criterion as opposed to a control meeting only some weaker goal such as stability. The problem is that one must solve the HJB PDE! We should remark that such HJB PDEs also arise in robust/H_∞ nonlinear filtering and robust/H_∞ control under partial information.

Various approaches have been taken for solution of the HJB PDE. First, note that it is a fully nonlinear, first-order PDE. Consequently, the solutions are generally nonsmooth (with the exception of the linear/quadratic case, of course), and one must use the theory of viscosity solutions [3], [10], [11], [12], [20]. One approach to the solution is through generalized characteristics (cf. [36], [37], as well as [15], [23] for classical treatments). This approach can obtain the solution very quickly at a single point *if* the solution is smooth. However, the nonsmoothness introduces tremendous difficulties, which appear, to the author, to be difficult to handle in an automated

*Received by the editors June 30, 2004; accepted for publication (in revised form) February 12, 2007; published electronically September 12, 2007. This research was partially supported by NSF grant DMS-0307229 and AFOSR grant FA9550-06-1-0238.

http://www.siam.org/journals/sicon/46-4/61083.html

†Department of Mechanical and Aerospace Engineering, and Department of Mathematics, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0411 (wmceneaney@ucsd.edu).

approach. In particular, the projections of the characteristics into the state space may cross and/or may not cover the entire state space (in analogy with shocks and rarefaction waves).

The most common methods by far fall into the class of grid-based methods (cf. [3], [4], [14], [20], [25] among many others). These require that one generate a grid over some bounded region of the state space. In this general class of methods, we include finite-difference methods, finite element methods, and those DP-based methods which map the continuum problem onto some discrete space. Although higher-order grid-based methods are being explored (cf. [6], [41], [16]), there are still hard lower limits to the computational growth as a function of the space dimension. In particular, suppose the region over which one constructs the grid is rectangular, say square, for simplicity. Further, suppose one uses 100 grid points per dimension. (Clearly 50 would be the minimum acceptable, and 100 could be a bit sparse.) If the state dimension is n, then one has 100^n grid points. Thus the computations grow exponentially in state-space dimension n. If the computations per grid point grew with a state-space dimension such as 2^n, then the computations would grow at a rate of $(200C)^n$ for some constant, C. For concreteness, we discuss only the steady-state PDE case here. If the state-space dimension is 3, it is feasible to solve these problems on current generation machinery. However, the computations will grow by more than 8×10^6 when going from a dimension 3 problem to a dimension 6 problem. Parallel algorithms can alleviate this problem to some extent (cf. [5]). However, there can be only rather limited improvement in the dimension of problems which can be handled by such techniques.

In recent years, an entirely new class of numerical methods for HJB PDEs has emerged [19], [34], [1], [22], [32], [31], [33], [30], [28]. These methods exploit the max-plus (or min-plus [9], [32]) linearity of the associated semigroup. They employ a max-plus basis function expansion of the solution, and the numerical methods obtain the coefficients in the basis expansion. We will refer to these methods as *max-plus basis methods.* Much of the previous work has concentrated on the (harder) steady-state HJB PDE class, where (for both max-plus basis and grid-based methods), one propagates forward in "time" to obtain the steady-state limit solution. With the max-plus basis methods, the number of basis functions required still typically grows exponentially with space dimension. For instance, one might use 25 basis functions per space dimension. Consequently, one still has the curse of dimensionality. With the max-plus basis methods, the "time-step" tends to be much larger than what can be used in grid-based methods (since it encapsulates the action of the semigroup propagation on each basis function), and so these methods can be quite fast on small problems. Even with a max-plus basis approach, the curse of dimensionality growth is so fast that one cannot expect to solve general problems of more than, say, dimension 5, on current machinery, and again, the computing machinery speed increases that are expected in the foreseeable future cannot do much to raise this.

Many researchers have noticed that the introduction of even a single, simple nonlinearity into an otherwise linear control problem of high dimensionality, say n, has disastrous computational repercussions. Specifically, one goes from solution of an n-dimensional Riccati equation to solution of a grid-based or max-plus basis method over a space of dimension n. While the Riccati equation may be "relatively" easily solved for large n, the max-plus and grid-based methods have no hope of obtaining solutions on general problems of dimension, say, $n \geq 6$. This has been a frustrating, counter-intuitive situation for decades.

This paper discusses an approach to certain nonlinear HJB PDEs which is not subject to the curse of dimensionality. Although this approach also utilizes the max-plus algebra, the method is largely unrelated to the max-plus basis approaches discussed above. In fact, for this new method, the computational growth in the state-space dimension is on the order of n^3. There is of course no "free lunch," and there is exponential computational growth in a certain measure of complexity of the Hamiltonian. Under this measure, the minimal complexity Hamiltonian is the linear/quadratic Hamiltonian—corresponding to a solution by a Riccati equation. If the Hamiltonian is given as a pointwise maximum or minimum of M linear/quadratic Hamiltonians, then one could say the *complexity* of the Hamiltonian is M. One could also apply this approach to a wider class of HJB PDEs with semiconvex Hamiltonians (by approximation of the Hamiltonian by a finite number of quadratic forms), but that is certainly beyond the scope of this paper.

The approach has been applied on some simple nonlinear problems. A steady-state HJB PDE comprised of 2 linear/quadratic components was solved in dimensions 2 and 3 in under 10 seconds on a standard PC, and in 20 seconds over $\mathbf{R}^4$. A few simple examples comprised of 3 linear/quadratic components were solved in 10–20 seconds over $\mathbf{R}^3$ and 10–45 seconds over $\mathbf{R}^4$. For these particular problems, the solution was obtained over the entire space (as opposed to a rectangular region) with the resulting errors in the *gradients* growing linearly in $|x|$. (See section 7 for more information on specific examples.) These speeds are of course unprecedented in standard general approaches to nonlinear PDEs. This code was not optimized, and there are many computational cost reduction methods that one could employ to further reduce computational growth. Further, the computational growth in going from $n = 4$ up to, say, $n = 6$ would be on the order of $6^3/4^3 \simeq 4$ as opposed to, say, more than 10^4 for a grid-based method.

We will be concerned here with HJB PDEs of the form $0 = \widetilde{H}(x, \operatorname{grad} V)$, where the Hamiltonians are given or approximated as

$$\widetilde{H}(x, \operatorname{grad} V) = \max_{m \in \{1,2,\ldots,M\}} \{H^m(x, \operatorname{grad} V)\}.$$

In order to make the problem tractable, we will concentrate on a single class of HJB PDEs: those for long-run average-cost-per-unit-time problems. However, the theory obviously can be expanded to a much larger class.

Since the development of the proposed method in the following sections takes quite a few pages, we briefly outline the main points here. First, recall that the solution of the above PDE is the eigenfunction of the corresponding semigroup, that is,

$$0 \otimes V = V = \widetilde{S}_\tau[V],$$

where $\oplus, \otimes$ denote max-plus addition and multiplication, and we note that $\widetilde{S}_\tau$ is max-plus linear (cf. [19], [27], [32]). The Legendre–Fenchel transform maps this into the dual-space eigenfunction problem

$$0 \otimes e = \widetilde{\mathcal{B}}_\tau \odot e,$$

where we use the $\odot$ notation to indicate $\widetilde{\mathcal{B}}_\tau \odot e \doteq \int^{\oplus}_{\mathbb{R}^n} \widetilde{\mathcal{B}}_\tau(x, y) \otimes e(y)\, dy$, where $\int^{\oplus}$ denotes max-plus integration (maximization). Then one approximates $\widetilde{\mathcal{B}}_\tau \simeq \bigoplus_{m \in \mathcal{M}} \mathcal{B}^m_\tau$, where $\mathcal{M} \doteq \{1, 2, \ldots, M\}$ and the $\mathcal{B}^m_\tau$ correspond to the H^m. The

max-plus power method [13], [24], [32] suggests that the solution is approximated by the form

$$e \simeq \lim_{N\to\infty} \left[\bigoplus_{m\in\mathcal{M}} \mathcal{B}_\tau^m\right]^N \odot 0 = \lim_{N\to\infty} \left[\bigoplus_{\{m_i\}_{i=1}^N} \mathcal{B}_\tau^{m_1} \otimes \mathcal{B}_\tau^{m_2} \otimes \cdots \mathcal{B}_\tau^{m_N}\right] \odot 0,$$

where the N superscript denotes the $\odot$ operation N times, and 0 represents the zero-function. Given linear/quadratic forms for each of the H^m, the $\mathcal{B}_\tau^m$ are obtained by Riccati equations. Let $e_N \doteq [\bigoplus_{m\in\mathcal{M}} \mathcal{B}_\tau^m]^N \odot 0$. Then $e_N \to e$. The convergence rate does not depend on the space dimension, but on the dynamics of the problem. There is no curse of dimensionality. The exponential growth is in $M = \#\mathcal{M}$. Given the solution of the Riccati equations for the H^m, the computation of each product, $\mathcal{B}_\tau^{m_1} \otimes \mathcal{B}_\tau^{m_2} \otimes \cdots \mathcal{B}_\tau^{m_N}$, is analytical, modulo $n \times n$ matrix inversions (and hence the n^3 computational growth rate).

In section 2, the class of control problems and HJB PDEs which we will use to demonstrate the theory will be given. We will also review the existing theory relevant to our problem there. In section 3 the relation between solution of the HJB PDEs and their corresponding semiconvex dual problems will be discussed. In section 4, a discrete-time approximation of the semigroup for the problem of interest will be introduced, and convergence of the solutions of the approximate problems to the original problem will be obtained. The algorithm itself will be developed in section 5. The basic algorithm is not subject to the curse of dimensionality. However, practical implementation requires some additional work; some initial remarks on this appear in section 6. The algorithm is applied to some simple examples in section 7. Finally, section 8 sketches some future directions.

2. Sample problem class and review of theory. There are certain conditions which must be satisfied for solutions to exist and the method to apply. In order that the assumptions are not completely abstract, we will work with a specific problem class: the infinite time-horizon H_∞ problem with fixed feedback. This class consists of long-term average-cost-per-unit-time problems. Moreover, it is a problem class in which there already exists a good deal of results, and so less analysis will be required for application of the new method.

As indicated above, we suppose the individual H^m are linear/quadratic Hamiltonians. Consequently, consider a finite set of linear systems

$$\dot{\xi}^m = A^m \xi^m + \sigma^m w, \quad \xi_0^m = x \in \mathbb{R}^n. \tag{1}$$

Let $w \in \mathcal{W} \doteq L_2^{loc}([0,\infty);\mathbb{R}^m)$, where we recall that $L_2^{loc}([0,\infty);\mathbb{R}^m) = \{w : [0,\infty) \to \mathbb{R}^m : \int_0^T |w_t|^2\,dt < \infty \text{ for all } T < \infty\}$. Let the cost functionals be

$$J^m(x,T;w) \doteq \int_0^T \frac{1}{2}\xi_t^m D^m \xi_t^m - \frac{\gamma^2}{2}|w_t|^2\,dt, \tag{2}$$

and let the value function (also known as the available storage in this context) be

$$V^m(x) = \sup_{w\in\mathcal{W}} \sup_{T<\infty} J^m(x,T;w) = \lim_{T\to\infty} \sup_{w\in\mathcal{W}} J^m(x,T;w). \tag{3}$$

We remark that a generalization of the second term in the integrand of the cost functional to $\frac{1}{2}w^T C^m w$ with C^m symmetric and positive definite is not needed since

this is equivalent to a change in σ^m in the dynamics (1). Obviously J^m and V^m require some assumptions in order to guarantee their existence. The assumptions will hold throughout the paper. Since these assumptions only appear together, we will refer to this entire set of assumptions as assumption block $(A.m)$, and these are as follows:

$(A.m)$ Assume that there exists $c_A \in (0,\infty)$ such that

$$x^T A^m x \le -c_A |x|^2 \qquad \forall\, x \in \mathbb{R}^n,\ m \in \mathcal{M}.$$

Assume that all D^m are positive definite and symmetric, and let c_D be such that

$$x^T D^m x \le c_D |x|^2 \qquad \forall\, x \in \mathbb{R}^n,\ m \in \mathcal{M}$$

(which is obviously equivalent to all eigenvalues of the D^m being no greater than c_D). Lastly, assume that $\gamma^2/c_\sigma^2 > c_D/c_A^2$, where $c_\sigma \ge \sigma^m$.

Note that these assumptions guarantee the existence of the V^m as locally bounded functions which are zero at the origin (cf. [35]). (These assumptions could be weakened by using the specific linear/quadratic structure, but that would distract from the goal of this paper.) The corresponding HJB PDEs are

$$\begin{aligned} (4)\qquad 0 &= -H^m(x, \operatorname{grad} V) \\ &= -\left\{ \frac{1}{2} x^T D^m x + (A^m x)^T \operatorname{grad} V + \max_{w \in \mathbb{R}^m} \left[(\sigma^m w)^T \operatorname{grad} V - \frac{\gamma^2}{2} |w|^2 \right] \right\} \\ &= -\left\{ \frac{1}{2} x^T D^m x + (A^m x)^T \operatorname{grad} V + \frac{1}{2} \operatorname{grad} V^T \Sigma^m \operatorname{grad} V \right\} \\ V(0) &= 0, \end{aligned}$$

where $\Sigma^m \doteq \frac{1}{\gamma^2} \sigma^m (\sigma^m)^T$. Let $\mathbb{R}^- \doteq \mathbb{R} \cup \{-\infty\}$. Recall that a function, $\phi : \mathbb{R}^n \to \mathbb{R}^-$ is semiconvex if given any $R \in (0,\infty)$ there exists $k_R \in \mathbb{R}$ such that $\phi(x) + \frac{k_R}{2}|x|^2$ is convex over $\overline{B}_R(0) = \{x \in \mathbb{R}^n : |x| \le R\}$. For a fixed choice of c_A, c_σ, $\gamma > 0$ satisfying the above assumptions, and for any $\delta \in (0, \gamma)$, we define

$$\mathcal{G}_\delta = \left\{ V : \mathbb{R}^n \to [0,\infty) \,\middle|\, V \text{ is semiconvex and } V(x) \le \frac{c_A(\gamma-\delta)^2}{c_\sigma^2} |x|^2 \ \forall x \in \mathbb{R}^n \right\}.$$

From [35] (undoubtedly among many others), each value function (3) is the unique viscosity solution of its corresponding HJB PDE (4) in the class $\mathcal{G}_\delta$ for sufficiently small $\delta > 0$.

From the structure of the running cost and dynamics, it is easy to see (cf. [42], [35]) that each V^m satisfies

$$(5)\quad V^m(x) = \sup_{T<\infty} \sup_{w \in \mathcal{W}} J^m(x,T;w) = \lim_{T\to\infty} \sup_{w\in\mathcal{W}} J^m(x,T;w) \doteq \lim_{T\to\infty} V^{m,f}(x,T),$$

and that each $V^{m,f}$ is the unique continuous viscosity solution of (cf. [3], [20])

$$(6)\qquad 0 = V_T - H^m(x, \operatorname{grad} V), \quad V(0,x) = 0.$$

It is easy to see that these solutions have the form $V^{m,f}(x,t) = \frac{1}{2} x^T P_t^{m,f} x$, where each $P^{m,f}$ satisfies the differential Riccati equation

$$(7)\qquad \dot{P}^{m,f} = (A^m)^T P^{m,f} + P^{m,f} A^m + D^m + P^{m,f} \Sigma^m P^{m,f}, \quad P_0^{m,f} = 0.$$

By (5) and (7), the V^m take the form $V(x) = \frac{1}{2}x^T P^m x$, where $P^m = \lim_{t\to\infty} P_t^{m,f}$. With this form and (4) (or (7)), we see that the P^m satisfy the algebraic Riccati equations

$$0 = (A^m)^T P^m + P^m A^m + D^m + P^m \Sigma^m P^m. \tag{8}$$

Combining this with the above, one has the following.

THEOREM 2.1. *Each value function* (3) *is the unique classical solution of its corresponding HJB PDE* (4) *in the class* $\mathcal{G}_\delta$ *for sufficiently small* $\delta > 0$. *Further,* $V^m(x) = \frac{1}{2}x^T P^m x$, *where* P^m *is the smallest symmetric, positive definite solution of* (8).

The duality between viscosity (and/or classical) solutions of the HJB PDEs and the corresponding value functions is certainly very important. However, the method we will use to obtain these value functions/HJB PDE solutions will be through the associated semigroups. These semigroups are equivalent to dynamic programming principles (DPPs). Consequently, for each m we define the semigroup

$$S_T^m[\phi] \doteq \sup_{w\in\mathcal{W}} \left[\int_0^T \frac{1}{2}(\xi_t^m)^T D^m \xi_t^m - \frac{\gamma^2}{2}|w_t|^2\,dt + \phi(\xi_T^m)\right], \tag{9}$$

where ξ^m satisfies (1). By [35], the domain of S_T^m includes $\mathcal{G}_\delta$ for all $\delta > 0$. The following result is similar to that in [32]; the only significant difference is that, in this case, $V^m(x) = \frac{1}{2}x^T P^m x$ is smooth.

THEOREM 2.2. *Fix any* $T > 0$. *Each value function,* V^m, *is the unique smooth solution of* $V = S_T^m[V]$ *in the class* $\mathcal{G}_\delta$ *for sufficiently small* $\delta > 0$. *Further, given any* $V \in \mathcal{G}_\delta$, $\lim_{T\to\infty} S_T^m[V](x) = V^m(x)$ *for all* $x \in \mathbb{R}^n$ *(uniformly on compact sets).*

Recall that the HJB PDE problem of interest is

$$0 = -\widetilde{H}(x, \operatorname{grad} V) \doteq -\max_{m\in\mathcal{M}} H^m(x, \operatorname{grad} V), \quad V(0) = 0. \tag{10}$$

The corresponding value function is

$$\widetilde{V}(x) = \sup_{w\in\mathcal{W}} \sup_{\mu\in\mathcal{D}_\infty} \widetilde{J}(x, w, \mu) \doteq \sup_{w\in\mathcal{W}} \sup_{\mu\in\mathcal{D}_\infty} \sup_{T<\infty} \int_0^T l^{\mu_t}(\xi_t) - \frac{\gamma^2}{2}|w_t|^2\,dt, \tag{11}$$

where $l^{\mu_t}(x) = \frac{1}{2}x^T D^{\mu_t} x$, $\mathcal{D}_\infty = \{\mu : [0,\infty) \to \mathcal{M} :\ \text{measurable}\}$, and ξ satisfies

$$\dot{\xi} = A^{\mu_t}\xi + \sigma^{\mu_t} w_t, \quad \xi_0 = x. \tag{12}$$

THEOREM 2.3. *Value function* $\widetilde{V}$ *is the unique viscosity solution of* (10) *in the class* $\mathcal{G}_\delta$ *for sufficiently small* $\delta > 0$.

Remark 2.4. The proof of Theorem 2.3 is nearly identical to the proofs of Theorems 2.5 and 2.6 from [35], with only trivial changes, and so is not included. In particular, rather than choosing any $w \in \mathcal{W}$, one chooses both any $w \in \mathcal{W}$ and any $\mu \in \mathcal{D}_\infty$.

Define the semigroup

$$\widetilde{S}_T[\phi] = \sup_{w\in\mathcal{W}} \sup_{\mu\in\mathcal{D}_T} \left[\int_0^T l^{\mu_t}(\xi_t) - \frac{\gamma^2}{2}|w_t|^2\,dt + \phi(\xi_T)\right], \tag{13}$$

where $\mathcal{D}_T = \{\mu : [0,T) \to \mathcal{M} :\ \text{measurable}\}$. In analogy with Theorem 2.2, one has the following.

THEOREM 2.5. *Fix any $T > 0$. Value function $\widetilde{V}$ is the unique continuous solution of $V = \widetilde{S}_T[V]$ in the class $\mathcal{G}_\delta$ for sufficiently small $\delta > 0$. Further, given any $V \in \mathcal{G}_\delta$, $\lim_{T\to\infty} \widetilde{S}_T[V](x) = \widetilde{V}(x)$ for all $x \in \mathbb{R}^n$ (uniformly on compact sets).*

The proof is nearly identical to the proof of a similar result in [32] and so is not included. In particular, the only change is the addition of the supremum over $\mathcal{D}_T$—which makes no substantial change in the proof. More important, we also have the following.

THEOREM 2.6. *There exists $c_V > 0$ such that $\widetilde{V}(x) - \frac{1}{2}c_V|x|^2$ is strictly convex.*

Proof. Fix any $x, \nu \in \mathbb{R}^n$ with $|\nu| = 1$ and any $\delta > 0$. Let $\varepsilon > 0$. Given x, let $w^\varepsilon \in \mathcal{W}$, $\mu^\varepsilon \in \mathcal{D}_\infty$ be ε-optimal for $\widetilde{V}(x)$. Then

$$\begin{aligned}&\widetilde{V}(x-\delta\nu) - 2\widetilde{V}(x) + \widetilde{V}(x+\delta\nu)\\ &\quad\geq \widetilde{J}(x-\delta\nu, w^\varepsilon, \mu^\varepsilon) - 2\widetilde{J}(x, w^\varepsilon, \mu^\varepsilon) + \widetilde{J}(x+\delta\nu, w^\varepsilon, \mu^\varepsilon) - 2\varepsilon.\end{aligned} \tag{14}$$

Let $\xi^\delta, \xi^0, \xi^{-\delta}$ be solutions of dynamics (12), but with initial conditions $\xi_0^\delta = x + \delta\nu$, $\xi_0^0 = x$, and $\xi_0^{-\delta} = x - \delta\nu$, respectively, where the inputs are w^ε and μ^ε for all three processes. Then

$$\dot{\xi}^\delta - \dot{\xi}^0 = A^{\mu_t^\varepsilon}[\xi^\delta - \xi^0] \;\text{ and }\; \dot{\xi}^0 - \dot{\xi}^{-\delta} = A^{\mu_t^\varepsilon}[\xi^0 - \xi^{-\delta}]. \tag{15}$$

Letting $\Delta_t^+ \doteq \xi_t^\delta - \xi_t^0$, one also has $\xi_t^0 - \xi_t^{-\delta} = \Delta_t^+$, and by linearity one finds $\dot{\Delta}^+ = A^{\mu_t^\varepsilon}\Delta^+$. Also, using (14) and (11), we have

$$\widetilde{V}(x-\delta\nu) - 2\widetilde{V}(x) + \widetilde{V}(x+\delta\nu) \geq \int_0^\infty (\Delta^+)^T D^{\mu_t^\varepsilon}\Delta^+\,dt - 2\varepsilon. \tag{16}$$

Also, by the finiteness of $\mathcal{M}$, there exists $K < \infty$ such that

$$\frac{d}{dt}|\Delta^+|^2 = 2(\Delta^+)^T A^{\mu_t^\varepsilon}\Delta^+ \geq -K|\Delta^+|^2,$$

which implies

$$|\Delta^+|^2 \geq e^{-Kt}\delta^2 \qquad \forall\, t \geq 0. \tag{17}$$

Let $\lambda_D \doteq \min\{\lambda \in \mathbb{R} :\ \lambda \text{ is an eigenvalue of a } D^m\}$. By the positive definiteness of the D^m and finiteness of $\mathcal{M}$, $\lambda_D > 0$. Consequently, by (16), and then (17),

$$\widetilde{V}(x-\delta\nu) - 2\widetilde{V}(x) + \widetilde{V}(x+\delta\nu) \geq \int_0^\infty \lambda_D|\Delta^+|^2\,dt - 2\varepsilon \geq \frac{\lambda_D}{K}\delta^2 - 2\varepsilon. \tag{18}$$

Since $\varepsilon > 0$ and $|\nu| = 1$ were arbitrary, one obtains the result. ∎

3. Max-plus spaces and dual operators. Again, recall that a function $\phi : \mathbb{R}^n \to \mathbb{R}^-$ is semiconvex if, given any $R \in (0, \infty)$, there exists $\beta_R \in \mathbb{R}$ such that $\phi(x) + \frac{\beta_R}{2}|x|^2$ is convex over $\overline{B}_R(0) = \{x \in \mathbb{R}^n : |x| \leq R\}$. We will modify this definition by allowing the β_R to be $n \times n$, symmetric, and positive or negative definite matrices. We will denote the set of such matrices as $\mathcal{D}_n$. We say ϕ is uniformly semiconvex with (symmetric, definite matrix) constant $\beta \in \mathcal{D}_n$ if $\phi(x) + \frac{1}{2}x^T\beta x$ is convex over $\mathbb{R}^n$. Let $\mathcal{S}_\beta = \mathcal{S}_\beta(\mathbb{R}^n)$ be the set of functions mapping $\mathbb{R}^n$ into $\mathbb{R}^-$ which are uniformly semiconvex with (symmetric, definite matrix) constant β. (A negative definite semiconvexity constant corresponds to functions which are still convex after subtracting a convex quadratic.) Also note that $\mathcal{S}_\beta$ is a max-plus vector space (also

known as a moduloid) [19], [32], [2], [8], [26]. For instance, $\alpha_1 \otimes \phi_1 \oplus \alpha_2 \otimes \phi_2 \in \mathcal{S}_\beta$ for all $\alpha_1, \alpha_2 \in \mathbb{R}^-$ and all $\phi_1, \phi_2 \in \mathcal{S}_\beta$. Combining Theorems 2.1 and 2.6, we have the following.

THEOREM 3.1. *There exists $\overline{\beta} \in \mathcal{D}_n$ such that given any β such that $\beta - \overline{\beta} > 0$ (i.e., $\beta - \overline{\beta}$ positive definite), $\widetilde{V} \in \mathcal{S}_\beta$ and $V^m \in \mathcal{S}_\beta$ for all $m \in \mathcal{M}$. Further, one may take β negative definite (i.e., $\widetilde{V}, V^m$ are convex).*

We henceforth assume we have chosen β such that $\beta - \overline{\beta} > 0$.

Throughout the remainder, we will employ certain transform kernel functions, $\psi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, which take the form

$$\psi(x, z) = \tfrac{1}{2}(x - z)^T C(x - z)$$

with nonsingular, symmetric C satisfying $C + \beta < 0$ (i.e., $C + \beta$ negative definite). The following semiconvex duality result [19], [31], [32] requires only a small modification of convex duality and Legendre–Fenchel transform results [38], [39].

THEOREM 3.2. *Let $\phi \in \mathcal{S}_\beta$. Let C and ψ be as above. Then, for all $x \in \mathbb{R}^n$,*

$$\phi(x) = \max_{z \in \mathbb{R}^n} [\psi(x, z) + a(z)] \tag{19}$$

$$= \int_{\mathbb{R}^n}^{\oplus} \psi(x, z) \otimes a(z)\, dz = \psi(x, \cdot) \odot a(\cdot), \tag{20}$$

where for all $z \in \mathbb{R}^n$,

$$a(z) = -\max_{x \in \mathbb{R}^n} [\psi(x, z) - \phi(x)] \tag{21}$$

$$= -\int_{\mathbb{R}^n}^{\oplus} \psi(x, z) \otimes [-\phi(x)]\, dx = -\left\{\psi(\cdot, z) \odot [-\phi(\cdot)]\right\}, \tag{22}$$

which, using the notation of [8],

$$= \left\{\psi(\cdot, z) \odot [\phi^-(\cdot)]\right\}^-. \tag{23}$$

We will refer to a as the *semiconvex dual* of ϕ (with respect to ψ).

Remark 3.3. We note that $\phi \in \mathcal{S}_\beta$ implies that ϕ is locally Lipschitz (cf. [18]). We also note that if $\phi \in \mathcal{S}_\beta$ and if there is any $x \in \mathbb{R}^n$ such that $\phi(x) = -\infty$, then $\phi \equiv -\infty$. Henceforth, we will ignore the special case of $\phi \equiv -\infty$ and assume that all functions are real-valued.

Semiconcavity is the obvious analogue of semiconvexity. In particular, a function, $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, is uniformly semiconcave with constant $\beta \in \mathcal{D}_n$ if $\phi(x) - \frac{1}{2}x^T \beta x$ is concave over $\mathbb{R}^n$. Let $\mathcal{S}_\beta^-$ be the set of functions mapping $\mathbb{R}^n$ into $\mathbb{R} \cup \{+\infty\}$ which are uniformly semiconcave with constant β.

LEMMA 3.4. *Let $\phi \in \mathcal{S}_\beta$ (still with $C + \beta < 0$), and let a be the semiconvex dual of ϕ. Then $a \in \mathcal{S}_d^-$ for some $d \in \mathcal{D}_n$ such that $C + d < 0$.*

Proof. A proof only in the case $\phi \in C^2$ is provided; in the more general case, a mollification argument can be employed.

Noting that $\phi \in \mathcal{S}_\beta$ and $-C - \beta > 0$, there exists a unique minimizer,

$$\overline{x}(z) = \operatorname*{argmin}_{x \in \mathbb{R}} [\phi(x) - \psi(x, z)],$$

and one has

$$a(z) = \phi(\overline{x}(z)) - \psi(\overline{x}(z), z). \tag{24}$$

Fix any $z, \nu \in \mathbb{R}^n$ with $|\nu| = 1$. Define $a^s : \mathbb{R} \to \mathbb{R}$ and $\overline{x}^s : \mathbb{R} \to \mathbb{R}^n$ by

$$a^s(\delta) \doteq a(z + \delta\nu) \quad \text{and} \quad \overline{x}^s(\delta) = \overline{x}(z + \delta\nu). \tag{25}$$

We will obtain a lower bound on the second derivative of a^s, and this will prove the result. Differentiating a^s, one has

$$\begin{aligned}\left.\frac{da^s}{d\delta}\right|_{\delta=0} &= \frac{d}{d\delta}\left[\phi(\overline{x}(z+\delta\nu)) - \psi(\overline{x}(z+\delta\nu), z+\delta\nu)\right] \\ &= \operatorname{grad}_x \phi(\overline{x}(z)) \cdot \frac{d\overline{x}^s}{d\delta} - \operatorname{grad}_x \psi(\overline{x}(z), z) \cdot \frac{d\overline{x}^s}{d\delta} - \operatorname{grad}_z \psi(\overline{x}(z), z) \cdot \nu,\end{aligned}$$

which, using the fact that $\operatorname{grad}_x \phi(\overline{x}(z)) - \operatorname{grad}_x \psi(\overline{x}(z), z) = 0$,

$$= -\operatorname{grad}_z \psi(\overline{x}(z), z) \cdot \nu.$$

Differentiating again, one finds

$$\left.\frac{d^2a^s}{d\delta^2}\right|_{\delta=0} = -\sum_{i=1}^{n}\left\{\sum_{j=1}^{n}\psi_{z_i x_j}(\overline{x}(z), z)\frac{d\overline{x}_j^s}{d\delta}\nu_i + \sum_{k=1}^{n}\psi_{z_i z_k}(\overline{x}(z), z)\nu_k\nu_i\right\} \tag{26}$$

$$= -\nu^T C\nu + \nu^T C\left.\frac{d\overline{x}^s}{d\delta}\right|_{\delta=0}. \tag{27}$$

Now, differentiating both sides of $\operatorname{grad}_x \phi(\overline{x}(z+\delta\nu)) - \operatorname{grad}_x \psi(\overline{x}(z+\delta\nu), z+\delta\nu) = 0$ yields

$$\sum_{j=1}^{n}\phi_{x_i x_j}\frac{d\overline{x}_j^s}{d\delta} - \sum_{k=1}^{n}\psi_{x_i x_k}\frac{d\overline{x}_k^s}{d\delta} - \sum_{l=1}^{n}\psi_{x_i z_l}\nu_l = 0 \qquad \forall\, i,$$

which yields

$$\frac{d\overline{x}^s}{d\delta} = -\left[\phi_{xx}(\overline{x}(z), z) - C\right]^{-1} C\nu. \tag{28}$$

Substituting (28) into (27), one obtains

$$\frac{d^2a^s}{d\delta^2} = \nu^T\left\{-C + C\left[C - \phi_{xx}(\overline{x}(z), z)\right]^{-1} C\right\}\nu. \tag{29}$$

Now, $\phi \in \mathcal{S}_\beta$ implies $-\phi_{xx}(\overline{x}(z)) - \beta < 0$, which implies that there exists $k_0 > 0$ such that

$$\nu^T\left[-\beta - \phi_{xx}(\overline{x}(z))\right]\nu < -k_0|\nu|^2 \qquad \forall\, \nu.$$

Combining this with the fact that $C + \beta < 0$ implies that

$$\nu^T[C - \phi_{xx}(\overline{x}(z))]\nu < -k_0|\nu|^2 \qquad \forall\, \nu.$$

Now, $C - \phi_{xx}(\overline{x}(z))$ being symmetric, negative definite implies that $C - \phi_{xx}(\overline{x}(z)) = U\Lambda U^T$ for some diagonal Λ (with all diagonal entries negative, of course) and some

real, unitary U. Consequently, $[C-\phi_{xx}(\overline{x}(z))]^{-1} = U\Lambda^{-1}U^T < 0$, which is negative definite. Then, since $\zeta^T[C-\phi_{xx}(\overline{x}(z))]^{-1}\zeta < 0$ for all $\zeta \in \mathbb{R}^n$, $\zeta \neq 0$, one sees that

$$\nu^T C[C-\phi_{xx}(\overline{x}(z))]^{-1}C\nu < 0$$

for all $\nu \in \mathbb{R}^n$, $\nu \neq 0$, and so

$$C[C-\phi_{xx}(\overline{x}(z))]^{-1}C < 0. \tag{30}$$

Let $d \doteq -C+\frac{1}{2}C[C-\phi_{xx}(\overline{x}(z))]^{-1}C$. Then, by (30), $C+d = \frac{1}{2}C[C-\phi_{xx}(\overline{x}(z))]^{-1}C < 0$. (Note that if d is not definite, then by addition of εI for arbitrarily small ε, one can make d definite without violating the inequalities.) Further, by (29) and (30),

$$\left.\frac{d^2a^s}{d\delta^2}\right|_{\delta=0} = \nu^T\left[d+\frac{1}{2}C[C-\phi_{xx}(\overline{x}(z))]^{-1}C\right]\nu < \nu^T d\nu,$$

which yields the result. □

Remark 3.5. Fix any $\delta > 0$ such that $\widetilde{V} \in \mathcal{G}_\delta$, and let $K_\delta = 2\frac{c_A(\gamma-\delta)^2}{c_\sigma^2}$ so that $0 \leq \widetilde{V}(x) \leq \frac{K_\delta}{2}|x|^2$. Then, using Lemma 3.4 and the monotonicity of the dual operations, the semiconvex dual, $\widetilde{a}$, of $\widetilde{V}$ is in $\mathcal{S}_d^- \cap \mathcal{G}_\delta^-$ for some $d \in \mathcal{D}_n$ such that $C+d<0$, where $\mathcal{G}_\delta^-$ is the space of semiconcave functions satisfying

$$0 \leq \widetilde{a}(z) \leq \tfrac{1}{2}z^T Q_\delta^- z,$$

where $Q_\delta^- \doteq C(C-K_\delta I)^{-1}K_\delta(C-K_\delta I)^{-1}C - K_\delta^2(C-K_\delta I)^{-1}C(C-K_\delta I)^{-1}$, and where the last term on the right is the dual of $\frac{K_\delta}{2}|x|^2$. Further, by the monotonicity of the dual operations, any $a \in \mathcal{G}_\delta^-$ has dual $V \in \mathcal{G}_\delta$.

LEMMA 3.6. *Let $\phi \in \mathcal{S}_\beta$ with semiconvex dual a. Suppose $b \in \mathcal{S}_d^-$ with $C+d<0$ is such that $\phi = \psi(x,\cdot)\odot b(\cdot)$. Then $b=a$.*

Proof. Note that $-b \in \mathcal{S}_d$. Therefore, for all $y \in \mathbb{R}^n$, we have $-b(y) = \max_{\zeta\in\mathbb{R}^n}[\psi(y,\zeta)+\alpha(\zeta)]$, or equivalently,

$$b(y) = -\max_{\zeta\in\mathbb{R}^n}[\psi(y,\zeta)+\alpha(\zeta)], \tag{31}$$

where for all $\zeta \in \mathbb{R}^n$,

$$\alpha(\zeta) = -\max_{y\in\mathbb{R}^n}[\psi(y,\zeta)+b(y)],$$

which by assumption

$$= -\phi(\zeta). \tag{32}$$

Combining (31) and (32), and then using (21), one obtains

$$b(y) = -\max_{\zeta\in\mathbb{R}^n}[\psi(y,\zeta)-\phi(\zeta)] = a(y) \qquad \forall\, y \in \mathbb{R}^n. \qquad \square$$

We will hereafter refer to the uniqueness of the semiconvex dual in the sense of Lemma 3.6 simply as the uniqueness of the semiconvex dual. It will be critical to the method that the functions obtained by application of the semigroups to the $\psi(\cdot,z)$ be semiconvex with less concavity than the $\psi(\cdot,z)$ themselves. In other words, we will want, for instance, $\widetilde{S}_\tau[\psi(\cdot,z)] \in \mathcal{S}_{-(c+\varepsilon I)}$ for some $\varepsilon > 0$. This is the subject of the

next theorem. Also, in order to keep the theorem statement clean, we will first state some definitions. Define

$$\lambda_D \doteq \min\{\lambda \in \mathbb{R} : \lambda \text{ is an eigenvalue of } D^m,\ m \in \mathcal{M}\}.$$

Note that the finiteness of $\mathcal{M}$ and positive definiteness of the D^m imply that $\lambda_D > 0$. Let

$$I_C \doteq \left\{ C \in \mathcal{D}_n \;\middle|\; \min_{|\nu|=1} \min_{m\in\mathcal{M}} \nu^T \left[A^{m\,T} C + C A^m \right] \nu \geq -\lambda_D/4 \right\}.$$

THEOREM 3.7. *Let $C \in I_C$. Then there exists $\overline{\tau} > 0$ and $\eta > 0$ such that for all $\tau \in [0, \overline{\tau}]$,*

$$\widetilde{S}_\tau[\psi(\cdot, z)],\ S^m_\tau[\psi(\cdot, z)] \in \mathcal{S}_{-(C+\eta I\tau)}.$$

Remark 3.8. If we restrict our attention to $C = cI$ for some $c \in \mathbb{R}$, then $C \in I_C$ if one takes $|c| \leq \lambda_D/[8 \max_{m\in\mathcal{M}} |A^m|]$, $c \neq 0$, and so the theorem condition can be satisfied.

Proof. We prove the result only for $\widetilde{S}_\tau$. The proof for S^m_τ is nearly identical and slightly simpler.

The first portion of the proof is similar to the proof of Theorem 2.6. Again, fix any $x, \nu \in \mathbb{R}^n$ with $|\nu| = 1$ and any $\delta > 0$. Fix $\tau > 0$ (to be specified below), and let $\varepsilon > 0$. Let $w^\varepsilon, \mu^\varepsilon$ be ε-optimal for $\widetilde{S}_\tau[\psi(\cdot, z)](x)$. Specifically, suppose $\widehat{\mathcal{I}}^\psi(x, \tau, w^\varepsilon, \mu^\varepsilon) \geq \widetilde{S}_\tau[\psi(\cdot, z)](x) - \varepsilon$, where

$$\widehat{\mathcal{I}}^\psi(x, \tau, w, \mu) \doteq \int_0^\tau l^{\mu_t}(\xi_t) - \frac{\gamma^2}{2}|w_t|^2\, dt + \psi(\xi_\tau, z) \tag{33}$$

and ξ_t satisfies (12). For simplicity of notation, let $\widehat{V}^{\tau,\psi} = \widetilde{S}_\tau[\psi(\cdot, z)]$. Then

$$\begin{aligned} &\widehat{V}^{\tau,\psi}(x - \delta\nu) - 2\widehat{V}^{\tau,\psi}(x) + \widehat{V}^{\tau,\psi}(x + \delta\nu) \\ &\quad \geq \widehat{\mathcal{I}}^\psi(x - \delta\nu, \tau, w^\varepsilon, \mu^\varepsilon) - 2\widehat{\mathcal{I}}^\psi(x, \tau, w^\varepsilon, \mu^\varepsilon) + \widehat{\mathcal{I}}^\psi(x + \delta\nu, \tau, w^\varepsilon, \mu^\varepsilon) - 2\varepsilon. \end{aligned} \tag{34}$$

Let $\xi^\delta, \xi^0, \xi^{-\delta}, \Delta^+$ be as given in the proof of Theorem 2.6. Note that

$$\psi(\xi^\delta_\tau, z) - 2\psi(\xi^0_\tau, z) + \psi(\xi^{-\delta}_\tau, z) = (\Delta^+_\tau)^T C \Delta^+_\tau. \tag{35}$$

Note also that as in the proof of Theorem 2.6,

$$\frac{1}{2}\left[\xi^\delta_t D^{\mu^\varepsilon_t} \xi^\delta_t - 2\xi^0_t D^{\mu^\varepsilon_t} \xi^0_t + \xi^{-\delta}_t D^{\mu^\varepsilon_t} \xi^{-\delta}_t \right] = (\Delta^+_t)^T D^{\mu^\varepsilon_t} \Delta^+_t. \tag{36}$$

Combining (33), (34), (35), and (36), one obtains

$$\begin{aligned} &\widehat{V}^{\tau,\psi}(x - \delta\nu) - 2\widehat{V}^{\tau,\psi}(x) + \widehat{V}^{\tau,\psi}(x + \delta\nu) \\ &\quad \geq \int_0^\tau (\Delta^+_t)^T D^{\mu^\varepsilon_t} \Delta^+_t\, dt + (\Delta^+_\tau)^T C \Delta^+_\tau - 2\varepsilon. \end{aligned} \tag{37}$$

Further, noting as before that $\dot{\Delta}^+ = A^{\mu^\varepsilon_t} \Delta^+$, one has

$$\Delta^+_t = \exp\left\{ \int_0^t A^{\mu^\varepsilon_r}\, dr \right\} \delta\nu \doteq \Lambda^\varepsilon_t \delta\nu. \tag{38}$$

Combining (37) and (38), one has

$$\widehat{V}^{\tau,\psi}(x-\delta\nu) - 2\widehat{V}^{\tau,\psi}(x) + \widehat{V}^{\tau,\psi}(x+\delta\nu)$$
$$\geq \delta^2\left\{\int_0^\tau \nu^T(\Lambda_t^\varepsilon)^T D^{\mu_t^\varepsilon}\Lambda_t^\varepsilon\nu\,dt + \nu^T(\Lambda_\tau^\varepsilon)^T C\Lambda_\tau^\varepsilon\nu\right\} - 2\varepsilon.$$

However, since $\lambda_D > 0$ and $\Lambda_0^\varepsilon = I$, there exists $\bar\tau > 0$ such that for all $\tau \in (0,\bar\tau)$,

$$\geq \delta^2\left[\frac{\lambda_D}{2}\tau + \nu^T C\nu\right] + \delta^2\left[\nu^T(\Lambda_\tau^\varepsilon)^T C\Lambda_\tau^\varepsilon\nu - \nu^T C\nu\right] - 2\varepsilon. \tag{39}$$

Now define $g_t^\nu \doteq \nu^T(\Lambda_t^\varepsilon)^T C\Lambda_t^\varepsilon\nu - \nu^T C\nu$. Noting that $\frac{d}{dt}[\Lambda_t^\varepsilon] = A^{\mu_t^\varepsilon}\Lambda_t^\varepsilon$, one obviously has

$$\frac{dg^\nu}{dt} = \nu^T\left[(\Lambda_t^\varepsilon)^T(A^{\mu_t^\varepsilon})^T C\Lambda_t^\varepsilon + (\Lambda_t^\varepsilon)^T C A^{\mu_t^\varepsilon}\Lambda_t^\varepsilon\right]\nu,$$

and consequently,

$$g_t^\nu = \int_0^t \nu^T\left[(\Lambda_r^\varepsilon)^T(A^{\mu_r^\varepsilon})^T C\Lambda_r^\varepsilon + (\Lambda_r^\varepsilon)^T C A^{\mu_r^\varepsilon}\Lambda_r^\varepsilon\right]\nu\,dr. \tag{40}$$

Also, define

$$\bar g_t^\nu \doteq \int_0^t \nu^T\left[(A^{\mu_r^\varepsilon})^T C + C A^{\mu_r^\varepsilon}\right]\nu\,dr.$$

Noting that Λ_t^ε is continuous, and that $\Lambda_0^\varepsilon = I$, one sees that there exist $\hat\delta > 0$ and $\hat\tau > 0$ such that for all $\tau \in (0,\hat\tau)$,

$$|g_t^\nu - \bar g_t^\nu| \leq \frac{\hat\delta}{2}t^2 \qquad \forall\, t \in (0,\hat\tau). \tag{41}$$

Let $\tilde\tau = \min\{\bar\tau, \hat\tau, \frac{\lambda_D}{4\hat\delta}\}$. By (39) and the definition of g^ν,

$$\widehat{V}^{\tau,\psi}(x-\delta\nu) - 2\widehat{V}^{\tau,\psi}(x) + \widehat{V}^{\tau,\psi}(x+\delta\nu) \geq \delta^2\nu^T C\nu + \delta^2\left[\frac{\lambda_D}{2}\tau + \bar g_\tau^\nu - |g_\tau^\nu - \bar g_\tau^\nu|\right] - 2\varepsilon,$$

which, by the definition of $\bar g^\nu$ and (41)

$$\geq \delta^2\nu^T C\nu + \delta^2\left[\frac{\lambda_D}{2}\tau + \int_0^\tau \nu^T\left[(A^{\mu_r^\varepsilon})^T C + C A^{\mu_r^\varepsilon}\right]\nu\,dr - \frac{\hat\delta}{2}\tau^2\right] - 2\varepsilon,$$

which, by the definition of $\tilde\tau$ and the assumption that $C \in \mathcal{I}_C$,

$$\geq \delta^2\nu^T C\nu + \delta^2\frac{\lambda_D\tau}{8} - 2\varepsilon \qquad \forall\,\tau \in (0,\tilde\tau).$$

Since this is true for all $\varepsilon > 0$, letting $\eta = \lambda_D/8$, one has

$$\widehat{V}^{\tau,\psi}(x-\delta\nu) - 2\widehat{V}^{\tau,\psi}(x) + \widehat{V}^{\tau,\psi}(x+\delta\nu) \geq \delta^2\nu^T[C + \eta I\tau]\nu \qquad \forall\,\tau \in (0,\tilde\tau). \qquad \square$$

COROLLARY 3.9. *We may choose $C \in \mathcal{D}_n$ such that $\widetilde{V}, V^m \in \mathcal{S}_{-C}$, and such that with $\psi, \overline{\tau}, \eta$ as in the statement of Theorem* 3.7,

$$\widetilde{S}_\tau[\psi(\cdot,z)], S_\tau^m[\psi(\cdot,z)] \in \mathcal{S}_{-(C+\eta I\tau)} \qquad \forall\,\tau \in [0,\overline{\tau}].$$

Henceforth, we suppose C chosen so that the results of Corollary 3.9 hold. We also suppose τ, η chosen according to the corollary as well.

Now for each $z \in \mathbb{R}^n$, $\widetilde{S}_\tau[\psi(\cdot, z)] \in \mathcal{S}_{-(C+\eta I\tau)}$. Therefore, by Theorem 3.2,

$$\widetilde{S}_\tau[\psi(\cdot,z)](x) = \int_{\mathbb{R}^n}^{\oplus} \psi(x,y) \otimes \widetilde{\mathcal{B}}_\tau(y,z)\, dy = \psi(x,\cdot) \odot \widetilde{\mathcal{B}}_\tau(\cdot,z), \tag{42}$$

where for all $y \in \mathbb{R}^n$,

$$\widetilde{\mathcal{B}}_\tau(y,z) = -\int_{\mathbb{R}^n}^{\oplus} \psi(x,y) \otimes \big\{-\widetilde{S}_\tau[\psi(\cdot,z)](x)\big\}\, dx = \big\{\psi(\cdot,y) \odot [\widetilde{S}_\tau[\psi(\cdot,z)](\cdot)]^-\big\}^-. \tag{43}$$

It is handy to define the max-plus linear operator with "kernel" $\widetilde{\mathcal{B}}_\tau$ (where we do not rigorously define the term kernel, as it will not be needed here) as $\widehat{\widetilde{\mathcal{B}}}_\tau[\alpha](z) \doteq \widetilde{\mathcal{B}}_\tau(z,\cdot) \odot \alpha(\cdot)$ for all $\alpha \in \mathcal{S}_{-C}$.

PROPOSITION 3.10. *Let $\phi \in \mathcal{S}_{-C}$ with the semiconvex dual denoted by a. Define $\phi^1 = \widetilde{S}_\tau[\phi]$. Then $\phi^1 \in \mathcal{S}_{-(C+\eta I\tau)}$, and*

$$\phi^1(x) = \psi(x,\cdot) \odot a^1(\cdot),$$

where

$$a^1(x) = \widetilde{\mathcal{B}}_\tau(x,\cdot) \odot a(\cdot).$$

Proof. The proof that $\phi^1 \in \mathcal{S}_{-(C+\eta I\tau)}$ is similar to the proof of Theorem 3.7. Consequently, we prove only the second assertion:

$$\begin{aligned}
\phi^1(x) &= \sup_{w\in\mathcal{W}} \sup_{\mu\in\mathcal{D}_\infty} \left[\int_0^\tau l^{\mu_t}(\xi_t) - \frac{\gamma^2}{2}|w_t|^2\, dt + \phi(\xi_\tau)\right] \\
&= \sup_{w\in\mathcal{W}} \sup_{\mu\in\mathcal{D}_\infty} \max_{z\in\mathbb{R}^n} \left[\int_0^\tau l^{\mu_t}(\xi_t) - \frac{\gamma^2}{2}|w_t|^2\, dt + \psi(\xi_\tau,z) + a(z)\right] \\
&= \max_{z\in\mathbb{R}^n} \Big\{\widetilde{S}_\tau[\psi(\cdot,z)](x) + a(z)\Big\},
\end{aligned}$$

which by (42),

$$\begin{aligned}
&= \max_{z\in\mathbb{R}^n} \max_{y\in\mathbb{R}^n} \Big\{\psi(x,y) + \widetilde{\mathcal{B}}_\tau(y,z) + a(z)\Big\} \\
&= \int_{y\in\mathbb{R}^n}^{\oplus} \int_{z\in\mathbb{R}^n}^{\oplus} \widetilde{\mathcal{B}}_\tau(y,z) \otimes a(z)\, dz \otimes \psi(x,y)\, dy \\
&= \int_{y\in\mathbb{R}^n}^{\oplus} a^1(x) \otimes \psi(x,y)\, dy. \qquad \square
\end{aligned}$$

THEOREM 3.11. *Let $V \in \mathcal{S}_{-C}$, let a be its semiconvex dual (with respect to ψ), and suppose $\widetilde{\mathcal{B}}_\tau(z,\cdot) \odot a(\cdot) \in \mathcal{S}_d^-$ with $C + d < 0$. Then $V = \widetilde{S}_\tau[V]$ if and only if*

$$a(z) = \max_{y\in\mathbb{R}^n} \big[\widetilde{\mathcal{B}}_\tau(z,y) + a(y)\big],$$

which of course

$$= \int_{\mathbb{R}^n}^{\oplus} \widetilde{\mathcal{B}}_\tau(z,y) \otimes a(y)\, dy = \widetilde{\mathcal{B}}_\tau(z,\cdot) \odot a(\cdot) = \widehat{\widetilde{\mathcal{B}}}_\tau[a](z) \qquad \forall\, z \in \mathbb{R}^n.$$

Proof. Since a is the semiconvex dual of V, for all $x \in \mathbb{R}^n$,

$$\begin{aligned}
\psi(x,\cdot) \odot a(\cdot) &= V(x) = \widetilde{S}_\tau[V](x) \\
&= \widetilde{S}_\tau \left[\max_{z\in\mathbb{R}^n} \{\psi(\cdot, z) + a(z)\} \right](x) \\
&= \sup_{w\in\mathcal{W}} \sup_{\mu\in\mathcal{D}_\infty} \left[\int_0^\tau l^{\mu_t}(\xi_t) - \frac{\gamma^2}{2}|w_t|^2\, dt + \max_{z\in\mathbb{R}^n} \{\psi(\xi_\tau, z) + a(z)\} \right] \\
&= \max_{z\in\mathbb{R}^n} \left[a(z) + \sup_{w\in\mathcal{W}} \sup_{\mu\in\mathcal{D}_\infty} \left\{ \int_0^\tau l^{\mu_t}(\xi_t) - \frac{\gamma^2}{2}|w_t|^2\, dt + \psi(\xi_\tau, z) \right\} \right] \\
&= \max_{z\in\mathbb{R}^n} \left\{ a(z) + \widetilde{S}_\tau[\psi(\cdot, z)](x) \right\} \\
&= \int_{\mathbb{R}^n}^{\oplus} a(z) \otimes \widetilde{S}_\tau[\psi(\cdot, z)](x)\, dz,
\end{aligned}$$

which by (42)

$$\begin{aligned}
&= \int_{\mathbb{R}^n}^{\oplus} a(z) \otimes \int_{\mathbb{R}^n}^{\oplus} \widetilde{\mathcal{B}}_\tau(y, z) \otimes \psi(x, y)\, dy\, dz \\
&= \int_{\mathbb{R}^n}^{\oplus} \int_{\mathbb{R}^n}^{\oplus} \widetilde{\mathcal{B}}_\tau(y, z) \otimes a(z) \otimes \psi(x, y)\, dy\, dz \\
&= \int_{\mathbb{R}^n}^{\oplus} \left[\int_{\mathbb{R}^n}^{\oplus} \widetilde{\mathcal{B}}_\tau(y, z) \otimes a(z)\, dz \right] \otimes \psi(x, y)\, dy \\
&= \left[\int_{\mathbb{R}^n}^{\oplus} \widetilde{\mathcal{B}}_\tau(\cdot, z) \otimes a(z)\, dz \right] \odot \psi(x, \cdot),
\end{aligned}$$

where by Proposition 3.10, the first term is in $\mathcal{S}_{-(C+\eta I\tau)}$. Combining this with Lemma 3.6, one has

$$a(y) = \int_{\mathbb{R}^n}^{\oplus} \widetilde{\mathcal{B}}_\tau(\cdot, z) \otimes a(z)\, dz = \widetilde{\mathcal{B}}_\tau(y, \cdot) \odot a(\cdot) \qquad \forall\, y \in \mathbb{R}^n.$$

The reverse implication follows by supposing $a(\cdot) = \widetilde{\mathcal{B}}_\tau(y, \cdot) \odot a(\cdot)$ and reordering the above argument. □

COROLLARY 3.12. *Value function $\widetilde{V}$ is given by $\widetilde{V}(x) = \psi(x, \cdot) \odot \widetilde{a}(\cdot)$, where $\widetilde{a}$ is the unique solution of*

$$\widetilde{a}(y) = \widetilde{\mathcal{B}}_\tau(y, \cdot) \odot \widetilde{a}(\cdot) \qquad \forall\, y \in= \mathbb{R}^n,$$

or equivalently, $\widetilde{a} = \widehat{\widetilde{\mathcal{B}}}_\tau[\widetilde{a}]$.

Proof. Combining Theorems 2.5 and 3.11 yields the assertion that $\widetilde{V}$ has this representation. The uniqueness follows from the uniqueness assertion of Theorem 2.5 and Lemma 3.6. □

Similarly, for each $m \in \mathcal{M}$ and $z \in \mathbb{R}^n$, we have $S_\tau^m[\psi(\cdot, z)] \in \mathcal{S}_{-(C+\eta I\tau)}$ and

$$S_\tau^m[\psi(\cdot, z)](x) = \psi(x, \cdot) \odot \mathcal{B}_\tau^m(\cdot, z) \qquad \forall\, x \in \mathbb{R}^n,$$

where

$$\mathcal{B}_\tau^m(y, z) = \left\{ \psi(\cdot, y) \odot \left[S_\tau^m[\psi(\cdot, z)] \right]^-(\cdot) \right\}^- \qquad \forall\, y \in \mathbb{R}^n.$$

As before, it will be handy to define the max-plus linear operator with "kernel" $\mathcal{B}_\tau^m$ as $\widehat{\mathcal{B}}_\tau^m[a](z) \doteq \mathcal{B}_\tau^m(z, \cdot) \odot a(\cdot)$ for all $a \in \mathcal{S}_{-C}$. Further, one also obtains analogous results (by similar proofs). In particular, one has the following.

THEOREM 3.13. *Let $V \in \mathcal{S}_{-C}$, and let a be its semiconvex dual (with respect to ψ). Then $V = S_\tau^m[V]$ if and only if*

$$a(z) = \mathcal{B}_\tau^m(z,\cdot) \odot a(\cdot) \qquad \forall\, z \in \mathbb{R}^n.$$

COROLLARY 3.14. *Each value function V^m is given by $V^m(x) = \psi(x,\cdot) \odot a^m(\cdot)$, where each a^m is the unique solution of the problem $a^m(y) = \mathcal{B}_\tau^m(y,\cdot) \odot a^m(\cdot)$ for all $y \in \mathbb{R}^n$.*

4. Discrete-time approximation. The method developed here will not involve any discretization over space. Of course this is obvious since otherwise one could not avoid the curse of dimensionality. The discretization will be over time where approximate μ processes will be constant over the length of each time-step.

We define the operator $\bar{S}_\tau$ on $\mathcal{G}_\delta$ by

$$\begin{aligned}\bar{S}_\tau[\phi](x) &= \sup_{w\in\mathcal{W}} \max_{m\in\mathcal{M}} \left[\int_0^\tau l^m(\xi_t^m) - \frac{\gamma^2}{2}|w_t|^2\, dt + \phi(\xi_\tau^m)\right](x)\\ &= \max_{m\in\mathcal{M}} S_\tau^m[\phi](x),\end{aligned}$$

where ξ^m satisfies (1). Let

$$\overline{\mathcal{B}}_\tau(y,z) \doteq \max_{m\in\mathcal{M}} \mathcal{B}_\tau^m(y,z) = \bigoplus_{m\in\mathcal{M}} \mathcal{B}_\tau^m(y,z) \qquad \forall\, y,z \in \mathbb{R}^n.$$

The corresponding max-plus linear operator is

$$\widehat{\overline{\mathcal{B}}}_\tau = \bigoplus_{m\in\mathcal{M}} \widehat{\mathcal{B}}_\tau^m.$$

LEMMA 4.1. *For all $z \in \mathbb{R}^n$, we have $\bar{S}_\tau[\psi(\cdot,z)] \in \mathcal{S}_{-(C+\eta I\tau)}$. Further,*

$$\bar{S}_\tau[\psi(\cdot,z)](x) = \psi(x,\cdot) \odot \overline{\mathcal{B}}_\tau(\cdot,z) \qquad \forall\, x \in \mathbb{R}^n. \tag{44}$$

Proof. We provide the proof of the last statement as follows:

$$\begin{aligned}\bar{S}_\tau[\psi(\cdot,z)](x) &= \max_{m\in\mathcal{M}} S_\tau^m[\psi(\cdot,z)](x) = \max_{m\in\mathcal{M}} \psi(x,\cdot) \odot \mathcal{B}_\tau^m(\cdot,z)\\ &= \max_{m\in\mathcal{M}} \max_{y\in\mathbb{R}^n} [\psi(x,y) + \mathcal{B}_\tau^m(y,z)] = \max_{y\in\mathbb{R}^n}\Big[\psi(x,y) + \max_{m\in\mathcal{M}} \mathcal{B}_\tau^m(y,z)\Big]\\ &= \psi(x,\cdot) \odot \Big[\max_{m\in\mathcal{M}} \mathcal{B}_\tau^m(\cdot,z)\Big]. \qquad \square\end{aligned}$$

We remark that, parameterized by τ, the operators $\bar{S}_\tau$ do not necessarily form a semigroup, although they do form a sub-semigroup (i.e., $\bar{S}_{\tau_1+\tau_2}[\phi](x) \le \bar{S}_{\tau_1}\bar{S}_{\tau_2}[\phi](x)$ for all $x \in \mathbb{R}^n$ and all $\phi \in \mathcal{S}_{-C}$). In spite of this, one does have $S_\tau^m \le \bar{S}_\tau \le \widetilde{S}_\tau$ for all $m \in \mathcal{M}$.

With τ acting as a time-discretization step size, let

$$\begin{aligned}\mathcal{D}_\infty^\tau = \Big\{&\mu : [0,\infty) \to \mathcal{M} \mid \text{for each } n \in \mathbf{N}\cup\{0\}, \text{ there exists } m_n \in \mathcal{M}\\ &\text{such that } \mu(t) = m_n \ \forall\ t \in [n\tau,(n+1)\tau)\Big\},\end{aligned}$$

and for $T = \bar{n}\tau$ with $\bar{n} \in \mathbf{N}$ define $\mathcal{D}_T^\tau$ similarly but with domain $[0, T)$ rather than $[0, \infty)$. Let $\mathcal{M}^{\bar{n}}$ denote the outer product of $\mathcal{M}$, $\bar{n}$ times. Let $T = \bar{n}\tau$, and define

$$\bar{\bar{S}}_T^\tau[\phi](x) = \max_{\{m_k\}_{k=0}^{\bar{n}-1} \in \mathcal{M}^{\bar{n}}} \left\{ \prod_{k=0}^{\bar{n}-1} S_\tau^{m_k} \right\} [\phi](x) = (\bar{S}_\tau)^{\bar{n}}[\phi](0),$$

where the Π notation indicates operator composition, and the superscript in the last expression indicates repeated application of $\bar{S}_\tau$, $\bar{n}$ times.

We will approximate $\widetilde{V}$ by solving $V = \bar{S}_\tau[V]$ via its dual problem $a = \widehat{\bar{\mathcal{B}}}_\tau[a]$ for small τ. Consequently, we will need to show that there exists a solution to $V = \bar{S}_\tau[V]$, that the solution is unique, and that it can be found by solving the dual problem. We begin with existence.

THEOREM 4.2. *Let*

$$\overline{V}(x) \doteq \lim_{N \to \infty} \bar{\bar{S}}_{N\tau}^\tau[0](x) \tag{45}$$

for all $x \in \mathbb{R}^n$, where 0 here represents the zero-function. Then, $\overline{V}$ satisfies

$$V = \bar{S}_\tau[V], \qquad V(0) = 0. \tag{46}$$

Further, $0 \le V^m \le \overline{V} \le \widetilde{V}$ for all $m \in \mathcal{M}$, and consequently, $\overline{V} \in \mathcal{G}_\delta$.

Proof. Note that

$$\begin{aligned} V^m(x) &= \lim_{N \to \infty} S_{N\tau}^m[0](x) \le \limsup_{N \to \infty} \bar{\bar{S}}_{N\tau}^\tau[0] \\ &\le \lim_{N \to \infty} \widetilde{S}_{N\tau}[0](x) = \widetilde{V}(x) \qquad \forall\, x \in \mathbb{R}^n. \end{aligned} \tag{47}$$

Also,

$$\begin{aligned} \bar{\bar{S}}_{(N+1)\tau}^\tau[0](x) &= \bar{\bar{S}}_{N\tau}^\tau[\bar{S}_\tau[0](\cdot)](x) \\ &= \sup_{\hat{w} \in \mathcal{W}} \sup_{\hat{\mu} \in \mathcal{D}_{N\tau}} \int_0^{N\tau} l^{\hat{\mu}_t}(\xi_t) - \frac{\gamma^2}{2} |\hat{w}_t|^2 \, dt \\ &\quad + \sup_{w \in \mathcal{W}} \max_{m \in \mathcal{M}} \int_{N\tau}^{(N+1)\tau} l^m(\xi_t) - \frac{\gamma^2}{2} |w_t|^2 \, dt, \end{aligned} \tag{48}$$

which by taking $w \equiv 0$

$$\ge \sup_{\hat{w} \in \mathcal{W}} \sup_{\hat{\mu} \in \mathcal{D}_{N\tau}} \int_0^{N\tau} l^{\hat{\mu}_t}(\xi_t) - \frac{\gamma^2}{2} |\hat{w}_t|^2 \, dt = \bar{\bar{S}}_{N\tau}^\tau[0](x), \tag{49}$$

which implies that $\bar{\bar{S}}_{N\tau}^\tau[0](x)$ is a monotonically increasing function of N. Since it is also bounded from above (by (47)), one finds

$$V^m(x) \le \lim_{N \to \infty} \bar{\bar{S}}_{N\tau}^\tau[0](x) \le \widetilde{V}(x) \qquad \forall\, x \in \mathbb{R}^n, \tag{50}$$

which also justifies the use of the limit definition of $\overline{V}$ in the statement of the theorem. In particular, one has $0 \le V^m \le \overline{V} \le \widetilde{V}$, and so $\overline{V} \in \mathcal{G}_\delta$.

Fix any $x \in \mathbb{R}^n$, and suppose there exists $\delta > 0$ such that

$$\overline{V}(x) \le \bar{S}_\tau[\overline{V}](x) - \delta. \tag{51}$$

However, by the definition of $\overline{V}$, given any $y \in \mathbb{R}^n$, there exists $N_\delta < \infty$ such that for all $N \geq N_\delta$,

$$\overline{V}(y) \leq \bar{\bar{S}}^{\tau}_{N_\delta \tau}[0](y) + \delta/4. \tag{52}$$

Combining (51) and (52), one finds after a small bit of work that

$$\overline{V}(x) \leq \bar{S}_\tau \big[\bar{\bar{S}}^{\tau}_{N_\delta \tau}[0] + \delta/2\big](x) - \delta,$$

which, using the max-plus linearity of $\bar{S}_\tau$,

$$= \bar{\bar{S}}^{\tau}_{(N_\delta+1)\tau}[0](x) - \delta/2$$

for all $N \geq N_\delta$. Consequently, $\overline{V}(x) \leq \lim_{N\to\infty} \bar{\bar{S}}^{\tau}_{N\tau}[0](x) - \delta/2$, which is a contradiction. Therefore, $\overline{V}(x) \geq \bar{S}_\tau[\overline{V}](x)$ for all $x \in \mathbb{R}^n$. The reverse inequality follows in a similar way. Specifically, fix $x \in \mathbb{R}^n$ and suppose there exists $\delta > 0$ such that

$$\overline{V}(x) \geq \bar{S}_\tau[\overline{V}](x) + \delta. \tag{53}$$

By the monotonicity of $\bar{\bar{S}}^{\tau}_{N\tau}$ with respect to N, for any $N < \infty$,

$$\overline{V}(x) \geq \bar{\bar{S}}^{\tau}_{N\tau}[0](x) \qquad \forall\, x \in \mathbb{R}^n.$$

By the monotonicity of $\bar{S}_\tau$ with respect to its argument (i.e., $\phi_1(x) \leq \phi_2(x)$ for all x implying $\bar{S}_\tau[\phi_1](x) \leq \bar{S}_\tau[\phi_2](x)$ for all x), this implies

$$\bar{S}_\tau[\overline{V}] \geq \bar{\bar{S}}^{\tau}_{(N+1)\tau}[0] \qquad \forall\, x \in \mathbb{R}^n. \tag{54}$$

Combining (53) and (54) yields

$$\overline{V}(x) \geq \bar{\bar{S}}^{\tau}_{(N+1)\tau}[0](x) + \delta.$$

Letting $N \to \infty$ yields a contradiction, and so $\overline{V} \leq \bar{S}_\tau[\overline{V}]$. □

The following result is immediate.

THEOREM 4.3.

$$\overline{V}(x) = \sup_{\mu\in\mathcal{D}^{\tau}_{\infty}} \sup_{w\in\mathcal{W}} \sup_{T\in[0,\infty)} \left[\int_0^T l^{\mu_t}(\xi_t) - \frac{\gamma^2}{2}|w_t|^2\, dt\right],$$

where ξ_t satisfies (12).

THEOREM 4.4. *$\overline{V}(x) - \frac{1}{2}c_V|x|^2$ is strictly convex.*

Proof. The proof is identical to the proof of Theorem 2.6 with the exception that μ^ε is chosen from $\mathcal{D}^{\tau}_{\infty}$ instead of $\mathcal{D}_\infty$. □

Remark 4.5. From the choice of β in section 3, this immediately implies that $\overline{V} \in \mathcal{S}_\beta$, and of course since $C + \beta < 0$, that $\overline{V} \in \mathcal{S}_{-C}$.

We now address the uniqueness issue. Similar techniques to those used for V^m and $\widetilde{V}$ will prove uniqueness for (46) within $\mathcal{G}_\delta$. A slightly weaker type of result under weaker assumptions will be obtained first; this result is similar in form to that of [40].

Suppose $\overline{V}' \neq \overline{V}$, $\overline{V}' \in \mathcal{G}_\delta$ satisfies (46). This implies that for all $x \in \mathbb{R}^n$ and all $N < \infty$

$$\begin{aligned}\overline{V}'(x) &= \bar{\bar{S}}^{\tau}_{N\tau}[\overline{V}'](x)\\ &= \sup_{w\in\mathcal{W}}\sup_{\mu\in\mathcal{D}^{\tau}_{\infty}}\left\{\int_0^{N\tau} l^{\mu_t}(\xi_t)-\frac{\gamma^2}{2}|w_t|^2\,dt+\overline{V}'(\xi_{N\tau})\right\},\end{aligned}$$

which, by taking $w^0\equiv 0$ (with corresponding trajectory denoted by ξ^0),

$$\geq \overline{V}'(\xi^0_{N\tau}). \tag{55}$$

However, by (12), one has $\dot{\xi}^0 = A^{\mu_t}\xi^0$, and so $|\xi^0_t|\leq e^{-c_A t}|x|$ for all $t\geq 0$, which implies that $|\xi^0_{N\tau}|\to 0$ as $N\to\infty$. Consequently

$$\lim_{N\to\infty}\overline{V}'(\xi^0_{N\tau})=0. \tag{56}$$

Combining (55) and (56), one has

$$\overline{V}'(x)\geq 0 \qquad \forall\, x\in\mathbb{R}^n. \tag{57}$$

Also, by (46)

$$\overline{V}'(x)=\lim_{N\to\infty}\bar{\bar{S}}^{\tau}_{N\tau}[\overline{V}'](x) \qquad \forall\, x\in\mathbb{R}^n.$$

By (57) and the monotonicity of $\bar{\bar{S}}^{\tau}_{N\tau}$ with respect to its argument, this is

$$\geq \lim_{N\to\infty}\bar{\bar{S}}^{\tau}_{N\tau}[0](x)=\overline{V}(x). \tag{58}$$

By (57) and (59), one has the uniqueness result analogous to [40], which is as follows.

THEOREM 4.6. *$\overline{V}$ is the unique minimal, nonnegative solution to* (46).

The stronger uniqueness statement (making use of the quadratic bound on $l^{\mu_t}(x)$) is as follows. As with $V^m,\widetilde{V}$, the proof is similar to that in [35]. However in this case, there is a small difference in the proof, and this difference requires another lemma. Due to this difference in the case of $\overline{V}$, we include a sketch of the proof (but with the new lemma in full) in Appendix A.

THEOREM 4.7. *$\overline{V}$ is the unique solution of* (46) *within the class $\mathcal{G}_\delta$ for sufficiently small $\delta>0$. Further, given any $V\in\mathcal{G}_\delta$, we have $\lim_{N\to\infty}\bar{\bar{S}}^{\tau}_{N\tau}[V](x)=\overline{V}(x)$ for all $x\in\mathbb{R}^n$ (uniformly on compact sets).*

Henceforth, we let $\delta>0$ be sufficiently small such that $V^m,\widetilde{V},\overline{V}\in\mathcal{G}_\delta$ for all $m\in\mathcal{M}$.

THEOREM 4.8. *Let $V\in\mathcal{S}_{-C}$, and let a be its semiconvex dual. Then, if $\overline{\mathcal{B}}_\tau(y,\cdot)\odot a(\cdot)\in\mathcal{S}^-_d$, $V=\bar{S}_\tau[V]$ if and only if $a(y)=\overline{\mathcal{B}}_\tau(y,\cdot)\odot a(\cdot)$ for all $y\in\mathbb{R}^n$.*

Proof. By the semiconvex duality,

$$\begin{aligned}\psi(x,\cdot)\odot a(\cdot) &= V(x)=\bar{S}_\tau[V](x)\\ &= \bar{S}_\tau\Big[\max_{z\in\mathbb{R}^n}\{\psi(\cdot,z)+a(z)\}\Big](x),\end{aligned} \tag{59}$$

which, as in the first part of the proof of Theorem 3.11,

$$=\int^{\oplus}_{\mathbb{R}^n} a(z)\otimes\bar{S}_\tau[\psi(\cdot,z)](x)\,dz,$$

which, by Lemma 4.1,

$$= \int_{\mathbb{R}^n}^{\oplus} a(z) \otimes \int_{\mathbb{R}^n}^{\oplus} \psi(x,y) \otimes \overline{\mathcal{B}}_\tau(y,z)\, dy\, dz,$$

which, as in the latter part of the proof of Theorem 3.11,

$$= \left[\int_{\mathbb{R}^n}^{\oplus} \overline{\mathcal{B}}_\tau(\cdot,z) \otimes a(z)\, dz \right] \odot \psi(x,\cdot). \tag{60}$$

By Lemmas 3.4 and 3.6, this implies

$$a(y) = \overline{\mathcal{B}}_\tau(y,\cdot) \odot a(\cdot) \qquad \forall\, y \in \mathbb{R}^n.$$

Alternatively, if $a(y) = \overline{\mathcal{B}}_\tau(y,\cdot) \odot a(\cdot)$ for all y, then

$$V(x) = \psi(x,\cdot) \odot a(\cdot) = \left[\int_{\mathbb{R}^n}^{\oplus} \overline{\mathcal{B}}_\tau(\cdot,z) \otimes a(z)\, dz \right] \odot \psi(x,\cdot) \qquad \forall\, x \in \mathbb{R}^n,$$

which by (59)–(60) yields $V = \bar{S}_\tau[V]$. □

COROLLARY 4.9. *Value function $\overline{V}$ given by* (45) *is in $\mathcal{S}_\beta \subset \mathcal{S}_{-C}$ and has representation $\overline{V}(x) = \psi(x,\cdot) \odot \overline{a}(\cdot)$, where $\overline{a}$ is the unique solution in $\mathcal{S}_d^- \cap \mathcal{G}_\delta^-$ of*

$$\overline{a}(y) = \overline{\mathcal{B}}_\tau(y,\cdot) \odot \overline{a}(\cdot) \qquad \forall\, y \in \mathbb{R}^n, \tag{61}$$

or equivalently, $\overline{a} = \widehat{\overline{\mathcal{B}}}_\tau[\overline{a}]$.

Proof. The fact that $\overline{V} \in \mathcal{S}_\beta$ follows from Theorem 4.4 and the choice of β. By Theorem 4.7, $\overline{V} \in \mathcal{G}_\delta$ and is the unique solution of (46) in $\mathcal{G}_\delta$.

By Theorem 4.8, its semiconvex dual, $\overline{a}$, satisfies (61), and by Lemma 3.4, $\overline{a} \in \mathcal{S}_d^-$ for some $d \in \mathcal{D}_n$ such that $C + d < 0$. Suppose there is $\hat{a} \in \mathcal{S}_d^- \cap \mathcal{G}_\delta^-$ for some $\hat{d} \in \mathcal{D}_n$ such that $C + \hat{d} < 0$, and that $\hat{a}$ satisfies (61). Then, by Theorem 4.8 and Remark 3.5, its dual, $\widehat{V}$, is in $\mathcal{G}_\delta$ and $\widehat{V} = \bar{S}_\tau[\widehat{V}]$, $\widehat{V}(0) = 0$. By Theorem 4.7 then, $\widehat{V} = \overline{V}$. By Lemma 3.6, this implies that $\hat{a} = \overline{a}$. □

The following result on propagation of the semiconvex dual will also come in handy.

PROPOSITION 4.10. *Let $\phi \in \mathcal{S}_\beta \subset \mathcal{S}_{-C}$ with the semiconvex dual denoted by a. Define $\phi^1 = \bar{S}_\tau[\phi]$. Then $\phi^1 \in \mathcal{S}_{-(C+\eta I\tau)}$, and*

$$\phi^1(x) = \psi(x,\cdot) \odot a^1(\cdot),$$

where

$$a^1(y) = \overline{\mathcal{B}}_\tau(y,\cdot) \odot a(\cdot) \qquad \forall\, y \in \mathbb{R}^n.$$

Proof. The proof is similar to the proof of Proposition 3.10, and consequently some details are not included. To begin, as in the proof of Proposition 3.10, we note that the proof that $\phi^1 \in \mathcal{S}_{-(C+\eta I\tau)}$ is nearly identical to the proof of Theorem 3.7. In particular, fix any $x, \nu \in \mathbb{R}^n$ with $|\nu| = 1$ and any $\delta > 0$. Let $\overline{m} \in \mathcal{M}$ be optimal, and w^ε be ε-optimal, for $\bar{S}_\tau[\phi](x)$. That is, suppose $\overline{\mathcal{I}}^\phi(x,\tau,w^\varepsilon,\overline{m}) \geq \bar{S}_\tau[\phi](x) - \varepsilon$, where

$$\overline{\mathcal{I}}^\phi(x,\tau,w,m) \doteq \int_0^\tau l^{\overline{m}}(\xi_t) - \frac{\gamma^2}{2}|w_t|^2\, dt + \phi(\xi_\tau)$$

and ξ satisfies (1). Then

$$\begin{aligned} &\bar{S}_\tau[\phi](x-\delta\nu) - 2\bar{S}_\tau[\phi](x) + \bar{S}_\tau[\phi](x+\delta\nu) \\ &\quad \geq \overline{\mathcal{I}}^\phi(x-\delta\nu,\tau,w^\varepsilon,\overline{m}) - 2\overline{\mathcal{I}}^\phi(x,\tau,w^\varepsilon,\overline{m}) + \overline{\mathcal{I}}^\phi(x+\delta\nu,\tau,w^\varepsilon,\overline{m}) - 2\varepsilon. \end{aligned}$$

Let $\xi^\delta, \xi^0, \xi^{-\delta}$ satisfy the dynamics of (1) with inputs w^ε and $\overline{m}$, and with initial conditions $\xi_0^\delta = x + \delta\nu$, $\xi_0^0 = x$, and $\xi_0^{-\delta} = x - \delta\nu$, respectively. Letting $\Delta_t^+ \doteq \xi_t^\delta - \xi_t^0$, one finds

$$\phi(\xi_\tau^\delta) - 2\phi(\xi_\tau^0) + \phi(\xi_\tau^{-\delta}) \geq (\Delta_\tau^+)^T C \Delta_\tau^+$$

because $\phi \in \mathcal{S}_{-C}$. One then continues as in the proof of Theorem 3.7, but with $D^{\overline{m}}$ and $A^{\overline{m}}$ replacing $D^{\mu_t^\varepsilon}$ and $A^{\mu_t^\varepsilon}$, respectively. In particular, one has $\Lambda_t^\varepsilon = \exp\{A^{\overline{m}} t\}$. More important, one may use the same values of η and τ which were fixed in section 3.

Now we turn to the second assertion of the proposition. This follows exactly as in the proof of Proposition 3.10 with two minor exceptions: First, the supremum over $\mu \in \mathcal{D}_\infty$ is replaced by a maximum over $m \in \mathcal{M}$. Second, the use of (42) is replaced by the invocation of (44). □

We now show that one may approximate $\widetilde{V}$, the solution of $V = \widetilde{S}_\tau[V]$, to as accurate a level as one desires by solving $V = \bar{S}_\tau[V]$ for sufficiently small τ. Recall that if $V = \bar{S}_\tau[V]$, then it satisfies $V = \bar{\bar{S}}_{N\tau}^\tau[V]$ for all $N > 0$ (while $\widetilde{V}$ satisfies $V = \widetilde{S}_{N\tau}[V]$), and so this is essentially equivalent to introducing a discrete-time $\overline{\mu} \in \mathcal{D}_{N\tau}^\tau$ approximation to the μ process in $\widetilde{S}_{N\tau}$. The result will follow easily from the following technical lemma. The lemma uses the particular structure of our example class of problems as given by assumption block $(A.m)$. As the proof of the lemma is technical and long, it is delayed to Appendix B.

LEMMA 4.11. *Given $\hat{\varepsilon} \in (0, 1]$, $\overline{T} < \infty$, there exist $T \in [\overline{T}/2, \overline{T}]$ and $\tau > 0$ such that*

$$\widetilde{S}_T[V^m](x) - \bar{\bar{S}}_T^\tau[V^m](x) \leq \hat{\varepsilon}(1 + |x|^2) \qquad \forall\, x \in \mathbb{R}^n,\ \forall m \in \mathcal{M}.$$

We now obtain the main approximation result.

THEOREM 4.12. *Given $\overline{\varepsilon} > 0$ and $R < \infty$, there exists $\tau > 0$ such that*

$$\widetilde{V}(x) - \overline{\varepsilon} \leq \overline{V}(x) \leq \widetilde{V}(x) \qquad \forall\, x \in \overline{B}_R(0).$$

Proof. From Theorem 4.2, we have

$$0 \leq V^m(x) \leq \overline{V}(x) \leq \widetilde{V}(x) \leq \frac{c_A(\gamma - \delta)^2}{c_\sigma^2}|x|^2 \qquad \forall\, x \in \mathbb{R}^n. \tag{62}$$

Also, with $T = N\tau$ for any positive integer N,

$$\bar{\bar{S}}_{N\tau}^\tau[\phi] \leq \widetilde{S}_T[\phi] \qquad \forall\, \phi \in \mathcal{G}_\delta. \tag{63}$$

Further, by Theorem 2.5, given $\varepsilon > 0$ and $R < \infty$, there exists $\widehat{T} < \infty$ such that for all $T > \widehat{T}$ and all $m \in \mathcal{M}$,

$$\widetilde{S}_T[\widetilde{V}](x) - \varepsilon/2 \leq \widetilde{S}_T[V^m](x) \qquad \forall\, x \in \overline{B}_R(0). \tag{64}$$

By (64) and Lemma 4.11, given $\overline{\varepsilon} > 0$ and $R < \infty$, there exists $T \in [0, \infty)$, $\tau \in [0, T]$, where $T = N\tau$ for some integer N such that for all $|x| \leq R$,

$$\begin{aligned} \widetilde{V}(x) - \overline{\varepsilon} &= \widetilde{S}_T[\widetilde{V}](x) - \overline{\varepsilon} \\ &\leq \widetilde{S}_T[V^m](x) - \overline{\varepsilon}/2 \\ &\leq \bar{\bar{S}}_T^\tau[V^m](x), \end{aligned}$$

where $\hat{\varepsilon}(1 + R^2) = \overline{\varepsilon}/2$, and which, by (62) and the monotonicity of $\bar{\bar{S}}_T^\tau[\cdot]$,

$$\leq \bar{\widetilde{S}}^{\tau}_{T}[\overline{V}](x),$$

which, by (63),

$$\leq \widetilde{S}_{T}[\overline{V}](x),$$

which, by the monotonicity of $\widetilde{S}_{T}[\cdot]$,

$$\leq \widetilde{S}_{T}[\widetilde{V}](x) = \widetilde{V}(x).$$

Noting (from Theorem 4.7) that $\overline{V} = \bar{\widetilde{S}}^{\tau}_{T}[\overline{V}]$ completes the proof. □

Remark 4.13. For this class of systems (defined by assumption block $(A.m)$), we expect this result could be sharpened to

$$\widetilde{V}(x) \leq -\hat{\varepsilon}(1+|x|^2) \leq \overline{V}(x) \leq \widetilde{V}(x) \qquad \forall\, x \in \mathbb{R}^n$$

by sharpening Theorem 2.5. However, this type of result might be valid only for limited classes of systems, and so we have not pursued it here.

5. The algorithm. We now begin discussion of the actual algorithm.

Let $C \in \mathcal{I}_C$ such that $C - c_V I < 0$, and initialize with $\overline{V}^0(x) \doteq \frac{c_V}{2}|x|^2$. From Theorem 4.2, $\overline{V} = \lim_{N\to\infty} \bar{\widetilde{S}}^{\tau}_{N\tau}[\overline{V}^0]$. Given $\overline{V}^k$, let

$$\overline{V}^{k+1} \doteq \bar{S}_{\tau}[\overline{V}^k]$$

so that $\overline{V}^k = \bar{\widetilde{S}}^{\tau}_{k\tau}[\overline{V}^0]$ for all $k \geq 1$.

Let $\overline{a}^k$ be the semiconvex dual of $\overline{V}^k$ for all k. Since $\overline{V}^0 = \frac{c_V}{2}|x|^2$, one easily finds the quadratic $\overline{a}^0(\cdot)$. Note also that by Proposition 4.10,

$$\overline{a}^{k+1} = \overline{\mathcal{B}}_{\tau}(x,\cdot) \odot \overline{a}^k(\cdot) = \widehat{\overline{\mathcal{B}}}_{\tau}[\overline{a}^k]$$

for all $n \geq 0$.

Recall that

$$\overline{\mathcal{B}}_{\tau}(x,\cdot) \odot \overline{a}^k(\cdot) = \int^{\oplus}_{\mathbb{R}^n} \overline{\mathcal{B}}_{\tau}(x,y) \otimes \overline{a}^k(y)\, dy = \int^{\oplus}_{\mathbb{R}^n} \bigoplus_{m\in\mathcal{M}} \mathcal{B}^m_{\tau}(x,y) \otimes \overline{a}^k(y)\, dy$$

$$= \bigoplus_{m\in\mathcal{M}} \int^{\oplus}_{\mathbb{R}^n} \mathcal{B}^m_{\tau}(x,y) \otimes \overline{a}^k(y)\, dy = \bigoplus_{m\in\mathcal{M}} \left[\mathcal{B}^m_{\tau}(x,\cdot) \odot \overline{a}^k(\cdot)\right]. \tag{65}$$

By (65),

$$\overline{a}^1(x) = \bigoplus_{m\in\mathcal{M}} \widehat{a}^1_m(x), \quad \text{where} \tag{66}$$

$$\widehat{a}^1_m(x) \doteq \mathcal{B}^m_{\tau}(x,\cdot) \odot \overline{a}^0(\cdot) \qquad \forall\, m.$$

By (65) and (66),

$$\overline{a}^2(x) = \bigoplus_{m_2\in\mathcal{M}} \int^{\oplus}_{\mathbb{R}^n} \mathcal{B}^{m_2}_{\tau}(x,y) \otimes \left[\bigoplus_{m_1\in\mathcal{M}} \widehat{a}^1_{m_1}(y)\right] dy$$

$$= \bigoplus_{\{m_1,m_2\}\in\mathcal{M}\times\mathcal{M}} \int^{\oplus}_{\mathbb{R}^n} \mathcal{B}^{m_2}_{\tau}(x,y) \otimes \widehat{a}^1_{m_1}(y)\, dy.$$

Consequently,

$$\overline{a}^2(x) = \bigoplus_{\{m_1,m_2\}\in\mathcal{M}^2} \widehat{a}^2_{\{m_1,m_2\}}(x), \quad \text{where}$$

$$\widehat{a}^2_{\{m_1,m_2\}}(x) \doteq \mathcal{B}^{m_2}_\tau(x,\cdot)\odot \widehat{a}^1_{m_1}(\cdot) \qquad \forall\, m_1, m_2 \tag{67}$$

and $\mathcal{M}^2$ represents the outer product $\mathcal{M}\times\mathcal{M}$. Proceeding with this, one finds that in general,

$$\overline{a}^k(x) = \bigoplus_{\{m_i\}_{i=1}^k\in\mathcal{M}^k} \widehat{a}^k_{\{m_i\}_{i=1}^k}(x), \quad \text{where} \tag{68}$$

$$\widehat{a}^k_{\{m_i\}_{i=1}^k}(x) \doteq \mathcal{B}^{m_k}_\tau(x,\cdot)\odot \widehat{a}^{k-1}_{\{m_i\}_{i=1}^{k-1}}(\cdot) \qquad \forall\, \{m_i\}_{i=1}^k \in \mathcal{M}^k.$$

Of course one can obtain $\overline{V}^k$ from its dual as

$$\begin{aligned}
\overline{V}^k(x) &= \max_{y\in\mathbb{R}^n}[\psi(x,y)+\overline{a}^k(y)] \\
&= \max_{y\in\mathbb{R}^n}\left[\psi(x,y)+\max_{\{m_i\}_{i=1}^k\in\mathcal{M}^k}\widehat{a}^k_{\{m_i\}_{i=1}^k}(y)\right] \\
&= \max_{\{m_i\}_{i=1}^k\in\mathcal{M}^k}\left\{\max_{y\in\mathbb{R}^n}[\psi(x,y)+\widehat{a}^k_{\{m_i\}_{i=1}^k}(y)]\right\} \\
&\doteq \max_{\{m_i\}_{i=1}^k\in\mathcal{M}^k}\widehat{V}^k_{\{m_i\}_{i=1}^k}(x), \qquad (69)
\end{aligned}$$

where

$$\widehat{V}^k_{\{m_i\}_{i=1}^k} = \max_{y\in\mathbb{R}^n}[\psi(x,y)+\widehat{a}^k_{\{m_i\}_{i=1}^k}(y)] = \int_{\mathbb{R}^n}^{\oplus}\psi(x,y)\otimes\widehat{a}^k_{\{m_i\}_{i=1}^k}(y)\,dy. \tag{70}$$

The algorithm will consist of the forward propagation of the $\widehat{a}^k_{\{m_i\}_{i=1}^k}$ (according to (68)) from $k=0$ to some termination step $k=N$, followed by construction of the value as $\widehat{V}^k_{\{m_i\}_{i=1}^k}$ (according to (70)).

It is important to note that the computation of each $\widehat{a}^k_{\{m_i\}_{i=1}^k}$ is analytical. We will indicate the actual analytical computations.

By the linear/quadratic nature of the m-indexed systems, we find that the $S^m_\tau[\psi(\cdot,z)]$ take the form

$$S^m_\tau[\psi(\cdot,z)](x) = \tfrac{1}{2}(x-\Lambda^m_\tau z)^T P^m_\tau (x-\Lambda^m_\tau) + \tfrac{1}{2} z^T R^m_\tau z,$$

where the time-dependent $n\times n$ matrices P^m_t, Λ^m_t, and R^m_t satisfy $P^m_0 = C$, $\Lambda^m_0 = I$, $R^m_0 = 0$:

$$\begin{aligned}
\dot{P}^m &= (A^m)^T P^m + P^m A^m + D^m + P^m \Sigma^m P^m, \\
\dot{\Lambda}^m &= \left[(P^m)^{-1} D^m - A^m\right]\Lambda^m, \\
\dot{R}^m &= (\Lambda^m)^T D^m \Lambda^m.
\end{aligned} \tag{71}$$

We note that each of the $P^m_\tau, \Lambda^m_\tau, R^m_\tau$ need only be computed once.

Next, one computes each quadratic function $\mathcal{B}^m_\tau(x,z)$ (one time only) as follows. One has

$$\mathcal{B}^m_\tau = -\max_{y\in\mathbb{R}^n}\left\{\psi(y,x) - S^m_\tau[\psi(\cdot,z)](y)\right\},$$

which, by the above,

(72)
$$= \min_{y\in\mathbb{R}^n} \left\{ \tfrac{1}{2}(y-x)^T C(y-x) + \tfrac{1}{2}(y-\Lambda_\tau^m z)^T P_\tau^m (y-\Lambda_\tau^m z) + \tfrac{1}{2} z^T R_\tau^m z \right\}.$$

Recall that by Theorem 3.7, this has a finite minimum ($P^m - (C + \eta I\tau)$ positive definite). Taking the minimum in (73), one has

$$\mathcal{B}_\tau^m(x,z) = \tfrac{1}{2}\left[x^T M_{1,1}^m x + x^T M_{1,2}^m z + z^T (M_{1,2}^m)^T x + z^T M_{2,2}^m z\right],$$

where, with shorthand notation $D_\tau \doteq (P_\tau^m - C)^{-1}$,

(73)
$$M_{1,1}^m = \left[C D_\tau^{-1} P_\tau^m D_\tau^{-1} C - (D_\tau^{-1} C + I)^T C (D_\tau^{-1} C + I)\right],$$
(74)
$$M_{1,2}^m = \left[(D_\tau^{-1} C + I)^T C D_\tau^{-1} P_\tau^m - C D_\tau^{-1} P_\tau^m (D_\tau^{-1} P_\tau^m - I)\right] \Lambda_\tau^m,$$
(75)
$$M_{2,2}^m = (\Lambda_\tau^m)^T \left[(D_\tau^{-1} P_\tau^m - I)^T P_\tau^m (D_\tau^{-1} P_\tau^m - I) - P_\tau^m D_\tau^{-1} C D_\tau^{-1} P_\tau^m\right] \Lambda_\tau^m + R_\tau^m.$$

Note that given the $P_\tau^m, \Lambda_\tau^m, R_\tau^m$, the $\mathcal{B}_\tau^m$ are quadratic functions with analytical expressions for their coefficients. Also note that all the matrices in the definition of $\mathcal{B}_\tau^m$ may be precomputed.

Now let us write the (quadratic) $\widehat{a}^k_{\{m_i\}_{i=1}^k}$ in the form

$$\widehat{a}^k_{\{m_i\}_{i=1}^k}(x) = \tfrac{1}{2}\left(x - \widehat{z}^k_{\{m_i\}_{i=1}^k}\right)^T \widehat{Q}^k_{\{m_i\}_{i=1}^k}\left(x - \widehat{z}^k_{\{m_i\}_{i=1}^k}\right) + \widehat{r}^k_{\{m_i\}_{i=1}^k}.$$

Then, for each m_{k+1},

$$\begin{aligned}
\widehat{a}^{k+1}_{\{m_i\}_{i=1}^{k+1}} &= \max_{z\in\mathbb{R}^n}\left\{\mathcal{B}_\tau^{m_{k+1}}(x,z) + \widehat{a}^k_{\{m_i\}_{i=1}^k}(z)\right\} \\
&= \max_{z\in\mathbb{R}^n}\Bigg\{ \tfrac{1}{2}\left[x^T M_{1,1}^m x + x^T M_{1,2}^m z + z^T (M_{1,2}^m)^T x + z^T M_{2,2}^m z\right] \\
&\quad + \tfrac{1}{2}\left(x - \widehat{z}^k_{\{m_i\}_{i=1}^k}\right)^T \widehat{Q}^k_{\{m_i\}_{i=1}^k}\left(x - \widehat{z}^k_{\{m_i\}_{i=1}^k}\right) + \widehat{r}^k_{\{m_i\}_{i=1}^k}\Bigg\} \\
&= \tfrac{1}{2}\left(x - \widehat{z}^{k+1}_{\{m_i\}_{i=1}^{k+1}}\right)^T \widehat{Q}^{k+1}_{\{m_i\}_{i=1}^{k+1}}\left(x - \widehat{z}^{k+1}_{\{m_i\}_{i=1}^{k+1}}\right) + \widehat{r}^{k+1}_{\{m_i\}_{i=1}^{k+1}}, \qquad (76)
\end{aligned}$$

where

(77)
$$\begin{aligned}
\widehat{Q}^{k+1}_{\{m_i\}_{i=1}^{k+1}} &= M_{1,1}^{m_{k+1}} - M_{1,2}^{m_{k+1}} \widehat{D}\left(M_{1,2}^{m_{k+1}}\right)^T, \\
\widehat{z}^{k+1}_{\{m_i\}_{i=1}^{k+1}} &= -\left(\widehat{Q}^{k+1}_{\{m_i\}_{i=1}^{k+1}}\right)^{-1} M_{1,2}^{m_{k+1}} \widehat{E}, \\
\widehat{r}^{k+1}_{\{m_i\}_{i=1}^{k+1}} &= \widehat{r}^k_{\{m_i\}_{i=1}^k} + \tfrac{1}{2}\widehat{E}^T M_{2,2}^m \widehat{z}^k_{\{m_i\}_{i=1}^k} - \tfrac{1}{2}\left(\widehat{z}^{k+1}_{\{m_i\}_{i=1}^{k+1}}\right)^T \widehat{Q}^{k+1}_{\{m_i\}_{i=1}^{k+1}} \widehat{z}^{k+1}_{\{m_i\}_{i=1}^{k+1}}, \\
\widehat{D} &= \left(M_{2,2}^{m_{k+1}} + \widehat{Q}^k_{\{m_i\}_{i=1}^k}\right)^{-1}, \\
\widehat{E} &= \widehat{D}\widehat{Q}^k_{\{m_i\}_{i=1}^k} \widehat{z}^k_{\{m_i\}_{i=1}^k}.
\end{aligned}$$

Thus we have the analytical expression for the propagation of each (quadratic) $\widehat{a}^k_{\{m_i\}_{i=1}^k}$ function. Specifically, we see that the propagation of each $\widehat{a}^k_{\{m_i\}_{i=1}^k}$ amounts to a set of matrix multiplications (and an inverse). Note that for the purely quadratic constituent Hamiltonians considered here (without terms that are linear or constant in

the state and gradient variables), one will have $\widehat{z}^k_{\{m_i\}_{i=1}^k} = 0$ and $\widehat{r}^k_{\{m_i\}_{i=1}^k} = 0$, and so computation of these terms is not necessary (unless one adds linear and/or constant terms).

At each step k, the semiconvex dual $\overline{a}^k$ of $\overline{V}^k$ is represented as the finite set of functions

$$\widehat{\mathcal{A}}_k \doteq \left\{ \widehat{a}^k_{\{m_i\}_{i=1}^k} \,|\, m_i \in \mathcal{M} \ \forall i \in \{1, 2, \ldots, k\} \right\},$$

where this is equivalently represented as the set of triples

$$\widehat{\mathcal{Q}}_k \doteq \left\{ \left(\widehat{Q}^k_{\{m_i\}_{i=1}^k}, \widehat{z}^k_{\{m_i\}_{i=1}^k}, \widehat{r}^k_{\{m_i\}_{i=1}^k} \right) | \, m_i \in \mathcal{M} \ \forall i \in \{1, 2, \ldots, k\} \right\}.$$

At any desired stopping time, one can recover a representation of $\overline{V}^k$ as

$$\widehat{\mathcal{V}}_k \doteq \left\{ \widehat{V}^k_{\{m_i\}_{i=1}^k} \,|\, m_i \in \mathcal{M} \ \forall i \in \{1, 2, \ldots, k\} \right\},$$

where these $\widehat{V}^k_{\{m_i\}_{i=1}^k}$ are also quadratics. In fact, recall

$$\begin{aligned}
\overline{V}^k(x) &= \max_{z \in \mathbb{R}^n} [\overline{a}^k(z) + \psi(x, z)] \\
&= \max_{\{m_i\}_{i=1}^k} \max_{z \in \mathbb{R}^n} \left[\tfrac{1}{2} (z - \widehat{z}^k_{\{m_i\}_{i=1}^k})^T \widehat{Q}^k_{\{m_i\}_{i=1}^k} (z - \widehat{z}^k_{\{m_i\}_{i=1}^k}) + \widehat{r}^k_{\{m_i\}_{i=1}^k} + \tfrac{c}{2} |x - z|^2 \right] \\
&\doteq \max_{\{m_i\}_{i=1}^k} \tfrac{1}{2} (x - \widehat{x}^k_{\{m_i\}_{i=1}^k})^T \widehat{P}^k_{\{m_i\}_{i=1}^k} (x - \widehat{x}^k_{\{m_i\}_{i=1}^k}) + \widehat{\rho}^k_{\{m_i\}_{i=1}^k} \\
&\doteq \bigoplus_{\{m_i\}_{i=1}^k} \widehat{V}^k_{\{m_i\}_{i=1}^k}(x),
\end{aligned}$$

where with $C \doteq cI$,

(78)

$$\begin{aligned}
\widehat{P}^k_{\{m_i\}_{i=1}^k} &= C\widehat{F}\widehat{Q}^k_{\{m_i\}_{i=1}^k}\widehat{F}C + (\widehat{F}C - I)^T C(\widehat{F}C - I), \\
\widehat{x}^k_{\{m_i\}_{i=1}^k} &= -\left(\widehat{P}^k_{\{m_i\}_{i=1}^k}\right)^{-1} \left[C\widehat{F}\widehat{Q}^k_{\{m_i\}_{i=1}^k}\widehat{G} + (\widehat{F}C - I)^T C\widehat{F}\widehat{Q}^k_{\{m_i\}_{i=1}^k} \right] \widehat{z}^k_{\{m_i\}_{i=1}^k}, \\
\widehat{\rho}^k_{\{m_i\}_{i=1}^k} &= \widehat{r}^k_{\{m_i\}_{i=1}^k} + \tfrac{1}{2} \left(\widehat{z}^k_{\{m_i\}_{i=1}^k}\right)^T \left[\widehat{G}^T \widehat{Q}^k_{\{m_i\}_{i=1}^k} \widehat{G} + \widehat{Q}^k_{\{m_i\}_{i=1}^k} \widehat{F}C\widehat{F}\widehat{Q}^k_{\{m_i\}_{i=1}^k} \right] \widehat{z}^k_{\{m_i\}_{i=1}^k}, \\
\widehat{F} &\doteq (\widehat{Q}^k_{\{m_i\}_{i=1}^k} + C)^{-1},
\end{aligned}$$

and

$$\widehat{G} \doteq (\widehat{F}\widehat{Q}^k_{\{m_i\}_{i=1}^k} - I).$$

Thus, $\overline{V}^k$ has the representation as the set of triples

$$\mathcal{P}_k \doteq \left\{ \left(\widehat{P}^k_{\{m_i\}_{i=1}^k}, \widehat{x}^k_{\{m_i\}_{i=1}^k}, \widehat{\rho}^k_{\{m_i\}_{i=1}^k} \right) | \, m_i \in \mathcal{M} \ \forall i \in \{1, 2, \ldots, k\} \right\}. \tag{79}$$

We note that the triples which comprise $\mathcal{P}_k$ are obtained from the triples

$$\left(\widehat{Q}^k_{\{m_i\}_{i=1}^k}, \widehat{z}^k_{\{m_i\}_{i=1}^k}, \widehat{r}^k_{\{m_i\}_{i=1}^k} \right)$$

by matrix multiplications and an inverse. The transference from triples

$$\left(\widehat{Q}^k_{\{m_i\}_{i=1}^k}, \widehat{z}^k_{\{m_i\}_{i=1}^k}, \widehat{r}^k_{\{m_i\}_{i=1}^k}\right)$$

to triples $(\widehat{P}^k_{\{m_i\}_{i=1}^k}, \widehat{x}^k_{\{m_i\}_{i=1}^k}, \widehat{\rho}^k_{\{m_i\}_{i=1}^k})$ need only be done once, which is at the termination of the algorithm propagation. Again, in the purely quadratic class of problems addressed here, and with the pure quadratic initialization, the $\widehat{x}^k_{\{m_i\}_{i=1}^k}$ and $\widehat{\rho}^k_{\{m_i\}_{i=1}^k}$ terms will be zero. We note that (79) is our approximate solution of the original control problem/HJB PDE.

The errors are due to our approximation of $\widetilde{V}$ by $\overline{V}$ (see Theorem 4.12 and Remark 4.13) and to the approximation of $\overline{V}$ by the prelimit $\overline{V}^N$ for stopping time $k = N$. Neither of these errors is related to the space dimension. The errors in $|\widetilde{V} - \overline{V}|$ are dependent on the step size τ. The errors in $|\overline{V}^N - \overline{V}| = |\bar{\bar{S}}^\tau_{N\tau}[0] - \overline{V}|$ are due to premature termination in the limit $\overline{V} = \lim_{N\to\infty} \bar{\bar{S}}^\tau_{N\tau}[0]$. The computation of each triple $(\widehat{P}^k_{\{m_i\}_{i=1}^k}, \widehat{x}^k_{\{m_i\}_{i=1}^k}, \widehat{\rho}^k_{\{m_i\}_{i=1}^k})$ grows as the cube of the space dimension (due to the matrix operations). Thus one avoids the curse of dimensionality. Of course if one then chooses to compute $\overline{V}^N(x)$ for all x on some grid over, say, a rectangular region in $\mathbb{R}^n$, then by definition one has exponential growth in this computation as the space dimension increases. We stress that one does not need to compute $\overline{V}^N \simeq \widetilde{V}$ at each such point.

However, the curse of dimensionality is replaced by another type of rapid computational cost growth. Here, we refer to this as the curse of complexity. If $\#\mathcal{M} = 1$, then all the computations of our algorithm (excepting the solution of the Riccati equation) are unnecessary, and we *informally* refer to this as complexity one. When there are $M = \#\mathcal{M}$ such quadratics in the Hamiltonian, $\widetilde{H}$, we say it has complexity M. Note that

$$\#\left\{\widehat{V}^k_{\{m_i\}_{i=1}^k} \,|\, m_i \in \mathcal{M}\ \forall i \in \{1, 2, \ldots, k\}\right\} \sim M^N.$$

For large N, this is indeed a large number. (We very briefly discuss means for reducing this in the next section.) Nevertheless, for small values of M, we obtain a very rapid solution of such nonlinear HJB PDEs, as will be indicated in the examples to follow. Further, the computational cost growth in space dimension n is limited to cubic growth. We emphasize that the existence of an algorithm avoiding the curse of dimensionality is significant regardless of the practical issues.

6. Practical issues. The bulk of this paper develops an algorithm which avoids the curse of dimensionality. However, the curse of complexity is also a formidable barrier. The purpose of the paper is to bring to light the existence of this class of algorithms. Considering the long development of finite element methods, it is clear that the development of highly efficient methods from this new class could be a further substantial achievement. (Nevertheless, some impressive computational times are indicated in the next section.) In this section, we briefly indicate some practical heuristics that have been helpful and outline the actual steps in an implementation of the basic algorithm.

6.1. Pruning. The number of quadratics in $\mathcal{Q}_k$ grows exponentially in k. However, in practice (for the cases we have tried) we have found that relatively few of these actually contribute to $\overline{V}^k$. Thus it would be very useful to prune the set.

Note that if

$$\widehat{a}^k_{\{\hat{m}_i\}_{i=1}^k}(x) \le \bigoplus_{\{m_i\}_{i=1}^k \neq \{\hat{m}_i\}_{i=1}^k} \widehat{a}^k_{\{m_i\}_{i=1}^k}(x) \qquad \forall\, x \in \mathbb{R}^n, \tag{80}$$

then

$$\int_{\mathbb{R}^n}^{\oplus} \overline{\mathcal{B}}_\tau(x,z) \otimes \overline{a}^k(z)\, dz \le \int_{\mathbb{R}^n}^{\oplus} \overline{\mathcal{B}}_\tau(x,z) \otimes \left[\bigoplus_{\{m_i\}_{i=1}^k \neq \{\hat{m}_i\}_{i=1}^k} \widehat{a}^k_{\{m_i\}_{i=1}^k}(z) \right] dz.$$

Consequently $\widehat{a}^k_{\{\hat{m}_i\}_{i=1}^k}$ will play no role whatsoever in the computation of $\overline{V}^k$. Further, it is easy to show that the progeny of $\widehat{a}^k_{\{\hat{m}_i\}_{i=1}^k}$ (i.e., those $\widehat{a}^{k+j}_{\{m_i\}_{i=1}^{k+j}}$ for which $\{m_i\}_{i=1}^k = \{\hat{m}_i\}_{i=1}^k$) never contribute either. Thus, one may prune such $\widehat{a}^k_{\{\hat{m}_i\}_{i=1}^k}$ without any loss of accuracy. This shrinks not only the current $\mathcal{Q}_k$, but also the growth of the future $\mathcal{Q}_{k+j}$.

In the examples to follow, we pruned $\widehat{a}^k_{\{\hat{m}_i\}_{i=1}^k}$ if there existed a single sequence $\{\tilde{m}_i\}_{i=1}^k$ such that $\widehat{a}^k_{\{\hat{m}_i\}_{i=1}^k}(x) \le \widehat{a}^k_{\{\tilde{m}_i\}_{i=1}^k}(x)$ for all x. This significantly reduced the growth in the size of $\mathcal{Q}_k$. However, it clearly failed to prune anywhere near the number of elements that could be pruned according to condition (80), and thus much greater computational reduction might be possible. This would require an ability to determine when a quadratic was dominated by the maximum of a set of other quadratic functions.

Also in the examples to follow, an additional heuristic pruning technique was applied for a number of iterations to delay hitting the curse of complexity growth rate. A function $\widehat{a}^k_{\{m_i\}_{i=1}^k}$ was pruned if it did not dominate at least one of the corners of the unit cube. Specifically, let $\mathcal{C} = \{x^j\}$ be the corners of the unit cube. The set of functions was pruned down to a subset of $L \le 2^n$ functions, $\{\widehat{a}^k_{\{\hat{m}_i^l\}_{i=1}^k} \,|\, l \le L\}$, such that $\overline{a}^k(x^j) = \max_{l \le L} \widehat{a}^k_{\{\hat{m}_i^l\}_{i=1}^k}(x^j)$ for all $x^j \in \mathcal{C}$. This introduces a component of the calculations which is subject to curse of dimensionality growth, but in the examples run so far it reduced the computations over what they were needed without the heuristic. (Also, the curse of dimensionality growth due to this heuristic is 2^n rather than on the order of 200^n, as in the discussion of other methods in section 1.)

6.2. Initialization. It is also easy to see that one may initialize with an arbitrary quadratic function less than an $\overline{a}^k(x)$ rather than with $\overline{a}^0 \equiv 0$. Significant savings were obtained by initializing with a set of $M = \#\mathcal{M}$ quadratics, $\{a^m(x)\}$ where the a^m were the convex duals of the V^m (which were each obtained by the solution of the corresponding Riccati equation). With $\overline{a}^0(z) \doteq \bigoplus_{m \in \mathcal{M}} a^m(z)$, one starts much closer to the final solution, and so the number of steps where one is encountering the curse of complexity is greatly reduced.

6.3. Pseudocode for the algorithm. In this short section, we briefly indicate the actual steps that one would code in an instantiation of the algorithm.

1. Choose a time-step size, τ, and number of steps, K. (We do not address error analysis and stopping-time criteria in this paper.)
2. For each $m \in \mathcal{M}$, compute P^m_τ from (71). Next, for each $m \in \mathcal{M}$, compute $M^m_{1,1}$, $M^m_{1,2}$, and $M^m_{2,2}$ from (73), (74), and (75), respectively. These are used in each iteration update below.

3. Initialize the iteration. One may initialize with $\overline{a}^0(x) \doteq 0$, which is $\widehat{\mathcal{Q}}_0 = \{\widehat{Q}_1^0\}$ with $\widehat{Q}_1^0 = 0$ (the $n \times n$ matrix of zeros). Note that in this pseudocode, we will index the $\widehat{Q}^k$ by a generic subscript rather than by the sequences $\{m_i\}_{i=1}^k$, as this is more convenient in software. Although this is a simple initialization, the computational time is hugely improved through the use of the initialization described in section 6.2. In this latter case, we first compute (approximately) the $P_\infty^m \doteq \lim_{t\to\infty} P_t^m$ from (71). The initialization is then $\widehat{\mathcal{Q}}_0 = \{\widehat{Q}_j^0\}_{j=1}^M$, where each $\widehat{Q}_j^0$ is obtained from the corresponding P_∞^j, by the dual operation, and in particular is given by

$$\widehat{Q}_j^0 = C(C - P_\infty^j)^{-1} P_\infty^j (C - P_\infty^j)^{-1} C - P_\infty^j (C - P_\infty^j)^{-1} C (C - P_\infty^j)^{-1} P_\infty^j.$$

4. Perform the basic iteration step. That is, given $\widehat{\mathcal{Q}}_k = \{\widehat{Q}_j^k\}_{j=1}^{J_k}$, compute $\widehat{\mathcal{Q}}_{k+1}$ as follows:
   (a) Start with $j = 1$ and $m = 1$. Let $\ell = 1$.
   (b) (iteration-subloop): Obtain $\widehat{Q}_\ell^{k+1}$ from update equation (78), that is,

$$\widehat{Q}_\ell^{k+1} = M_{1,1}^m - M_{1,2}^m \left(M_{2,2}^m + \widehat{Q}_j^k \right)^{-1} \left(M_{1,2}^m \right)^T.$$

   (c) Let $\ell = \ell + 1$. If $m < M$, set $m = m + 1$, and go to step 4(b). If $m = M$ and $j < J_k$, set $m = 1$, $j = j + 1$, and go to step 4(b). If $m = M$ and $j = J_k$, set $J_{K+1} = \ell - 1$; we are done with the iteration step.
5. Repeat step 4 K times.
6. Recover the solution approximation from the dual matrices. That is, given $\widehat{\mathcal{Q}}_K = \{\widehat{Q}_j^K\}_{j=1}^{J_K}$, compute $\mathcal{P}_k = \{\widehat{P}_j^k\}_{j=1}^{J_K}$ from (78). The solution approximation is the pointwise maximum $\overline{V}(x) = \max_{j \le J_K} \frac{1}{2} x^T \widehat{P}_j^K x$.

Remark 6.1. We emphasize that pruning techniques, such as those of section 6.1 are critical to rapid computational rates, but this is still an open area of research, and we leave instantiation of such to the intrepid researcher.

7. Examples. A number of examples have so far been tested. In these tests, the computational speeds were very great. This is due to the fact that $M = \#\mathcal{M}$ was small. The algorithm as described above was coded in MATLAB. This includes the very simple pruning technique and initialization discussed in the previous section. The quoted computational times were obtained with a standard 2001 PC. The times correspond to the times to compute $\mathcal{V}_N$ or, equivalently, $\mathcal{P}_N$. The plots below require one to compute the value function and/or gradients pointwise on planes in the state space. These plotting computations are not included in the quoted computational times.

We will briefly indicate the results of three similar examples with state-space dimensions of 2, 3, and 4. The number of constituent linear/quadratic Hamiltonians for each of them is 3. The structures of the dynamics are similar for each of them so as to focus on the change in dimension.

Example 1. The first case has constituent Hamiltonians with the A^m given by

$$A^1 = \begin{bmatrix} -1.0 & 0.5 \\ 0.1 & -1.0 \end{bmatrix}, \quad A^2 = (A^1)^T, \quad A^3 = \begin{bmatrix} -1.0 & 0.5 \\ 0.5 & -1.9 \end{bmatrix}.$$

The D^m and Σ^m are simply

$$D^1 = D^2 = D^3 = \begin{bmatrix} 1.5 & 0.2 \\ 0.2 & 1.5 \end{bmatrix}$$

and

$$\Sigma^1 = \Sigma^2 = \Sigma^3 = \begin{bmatrix} 0.27 & -0.01 \\ -0.01 & 0.27 \end{bmatrix}.$$

Figure 1 depicts the value function and first partial derivative (computed by a simple first-difference on the grid points) over the region $[-1, 1] \times [-1, 1]$. Note the discontinuity in the first partial along one of the diagonals. Figure 2 depicts the second partial and a backsubstitution error over the same region. The second partial also has a discontinuity along the same diagonal as the first. The error plot has been rotated for better viewing due to the high error along the discontinuity in the gradient. The backsubstitution error is computed by taking these approximate partials and substituting them back into the original HJB PDE. Consequently, the depicted errors contain components due to the approximate gradient dotted with the dynamics, and the term with the square in the gradient in the Hamiltonian. Perhaps it should be noted that the solutions of such problems *cannot* be obtained by patching together the quadratic functions corresponding to solutions of the corresponding algebraic Riccati equations. The computations required slightly less than 10 seconds.

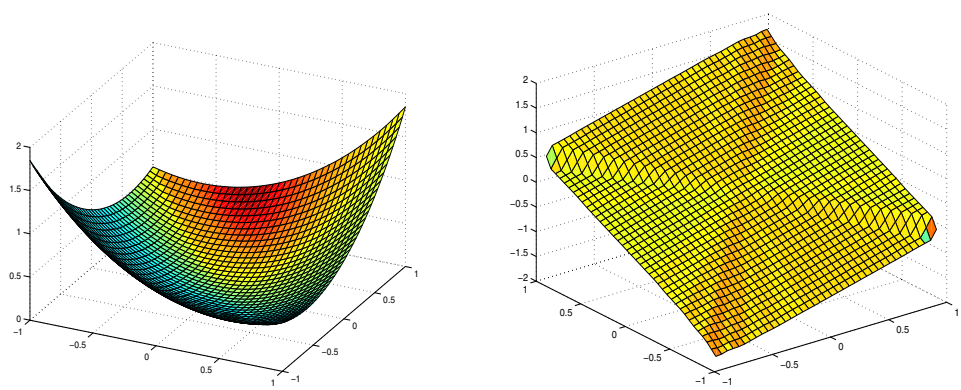


FIG. 1. *Value function and first partial (two-dimensional case).*

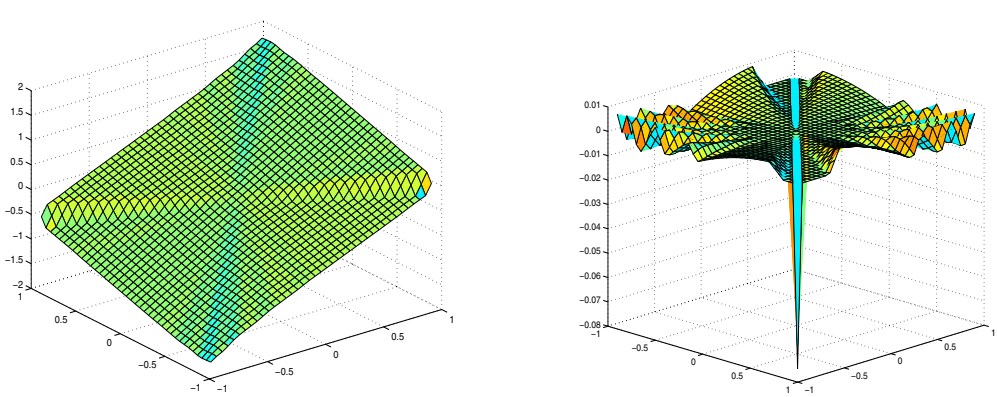


FIG. 2. *Second partial and backsubstitution error (two-dimensional case).*

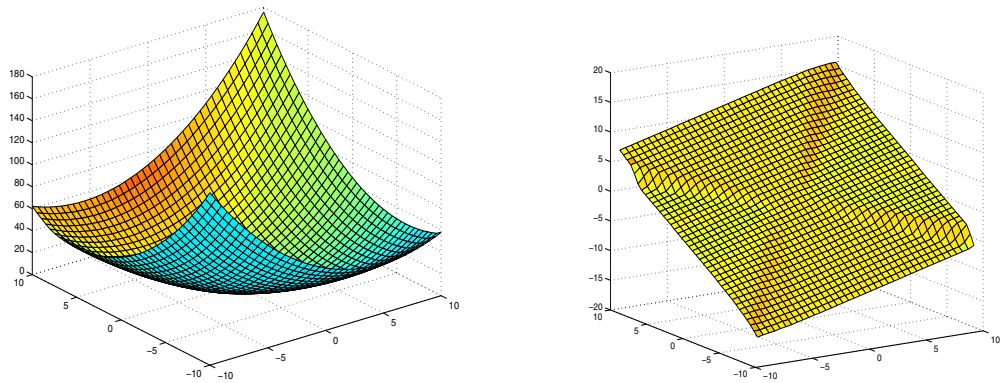


FIG. 3. *Value function and first partial (three-dimensional case).*

Example 2. We now consider a case where the A^m are given by

$$A^1 = \begin{bmatrix} -1.0 & 0.5 & 0.0 \\ 0.1 & -1.0 & 0.2 \\ 0.2 & 0.0 & -1.5 \end{bmatrix}, \quad A^2 = (A^1)^T, \quad A^3 = \begin{bmatrix} -1.0 & 0.5 & 0.0 \\ 0.1 & -1.0 & 0.2 \\ 0.2 & 0.0 & -1.5 \end{bmatrix},$$

the D^m are

$$D^1 = \begin{bmatrix} 1.5 & 0.2 & 0.1 \\ 0.2 & 1.5 & 0.0 \\ 0.1 & 0.0 & 1.5 \end{bmatrix}, \quad D^2 = \begin{bmatrix} 1.6 & 0.2 & 0.1 \\ 0.2 & 1.6 & 0.0 \\ 0.1 & 0.0 & 1.6 \end{bmatrix}, \quad D^3 = D^1,$$

and the Σ^m are

$$\Sigma^1 = \begin{bmatrix} 0.2 & -0.01 & 0.02 \\ -0.01 & 0.2 & 0.0 \\ 0.02 & 0.0 & 0.25 \end{bmatrix}, \quad \Sigma^2 = \begin{bmatrix} 0.16 & -0.005 & 0.015 \\ -0.005 & 0.16 & 0.0 \\ 0.015 & 0.0 & 0.2 \end{bmatrix}, \quad \Sigma^3 = \Sigma^1.$$

The results of this three-dimensional example appear in Figures 3–5. In this case, the results have been plotted over the region of the affine plane $x_3 = 3$ given by $x_1 \in [-10, 10]$ and $x_2 \in [-10, 10]$. The backsubstitution error has been scaled by dividing by $|x|^2 + 10^{-5}$. Note that the scaled backsubstitution errors (away from the discontinuity in the gradient) grow only slowly or are possibly bounded with increasing $|x|$. (Recall that the approximate solution is obtained over the whole space.) Since the gradient errors are multiplied by the nominal dynamics in one component of this term (as well as being squared in another), this indicates that the errors in the gradient itself likely grow only linearly (or nearly linearly) with increasing $|x|$. The computations required approximately 13 seconds.

Example 3. The four-dimensional example has constituent Hamiltonians with the A^m, D^m, and Σ^m given by

$$A^1 = \begin{bmatrix} -1.0 & 0.5 & 0.0 & 0.1 \\ 0.1 & -1.0 & 0.2 & 0.0 \\ 0.2 & 0.0 & -1.5 & 0.1 \\ 0.0 & -0.1 & 0.0 & -1.5 \end{bmatrix}, \quad A^2 = (A^1)^T,$$

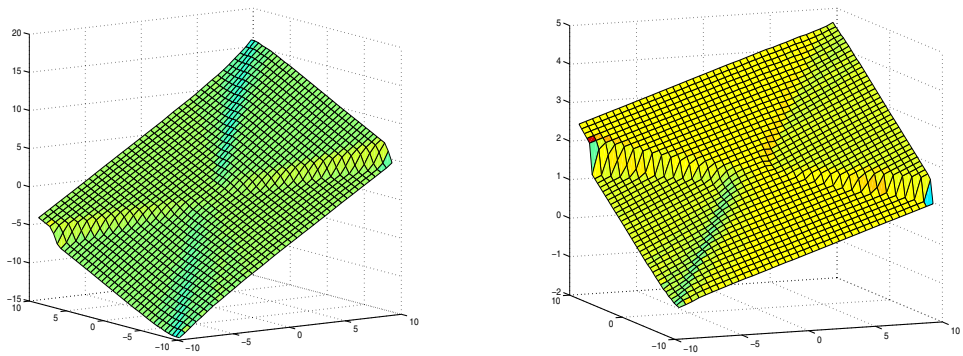


FIG. 4. *Second and third partials (three-dimensional case).*

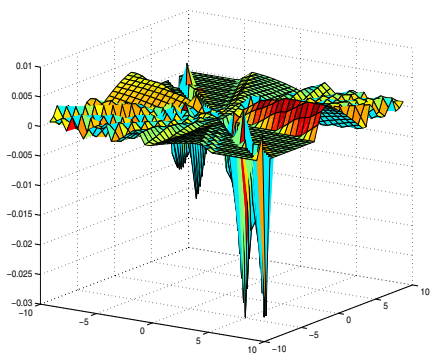


FIG. 5. *Scaled backsubstitution error (three-dimensional case).*

$$A^3 = \begin{bmatrix} -1.0 & 0.5 & 0.0 & 0.1 \\ 0.1 & -1.0 & 0.2 & 0.0 \\ 0.2 & 0.0 & -1.6 & -0.1 \\ 0.0 & -0.05 & 0.1 & -1.5 \end{bmatrix},$$

$$D^1 = D^2 = D^3 = \begin{bmatrix} 1.5 & 0.2 & 0.1 & 0.0 \\ 0.2 & 1.5 & 0.0 & 0.1 \\ 0.1 & 0.0 & 1.5 & 0.0 \\ 0.0 & 0.1 & 0.0 & 1.5 \end{bmatrix},$$

and

$$\Sigma^1 = \Sigma^2 = \Sigma^3 = \begin{bmatrix} 0.2 & -0.01 & 0.02 & 0.01 \\ -0.01 & 0.2 & 0.0 & 0.0 \\ 0.02 & 0.0 & 0.25 & 0.0 \\ 0.01 & 0.0 & 0.0 & 0.25 \end{bmatrix}.$$

The results for this example appear in Figures 6–8. In this case, the results have been plotted over the region of the affine plane $x_3 = 3$, $x_4 = -0.5$ given by $x_1 \in [-10, 10]$ and $x_2 \in [-10, 10]$. The backsubstitution error has again been scaled by dividing by $|x|^2 + 10^{-5}$. The computations required approximately 40 seconds. We remark that one cannot change dimension independent of dynamics (except in

the trivial case, where each component of the system has exactly the same dynamics of the other components with no interdependence), and so one cannot directly compare the computation times of these three examples. However, it is easy to see that the computation time increases are on the order of square to cubic in space dimension, rather than being subject to curse-of-dimensionality-type growth.

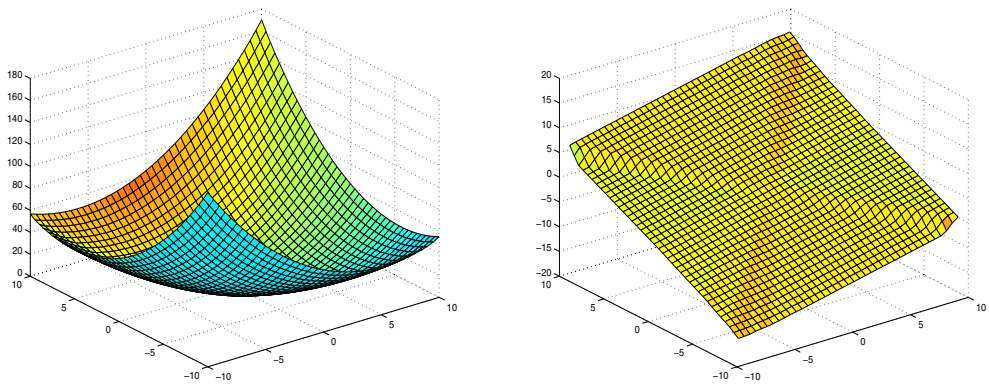


FIG. 6. *Value function and first partial (four-dimensional case).*

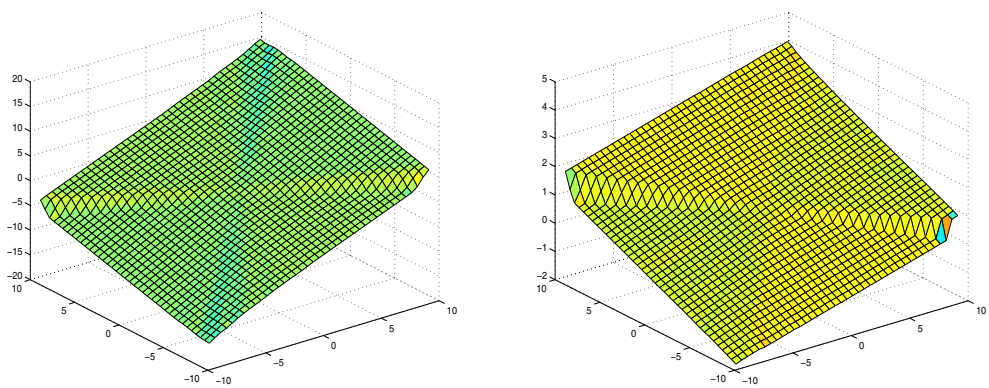


FIG. 7. *Second and third partials (four-dimensional case).*

8. Future directions.

Pruning. In order to make these methods more practical, algorithms need to be developed for determining when a quadratic function is dominated by the function, which is the pointwise maximum of a set of quadratic functions. This has the potential for greatly reducing the effects of the curse of complexity, and consequently greatly decreasing computational times.

Constant/linear terms. An instantiation of this class of methods was developed here for a very particular type of Hamiltonian, $\widetilde{H}(x,p) = \max_m\{H^m(x,p)\}$, where the H^m corresponded to a very specific type of linear/quadratic problem. One

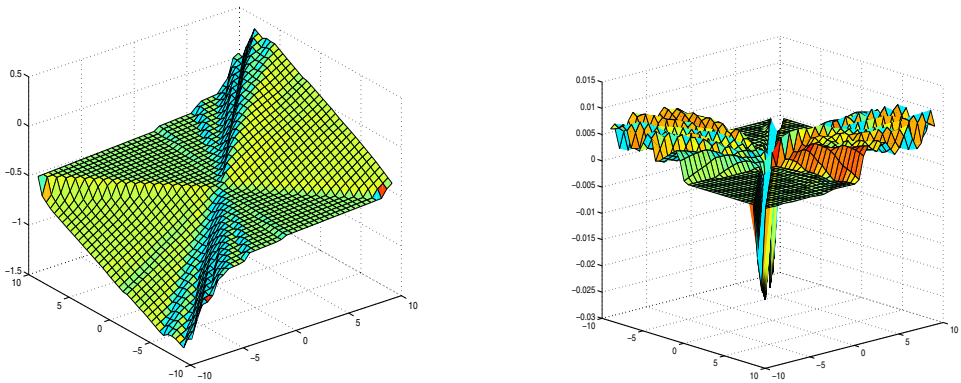


FIG. 8. *Fourth partial and scaled backsubstitution error (four-dimensional case).*

would like to generalize the H^m to, say,

$$H^m(x,p) = \tfrac{1}{2}x^T D^m x + \tfrac{1}{2}p^T \Sigma^m p + (A^m x)^T p + (l_1^m)^T x + (l_2^m)^T p + \alpha^m.$$

Clearly certain conditions on $\widetilde{H}(x,p) = \max_m\{H^m(x,p)\}$ would be necessary. It is not obvious that these conditions would need to apply to each of the constituent H^m individually. In the work here, the H^m corresponded to linear/quadratic problems with maximizing controllers/disturbances. It is not clear that the constituent linear/quadratic problems need to be constricted in this way either. For instance, could some or all of the H^m correspond to, say, game problems?

Convergence/error analysis. Only convergence of the approximation to the solution was obtained here. Estimates of error size and convergence rate need to be determined. For instance, it was hypothesized (and observed in the examples) that one obtains the solution over the whole state space with linear growth rate in the errors in the gradient. Is this true in any generality?

Nonergodic problem. The algorithm was developed for an infinite time-horizon problem, where the dynamics were stable to the origin. One expects the approach would also be applicable to discounted cost problems and exit problems. One would also expect that a similar theory could be developed for finite time-horizon problems such as robust filtering. Max-plus methods have also been discussed for problems corresponding to variational inequalities [30]. The analysis and algorithm necessary for a variational inequality would be of interest.

Other nonlinearities. This work concentrated only on the case of a nonlinearity due to taking the maximum of a set of Hamiltonians for linear/quadratic problems. An obvious question is how well this approach might work for other classes of nonlinearities. What classes of nonlinear HJB PDEs could be best approximated by maxima over reasonably small numbers of linear/quadratic HJB PDEs? Perhaps a single nonlinearity in only one variable (possibly appearing in multiple places) would be the most tractable?

Appendix A. Sketch of proof of Theorem 4.7. Fix $\delta > 0$ (used in the definition of $\mathcal{G}_\delta$). Suppose $\overline{V}' \in \mathcal{G}_\delta$ satisfies (46). Then,

$$\begin{aligned}\overline{V}'(x) &= \bar{\bar{S}}^\tau_{N\tau}[\overline{V}'](x)\\ &= \sup_{w\in\mathcal{W}}\sup_{\mu\in\mathcal{D}^\tau_\infty}\left\{\int_0^{N\tau} l^{\mu_t}(\xi_t) - \frac{\gamma^2}{2}|w_t|^2\,dt + \overline{V}'(\xi_{N\tau})\right\} \qquad \forall\, x\in\mathbb{R}^n,\end{aligned}$$

where ξ satisfies (12). Fix $x\in\mathbb{R}^n$, and let $\mu^\varepsilon \in \mathcal{D}^\tau_\infty$, $w^\varepsilon \in \mathcal{W}$ be ε-optimal , i.e.,

$$\overline{V}'(x) \le \int_0^{N\tau} l^{\mu^\varepsilon_t}(\xi^\varepsilon_t) - \frac{\gamma^2}{2}|w^\varepsilon_t|^2\,dt + \overline{V}'(\xi^\varepsilon_{N\tau}) + \varepsilon,$$

where ξ^ε satisfies (12) with inputs $\mu^\varepsilon, w^\varepsilon$.

Following the same steps as in [35], one obtains the same lemmas.

LEMMA A.1. *For any $N<\infty$, $\|w^\varepsilon\|^2_{L_2(0,N\tau)} \le \frac{\varepsilon}{\delta} + \frac{1}{\delta}\left[\frac{c_A\gamma^2}{c_\sigma^2}e^{-c_AN\tau} + \frac{c_D}{c_A}\right]|x|^2$.*

LEMMA A.2. *For any $N<\infty$,*

$$\int_0^{N\tau}|\xi^\varepsilon_t|^2\,dt \le \frac{\varepsilon}{\delta}\frac{c_\sigma^2}{c_A} + \frac{c_\sigma^2}{\delta}\left[\left(\frac{c_D}{c_A^2}+\frac{\gamma^2}{c_\sigma^2}\right)+\frac{1}{c_A}\right]|x|^2.$$

LEMMA A.3. *If $w^\varepsilon, \mu^\varepsilon$ are ε-optimal over $[0, N\tau)$, then they are also ε-optimal over $[0, n\tau)$ for all $n \le N$, i.e.,*

$$\int_0^{n\tau} l^{\mu^\varepsilon_t}(\xi^\varepsilon_t) - \frac{\gamma^2}{2}|w^\varepsilon_t|^2\,dt + \overline{V}'(\xi^\varepsilon_{n\tau}) \ge \overline{V}'(x) - \varepsilon.$$

The independence of the above bounds with respect to N is important. Specifically, since there is a finite bound on the energy (the bound on w^ε) coming into the trajectories, roughly speaking the ξ^ε "tend" toward the origin.

Now we need a lemma which will replace equation (20) in [35].

Lemma A.4. *For any $N<\infty$,*

$$\sum_{n=1}^N |\xi^\varepsilon_{n\tau}|^2 \le \frac{1}{1-e^{-c_A\tau}}\left[|x|^2 + \left(\frac{c_\sigma}{c_A^2}\right)\|w^\varepsilon\|^2_{L_2(0,N\tau)}\right].$$

Proof. Note that $\frac{d}{dt}|\xi^\varepsilon|^2 \le -c_A|\xi^\varepsilon| + \hat{d}|w^\varepsilon|^2$ with $\hat{d} = c_\sigma^2/c_A$. Solving this on intervals of the form $[n\tau,(n+1)\tau)$, one finds

$$\begin{aligned}|\xi^\varepsilon_\tau|^2 &\le |x|^2e^{-c_A\tau} + \hat{d}\|w^\varepsilon\|^2_{L_2(0,\tau)},\\ |\xi^\varepsilon_{2\tau}|^2 &\le |\xi^\varepsilon_\tau|^2e^{-c_A\tau} + \hat{d}\|w^\varepsilon\|^2_{L_2(\tau,2\tau)},\end{aligned}$$

and so on. Continuing this process, and combining the inequalities, yields

$$\sum_{n=1}^N |\xi^\varepsilon_{n\tau}|^2 \le \left(\sum_{n=1}^N e^{-nc_A\tau}\right)|x|^2 + \hat{d}\sum_{n=1}^N\left[\left(\sum_{j=0}^{N-n} e^{-jc_A\tau}\right)\|w^\varepsilon\|^2_{L_2((n-1)\tau,n\tau)}\right].$$

Using the standard geometric series limit yields the result. ∎

Combining Lemmas A.2 and A.4, one obtains a bound on $\sum_{n=1}^N |\xi^\varepsilon_{n\tau}|^2$ which is independent of N. Consequently, at least some of the $|\xi^\varepsilon_{n\tau}|$ can be guaranteed to

be arbitrarily small for large N. The remainder of the proof (of Theorem 4.7) then follows as in equations (24)–(28) in [35], but with $N\tau$ replacing T, and $n\tau$ replacing τ. This completes the sketch of the proof. □

Appendix B. Sketch of proof of Lemma 4.11. Fix $\delta > 0$ (used in the definition of $\mathcal{G}_\delta$). Fix $m \in \mathcal{M}$. Fix any $T < \infty$ and $x \in \mathbb{R}^n$. Let $\varepsilon = (\hat{\varepsilon}/2)(1 + |x|^2)$. Let $w^\varepsilon \in \mathcal{W}$, $\mu^\varepsilon \in \mathcal{D}_\infty$ be ε-optimal for $\widetilde{S}_T[V^m](x)$, i.e.,

$$\widetilde{S}_T[V^m](x) - \left[\int_0^T l^{\mu^\varepsilon_t}(\xi^\varepsilon{}_t) - \frac{\gamma^2}{2}|w^\varepsilon_t|^2\,dt + V^m(\xi^\varepsilon{}_T)\right] \le \varepsilon = \frac{\hat{\varepsilon}}{2}(1 + |x|^2), \tag{81}$$

where ξ^ε satisfies (12) with inputs w^ε, μ^ε.

We will let $\overline{\xi}^\varepsilon$ satisfy (12) with inputs w^ε and a $\overline{\mu}^\varepsilon \in \mathcal{D}^\tau_\infty$ (where τ has yet to be chosen). Solving (12), one has

$$\begin{aligned}
\xi^\varepsilon_t &= \exp\left[\int_0^t A^{\mu^\varepsilon_r}\,dr\right]x + \int_0^t \exp\left[\int_r^t A^{\mu^\varepsilon_\rho}\,d\rho\right]\sigma^{\mu^\varepsilon_r} w^\varepsilon_r\,dr,\\
\overline{\xi}^\varepsilon_t &= \exp\left[\int_0^t A^{\overline{\mu}^\varepsilon_r}\,dr\right]x + \int_0^t \exp\left[\int_r^t A^{\overline{\mu}^\varepsilon_\rho}\,d\rho\right]\sigma^{\overline{\mu}^\varepsilon_r} w^\varepsilon_r\,dr.
\end{aligned}$$

Consequently,

$$\begin{aligned}
|\xi^\varepsilon_t - \overline{\xi}^\varepsilon_t| \le{}& \left|\exp\left[\int_0^t A^{\mu^\varepsilon_r}\,dr\right] - \exp\left[\int_0^t A^{\overline{\mu}^\varepsilon_r}\,dr\right]\right| |x|\\
&+ \left\{\int_0^t \left|\exp\left[\int_r^t A^{\mu^\varepsilon_\rho}\,d\rho\right]\sigma^{\mu^\varepsilon_r} - \exp\left[\int_r^t A^{\overline{\mu}^\varepsilon_\rho}\,d\rho\right]\sigma^{\overline{\mu}^\varepsilon_r}\right|^2 dr\right\}^{1/2} \|w^\varepsilon\|_{L_2(0,t)}.
\end{aligned} \tag{82}$$

We now simply show that this can be made arbitrarily small by taking τ small. We will use the boundedness of $\|w^\varepsilon\|$ and $\|\xi^\varepsilon\|$ which are independent of t for this class of systems [35].

Consider the first term on the right in (82). Note that

$$\begin{aligned}
&\left|\exp\left[\int_0^t A^{\mu^\varepsilon_r}\,dr\right] - \exp\left[\int_0^t A^{\overline{\mu}^\varepsilon_r}\,dr\right]\right|\\
&\quad= \left|\exp\left[\int_0^t A^{\mu^\varepsilon_r}\,dr\right]\right| \left|1 - \exp\left[\int_0^t A^{\overline{\mu}^\varepsilon_r}\,dr - \int_0^t A^{\mu^\varepsilon_r}\,dr\right]\right|.
\end{aligned} \tag{83}$$

Fix $\tau > 0$. For any subset of $\mathbb{R}$, $\mathcal{I}$, let $\mathcal{L}(\mathcal{I})$ be the Lebesgue measure of $\mathcal{I}$. Let N be the largest integer such that $N\tau \le t$. Given $m \in \mathcal{M}$, let

$$\mathcal{I}^m = \{r \in [0, N\tau) \mid A^{\mu^\varepsilon_r} = A^m\} \quad \text{and} \quad \lambda^m = \mathcal{L}(\mathcal{I}^m).$$

Let $n_0 = 0$. For $1 \le k < M = \#\mathcal{M}$, let n_k be the largest integer such that $n_k\tau \le \lambda^k + n_{k-1}\tau$. For $m < M$, let

$$\overline{\mu}^\varepsilon_r = m \qquad \forall\, t \in [n_{m-1}\tau, n_m\tau).$$

Let $\overline{\mu}^\varepsilon{}_r = M$ for all $t \in [n_{M-1}\tau, t) = [n_{M-1}\tau, N\tau) \cup [N\tau, t)$. With this choice of $\overline{\mu}^\varepsilon$, one finds

$$\left|1 - \exp\left[\int_0^t A^{\overline{\mu}^\varepsilon_r}\,dr - \int_0^t A^{\mu^\varepsilon_r}\,dr\right]\right| < \beta^1_\tau, \tag{84}$$

where $\beta_\tau^1 \to 0$ as $\tau \to 0$ independent of t. We skip the details.

Let $y \in \mathbb{R}^n$. Define $F_t = \exp\bigl[\int_0^t A^{\mu_r^\varepsilon}\,dr\bigr]$. Then, using assumption block $(A.m)$,

$$\begin{aligned}\frac{d}{dt}\bigl[y^T F_t^T F_t y\bigr] &= y^T\left[F_t^T \dot F_t + \dot F_t^T F_t\right] y = 2y^T\left[F_t^T A^{\mu_t^\varepsilon} F_t\right] y \\ &= 2(F_t y)^T A^{\mu_t^\varepsilon}(F_t y) \le -2c_A|F_t y|^2 = -2c_A\bigl[y^T F_t^T F_t y\bigr].\end{aligned}$$

Solving this ordinary differential inequality, one finds $[y^T F_t^T F_t y] \le |y|^2 e^{-2c_A t}$. Since this is true for all $y \in \mathbb{R}^n$, we have

$$\left|\exp\left[\int_0^t A^{\mu_r^\varepsilon}\,dr\right]\right| \le e^{-c_A t} \qquad \forall\, t \ge 0. \tag{85}$$

By (83), (84), and (85),

$$\left|\exp\left[\int_0^t A^{\mu_r^\varepsilon}\,dr\right] - \exp\left[\int_0^t A^{\overline{\mu}_r^\varepsilon}\,dr\right]\right| \le \beta_\tau^1 e^{-c_A t} \qquad \forall\, t \ge 0. \tag{86}$$

We now turn to the second term on the right-hand side of (82). Note that

$$\begin{aligned}&\left\{\int_0^t \left|\exp\left[\int_r^t A^{\mu_\rho^\varepsilon}\,d\rho\right]\sigma^{\mu_r^\varepsilon} - \exp\left[\int_r^t A^{\overline{\mu}_\rho^\varepsilon}\,d\rho\right]\sigma^{\overline{\mu}_r^\varepsilon}\right|^2 dr\right\}^{1/2} \\ &\quad\le \left\{2\int_0^t \left|\exp\left[\int_r^t A^{\mu_\rho^\varepsilon}\,d\rho\right]\right|^2 \left|\sigma^{\mu_r^\varepsilon} - \sigma^{\overline{\mu}_r^\varepsilon}\right|^2 dr \right. \\ &\qquad \left. +2\int_0^t \left|\exp\left[\int_r^t A^{\mu_\rho^\varepsilon}\,d\rho\right] - \exp\left[\int_r^t A^{\overline{\mu}_\rho^\varepsilon}\,d\rho\right]\right|^2 \left|\sigma^{\overline{\mu}_r^\varepsilon}\right|^2 dr\right\}^{1/2},\end{aligned}$$

and proceeding as above,

$$\begin{aligned}&\le \left\{2\int_0^t e^{-2c_A(t-r)}\left|\sigma^{\mu_r^\varepsilon} - \sigma^{\overline{\mu}_r^\varepsilon}\right|^2 dr + 2\beta_\tau^1\int_0^t e^{-2c_A(t-r)}\left|\sigma^{\overline{\mu}_r^\varepsilon}\right|^2 dr\right\}^{1/2} \\ &\le \left\{2\left[\int_0^t e^{-4c_A(t-r)}\,dr\right]^{1/2}\left[\int_0^t \left|\sigma^{\mu_r^\varepsilon} - \sigma^{\overline{\mu}_r^\varepsilon}\right|^4 dr\right]^{1/2} + 2\beta_\tau^1 c_\sigma^2\int_0^t e^{-2c_A(t-r)}\,dr\right\}^{1/2}.\end{aligned}$$

Further, there exists β_τ^2 such that $[\int_0^t |\sigma^{\mu_r^\varepsilon} - \sigma^{\overline{\mu}_r^\varepsilon}|^4\,dr]^{1/2} \le \beta_\tau^2$, where $\beta_\tau^2 \to 0$ as $\tau \to 0$, and we skip the obvious, but technical, proof. Consequently,

$$\begin{aligned}&\left\{\int_0^t \left|\exp\left[\int_r^t A^{\mu_\rho^\varepsilon}\,d\rho\right]\sigma^{\mu_r^\varepsilon} - \exp\left[\int_r^t A^{\overline{\mu}_\rho^\varepsilon}\,d\rho\right]\sigma^{\overline{\mu}_r^\varepsilon}\right|^2 dr\right\}^{1/2} \\ &\qquad \le \left\{2\beta_\tau^2(4c_A)^{-1/2} + 2\beta_\tau^1 c_\sigma^2(2c_A)^{-1}\right\}^{1/2} \le \beta_\tau^3,\end{aligned} \tag{87}$$

where $\beta_\tau^3 \to 0$ as $\tau \to 0$ (independent of t).

Combining (82), (86), and (87), one has

$$|\xi_t^\varepsilon - \overline{\xi}_t^\varepsilon| \le \beta_\tau^1 e^{-c_A t}|x| + \beta_\tau^3\|w^\varepsilon\|_{L_2(0,t)}. \tag{88}$$

Now, by the system structure given by assumption block $(A.m)$ and by the fact that the V^m are in $\mathcal{G}_\delta$, one obtains the following lemmas exactly as in [35]. These are also analogous to their counterparts in Appendix A.

LEMMA B.1. *For any* $t < \infty$, $\|w^\varepsilon\|^2_{L_2(0,t)} \le \frac{\varepsilon}{\delta} + \frac{1}{\delta}\left[\frac{c_A\gamma^2}{c_\sigma^2}e^{-c_A N\tau} + \frac{c_D}{c_A}\right]|x|^2$.

LEMMA B.2. *For any* $t < \infty$,

$$\int_0^t |\xi_r^\varepsilon|^2\,dt \le \frac{\varepsilon}{\delta}\frac{c_\sigma^2}{c_A} + \frac{c_\sigma^2}{\delta}\left[\left(\frac{c_D}{c_A^2} + \frac{\gamma^2}{c_\sigma^2}\right) + \frac{1}{c_A}\right]|x|^2.$$

Let $c_1 \doteq \frac{\varepsilon}{\delta}$ and $c_2 \doteq \frac{1}{\delta}[\frac{c_A\gamma^2}{c_\sigma^2}e^{-c_A N\tau} + \frac{c_D}{c_A}]$. By Lemma B.1 and (88), for all $t < \infty$ one has

$$|\xi_t^\varepsilon - \overline{\xi}_t^\varepsilon| \le \beta_\tau^1 e^{-c_A t}|x| + \beta_\tau^3(c_1 + c_2|x|^2)^{1/2},$$

and by proper choice of β_τ^4,

$$\le \beta_\tau^4(1 + |x|), \tag{89}$$

where $\beta_\tau^4 \to 0$ as $\tau \to 0$ (independent of $t > 0$).

Now,

$$\begin{aligned}&\int_0^T l^{\mu_t^\varepsilon}(\xi_t^\varepsilon) - \frac{\gamma^2}{2}|w_t^\varepsilon|^2\,dt + V^m(\xi_T^\varepsilon) - \int_0^T l^{\overline{\mu}_t^\varepsilon}(\overline{\xi}_t^\varepsilon) - \frac{\gamma^2}{2}|w_t^\varepsilon|^2\,dt + V^m(\overline{\xi}_T^\varepsilon)\\ &= \int_0^T \xi_t^\varepsilon D^{\mu_t^\varepsilon}\xi_t^\varepsilon - \overline{\xi}_t^\varepsilon D^{\overline{\mu}_t^\varepsilon}\overline{\xi}_t^\varepsilon\,dt + (\xi_T^\varepsilon)^T P^m \xi_T^\varepsilon - (\overline{\xi}_T^\varepsilon)^T P^m \overline{\xi}_T^\varepsilon.\end{aligned} \tag{90}$$

Note that the integral term on the right-hand side in (90) is

$$\begin{aligned}&\int_0^T (\xi_t^\varepsilon)^T D^{\mu_t^\varepsilon}(\xi_t^\varepsilon - \overline{\xi}_t^\varepsilon) + (\xi_t^\varepsilon)^T\left(D^{\mu_t^\varepsilon} - D^{\overline{\mu}_t^\varepsilon}\right)\overline{\xi}_t^\varepsilon + (\xi_t^\varepsilon - \overline{\xi}_t^\varepsilon)^T D^{\overline{\mu}_t^\varepsilon}\overline{\xi}_t^\varepsilon\,dt\\ &\le \beta_\tau^4(1 + |x|)\int_0^T \left(|D^{\mu_t^\varepsilon}|\,|\xi_t^\varepsilon| + |D^{\overline{\mu}_t^\varepsilon}|\,|\overline{\xi}_t^\varepsilon|\right)dt + \beta_\tau^5\int_0^T |\xi_t^\varepsilon|\,|\overline{\xi}_t^\varepsilon|\,dt\end{aligned}$$

for appropriate $\beta_\tau^5 \to 0$ as $\tau \to 0$, which, after some work,

$$\le \beta_\tau^6(1 + |x|^2)(1 + \sqrt{T}) \tag{91}$$

for an appropriate choice of $\beta_\tau^6 \to 0$ as $\tau \to 0$ (independent of T).

Similarly, the last two terms on the right-hand side in (90) are

$$\begin{aligned}{\xi_T^\varepsilon}^T P^m \xi_T^\varepsilon - {\overline{\xi}_T^\varepsilon}^T P^m \overline{\xi}_T^\varepsilon &= (\xi_T^\varepsilon + \overline{\xi}_T^\varepsilon)^T P^m(\xi_T^\varepsilon - \overline{\xi}_T^\varepsilon)\\ &\le |P^m|\left[|\xi_T^\varepsilon - \overline{\xi}_T^\varepsilon|^2 + 2|\xi_T^\varepsilon|\,|\xi_T^\varepsilon - \overline{\xi}_T^\varepsilon|\right],\end{aligned}$$

which, by (89),

$$\le \beta_\tau^7(1 + |x|^2) + \beta_\tau^8|\xi_T^\varepsilon|(1 + |x|), \tag{92}$$

where $\beta_\tau^7, \beta_\tau^8 \to 0$ as $\tau \to 0$.

We also need the following lemma which is obtained in [35].

LEMMA B.3. *Given* $\overline{T} < \infty$, *there exist* $T \in [\overline{T}/2, \overline{T}]$ *and* ε*-optimal* $w^\varepsilon \in \mathcal{W}$, $\mu^\varepsilon \in \mathcal{D}_\infty$ *for* $\widetilde{S}_T[V^m]$ *such that*

$$|\xi_T^\varepsilon|^2 \le \frac{1}{\overline{T}}\left\{\frac{\varepsilon}{\delta}\frac{c_\sigma^2}{c_A} + \frac{c_\sigma^2}{\delta}\left[\left(\frac{c_D}{c_A^2} + \frac{\gamma^2}{c_\sigma^2}\right) + \frac{1}{c_A}\right]|x|^2\right\}.$$

Combining (92) and Lemma B.3, one finds that there exist $c_3, c_4 < \infty$ such that

$$\xi_T^{\varepsilon\,T} P^m \xi_T^\varepsilon - \overline{\xi}_T^{\varepsilon\,T} P^m \overline{\xi}_T^\varepsilon \le \beta_\tau^9 (1 + |x|^2), \tag{93}$$

where $\beta_\tau^9 \to 0$ as $\tau \to 0$ (independent of T).

Combining (90), (91), and (93),

$$\begin{aligned} &\int_0^T l^{\mu_t^\varepsilon}(\xi_t^\varepsilon) - \frac{\gamma^2}{2}|w_t^\varepsilon|^2\,dt + V^m(\xi_T^\varepsilon) - \int_0^T l^{\overline{\mu}_t^\varepsilon}(\overline{\xi}_t^\varepsilon) - \frac{\gamma^2}{2}|w_t^\varepsilon|^2\,dt + V^m(\overline{\xi}_T^\varepsilon) \\ &\qquad \le \beta_\tau^{10}(1 + |x|^2)(1 + \sqrt{T}), \end{aligned} \tag{94}$$

where $\beta_\tau^{10} \to 0$ as $\tau \to 0$ (independent of T).

Combining (81) and (94), one has

$$\widetilde{S}_T[V^m](x) - \int_0^T l^{\overline{\mu}_t^\varepsilon}(\overline{\xi}_t^\varepsilon) - \frac{\gamma^2}{2}|w_t^\varepsilon|^2\,dt + V^m(\overline{\xi}_T^\varepsilon) \le \frac{\varepsilon}{2}(1 + |x|^2) + \beta_\tau^{10}(1 + |x|^2)(1 + \sqrt{T}),$$

which, for τ sufficiently small (depending on T now),

$$\le \varepsilon(1 + |x|^2).$$

This completes the proof of Lemma 4.11. ∎

Acknowledgment. The author wishes to thank Prof. J. William Helton for helpful discussions, without which this direction may never have been explored.

REFERENCES

[1] M. Akian, S. Gaubert, and A. Lakhoua, *A max-plus finite element method for solving finite horizon determinsitic optimal control problems,* in Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS), Leuven, 2004.

[2] F. L. Baccelli, G. Cohen, G. J. Olsder, and J.-P. Quadrat, *Synchronization and Linearity*, John Wiley, New York, 1992.

[3] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser Boston, Boston, MA, 1997.

[4] M. Boué and P. Dupuis, *Markov chain approximations for deterministic control problems with affine dynamics and quadratic cost in the control,* SIAM J. Numer. Anal., 36 (1999), pp. 667–695.

[5] F. Camilli, M. Falcone, P. Lanucara, and A. Seghini, *A domain decomposition method for Bellman equations,* Contemp. Math., 180 (1994), pp. 477–483.

[6] E. Carlini, M. Falcone, and R. Ferretti, *An efficient algorithm for Hamilton–Jacobi equations in high dimensions,* Comput. Vis. Sci., 7 (2004), pp. 15–29.

[7] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[8] G. Cohen, S. Gaubert, and J.-P. Quadrat, *Duality and separation theorems in idempotent semimodules*, Linear Algebra Appl., 379 (2004), pp. 395–422.

[9] G. Collins and W. M. McEneaney, *Min-plus eigenvector methods for nonlinear H_∞ problems with active control,* in Optimal Control, Stabilization and Nonsmooth Analysis, Lecture Notes in Control and Inform. Sci. 301, M. S. de Queiroz, M. Malisoff, and P. Wolenski, eds., Springer, Berlin, 2004, pp. 101–120.

[10] M. G. Crandall, H. Ishii, and P.-L. Lions, *Uniqueness of viscosity solutions of Hamilton–Jacobi equations revisited,* J. Math. Soc. Japan, 39 (1987), pp. 581–595.

[11] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton–Jacobi equations,* Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[12] M. G. Crandall and P.-L. Lions, *On existence and uniqueness of solutions of Hamilton–Jacobi equations,* Nonlinear Anal. Theory Appl., 10 (1986), pp. 353–370.

[13] R. A. Cuninghame-Green, *Minimax Algebra*, Lecture Notes in Econom. and Math. Systems 166, Springer-Verlag, Berlin, New York, 1979.

[14] P. Dupuis and A. Szpiro, *Convergence of the optimal feedback policies in a numerical method for a class of deterministic optimal control problems,* SIAM J. Control Optim., 40 (2001), pp. 393–420.
[15] L. C. Evans, *Partial Differential Equations*, AMS, New York, 1998.
[16] M. Falcone and R. Ferretti, *Convergence analysis for a class of high-order semi-Lagrangian advection schemes,* SIAM J. Numer. Anal., 35 (1998), pp. 909–940.
[17] W. H. Fleming, *Max-plus stochastic processes*, Appl. Math. Optim., 49 (2004), pp. 159–181.
[18] W. H. Fleming, *Functions of Several Variables*, 2nd ed., Springer-Verlag, New York, 1977.
[19] W. H. Fleming and W. M. McEneaney, *A max-plus-based algorithm for a Hamilton–Jacobi–Bellman equation of nonlinear filtering,* SIAM J. Control Optim., 38 (2000), pp. 683–710.
[20] W. H. Fleming and H. M. Soner, *Controlled Markov Process and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[21] J. W. Helton and M. R. James, *Extending H^∞ Control to Nonlinear Systems: Control of Nonlinear Systems to Achieve Performance Objectives*, Adv. Des. Control 1, SIAM, Philadelphia, 1999.
[22] M. R. James and W. M. McEneaney, *Max–plus approximation methods in partially observed H_∞ control,* in Proceedings of the 38th IEEE Conference on Decision and Control, 1999, pp. 3011–3016.
[23] F. John, *Partial Differential Equations*, Springer-Verlag, New York, 1978.
[24] V. N. Kolokoltsov and V. P. Maslov, *Idempotent Analysis and Its Applications,* Kluwer, Dordrecht, 1997.
[25] H. J. Kushner and P. Dupuis, *Numerical methods for stochastic control problems in continuous time*, Springer-Verlag, New York, 1992.
[26] G. L. Litvinov, V. P. Maslov, and G. B. Shpiz, *Idempotent functional analysis: An algebraic approach,* Math. Notes, 69 (2001), pp. 696–729.
[27] V. P. Maslov, *On a new principle of superposition for optimization problems,* Russian Math. Surveys, 42 (1987), pp. 43–54.
[28] W. M. McEneaney, *Legendre transforms on max-plus spaces as a tool for nonlinear control problems,* in Proceedings of the International Symposium on Mathematical Theory of Networks and Systems (MTNS), 2004.
[29] W. M. McEneaney, *Max-plus summation of Fenchel-transformed semigroups for solution of nonlinear Bellman equations,* Systems Control Lett., 56 (2007), pp. 255–264.
[30] W. M. McEneaney and P. M. Dower, *A max-plus affine power method for approximation of a class of mixed l_∞/l_2 value functions,* in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, 2003, pp. 2573–2578.
[31] W. M. McEneaney, *Max-plus eigenvector methods for nonlinear H_∞ problems: Error analysis,* SIAM J. Control Optim., 43 (2004), pp. 379–412.
[32] W. M. McEneaney, *Max-plus eigenvector representations for solution of nonlinear H_∞ problems: Basic concepts,* IEEE Trans. Automat. Control, 48 (2003), pp. 1150–1163.
[33] W. M. McEneaney, *Error analysis of a max-plus algorithm for a first-order HJB equation,* in Stochastic Theory and Control, Proceedings of a Workshop held in Lawrence, KS, October 18–20, 2001, B. Pasik-Duncan, ed., Lecture Notes in Control and Inform. Sci., Springer-Verlag, New York, 2002, pp. 335–352.
[34] W. M. McEneaney and M. Horton, *Max-plus eigenvector representations for nonlinear H_∞ value functions,* in Proceedings of the 37th IEEE Conference on Decision and Control (Tampa, FL), 1998, pp. 3506–3511.
[35] W. M. McEneaney, *A uniqueness result for the Isaacs equation corresponding to nonlinear H_∞ control,* Math. Control Signals Systems, 11 (1998), pp. 303–334.
[36] W. M. McEneaney and M. V. Day, *Characteristic characterization of viscosity supersolutions corresponding to nonlinear H_∞ control,* in Proceedings of the 13th World Congress, International Federation of Automatic Control, Pergamon Press, Oxford, 1996, pp. 401–406.
[37] A. A. Melikyan, *Generalized Characteristics of First Order PDEs: Applications in Optimal Control and Differential Games*, Birkhäuser Boston, Boston, MA, 1998.
[38] R. T. Rockafellar, *Conjugate Duality and Optimization*, CBMS-NSF Reg. Conf. Ser. Appl. Math. 16, SIAM, Philadelphia, 1974.
[39] R. T. Rockafellar and R. J. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[40] P. Soravia, *H_∞ control of nonlinear systems: Differential games and viscosity solutions,* SIAM J. Control Optim., 34 (1996), pp. 1071–1097.
[41] A. Szpiro and P. Dupuis, *Second order numerical methods for first order Hamilton–Jacobi equations,* SIAM J. Numer. Anal., 40 (2002), pp. 1136–1183.
[42] A. J. van der Schaft, *L_2-gain analysis of nonlinear systems and nonlinear state feedback H_∞ control,* IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1277–1298

FEEDBACK SOLUTIONS OF OPTIMAL CONTROL PROBLEMS WITH DAE CONSTRAINTS*

GALINA A. KURINA† AND ROSWITHA MÄRZ‡

Abstract. An optimal feedback control has been obtained for linear-quadratic optimal control problems with constraints described by differential-algebraic equations (DAEs). For that purpose, a new implicit Riccati equation (Riccati differential-algebraic system) is provided, and its solvability is investigated. It is shown that one can do without the strong consistency conditions as used in several previous papers. Furthermore, the solvability of the resulting closed loop system is considered and the relations between Riccati equations and Hamiltonian systems are elucidated.

Key words. linear-quadratic optimal control problems, feedback control, differential-algebraic equations, descriptor systems, Riccati equations, Hamiltonian systems

AMS subject classification. 49N10, 49J15, 49N35, 34A09

DOI. 10.1137/050637352

1. Introduction. Feedback solutions via Riccati differential equations are a known and proven tool for solving linear-quadratic optimal control problems given by the cost

$$J(u,x) := \frac{1}{2}\langle x(T), Vx(T)\rangle + \frac{1}{2}\int_0^T \left\langle \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}, \begin{bmatrix} W(t) & S(t) \\ S(t)^* & R(t) \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \right\rangle dt \tag{1.1}$$

and the side conditions

$$x'(t) = C(t)x(t) + D(t)u(t), \quad t \in [0,T], \tag{1.2}$$

$$x(0) = z_0. \tag{1.3}$$

The superscript * denotes the transpose.

Let all coefficients be continuous and certain standard conditions be fulfilled (cf. section 2). In the following, the argument t is dropped almost everywhere, and the given relations are meant pointwise for all $t \in [0,T]$.

The terminal value problem for the relevant Riccati matrix differential equation with the symmetric solution Y is (see, e.g., [20] for $S = 0$)

$$Y' = -YC - C^*Y + (S + YD)R^{-1}(S^* + D^*Y) - W, \tag{1.4}$$

$$Y(T) = V. \tag{1.5}$$

*Received by the editors August 2, 2005; accepted for publication (in revised form) February 22, 2007; published electronically September 12, 2007. This research was supported by the DFG Research Center MATHEON "Mathematics for key technologies" and partially supported by Russian Fundamental Research Foundation Project 06-01-00296.

http://www.siam.org/journals/sicon/46-4/63735.html

†Voronezh State Forestry Academy, Timirjazeva 8, 394618 Voronezh, Russia (kurina@kma.vsu.ru).

‡Department of Mathematics, Humboldt University, 10099 Berlin, Germany (maerz@mathematik.hu-berlin.de).

If the explicit ordinary differential equation (ODE) in (1.2) is replaced by a differential-algebraic equation (DAE)

$$Ex' = Cx + Du, \tag{1.6}$$

with E being a singular constant square matrix, the situation becomes much more complex, and several different generalizations of the Riccati-ansatz are possible. For this, quite a lot of references are available (in particular for the case of constant coefficients); however, we can mention here only part of them. We refer to [5], [18], and [14] for further sources.

In [5] it was first noted that the modification

$$E^*Y'E = -E^*YC - C^*YE + (S + E^*YD)R^{-1}(S^* + D^*YE) - W, \tag{1.7}$$

which is considered to be *obvious*, leads to unacceptable solvability conditions. Consequently, more specific Riccati approaches that skillfully make use of the inherent structures find favor with [5]. Starting from a singular value decomposition $UEV = \operatorname{diag}(\Sigma, 0)$ and certain rank conditions, lower dimensional Riccati equations of the form $\Sigma Y'\Sigma = \dots$ are introduced. From the point of view of DAE theory the rank conditions used in [5] imply that the related Hamilton–Lagrange system is a regular DAE with tractability index one (cf. [3]).

In [15], [16], [17] (in a more general Hilbert space setting, with $S = 0$) a different ansatz was followed with Riccati equations of the form

$$E^*Y' = -Y^*C - C^*Y + (S + Y^*D)R^{-1}(S^* + D^*Y) - W. \tag{1.8}$$

The solutions of the terminal value problem for (1.8) with the condition

$$E^*Y(T) = V \tag{1.9}$$

have the symmetry property $E^*Y = Y^*E$. Like (1.7), (1.8) also is primarily a matrix-DAE; however, (1.8) has much better solvability properties than (1.7). In [15], [16], a decoupling into characteristic components is not used for the ansatz of the Riccati equation itself, as was done in [5], but for proving the solvability of the Riccati terminal value problem (1.8), (1.9).

Kunkel and Mehrmann [14] consider the Riccati DAE

$$(E^*YE)' = -E^*YC - C^*YE + (S + E^*YD)R^{-1}(S^* + D^*YE) - W, \tag{1.10}$$

which generalizes (1.7) to allow for time-dependent coefficients E. However, this equation is as unsuitable as its time-invariant version (1.7), and the authors have to admit that, *unfortunately, this approach can be used only in very special cases since, for $E(t)$ singular, the solutions of* (1.10) *and the Euler–Lagrange equation are not related via $u = -R^{-1}(S + D^*YE)x$, as in the case of nonsingular $E(t)$*. The difficulties with (1.10) are illustrated in [14] by means of a small academic problem. Below we shall resume this special problem to show that things work well when using more appropriate Riccati DAE systems.

If, in (1.6), there is no constant matrix E in front of the derivative but a time-dependent matrix, it makes sense to change to a DAE with a *properly formulated leading term*, i.e., instead of (1.6), using

$$A(Bx)' = Cx + Du, \tag{1.11}$$

with well-matched A and B (cf. [4]). The corresponding initial condition is

$$A(0)B(0)x(0) = z_0 \tag{1.12}$$

with $z_0 \in \mathrm{im}(A(0)B(0))$. For arguments that state the leading term in this way we refer to [4], [22]. Notice that, in particular, such equations arise in circuit simulation via modified nodal analysis (see, e.g., [9], [11]).

Under the assumption that B is continuously differentiable, the terminal problem for the Riccati equation

$$\begin{aligned}(B^*A^*Y)' &= -Y^*(C - AB') - (C^* - B^{*\prime}A^*)Y + (S + Y^*D)R^{-1}(S^* + D^*Y) - W,\\ B(T)^*A(T)^*Y(T) &= V,\end{aligned} \tag{1.13}$$

with the symmetry property $B^*A^*Y = Y^*AB$, is proved to be relevant in [12] (in a more general Hilbert space setting).

The Riccati DAE

$$B^*(A^*Y)'B = -B^*Y^*C - C^*YB + (S + B^*Y^*D)R^{-1}(S^* + D^*YB) - W, \tag{1.14}$$

with B being just continuous, is investigated in [8]. This equation is, however, also a sort of generalization of (1.7) and adopts the bad solvability properties of (1.7). Under the condition that $\ker B^* = 0$, if there exists a solution Y of (1.14) that satisfies the terminal condition $B(T)^*A(T)^*Y(T)B(T) = V$, then this solution has the symmetry property $A^*Y = Y^*A$, and the ansatz $u = -R^{-1}(S^* + D^*YB)x$ actually leads to an optimal feedback control.

In this paper we work with the Riccati DAE

$$B^*(A^*YB^-)'B = -Y^*C - C^*Y + (S + Y^*D)R^{-1}(S^* + D^*Y) - W \tag{1.15}$$

and the terminal value condition

$$A(T)^*Y(T)B(T)^- = B(T)^{-*}VB(T)^-, \tag{1.16}$$

where B is assumed to be just continuous. Here, the solutions meet the symmetry condition $A^*YB^- = B^{-*}Y^*A$ (B^- is a special, generalized inverse).

While, e.g., in [5], [8] the Riccati-type DAEs are constructed to solve a linear boundary value problem serving as an extremal condition for the linear-quadratic optimal control problem, we justify the Riccati DAE (1.15) by a direct optimality proof (Theorem 2.5).

Notice that (1.11) is no longer necessarily square but may contain k equations, while $x(t)$ has m components.

In section 2 it is shown that, for the linear-quadratic optimal control problem (1.1), (1.11), (1.12), analogously to the classical case (1.1)–(1.5), optimal feedback controls can be established from the solutions of (1.15), (1.16). The main result in this respect is Theorem 2.5.

Section 3 investigates the solvability of the Riccati equation (1.15), generalizing the positive results from [16], [17]. The new solvability statements are provided in Theorem 3.4.

In section 4 we show that the solvability assumptions from Theorem 3.4 simultaneously imply the solvability of the closed loop initial value problems (IVPs).

In section 5 we work the same example as discussed in [14]. This example was used in [14] merely to demonstrate drawbacks with (1.10). In the present paper, we use this task to confirm the better solvability properties of the new Riccati equation (1.15).

Finally, section 6 elucidates the relation between solutions of the Riccati equation (1.15) and solutions of the corresponding implicit Hamiltonian system.

2. Optimal feedback control. We deal with the quadratic cost functional

$$\begin{aligned} J(u,x) :=& \frac{1}{2}\langle x(T), Vx(T)\rangle \\ &+\frac{1}{2}\int_0^T \{\langle x(t), W(t)x(t)\rangle + 2\langle x(t), S(t)u(t)\rangle + \langle u(t), R(t)u(t)\rangle\}dt \end{aligned} \tag{2.1}$$

to be minimized on pairs $(u,x) \in \mathcal{C}\times\mathcal{C}^1_B$ satisfying the IVP

$$A(t)(B(t)x(t))' = C(t)x(t) + D(t)u(t), \quad t\in[0,T], \tag{2.2}$$

$$A(0)B(0)x(0) = z_0. \tag{2.3}$$

The coefficients in (2.1), (2.2) are matrices $W(t)\in L(\mathbb{R}^m,\mathbb{R}^m)$, $R(t)\in L(\mathbb{R}^l,\mathbb{R}^l)$, $S(t)\in L(\mathbb{R}^l,\mathbb{R}^m)$, $A(t)\in L(\mathbb{R}^n,\mathbb{R}^k)$, $B(t)\in L(\mathbb{R}^m,\mathbb{R}^n)$, $C(t)\in L(\mathbb{R}^m,\mathbb{R}^k)$, $D(t)\in L(\mathbb{R}^l,\mathbb{R}^k)$, $t\in[0,T]$, which depend continuously on t, and $V\in L(\mathbb{R}^m,\mathbb{R}^m)$.

The value $z_0\in \mathrm{im}(A(0)B(0))$ is given. The leading term of the DAE (2.2) is assumed to be properly stated in the sense that the decomposition

$$\ker A(t)\oplus \mathrm{im}\, B(t) = \mathbb{R}^n, \quad t\in[0,T], \tag{2.4}$$

holds true, and both subspaces forming this direct sum have constant dimensions and are spanned by continuously differentiable on $[0,T]$ functions (cf. [4]).

We use the symbols $\mathcal{C}$ and $\mathcal{C}^1$ for continuous and continuously differentiable function spaces, respectively (functions defined on $[0,T]$, with values in $\mathbb{R}^l,\mathbb{R}^m,\mathbb{R}^k$, or $\mathbb{R}^n$ as given by the context), and

$$\mathcal{C}^1_B := \{x\in\mathcal{C} : Bx\in\mathcal{C}^1\}, \ \ \mathcal{C}^1_{A^*} := \{\psi\in\mathcal{C} : A^*\psi\in\mathcal{C}^1\}.$$

The coefficients determining the cost (2.1) satisfy the following standard assumptions: $W(t), R(t)$, and V are symmetric, $R(t)$ is positive definite, and $\left[\begin{smallmatrix} W(t) & S(t)\\ S(t)^* & R(t)\end{smallmatrix}\right]$ is positive semidefinite, $t\in[0,T]$.

A pair $(u,x)\in\mathcal{C}\times\mathcal{C}^1_B$ satisfying the IVP (2.2), (2.3) is said to be *admissible*.

Let $K(t)\in L(\mathbb{R}^n,\mathbb{R}^n)$ denote the projector that realizes decomposition (2.4), $\ker K(t) = \ker A(t)$, $\mathrm{im} K(t) = \mathrm{im} B(t)$, $t\in[0,T]$. Since these subspaces are continuously differentiable, so is the projector function $K:[0,T]\to L(\mathbb{R}^n,\mathbb{R}^n)$.

In addition to $K(t)$ we introduce $Q(t)\in L(\mathbb{R}^m,\mathbb{R}^m)$, $Q_*(t)\in L(\mathbb{R}^k,\mathbb{R}^k)$, which are the orthoprojectors onto $\ker(A(t)B(t))$ and $\ker(B(t)^*A(t)^*) = \mathrm{im}(A(t)B(t))^\perp$, respectively; furthermore, $P(t) := I - Q(t), P_*(t) := I - Q_*(t)$, $t\in[0,T]$. The projector functions Q, P, Q_*, and P_* are continuous.

It is natural to assume that $V = VP(T)$ (see, e.g., [13]).

Having the projectors K, P, and P_*, we introduce the generalized inverses B^- of B and A^{*-} of A^* by

$$\begin{aligned} B^-BB^- &= B^-, \quad & BB^-B &= B, \quad & BB^- &= K, \quad & B^-B &= P,\\ A^{*-}A^*A^{*-} &= A^{*-}, \quad & A^*A^{*-}A^* &= A^*, \quad & A^*A^{*-} &= K^*, \quad & A^{*-}A^* &= P_*. \end{aligned} \tag{2.5}$$

Notice that B^- and A^{*-} are uniquely determined by (2.5) and continuous on $[0,T]$. It holds further that

$$B^-K = B^-, \quad A = AK, \quad A^* = K^*A^*, \quad B^{-*} = K^*B^{-*}. \tag{2.6}$$

Next we consider the terminal value problem

$$B^*(A^*YB^-)'B = -Y^*C - C^*Y + (S + Y^*D)R^{-1}(S^* + D^*Y) - W, \tag{2.7}$$

$$A(T)^*Y(T)B(T)^- = B(T)^{-*}VB(T)^-. \tag{2.8}$$

Equation (2.7) generalizes the (well-known for $A = I$, $B = I$) Riccati differential equation and may be understood as a Riccati DAE.

LEMMA 2.1. *If $Y : [0,T] \to L(\mathbb{R}^m, \mathbb{R}^k)$ is continuous with a continuously differentiable part A^*YB^-, and if it satisfies the terminal value problem* (2.7), (2.8), *then the symmetry relation*

$$A^*YB^- = B^{-*}Y^*A \tag{2.9}$$

becomes true.

Proof. Multiplying (2.7) by B^{-*} from the left, and by B^- from the right, leads to

$$K^*(A^*YB^-)'K = B^{-*}\{-Y^*C - C^*Y + (S + Y^*D)R^{-1}(S^* + D^*Y) - W\}B^- =: \mathfrak{A},$$

where $\mathfrak{A} = \mathfrak{A}^*$, and, further (cf. (2.6)),

$$(A^*YB^-)' = \mathfrak{A} + K^{*'}A^*YB^- + A^*YB^-K'.$$

It becomes clear that $U := A^*YB^-$ satisfies the ODE $U' = \mathfrak{A} + K^{*'}U + UK'$ as well as the condition $U(T) = B(T)^{-*}VB(T)^-$. Obviously, U^* is a further solution of the same final value problem; i.e., $U = U^*$ must be true. □

Remark 2.2. If Y solves (2.7), (2.8) and if, additionally, the condition $A^*YQ = 0$ is given, then it follows that

$$B^*A^*Y = Y^*AB \tag{2.10}$$

must hold. Conversely, relation (2.10) implies $A^*YQ = 0$.

Remark 2.3. If one has, instead of the continuous coefficient B considered here, a B that is continuously differentiable, one can consider

$$(B^*A^*Y)' = B^{*\prime}A^*Y + Y^*AB' - Y^*C - C^*Y + (S + Y^*D)R^{-1}(S^* + D^*Y) - W, \tag{2.11}$$

$$B(T)^*A(T)^*Y(T) = P(T)VP(T) = V \tag{2.12}$$

instead of (2.7), (2.8). All solutions of (2.11), (2.12) have the symmetry property (2.10). At the same time they are solutions of (2.7), (2.8) and satisfy the additional condition $A^*YQ = 0$. Observe that (2.11), (2.12) coincide with (1.13). At first glance this shows that, considering the Riccati DAE (2.7), we may expect positive solvability results as in [12] for (1.13).

Remark 2.4. Equation (2.11) was first considered in [12]. Special cases (resp., slight modifications of (2.11)) were discussed in [19] (where A or B is absent) and in [15], [16] (where A is absent, B is constant, and $S = 0$).

THEOREM 2.5. *Let Y be a solution of the terminal value problem* (2.7), (2.8), *and let the condition $A^*YQ = 0$ be fulfilled. Let $x_* \in \mathcal{C}_B^1$ be a solution of the IVP*

$$A(Bx)' = Cx - DR^{-1}(S^* + D^*Y)x, \quad A(0)B(0)x(0) = z_0, \tag{2.13}$$

and let

$$u_* := -R^{-1}(S^* + D^*Y)x_*. \tag{2.14}$$

Then it holds for each admissible pair $(u, x) \in \mathcal{C} \times \mathcal{C}_B^1$ that

$$J(u, x) \geq J(u_*, x_*) = \frac{1}{2}\langle z_0, A(0)^{*-}B(0)^{-*}Y(0)^* z_0\rangle;$$

i.e., (u_*, x_*) is an optimal pair and (2.14) *describes the optimal feedback.*

Proof. It holds that $A^*Y = A^*YP = A^*YB^-B$, and that $B^{-*}Y^*AB = A^*Y$. Given an admissible pair (u, x), we derive

$$\begin{aligned}
\frac{d}{dt}\langle Bx, A^*Yx\rangle &= \langle (Bx)', A^*Yx\rangle + \langle Bx, (A^*YB^-Bx)'\rangle \\
&= \langle (Bx)', A^*Yx\rangle + \langle Bx, (A^*YB^-)'Bx\rangle + \langle Bx, A^*YB^-(Bx)'\rangle \\
&= \langle (Bx)', A^*Yx\rangle + \langle Bx, (A^*YB^-)'Bx\rangle + \langle A^*Yx, (Bx)'\rangle \\
&= 2\langle (Bx)', A^*Yx\rangle + \langle x, B^*(A^*YB^-)'Bx\rangle \\
&= 2\langle A(Bx)', Yx\rangle + \langle x, B^*(A^*YB^-)'Bx\rangle.
\end{aligned}$$

Taking into account (2.2) and (2.7) we obtain the expression

$$\begin{aligned}
\frac{d}{dt}\langle Bx, A^*Yx\rangle = &-\{\langle Wx, x\rangle + 2\langle Su, x\rangle + \langle Ru, u\rangle\} \\
&+ \langle R(u + R^{-1}(S^*x + D^*Yx)), u + R^{-1}(S^*x + D^*Yx)\rangle.
\end{aligned}$$

By this we find

$$J(u, x) = \frac{1}{2}\langle x(T), Vx(T)\rangle - \frac{1}{2}\int_0^T \frac{d}{dt}\langle B(t)x(t), A(t)^*Y(t)x(t)\rangle dt + \mathfrak{B}(u, x),$$

$$\begin{aligned}
\mathfrak{B}(u, x) = \frac{1}{2}\int_0^T \langle R(t)(u(t) + R(t)^{-1}(S(t)^* + D(t)^*Y(t))x(t)), u(t) \\
+R(t)^{-1}(S(t)^* + D(t)^*Y(t))x(t)\rangle dt.
\end{aligned}$$

From the positive definiteness of $R(t)$ it follows that $\mathfrak{B}(u, x) \geq 0$.

Notice that $\mathfrak{B}(u_*, x_*) = 0$.

Compute further

$$\begin{aligned}
J(u, x) = &\frac{1}{2}\langle x(T), Vx(T)\rangle - \frac{1}{2}\langle B(T)x(T), A(T)^*Y(T)x(T)\rangle \\
&+\frac{1}{2}\langle B(0)x(0), A(0)^*Y(0)x(0)\rangle + \mathfrak{B}(u, x).
\end{aligned}$$

Using the conditions (2.3) and (2.8) as well as the relations $V = VP(T)$, $A^*Y = A^*YB^-B$, and (2.10) we arrive at

$$J(u,x) = \frac{1}{2}\langle z_0, A(0)^{*-}B(0)^{-*}Y(0)^*z_0\rangle + \mathfrak{B}(u,x).$$

Since the first term is independent of the admissible pair (u,x), we conclude that

$$J(u,x) \geq \frac{1}{2}\langle z_0, A(0)^{*-}B(0)^{-*}Y(0)^*z_0\rangle = J(u_*,x_*). \qquad \square$$

The linear-quadratic optimal control problem (2.1)–(2.3) is closely related to the boundary value problem (BVP)

$$\begin{bmatrix} A & 0 \\ 0 & -B^* \end{bmatrix} \frac{d}{dt}\left(\begin{bmatrix} B & 0 \\ 0 & A^* \end{bmatrix}\begin{bmatrix} x \\ \psi \end{bmatrix}\right) = \begin{bmatrix} C - DR^{-1}S^* & -DR^{-1}D^* \\ W - SR^{-1}S^* & C^* - SR^{-1}D^* \end{bmatrix}\begin{bmatrix} x \\ \psi \end{bmatrix}, \tag{2.15}$$

$$A(0)B(0)x(0) = z_0, \tag{2.16}$$

$$B(T)^*A(T)^*\psi(T) = Vx(T). \tag{2.17}$$

If this BVP has a solution pair x_*, ψ_*, then $u_* := -R^{-1}(S^*x_* + D^*\psi_*)$ is an optimal control. This can be realized by slightly modifying Proposition 3.2 in [7] or Lemma 2.2 in [13] . Conversely, if u_*, x_* is an optimal pair, and if the composed matrix function $[AB - CQ, D]$ has on $[0,T]$ full row rank, then there exists an adjoint function ψ_* such that x_*, ψ_* solve the BVP (2.15)–(2.17) (see [2]). If A and B are nonsingular, then the full rank condition is always given. For singular A and B, if the full rank condition fails to be valid, then it may happen (see [2]) that there is an optimal pair u_*, x_*, but an adjoint function to solve the BVP does not exist. Assuming the rank condition to be satisfied, we can use the BVP (2.15)–(2.17) as a sufficient and necessary optimality condition. For a further discussion of recent developments concerning extremal conditions for optimization problems involving linear and nonlinear DAEs we refer to [2].

In the case when $A = B = I$, system (2.15) is nothing else than the Hamiltonian ODE associated with the standard linear-quadratic optimal control problem (1.1)–(1.3). For singular A and B, (2.15) is a DAE with a properly stated leading term. We adopt the notion *Hamiltonian system* for this DAE. This is justified, since under certain conditions (cf. Remark 6.4) the dynamic part inherent in (2.15) actually shows a Hamiltonian flow (see [3]).

While, e.g., in [5], [8] the Riccati-type DAEs are constructed to solve the Hamiltonian system, here a direct optimality proof is applied to Theorem 2.5 and, at the same time, our new Riccati DAE system is justified. In section 6 we will elucidate relations between solutions of the Riccati DAE and solutions of the corresponding Hamiltonian system.

Remark 2.6. In [13] we dealt with linear-quadratic optimal control problems in a more general Hilbert space setting, where R is not necessarily invertible and the side conditions are given as $(Bx)' = Cx + Du$, $B(0)x(0) = z_0$. Sufficient solvability conditions are derived by investigating the structure as well as the inherent flow of a linear (abstract) descriptor system associated with a sufficient extremal condition.

3. Solvability of the Riccati DAE system. In this section we consider solutions of the system

$$B^*(A^*YB^-)'B = -Y^*C - C^*Y + (S + Y^*D)R^{-1}(S^* + D^*Y) - W, \tag{3.1}$$

$$P_*YQ = 0, \tag{3.2}$$

which satisfy the terminal condition

$$A(T)^*Y(T)B(T)^- = \widetilde{V} := B(T)^{-*}VB(T)^-. \tag{3.3}$$

Each solution Y that must be continuous with a continuously differentiable part A^*YB^- can be decomposed as

$$\begin{aligned} Y &= P_*YP + Q_*YP + Q_*YQ \\ &= A^{*-}A^*YB^-B + Q_*YP + Q_*YQ. \end{aligned}$$

We are going to show that the components

$$U := A^*YB^- \in \mathcal{C}^1, \quad \mathcal{V} := Q_*YP, \quad Z := Q_*YQ = YQ \in \mathcal{C} \tag{3.4}$$

satisfy a standard Riccati differential equation, a linear equation, and an algebraic Riccati equation, respectively.

Multiplying (3.1) by Q from the left and right, then by Q from the left and P from the right, and also by B^{-*} from the left and B^- from the right, we obtain the system

$$0 = -(YQ)^*CQ - QC^*YQ + (QS + (YQ)^*D)R^{-1}(S^*Q + D^*YQ) - QWQ, \tag{3.5}$$

$$0 = -(YQ)^*CP - QC^*YP + (QS + (YQ)^*D)R^{-1}(S^*P + D^*YP) - QWP, \tag{3.6}$$

$$\begin{aligned} K^*(A^*YB^-)'K &= -(YB^-)^*CB^- - B^{-*}C^*YB^- \\ &\quad + (B^{-*}S + (YB^-)^*D)R^{-1}(S^*B^- + D^*YB^-) - B^{-*}WB^-. \end{aligned} \tag{3.7}$$

Since multiplication of (3.1) by P from the left and Q from the right yields (3.6) once more, we know (3.1) to be equivalent to (3.5)–(3.7). Obviously, the component $Z = Q_*YQ = YQ$ satisfies (cf. (3.5)) the algebraic Riccati equation

$$0 = -Z^*Q_*CQ - QC^*Q_*Z + (QS + Z^*Q_*D)R^{-1}(S^*Q + D^*Q_*Z) - QWQ \tag{3.8}$$

and the trivial conditions $P_*Z = 0,\ ZP = 0$.

Next, from (3.6) we obtain a linear relation for the components Z, U, and $\mathcal{V}$, namely,

$$MQ_*\mathcal{V} + MP_*A^{*-}UB = -Z^*Q_*CP + (QS + Z^*Q_*D)R^{-1}S^*P - QWP,$$

where

$$M := QC^* - (QS + Z^*Q_*D)R^{-1}D^*, \quad M = QM. \tag{3.9}$$

Notice that, if the conditions

$$\mathrm{im}MQ_* = \mathrm{im}Q, \quad \ker M \cap \mathrm{im}Q_* = 0 \tag{3.10}$$

are fulfilled, we also have $\ker MQ_* = \ker Q_*$; further

$$(MQ_*)^+ MQ_* = Q_*, \quad MQ_*(MQ_*)^+ = Q, \tag{3.11}$$

and the resulting linear equation

$$MQ_*\mathcal{V} = -Z^*Q_*CP + (QS + Z^*Q_*D)R^{-1}S^*P - QWP - MP_*A^{*-}UB \tag{3.12}$$

determines $\mathcal{V}$ uniquely, depending on Z and U. Let us then write

$$\mathcal{V} = C_1 + C_2A^{*-}UB, \tag{3.13}$$

with

$$C_1 := (MQ_*)^+\{-Z^*Q_*CP + (QS + Z^*Q_*D)R^{-1}S^*P - QWP\},$$
$$C_2 := -(MQ_*)^+MP_*.$$

Notice that $(MQ_*)^+$ is continuous. It holds that $C_1 = Q_*C_1 = C_1P$, $C_2Q_*C_2 = C_2P_*$.

Finally, we turn to (3.7). Since K is continuously differentiable and $UK = U$, $K^*U = U$ hold true, we may write

$$K^*(A^*YB^-)'K = K^*U'K = U' - K^{*\prime}U - UK'.$$

Recall that U is symmetric due to Lemma 2.1. Using (3.13) we derive

$$\begin{aligned} YP &= Q_*YP + P_*YP = \mathcal{V} + A^{*-}UB \\ &= C_1 + C_2A^{*-}UB + A^{*-}UB, \end{aligned}$$

that is,

$$YP = C_1 + C_3A^{*-}UB, \quad C_3 := C_2 + P_*. \tag{3.14}$$

Thus we obtain, from (3.7), the following differential equation for U:

$$\begin{aligned} U' = {} & K^{*\prime}U + U^*K' - B^{-*}WB^- - B^{-*}(C_1 + C_3A^{*-}UB)^*CB^- \\ & - B^{-*}C^*(C_1 + C_3A^{*-}UB)B^- \\ & + B^{-*}(S + (C_1 + C_3A^{*-}UB)^*D)R^{-1}(S^* + D^*(C_1 + C_3A^{*-}UB))B^-, \end{aligned}$$

that is, considering that U is symmetric,

$$U' = -\widetilde{W} - \widetilde{C}^*U - U\widetilde{C} + U\widetilde{D}R^{-1}\widetilde{D}^*U, \tag{3.15}$$

where

$$\widetilde{C}^* := -K^{*\prime} + B^{-*}C^*C_3A^{*-} - B^{-*}(S + C_1^*D)R^{-1}D^*C_3A^{*-},$$
$$\widetilde{D}^* := D^*C_3A^{*-},$$
$$\begin{aligned} \widetilde{W} := {} & B^{-*}\{PWP + PC_1^*CP + PC^*C_1P \\ & -P(S + C_1^*D)R^{-1}(S^* + D^*C_1)P\}B^- = \widetilde{W}^*. \end{aligned}$$

LEMMA 3.1. *Let condition* (3.10) *be given, and additionally,*

$$\mathrm{im}Z = \mathrm{im}Q_*, \quad \ker Z = \ker Q. \tag{3.16}$$

Then, (3.15) *represents a standard Riccati differential equation with a symmetric, positive semidefinite coefficient* $\widetilde{W}$.

Proof. Condition (3.16) leads to $ZZ^+ = Q_*$, $Z^+Z = Q$, and Z^+ is continuous.

By construction of C_2, C_1 it holds that

$$MQ_*C_2 = -QMP_* = -MP_*,$$

$$MP_*A^{*-}UB = -MQ_*C_2A^{*-}UB = -MQ_*(\mathcal{V} - C_1) = MQ_*C_1 - MQ_*\mathcal{V}.$$

Taking this into account, from (3.12) we obtain the relation

$$0 = -QWP - Z^*Q_*CP + (QS + Z^*Q_*D)R^{-1}S^*P - MQ_*C_1. \tag{3.17}$$

Next we turn to

$$MQ_* = QC^*Q_* - (QS + Z^*Q_*D)R^{-1}D^*Q_*.$$

From (3.8) we derive the expression

$$\begin{aligned}(QS + Z^*Q_*D)R^{-1}D^*Q_* = {} & QWQZ^+ + Z^*Q_*CQZ^+ + QC^*Q_* \\ & -(QS + Z^*Q_*D)R^{-1}S^*QZ^+\end{aligned}$$

and put it into the formula for MQ_*, that is,

$$MQ_* = -QWQZ^+ - Z^*Q_*CQZ^+ + (QS + Z^*Q_*D)R^{-1}S^*QZ^+.$$

By this, (3.17) becomes

$$\begin{aligned}0 = {} & -QWP - Z^*Q_*CP + (QS + Z^*Q_*D)R^{-1}S^*P \\ & + QWQZ^+C_1 + Z^*Q_*CQZ^+C_1 - (QS + Z^*Q_*D)R^{-1}S^*QZ^+C_1,\end{aligned}$$

and hence, by multiplication from the left by $C_1^*Z^{+*}$,

$$\begin{aligned}0 = {} & -C_1^*Z^{+*}QWP - C_1^*Q_*CP + C_1^*Z^{+*}(QS + Z^*Q_*D)R^{-1}S^*(P - QZ^+C_1) \\ & + C_1^*Z^{+*}QWQZ^+C_1 + C_1^*Q_*CQZ^+C_1.\end{aligned}$$

This yields the expressions

$$\begin{aligned}C_1^*CP = C_1^*Q_*CP = {} & -C_1^*Z^{+*}QWP + C_1^*Q_*CQZ^+C_1 \\ & + C_1^*Z^{+*}QWQZ^+C_1 + C_1^*Z^{+*}(QS + Z^*Q_*D)R^{-1}S^*(P - QZ^+C_1),\end{aligned}$$

and, using properties of C_1,

$$\begin{aligned}
B^*\widetilde{W}B &= PWP + PC_1^*CP + PC^*C_1P - P(S + C_1^*D)R^{-1}(S^* + D^*C_1)P \\
&= PWP + C_1^*Q_*CP + PC^*Q_*C_1 - (PS + C_1^*D)R^{-1}(S^*P + D^*C_1) \\
&= PWP - PSR^{-1}S^*P - PSR^{-1}D^*C_1 - C_1^*DR^{-1}S^*P - C_1^*DR^{-1}D^*C_1 \\
&\quad - C_1^*Z^{+*}QWP + C_1^*Q_*CQZ^+C_1 + C_1^*Z^{+*}QWQZ^+C_1 \\
&\quad + C_1^*Z^{+*}QSR^{-1}S^*P - C_1^*Z^{+*}QSR^{-1}S^*QZ^+C_1 \\
&\quad + C_1^*Q_*DR^{-1}S^*P - C_1^*Q_*DR^{-1}S^*QZ^+C_1 \\
&\quad - PWQZ^+C_1 + C_1^*Z^{+*}QC^*Q_*C_1 + C_1^*Z^{+*}QWQZ^+C_1 \\
&\quad + PSR^{-1}S^*QZ^+C_1 - C_1^*Z^{+*}QSR^{-1}S^*QZ^+C_1 \\
&\quad + PSR^{-1}D^*Q_*C_1 - C_1^*Z^{+*}QSR^{-1}D^*Q_*C_1 \\
&= (P - C_1^*Z^{+*}Q)(W - SR^{-1}S^*)(P - QZ^+C_1) + \mathfrak{B}, \\
\mathfrak{B} &:= - C_1^*DR^{-1}D^*C_1 + C_1^*Q_*CQZ^+C_1 + C_1^*Z^{+*}QWQZ^+C_1 \\
&\quad - C_1^*Q_*DR^{-1}S^*QZ^+C_1 + C_1^*Z^{+*}QC^*Q_*C_1 \\
&\quad - C_1^*Z^{+*}QSR^{-1}S^*QZ^+C_1 - C_1^*Z^{+*}QSR^{-1}D^*Q_*C_1.
\end{aligned}$$

Taking into account that (cf. (3.8))

$$\begin{aligned}
&C_1^*Z^{+*}QWQZ^+C_1 + C_1^*Q_*CQZ^+C_1 + C_1^*Z^{+*}QC^*Q_*C_1 \\
&\qquad = C_1^*Z^{+*}(QS + Z^*Q_*D)R^{-1}(S^*Q + D^*Q_*Z)Z^+C_1
\end{aligned}$$

we find

$$\begin{aligned}
\mathfrak{B} &:= - C_1^*DR^{-1}D^*C_1 - C_1^*Q_*DR^{-1}S^*QZ^+C_1 - C_1^*Z^{+*}QSR^{-1}S^*QZ^+C_1 \\
&\quad - C_1^*Z^{+*}QSR^{-1}D^*Q_*C_1 \\
&\quad + C_1^*Z^{+*}QSR^{-1}S^*QZ^+C_1 + C_1^*Z^{+*}QSR^{-1}D^*Q_*C_1 \\
&\quad + C_1^*Q_*DR^{-1}S^*QZ^+C_1 + C_1^*Q_*DR^{-1}D^*Q_*C_1 = 0.
\end{aligned}$$

It results that

$$\widetilde{W} = B^{-*}(P - C_1^*Z^{+*}Q)(W - SR^{-1}S^*)(P - QZ^+C_1)B^-. \tag{3.18}$$

For $t \in [0,T]$ and all $x \in \mathbb{R}^m$ it holds that

$$\begin{aligned}
&\langle (W(t) - S(t)R(t)^{-1}S(t)^*)x, x\rangle \\
&\qquad = \left\langle \begin{bmatrix} W(t) & S(t) \\ S(t)^* & R(t) \end{bmatrix} \begin{bmatrix} x \\ -R(t)^{-1}S(t)^*x \end{bmatrix}, \begin{bmatrix} x \\ -R(t)^{-1}S(t)^*x \end{bmatrix} \right\rangle \geq 0,
\end{aligned}$$

i.e., $W(t) - S(t)R(t)^{-1}S(t)^*$ is positive semidefinite, and so is $\widetilde{W}(t)$. □

The following assertion reflects what we have derived.

THEOREM 3.2. *If Y is a solution of the Riccati-type terminal value problem* (3.1), (3.2), (3.3), *and if the conditions* (3.10) *and* (3.16) *are fulfilled, then the component $Z = Q_*YQ$ is a solution of the algebraic Riccati equation* (3.8), *$U = A^*YB^-$ is a solution of the standard Riccati differential equation* (3.15), *and $\mathcal{V} = Q_*YP$ satisfies* (3.12).

Conversely, considering now the following decoupled system for the unknown functions $Z, U, \mathcal{V}$ to be given (cf. (3.8), (3.2), (3.15), (2.8), (3.12)) as

$$0 = -Z^*Q_*CQ - QC^*Q_*Z + (QS + Z^*Q_*D)R^{-1}(S^*Q + D^*Q_*Z) - QWQ, \tag{3.19}$$

$$P_*Z = 0, \tag{3.20}$$

$$ZP = 0, \tag{3.21}$$

$$U' = -\widetilde{W} - \widetilde{C}^*U - U^*\widetilde{C} + U^*\widetilde{D}R^{-1}\widetilde{D}^*U, \tag{3.22}$$

$$U(T) = \widetilde{V} := B(T)^{-*}VB(T)^-, \tag{3.23}$$

$$MQ_*\mathcal{V} = -MP_*A^{*-}UB - QWP - Z^*Q_*CP + (QS + Z^*Q_*D)R^{-1}S^*P, \tag{3.24}$$

we may try to compose a solution Y of the original Riccati system (3.1)–(3.3) from the solutions $Z, U, \mathcal{V}$. Let us recall that the coefficients $\widetilde{C}, \widetilde{D}$, and M as defined above depend on Z.

If Z is a solution of the algebraic equation (3.19), then $Z + P_*\widetilde{Z}$, where $\widetilde{Z}$ is an arbitrary $k \times m$ matrix function, is also a solution of (3.19). By means of (3.20), the arbitrary solution part belonging to $\mathrm{im}P_*$ is fixed as zero.

By multiplication of (3.19) from both sides by Q we realize that, if Z solves (3.19), then ZQ also does. By means of condition (3.21) we pick up solutions with $Z = ZQ$. From (3.20), (3.21) we have $Z = Q_*ZQ$.

Obviously, (3.19) itself is symmetric, but Z is not so necessarily. Notice that Z has k rows and m columns. If $m = k$ and $Q_* = Q$ (i.e., $\ker AB = \ker(AB)^*$), then Z can be expected to be symmetric.

What we need is a continuous solution Z that satisfies the conditions

$$\mathrm{im}Z = \mathrm{im}Q_*, \quad \ker Z = \ker Q, \tag{3.25}$$

$$\mathrm{im}MQ_* = \mathrm{im}Q, \quad \ker MQ_* = \ker Q_*, \tag{3.26}$$

with $M = QC^* - (QS + Z^*Q_*D)R^{-1}D^*$.

These requirements ensure that the coefficients $\widetilde{W}$, $\widetilde{C}$, and $\widetilde{D}$ in (3.22) are well defined and continuous. Additionally, $\widetilde{W}$ is symmetric and positive semidefinite. It turns out that (3.22) is a standard Riccati differential equation, and the solution U of the terminal value problem (3.22), (3.23) is symmetric, $U = U^*$.

LEMMA 3.3. *We are given a continuous solution Z of* (3.19)–(3.21) *such that the conditions* (3.25), (3.26) *are fulfilled. Then, for the unique solution U of the resulting standard Riccati differential equation* (3.22), *which satisfies the terminal condition* (3.23), *the relations*

$$U = U^*, \quad U = UK, \quad U = K^*UK \tag{3.27}$$

hold true.

Proof. Let U be a solution of (3.22), (3.23). It remains to verify that $U = K^*UK$.

Inspecting the coefficients we find that $\widetilde{W}(I-K) = 0$, $\widetilde{D} = K\widetilde{D}$, $\widetilde{C} = -K' + K\widetilde{C}K$ must hold. Multiplying (3.22) by $(I-K)$ from the right-hand side, we derive

$$U'(I-K) = -\widetilde{C}^*U(I-K) - U^*(-K' + K\widetilde{C}K)(I-K) + U^*K\widetilde{D}R^{-1}\widetilde{D}^*U(I-K),$$

and hence, denoting $U(I-K) =: \widetilde{U}$, $UK =: U_K$ and taking into account that $U = U^*$,

$$\widetilde{U}' - U(I-K)' = -\widetilde{C}^*\widetilde{U} + UK'(I-K) + U_K\widetilde{D}R^{-1}\widetilde{D}^*\widetilde{U},$$

i.e.,

$$\begin{aligned}\widetilde{U}' &= U(I-K)' - UK(I-K)' + (U_K\widetilde{D}R^{-1}\widetilde{D}^* - \widetilde{C}^*)\widetilde{U}\\ &= \widetilde{U}(I-K)' + F\widetilde{U},\\ F &:= U_K\widetilde{D}R^{-1}\widetilde{D}^* - \widetilde{C}^*.\end{aligned}$$

It becomes clear that the function $\widetilde{U} = U(I-K)$ is the solution of the homogeneous linear terminal value problem $\widetilde{U}' = -\widetilde{U}K' + F\widetilde{U}$, $\widetilde{U}(T) = 0$, but then $\widetilde{U}$ vanishes identically. $\widetilde{U} = 0$ means $U = UK$; further, $U = U^* = K^*U = K^*UK$. ∎

Having the matrix functions U and Z, we compose

$$\mathcal{V} := (MQ_*)^+\{-MP_*A^{*-}UB + (QS + Z^*Q_*D)R^{-1}S^*P - QWP - Z^*Q_*CP\} \tag{3.28}$$

to satisfy (3.24) and, finally,

$$Y := A^{*-}UB + Z + \mathcal{V}. \tag{3.29}$$

Under the assumptions of Lemma 3.3, both $\mathcal{V}$ and Y are continuous. It holds that

$$Q_*YP = Q_*\mathcal{V}P = \mathcal{V}, \quad Q_*YQ = Q_*ZQ = Z, \quad A^*YB^- = K^*UK = U. \tag{3.30}$$

The component A^*YB^- of Y is continuously differentiable and symmetric. Straightforward calculations in the direction opposite to that which we realized provided system (3.19)–(3.24) will show Y to be a solution of our system (3.1)–(3.3). By this, the following assertion providing the solution Y for Theorem 2.5 is proved.

THEOREM 3.4. *Let the algebraic Riccati system* (3.19)–(3.21) *have a continuous solution Z that satisfies the conditions* (3.25) *and* (3.26).

Then, the original Riccati DAE system (3.1)–(3.3) *has a continuous solution Y whose component A^*YB^- is continuously differentiableand symmetric. Additionally, it holds that $A^*YQ = 0$.*

Remark 3.5. For special solvability assertions concerning algebraic Riccati equations as well as standard Riccati differential equations, we refer to [1].

Remark 3.6. As far as the practical potential of the new Riccati-type terminal value problem (3.1)–(3.3) is concerned, we believe that Theorem 3.2 and Lemma 3.1 offer good chances for effective use. The inherent dynamic part within the Riccati DAE system (3.1), (3.2) is just the standard Riccati matrix differential equation (3.15) that has a positive semidefinite coefficient $\widetilde{W}$. We expect this strong structural knowledge to be very helpful in view of applications and further research.

Remark 3.7. What concerns the numerical treatment of the terminal value problem (3.1)–(3.3) we recall that, in this work, is that the weight R is positive definite.

If R becomes singular, additional difficulties will arise. Further, we stress once more that the dynamic part contained in the Riccati DAE system (3.1), (3.2) is the standard Riccati matrix differential equation (3.15). It seems to be possible to figure out so-called *numerically qualified* versions (see, e.g., [11]) of the Riccati DAE system (in vectorized form). For this, following along the lines of [8] and [10] (resp., [11]), further careful investigations are to be carried out. To speculate upon this might lead to numerically qualified index-one DAEs. Since for those DAEs, discretization (by stiffly accurate Runge–Kutta methods and backward differentiation formulas) and theoretical structural decoupling commute, the integration should be as safe as for standard Riccati ODEs. Due to the index-one property and stability preservation (see [11]), it seems that difficulties as reported, for instance, in the earlier work [6], will be avoided. This is a topic of further research.

4. Solvability of the closed loop problem. To confirm the existence of an optimal control u_* with the minimal cost $J(u_*, x_*)$ from Theorem 2.5, in addition to the existence of a Riccati DAE solution Y, one necessarily needs to confirm the existence of a solution of the resulting closed loop DAE, that is (cf. (2.13)),

$$A(Bx)' = Cx - DR^{-1}(S^* + D^*Y)x, \tag{4.1}$$

which satisfies the initial condition

$$A(0)B(0)x(0) = z_0, \tag{4.2}$$

where $z_0 \in \operatorname{im}(A(0)B(0))$ is fixed but chosen arbitrarily.

Clearly, if A and B are nonsingular, then the IVP (4.1), (4.2) always has a uniquely determined solution for each arbitrary z_0. In the case of singular A and B the situation is different, and so for time-invariant descriptor systems (see, e.g., [5]) one takes care to obtain a closed loop system that has no so-called *impulsive behavior* for any z_0. Within the scope of DAE theory, this means that one should have closed loop systems (4.1) that are regular with tractability index one.

Below we shall provide conditions ensuring the index-one property for the DAE (4.1). To this end, we recall some basic information on the tractability index.

The tractability index generalizes the Kronecker index of matrix pencils to time-varying DAEs. The basic tools in this concept are special decoupling projectors computed from the coefficients of the given DAE and certain characteristic subspaces. Any regular linear DAE with a properly stated leading term (see [22]) can be decoupled into characteristic parts analogously to the decoupling of a regular matrix pencil into dynamic, nondynamic, and impulsive parts (see [23]). Since no so-called derivative arrays are used, the concept applies to continuous coefficient equations.

The tractability index is defined for linear DAEs with a time-varying constant-rank matrix E(t) in front of the derivative (cf. (1.6)), e.g., as in [21], and for DAEs with a properly stated leading term as in [4], [22]. A brief description is given in [3, pp. 293–296]. Notice that, for nonlinear DAEs, the tractability index works with linearizations (see, e.g., [24]).

Consider the DAE

$$A(t)(B(t)x(t))' = H(t)x(t) + q(t),\ t \in [0, T], \tag{4.3}$$

with k equations and the unknown function $x(.)$ with values in $\mathbb{R}^m$. Let (4.3) have a properly stated leading term as described in section 2. Introduce $G_0(t) := A(t)B(t)$, the nullspace $N_0(t) := \ker G_0(t)$, and the subspace $S_0(t) := \{z \in \mathbb{R}^m : H(t)z \in$

$\mathrm{im}G_0(t)\}$. Let $Q_0(t) := Q(t)$ be the orthoprojector onto $\ker G_0(t)$. Further, denote $G_1(t) := G_0(t) - H(t)Q_0(t)$, $N_1(t) := \ker G_1(t)$, and $S_1(t) := \{z \in \mathbb{R}^m : H(t)z \in \mathrm{im}G_1(t)\}$, $r_0 := \mathrm{rankG_0(t)}$, $\nu_0(t) := \dim(\mathrm{N_0(t)} \cap \mathrm{S_0(t)})$. By construction, $P_0(t)$ has constant rank r_0, and $Q_0(t)$ has rank $m - r_0$.

In this paper, we deal with DAEs up to index two only, and we quote just this definition from [4]. Note that a regular DAE with tractability index μ also has perturbation index μ.

DEFINITION 4.1. *Let the subspaces BN_1 and BS_1 be spanned by continuously differentiable basis functions. The DAE* (4.3) *is said to be regular with tractability index one, if $m = k$ and, for $t \in [0,T]$, it holds that $N_0(t) \cap S_0(t) = 0$. The DAE is regular with tractability index two if $m = k$, and if the last intersection has a constant positive dimension on $[0,T]$, but the intersection $N_1(t) \cap S_1(t)$ is trivial for all $t \in [0,T]$.*

It turns out (see [4]) that a regular DAE (4.3) with tractability index one has the solutions

$$x = (I + Q_0 G_1^{-1} H)B^- u + Q_0 G_1^{-1} q,$$

where $u = Bx$ is a solution of the explicit ODE (the so-called inherent explicit ODE)

$$u' = K'u + BG_1^{-1}HB^- u + BG_1^{-1}q,$$

the coefficients of which are uniquely determined by the coefficients A, B, H. This corresponds to the decoupling of the DAE (4.3) into two portions by using the decomposition $I = P_0 + Q_0$ and premultiplication by G_1^{-1}. The P_0-portion leads to the inherent ODE, which has dimension r_0, and the Q_0-portion is the algebraic constraint or nondynamic part with dimension $m - r_0$. With $Bx = BB^- u = u$, we realize that any IVP for (4.3) with continuous $q(.)$, and the initial condition

$$A(0)B(0)x(0) = z_0,\ z_0 \in \mathrm{im(A(0)B(0))},$$

has exactly one solution belonging to the function space $\mathcal{C}_B^1$.

A regular DAE with tractability index two is different. It decouples into three parts: the inherent explicit ODE that has the lower dimension $r_0 - \nu_0$, the algebraic part with dimension $m - r_0$, and an extra part containing a differentiation (an impulsive part in descriptor systems) which is associated with the intersection $N_0 \cap S_0$ and has dimension ν_0. By means of the further projectors Q_1 onto N_1 along S_1, $P_1 := I - Q_1$ we describe the initial conditions appropriate for unique solvability as [4]

$$A(0)B(0)P_1(0)x(0) = z_0,\ z_0 \in \mathrm{im(A(0)B(0)P_1(0))}.$$

IVPs with z_0 not belonging to $\mathrm{im(A(0)B(0)P_1(0))}$, in particular, those with $q = 0$ or smooth q, are no longer solvable in $\mathcal{C}_B^1$.

Let us now turn back to the closed loop DAE (4.1).

THEOREM 4.2. *Let the conditions of Theorem* 3.4 *be given, $m = k$, and Y be a solution of the Riccati DAE system* (3.1)–(3.3). *Then the DAE* (4.1) *is regular with tractability index one, and there is exactly one solution $x_* \in \mathcal{C}_B^1$ of the IVP* (4.1), (4.2).

Proof. As it is shown, e.g., in [4], the IVP-solvability is a consequence of the index-one property. Notice that a linear DAE with a properly stated leading term is

regular with index one if its adjoint equation is, and vice versa (see, e.g., [4]). The adjoint equation to (4.1) reads

$$-B^*(A^*\lambda)' = C^*\lambda - (Y^*D + S)R^{-1}D^*\lambda. \tag{4.4}$$

The DAE (4.4) is regular with index one if the subspaces $\ker B^*A^* = \mathrm{im}Q_*$ and $\ker Q\{C^* - (Y^*D + S)R^{-1}D^*\} =: S_*$ intersect trivially (see, e.g., [4]). Because of $QY^* = (YQ)^* = (ZQ)^* = QZ^*$ we have $S_* = \ker\{QC^* - (QZ^*D + QS)R^{-1}D^*\} = \ker M$. This means the DAE (4.4) is regular with index one if $\ker M$ and $\mathrm{im}Q_*$ intersect trivially, but this in turn is a consequence of condition (3.26). □

THEOREM 4.3. *Let the conditions of Theorem* 3.4 *be given, $m > k$, and Y be a solution of the Riccati DAE system* (3.1)–(3.3). *Then there are solutions $x_* \in \mathcal{C}_B^1$ of the IVP* (4.1), (4.2).

Proof. Compute $G_1 := AB - \{C - DR^{-1}(S^* + D^*Y)\}Q$ and ask whether this matrix function has full row rank k. Obviously, this is in fact the case if $Q_*\{CQ - DR^{-1}(S^*Q + D^*ZQ)\} = Q_*M^* = (MQ_*)^*$ has the same range as Q_*, i.e., if $\mathrm{im}(MQ_*)^* = \mathrm{im}Q_*$. However, this is ensured by (3.26).

Denote $F := C - DR^{-1}(S^* + D^*Y)$.

In [7, Proof of Proposition 3.2], a coordinate transform $x = H\bar{x}$ is applied to the DAE $A(Bx)' = Fx$ with full row rank $G_1 = AB - FQ$ such that the transformed variable has the structure

$$\bar{x} = \begin{pmatrix} z \\ v \end{pmatrix} \begin{matrix} \}k \\ \}m-k \end{matrix}$$

and the transformed IVP is of the form

$$A(\widetilde{B}z)' = \widetilde{F}_1 z + \widetilde{F}_2 v, \quad A(0)\widetilde{B}(0)z(0) = z_0, \tag{4.5}$$

while (4.5), with any given v, represents a regular index-one DAE for z. □

Remark 4.4. By fixing v in (4.5), the resulting IVP for z is uniquely solvable. How to choose the mentioned transformation H in practice is discussed in [7].

5. A case study. Here we deal with the very special case of $k = m = 2$, $n = 1$, $l = 1$, $T = 1$,

$$J(u,x) = \frac{1}{2}\int_0^1 (\alpha x_1(t)^2 + \beta x_2(t)^2 + u(t)^2)dt, \tag{5.1}$$

where $\alpha \geq 0$, $\beta \geq 0$, $W = \left(\begin{smallmatrix} \alpha & 0 \\ 0 & \beta \end{smallmatrix}\right)$, $R = 1$, $V = 0$, $S = 0$, and the DAE describing the side condition is

$$\begin{aligned} x_1'(t) &= c_{12}(t)x_2(t), \\ 0 &= c_{21}(t)x_1(t) + c_{22}(t)x_2(t) + u(t), \end{aligned} \tag{5.2}$$

i.e.,

$$A = \begin{bmatrix} 1 \\ 0 \end{bmatrix},\ B = [1\ \ 0],\ B^- = \begin{bmatrix} 1 \\ 0 \end{bmatrix},\ K = 1,\ D = \begin{bmatrix} 0 \\ 1 \end{bmatrix},\ C = \begin{bmatrix} 0 & c_{12} \\ c_{21} & c_{22} \end{bmatrix}.$$

The initial condition for (5.2) reads

$$x_1(0) = x_{10}. \tag{5.3}$$

We have taken this problem from [14] and will discuss the same three cases as considered there. In the first two cases, optimal controls exist, and we obtain them via our Riccati DAE system, while the Riccati DAE system used in [14] has no solutions. In the third case, if $x_{10} \neq 0$, there is no optimal control, which is reflected by the failure of our conditions (3.25) and (3.26). Notice that just in this case the Riccati DAE in [14] may have solutions. We should like also to point out that this little academic problem (namely, special cases and infinite-horizon modifications) was already used earlier in the literature for illustrative purposes (see, e.g., [5]).

Consider the Riccati DAE system (3.1)–(3.3) for the 2×2 matrix function $Y = \left[\begin{smallmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{smallmatrix}\right]$. We describe (3.1) by means of the following three equations (cf. (3.5), (3.6), (3.7)), taking into account that we have here $Q = Q_* = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, $P = P_* = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, and dropping the equations "$0 = 0$,"

$$0 = -\beta - (Y_{12}c_{12} + Y_{22}c_{22}) - (c_{12}Y_{12} + c_{22}Y_{22}) + Y_{22}^2, \tag{5.4}$$

$$0 = -c_{21}Y_{22} - (c_{12}Y_{11} + c_{22}Y_{21}) + Y_{22}Y_{21}, \tag{5.5}$$

$$Y_{11}' = -\alpha - c_{21}Y_{21} - c_{21}Y_{21} + Y_{21}^2. \tag{5.6}$$

The terminal value condition (3.3) is

$$Y_{11}(1) = 0, \tag{5.7}$$

and condition (3.2) here means that

$$Y_{12} = 0. \tag{5.8}$$

Applying (5.8), (5.4) simplifies to

$$0 = -\beta + (Y_{22} - c_{22})^2 - c_{22}^2. \tag{5.9}$$

This algebraic equation has the solutions

$$Y_{22} = c_{22} \pm \sqrt{\beta + c_{22}^2}, \tag{5.10}$$

and the resulting matrix functions $Z = Q_*YQ$ and MQ_* (cf. (3.9)) are

$$Z = \begin{bmatrix} 0 & 0 \\ 0 & Y_{22} \end{bmatrix}, \quad MQ_* = \begin{bmatrix} 0 & 0 \\ 0 & c_{22} - Y_{22} \end{bmatrix}.$$

In this case the conditions (3.25) and (3.26) are equivalent to the conditions that

$$Y_{22}(t) \quad \text{has no zeros} \tag{5.11}$$

and

$$Y_{22}(t) - c_{22}(t) = \pm\sqrt{\beta + c_{22}(t)^2} \quad \text{has no zeros, respectively.} \tag{5.12}$$

Case I. c_{12} and c_{21} vanish identically, and $\beta > 0$.

Here, both Y_{22} and $Y_{22} - c_{22}$ do not have zeros; i.e., the conditions (3.25) and (3.26) are fulfilled. Equation (5.5) is simply $0 = (Y_{22} - c_{22})Y_{21}$, which leads to $Y_{21} = 0$. Equation (5.6) yields $Y_{11}' = -\alpha$. Hence, in this case

$$Y(t) = \begin{bmatrix} -\alpha(t-1) & 0 \\ 0 & c_{22}(t) \pm \sqrt{\beta + c_{22}(t)^2} \end{bmatrix}$$

solves the system. The feedback optimal control is given by $u = -(c_{22} \pm \sqrt{\beta + c_{22}^2})x_2$. The optimal trajectory, i.e., the solution of the IVP (4.1), (4.2) (cf. (2.13)) is $x_*(t) \equiv \binom{x_{10}}{0}$ the optimal control is $u_* = 0$, and the optimal cost is $J(u_*, x_*) = \frac{1}{2}\alpha x_{10}^2$.

Case II. c_{22} vanishes identically, c_{12} and c_{21} have no zeros, and $\beta > 0$. Again, both $Y_{22} = \pm\sqrt{\beta}$ and $Y_{22} - c_{22} = Y_{22}$ have no zeros, and the conditions (3.25) and (3.26) are fulfilled. This time, (5.5) leads to

$$Y_{21} = c_{21} \pm \frac{1}{\sqrt{\beta}} c_{12} Y_{11}. \tag{5.13}$$

From (5.6) and (5.13) we derive the ODE

$$Y_{11}' = -\alpha - 2c_{21}\left(c_{21} \pm \frac{1}{\sqrt{\beta}} c_{12} Y_{11}\right) + \left(c_{21} \pm \frac{1}{\sqrt{\beta}} c_{12} Y_{11}\right)^2.$$

For example, for $c_{12} = c_{21} = 1,\ x_{10} = 1$, the result is that

$$Y_{11}' = -(\alpha + 1) + \frac{1}{\sqrt{\beta}} Y_{11}^2,$$

$$Y_{11}(t) = \beta\gamma \frac{1 - e^{2\gamma(t-1)}}{1 + e^{2\gamma(t-1)}} \qquad \text{with } \gamma = \sqrt{\frac{1+\alpha}{\beta}}.$$

Then, $u = (\mp\frac{1}{\sqrt{\beta}} Y_{11} - 1)x_1 \mp \sqrt{\beta} x_2$ is an optimal feedback control. The DAE (4.1) is of the form

$$x_1' = x_2, \ \ 0 = \frac{1}{\sqrt{\beta}} Y_{11} x_1 + \sqrt{\beta} x_2,$$

and the optimal pair (u_*, x_*) consists of

$$u_*(t) = -x_{*1}(t), \ \ x_{*1}(t) = \frac{e^{\gamma t} + e^{\gamma(2-t)}}{1 + e^{2\gamma}}, \ \ x_{*2}(t) = -\frac{1}{\beta} Y_{11}(t) x_{*1}(t).$$

The minimal cost is

$$J(u_*, x_*) = \frac{\beta\gamma}{2} \cdot \frac{1 - e^{-2\gamma}}{1 + e^{-2\gamma}}.$$

Case III. $\beta = 0,\ c_{22}$ vanishes identically, and c_{12}, c_{21} have no zeros. Here, (5.9) implies $Y_{22} = 0$, and hence, $Z = 0,\ MQ_* = 0$, and the conditions (3.25) and (3.26) fail to be valid. Equation (5.5) simplifies to $c_{12} Y_{11} = 0$, and hence $Y_{11} = 0$ must be true. By (5.6) we find $Y_{21} = c_{21} \pm \sqrt{\alpha + c_{21}^2}$. Therefore, the matrix function

$$Y = \begin{bmatrix} 0 & 0 \\ c_{21} \pm \sqrt{\alpha + c_{21}^2} & 0 \end{bmatrix}$$

solves the system (3.1)–(3.3); however, the conditions (3.25) and (3.26) no longer hold. The resulting closed loop DAE (4.1) is now $x_1' = c_{12} x_2,\ 0 = \sqrt{\alpha + c_{21}^2}\, x_1$, and it has only the trivial solution. Consequently, for $x_{10} \neq 0$, there is no solution of the IVP (4.1), (4.2). If $x_{10} = 0$, then the trivial pair $u_* = 0,\ x_* = 0$ is optimal in accordance with Theorem 2.5. If $x_{10} \neq 0$, then the linear-quadratic optimal control problem has no solution at all.

The Hamiltonian system (2.15) corresponding to the special problem (5.1)–(5.3) is the following:

$$\begin{aligned} x_1' &= c_{12}x_2, \\ 0 &= c_{21}x_1 + c_{22}x_2 - \psi_2, \\ -\psi_1' &= \alpha x_1 + c_{21}\psi_2, \\ 0 &= \beta x_2 + c_{12}\psi_1 + c_{22}\psi_2. \end{aligned} \tag{5.14}$$

For this system, the initial and terminal conditions

$$x_1(0) = x_{10}, \quad \psi_1(1) = 0 \tag{5.15}$$

have to be taken into account. This linear DAE with respect to x, ψ is regular with index one exactly if $\beta + c_{22}^2 \neq 0$. This index-one condition is valid in Cases I and II.

In Case III, the BVP (5.14), (5.15) has no solution for $x_{10} \neq 0$. For $x_{10} = 0$ it has the trivial solution. It may be checked that this DAE has index two.

Notice that, for the solvability of the corresponding Riccati DAE (1.10) treated in [14], it is necessary that $\beta = 0$ be given; i.e., unfortunately, this Riccati DAE is no longer solvable in the unproblematic cases I and II. In Case III the terminal value problem for the Riccati DAE (1.10) may or may not have solutions. From this point of view, the Riccati DAEs (1.7) or (1.10) seem not to be appropriate tools for constructing optimal feedback solutions, whereas the Riccati DAE (1.15) and the versions in [15], [16], [17], [12] are useful for this purpose.

6. Riccati equations and Hamiltonian systems.

THEOREM 6.1. *Given a solution Y of* (2.7), (2.8) *with $A^*YQ = 0$, if the continuous matrix function $X : [0,T] \to L(\mathbb{R}^p, \mathbb{R}^m)$, with a continuously differentiable part BX, satisfies the equation*

$$A(BX)' = (C - DR^{-1}S^* - DR^{-1}D^*Y)X, \tag{6.1}$$

then the pair $X, \Psi := YX$ forms a solution of the Hamiltonian system

$$A(BX)' = (C - DR^{-1}S^*)X - DR^{-1}D^*\Psi, \tag{6.2}$$

$$-B^*(A^*\Psi)' = (W - SR^{-1}S^*)X + (C^* - SR^{-1}D^*)\Psi. \tag{6.3}$$

Ψ is continuous with $A^*\Psi$ being continuously differentiable.

Proof. Equation (6.2) is a trivial consequence of (6.1).

Due to $A^*\Psi = A^*YX = A^*YB^-BX$, $A^*\Psi$ is continuously differentiable. We derive

$$\begin{aligned} B^*(A^*\Psi)' &= B^*(A^*YB^-)'BX + B^*A^*YB^-(BX)' \\ &= B^*(A^*YB^-)'BX + Y^*A(BX)' \\ &= -(W - SR^{-1}S^*)X - (C^* - SR^{-1}D^*)\Psi, \end{aligned}$$

and we are done. ∎

The above pair X, Ψ combines p columns of solutions of the differential-algebraic Hamiltonian system (cf. (2.15))

$$\begin{bmatrix} A & 0 \\ 0 & -B^* \end{bmatrix} \frac{d}{dt} \left(\begin{bmatrix} B & 0 \\ 0 & A^* \end{bmatrix} \begin{bmatrix} x \\ \psi \end{bmatrix} \right) = \begin{bmatrix} C - DR^{-1}S^* & -DR^{-1}D^* \\ W - SR^{-1}S^* & C^* - SR^{-1}D^* \end{bmatrix} \begin{bmatrix} x \\ \psi \end{bmatrix}. \tag{6.4}$$

If one tries to solve the system (6.2), (6.3), one is confronted by the index of the DAE (6.4). Equation (6.4) has a properly stated leading term since (2.2) has one. Equation (6.4) is a square system having $m+k$ equations and $m+k$ unknown functions, respectively.

THEOREM 6.2. *If A and B remain nonsingular,* (6.4) *represents an implicit regular ODE (regular DAE with tractability index zero). Otherwise, for the DAE* (6.4) *to be regular with tractability index one, it is necessary and sufficient that the following two conditions are satisfied:*

$$[AB - CQ,\ D] \quad \textit{has full row rank } k, \tag{6.5}$$

$$\operatorname{im}[Q(C^* - SR^{-1}D^*)Q_*,\ Q(W - SR^{-1}S^*)Q] = \operatorname{im} Q. \tag{6.6}$$

Proof. In the case of nonsingular A and B, the assertion is obvious. Let A, B be singular. In [3], the pair of conditions

$$[AB - CQ,\ D] \quad \text{has full row rank } k, \tag{6.7}$$

and

$$\begin{bmatrix} B^*A^* - C^*Q_* & WQ & S \\ -D^*Q_* & S^*Q & R \end{bmatrix} \quad \text{has full row rank } m + l \tag{6.8}$$

was shown to be necessary and sufficient for the DAE

$$\begin{bmatrix} A & 0 \\ 0 & -B^* \\ 0 & 0 \end{bmatrix} \frac{d}{dt} \left(\begin{bmatrix} B & 0 & 0 \\ 0 & A^* & 0 \end{bmatrix} \begin{bmatrix} x \\ \psi \\ u \end{bmatrix} \right) = \begin{bmatrix} C & 0 & D \\ W & C^* & S \\ S^* & D^* & R \end{bmatrix} \begin{bmatrix} x \\ \psi \\ u \end{bmatrix} \tag{6.9}$$

to be regular with tractability index one. Clearly, (6.9) is regular with index one if (6.4) is, and vice versa. Hence, the above two conditions are valid for (6.4), too. The condition (6.7) coincides with (6.5). Taking into account the invertibility of R, the second condition (6.8) is equivalent to the injectivity of

$$\begin{bmatrix} AB \\ Q_*(C - DR^{-1}S^*) \\ Q(W - SR^{-1}S^*) \end{bmatrix},$$

but this is equivalent to (6.6). (Notice that in [3] slightly more general problems with R positive semidefinite are considered.) □

Remark 6.3. In [5], descriptor systems (1.6) in an *SVD coordinate system* play a special role, and, in particular, the invertibility of a certain matrix $\bar{R}$ (cf. [5]) is a basic property assumed to be given in all four versions of the Riccati differential equations studied in [BeLa, section IV]. From the viewpoint of DAE theory, for those very special systems (6.9), the invertibility of $\bar{R}$ exactly means regularity with tractability index one (cf. [3]).

Remark 6.4. Recall from [3] that, if the system (6.9) is regular with tractability index one, and, additionally, $\ker B^* = 0$, then the so-called inherent regular ODEs of (6.9), and of (6.2), (6.3), actually have a Hamiltonian structure—a property that should be useful concerning the solvability of BVPs for the Hamiltonian system (6.2), (6.3). However, notice that in general it may happen that the so-called Hamiltonian system (6.2), (6.3) may lose the inherent Hamiltonian structure (cf. [3]).

We end up with an assertion saying that if the Hamiltonian system (6.2), (6.3) somehow has good solvability, then the Riccati DAE system (3.1), (3.2) is solvable at the same time.

THEOREM 6.5. *Let $X(t) \in L(\mathbb{R}^m, \mathbb{R}^m), \Psi(t) \in L(\mathbb{R}^m, \mathbb{R}^k)$ be continuous on [0,T], and such that their m columns belong to $\mathcal{C}_B^1$ and $\mathcal{C}_{A^*}^1$, respectively, and*

$$A(BX)' = (C - DR^{-1}S^*)X - DR^{-1}D^*\Psi, \tag{6.10}$$

$$-B^*(A^*\Psi)' = (W - SR^{-1}S^*)X + (C^* - SR^{-1}D^*)\Psi \tag{6.11}$$

is satisfied.

Let X be nonsingular and let $X^{-1}B^-$ belong to $\mathcal{C}^1$. Let $Y := \Psi X^{-1}$ be such that

$$P_*YQ = 0, \quad A^*YB^- = B^{-*}Y^*A.$$

Then, Y is continuous with a continuously differentiable part A^*YB^- and satisfies the Riccati DAE system (3.1), (3.2).

Proof. Here condition (3.2) is given, and $A^*YB^- = A^*YXX^{-1}B^- = A^*\Psi X^{-1}B^-$ belongs to $\mathcal{C}^1$. Derive from (6.11) that

$$B^*(A^*YX)'X^{-1} = -(W - SR^{-1}S^*) - (C^* - SR^{-1}D^*)Y.$$

By means of $B^*(A^*YX)'X^{-1} = B^*(A^*YB^-BX)'X^{-1} = B^*(A^*YB^-)'B + B^*A^*YB^-(BX)'X^{-1} = B^*(A^*YB^-)'B + Y^*A(BX)'X^{-1}$ and taking into account (6.10), (6.11), we obtain (3.1). ∎

If X, Ψ in Theorem 6.5 are chosen to meet the terminal conditions $B(T)^*A(T)^*\Psi(T) = V$, $A(T)B(T)X(T) = A(T)B(T)$, then it follows that $A(T)^*Y(T)B(T)^- = B(T)^{-*}VB(T)^-$; that is, the terminal condition (3.3) is satisfied.

7. Final remark. We have shown that optimal feedback controls of linear-quadratic optimal control problems with constraints described by general linear DAEs with variable coefficients can be obtained by suitably formulating a Riccati DAE system, similarly to the classical example in which the constraints are described by explicit ODEs. Compared to earlier papers and some less suitable Riccati DAEs, we could do without several restrictive assumptions.

Furthermore, we would like to stress that it is not necessary and probably not even reasonable to transform the DAE describing the constraints (descriptor system) or the DAE describing the Hamiltonian system with great expense into a special canonical form.

What is on the agenda is the development of feasible solution methods for the Riccati DAE (3.1), (3.2).

REFERENCES

[1] H. Abou-Kandil, G. Freiling, V. Ionescu, and G. Jank, *Matrix Riccati Equations in Control and Systems Theory*, Birkhäuser Verlag, Basel, Boston, Berlin, 2003.

[2] A. Backes, *Extremalbedingungen für Optimierungs-Probleme mit Algebro-Differentialgleichungen*, Dissertation, Mathematisch-Naturwissenschaftliche Fakultät II, Humboldt-Universität zu Berlin, Berlin, Germany, 2006.

[3] K. Balla, G. A. Kurina, and R. März, *Index criteria for differential-algebraic equations arising from linear-quadratic optimal control problems*, J. Dyn. Control Syst., 12 (2006), pp. 289–311.

[4] K. Balla and R. März, *A unified approach to linear differential algebraic equations and their adjoints*, Z. Anal. Anwend., 21(2002), pp. 783–802.

[5] D. J. Bender and A. J. Laub, *The linear-quadratic optimal regulator for descriptor systems*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 672–688.

[6] S. L. Campbell, *Singular Systems of Differential Equations* II, Pitman, San Francisco, London, Melbourne, 1982.

[7] S. L. Campbell and R. März, *Direct transcription solution of high index optimal control problems and regular Euler–Lagrange equations*, J. Comput. Appl. Math., 202 (2007), pp. 189–202.

[8] H. Döring, *Traktabilitätsindex und Eigenschaften von matrixwertigen Riccati-Typ Algebro-differentialgleichungen*, Diplomarbeit, Institut für Mathematik, Humboldt-Universität zu Berlin, Berlin, Germany, 2004.

[9] D. Estévez Schwarz and C. Tischendorf, *Structural analysis of electric circuits and consequences for MNA*, Internat. J. Circuit Theory Appl., 28 (2000), pp. 131–162.

[10] I. Higueras and R. März, *Differential algebraic equations with properly stated leading term*, Comput. Math. Appl., 48 (2004), pp. 215–235.

[11] I. Higueras, R. März, and C. Tischendorf, *Stability preserving integration of index-1 DAEs*, Appl. Numer. Math., 45 (2003), pp. 175–200.

[12] I. V. Klinskih and G. A. Kurina, *Feedback control for a class of descriptor systems*, in Theory of Evolution Equations, International Conference Fifth Bogolyubov's Reading, Abstracts of Reports, Kamyanets-Podilskyy, 2002, p. 85 (in Russian).

[13] G. A. Kurina and R. März, *On linear-quadratic optimal control problems for time-varying descriptor systems*, SIAM J. Control Optim., 42 (2004), pp. 2062–2077.

[14] P. Kunkel and V. Mehrmann, *The linear quadratic optimal control problem for linear descriptor systems with variable coefficients*, Math. Control Signals Systems, 10 (1997), pp. 247–264.

[15] G. A. Kurina, *Design of Feedback Control for Linear Control Systems Unresolved with Respect to Derivative*, unpublished paper N 3619-82, VINITI, Voronezh, 1982 (in Russian).

[16] G. A. Kurina, *Feedback control for linear systems unresolved with respect to derivative*, Avtomat. i Telemekh., 6 (1984), pp. 37–41 (in Russian); Automat. Remote Control, 45 (1984), pp. 713–717 (in English).

[17] G. A. Kurina, *On operator Riccati equation unresolved with respect to derivative*, Differencial. Uravnen., 22 (1986), pp. 1826–1829 (in Russian).

[18] G. A. Kurina, *Singular perturbations of control problems with equation of state not solved for the derivative (a survey)*, J. Comput. System Sci. Internat., 31 (1993), pp. 17–45.

[19] G. A. Kurina, *Feed-back control for time-varying descriptor systems*, Systems Sci., 26 (2000), pp. 47–59.

[20] E. B. Lee and L. Markus, *Foundations of optimal control theory*, John Wiley & Sons, Inc., New York, London, Sydney, 1967.

[21] R. März, *Numerical methods for differential algebraic equations*, in Acta Numerica, Cambridge University Press, Cambridge, UK, 1992, pp. 141–198.

[22] R. März, *The index of linear differential algebraic equations with properly stated leading terms*, Results in Math., 42 (2002), pp. 308–338.

[23] R. März, *Fine decouplings of regular differential algebraic equations*, Results in Math., 46 (2004), pp. 57–72.

[24] R. März, *Differential algebraic systems with properly stated leading term and MNA equations*, in Modeling, Simulation and Optimization of Integrated Circuits, K. Antreich, R. Bulirsch, A. Gilg, and P. Rentrop, eds., Birkhäuser Verlag, Basel, 2003, pp. 135–151.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1299–1322

PASSIVITY AND PASSIFICATION FOR NETWORKED CONTROL SYSTEMS*

HUIJUN GAO†, TONGWEN CHEN‡, AND TIANYOU CHAI§

Abstract. This paper investigates the problems of passivity analysis and passification for network-based linear control systems. A new sampled-data model is first formulated based on the updating instants of the ZOH (zeroth order hold), where the physical plant and the controller are, respectively, in continuous time and discrete time. In this model, network-induced delays, data packet dropouts, and signal measurement quantization have been taken into account. The measurement quantizer is assumed to be logarithmic, and the network-induced delays are assumed to have both a lower bound and an upper bound, which is more general than those assumptions used in the literature. The key idea is to transform the sampled-data model into a linear system with two successive delay components in the state. Then, by using a Lyapunov–Krasovskii approach plus the free weighting matrix technique, a passivity performance condition is formulated in the form of linear matrix inequalities (LMIs). Based on this condition, two procedures are proposed for designing passification controllers, which guarantee that the closed-loop networked control system (NCS) is passive. Finally, two illustrative examples are presented: one shows the advantage of introducing the lower bound of transmission delays and shows how much the quantization behavior affects the passivity performance; the other illustrates the applicability and effectiveness of the proposed passification results.

Key words. networked control systems, passivity and passification, quantization, sampled-data systems

AMS subject classification. 93A30

DOI. 10.1137/060655110

1. Introduction.

It is well known that in many practical systems, the physical plant, controller, sensor, and actuator are difficult to locate in the same place, and thus signals are required to be transmitted from one place to another. In modern industrial systems, these components are often connected over network media (typically digital band-limited serial communication channels), giving rise to the so-called networked control systems (NCSs). Compared with traditional feedback control systems, where these components are usually connected via point-to-point cables, the introduction of communication network media brings great advantages, such as low cost, reduced weight and power requirements, simple installation and maintenance, and high reliability [21]. Therefore, NCSs have received more and more attention and have become more and more popular in many practical applications in recent years. Modeling, analysis, and synthesis of network-based feedback systems with limited

*Received by the editors March 24, 2006; accepted for publication (in revised form) February 22, 2007; published electronically September 12, 2007. This work was partially supported by the Natural Sciences and Engineering Research Council of Canada, an Alberta Ingenuity Fellowship, an Honorary Izaak Walton Killam Memorial Postdoctoral Fellowship, National Natural Science Foundation of China (60528007, 60504008), Program for New Century Excellent Talents in University of China, and the Key Laboratory of Integrated Automation for the Process Industry (Northeastern University), Ministry of Education, China.

http://www.siam.org/journals/sicon/46-4/65511.html

†Space Control and Inertial Technology Research Center, Harbin Institute of Technology, Harbin, Heilongjiang Province, 150001, China (hjgao@hit.edu.cn).

‡Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2V4, Canada (tchen@ece.ualberta.ca).

§Research Center of Automation, Northeastern University, Shenyang 110004, Liaoning Province, China (tychai@mail.neu.edu.cn).

communication capability has emerged as a topic of significant interest in the control community, a topic which is highlighted in the recent special issue edited by Antsaklis and Baillieul [2]. From among the reported results on NCSs, we mention a few: the stability issue is investigated in [29, 36, 48], stabilizing controllers are designed in [43, 44, 47, 49], performance preserved control is studied in [23, 28, 33, 45], and moving horizon control is proposed in [17].

What makes an NCS distinct from traditional feedback control systems? To answer this question, we need to look at a few phenomena that typically exist for a network-based control system. Probably the most significant are the network-induced delays, which are usually caused by the limited bit rate of the communication channels, by a node waiting to send out a packet via a busy channel, or by signal processing and propagation. The existence of time delays generally brings negative effects on the stability and performance of NCSs. The time delays in a typical NCS usually take the forms of input delay and output delay, which are essentially different from the state-delayed models, for which a great number of results have been reported recently [18, 19]. The second interesting problem in an NCS is the packet dropout (or data missing [37, 38]) phenomenon, which is usually caused by unavoidable errors or losses in the transmitted packet. Though many NCSs employ automatic repeat request mechanisms, packet dropout phenomenon is still unavoidable. Moreover, packet dropouts may occur if one packet sampled at the sensor node reaches the destination later than its successors. In this situation, it is natural to use the most updated packet by dropping out the old ones, giving rise to packet dropout phenomenon. Another important issue in NCSs is the quantization effect. In NCSs, the measurement and command signals are usually quantized before being communicated, and the number of quantization levels is closely related to the information flow between the components of the control system and thus to the capacity required to transmit the information. The classical control theory, which is based on the standard assumption that data transmission required by the system can be performed with infinite precision, may not be valid in the presence of signal quantization or capacity-limited feedback, and therefore there is a need for developing tools for analysis and design of quantized feedback systems. Many important results on quantized control have been reported; see, for instance, [3, 6, 11, 12, 21, 22, 24, 25, 32, 34] and the references therein.

On the other hand, the notion of passivity plays an important role in the analysis and design of linear and nonlinear systems. In the first place, many systems need to be passive in order to attenuate noises effectively. In the second place, the robustness measure (such as robust stability or robust performance) of a system often reduces to a subsystem or a modified system that is passive. Passivity analysis is a major tool for studying stability of uncertain or nonlinear systems, especially for high-order systems, and thus the passivity analysis approach has been used in control problems for a long time to deal with robust stability problems for complex uncertain systems (see [5, 8, 27, 30, 35, 40] and the references therein). Apart from its direct applications, the notion of passivity is closely related to bounded realness, which is an equally important notion in control. In fact, it is well known that there is a one-to-one relationship between bounded realness and passivity [1]. Consequently, bounded realness analysis can be converted into passivity analysis and vice versa. Very recently, the passivity-based control has been investigated for a few classes of complex systems, including time-delay systems [31, 42], two-dimensional systems [41], fuzzy systems [7], and signal processing systems [39]. To the best of our knowledge, however, the problems of passivity analysis and passification for NCSs have not been investigated and still

remain challenging, which motivates the present study.

In this paper, we investigate the problems of passivity analysis and passification for network-based linear control systems. A new sampled-data model is first formulated based on the updating instants of the ZOH (zeroth order hold) (instead of the sampling instants), where the physical plant and the controller are, respectively, in continuous time and discrete time. In this model, network-induced delays, data packet dropouts, and signal measurement quantization have been taken into account. The measurement quantizer is assumed to be logarithmic, and the network-induced delays are assumed to have both a lower bound and an upper bound, which is more general than those assumptions used in the literature. The key idea is to transform the sampled-data model into a linear system with two successive delay components in the state. Then, by using a Lyapunov–Krasovskii approach plus the free weighting matrix technique, a passivity performance condition is formulated in the form of linear matrix inequalities (LMIs). Based on this condition, two procedures are proposed for designing passification controllers, which guarantee that the closed-loop NCS is passive. Finally, two illustrative examples are presented: one shows the advantage of introducing the lower bound of transmission delays, and how much the quantization behavior affects the passivity performance; the other illustrates the applicability and effectiveness of the proposed passification results.

The remainder of this paper is organized as follows. The problems of passivity analysis and passification for network-based linear control systems are formulated in section 2. Sections 3 and 4 present the main results on passivity analysis and passification, respectively. Section 5 gives two illustrative examples, and we conclude the paper in section 6.

Notation. The notation used throughout the paper is fairly standard. The superscript T stands for matrix transposition, $\mathbb{R}^n$ denotes the n-dimensional Euclidean space, and $P > 0$ (≥ 0) means that P is real symmetric and positive definite (semidefinite). In symmetric block matrices or complex matrix expressions, we use an asterisk ($*$) to represent a term that is induced by symmetry, and diag$\{\ldots\}$ stands for a block-diagonal matrix. Matrices, if their dimensions are not explicitly stated, are assumed to be compatible with algebraic operations.

2. Problem formulation. Consider a typical NCS, shown in Figure 1. Suppose the physical plant is given by the following linear system:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ z(t) &= Cx(t) + Du(t). \end{aligned} \tag{1}$$

Here $x(t) \in \mathbb{R}^n$ is the state vector; $u(t) \in \mathbb{R}^p$ is the input; $z(t) \in \mathbb{R}^q$ is the output; and A, B, C, D are system matrices with appropriate dimensions.

In Figure 1, it is assumed that the sampler is clock-driven, while the quantizer, controller, and actuator are event-driven. The sampling period is assumed to be h, where h is a positive real constant, and we denote the *sampling instant* of the sampler as s_k, $k = 1, \ldots, \infty$. In addition, it is assumed that the state variable $x(t)$ is online measurable, and the measurements of $x(t)$ are first quantized via a quantizer and then transmitted with a single packet. The quantizer is denoted as $f(\cdot) = \left[\begin{array}{cccc} f_1(\cdot) & f_2(\cdot) & \cdots & f_n(\cdot) \end{array}\right]^T$, which is assumed to be symmetric, that is, $f_j(-v) = -f_j(v)$, $j = 1, \ldots, n$. In this paper, we are interested in the logarithmic static and time-invariant quantizer. For each $f_j(\cdot)$, the set of quantized levels is

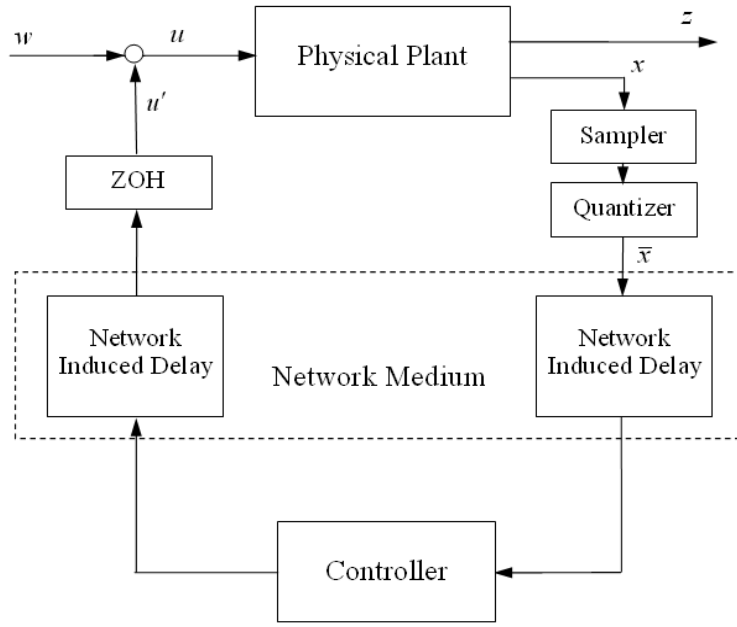


FIG. 1. *An NCS.*

described by

$$\mathcal{U}_j = \left\{ \pm u_i^{(j)}, \quad i = 0, \pm 1, \pm 2, \ldots \right\} \cup \{0\}. \tag{2}$$

According to [10, 13], a quantizer is called *logarithmic* if the set of quantized levels is characterized by

$$\mathcal{U}_j = \left\{ \pm u_i^{(j)},\ u_i^{(j)} = \rho_j^i u_0^{(j)},\ i = \pm 1, \pm 2, \ldots \right\} \cup \left\{ \pm u_0^{(j)} \right\} \cup \{0\}, \quad 0 < \rho_j < 1, \quad u_0^{(j)} > 0. \tag{3}$$

Each of the quantization levels $u_i^{(j)}$ corresponds to a segment such that the quantizer maps the whole segment to this quantization level. In addition, these segments form a partition of $\mathbb{R}$; that is, they are disjoint and their union equals $\mathbb{R}$. For the logarithmic quantizer, the associated quantizer $f_j(\cdot)$ is defined as follows:

$$f_j(v) = \begin{cases} u_i^{(j)} & \text{if } \frac{1}{1+\sigma_j} u_i^{(j)} < v \le \frac{1}{1-\sigma_j} u_i^{(j)}, \quad v > 0, \\ 0 & \text{if } v = 0, \\ -f_j(-v) & \text{if } v < 0, \end{cases} \tag{4}$$

where

$$\sigma_j = \frac{1-\rho_j}{1+\rho_j}. \tag{5}$$

Then, at the sampling instant s_k, we have

$$\bar{x}(s_k) = f(x(s_k)) = \left[\begin{array}{cccc} f_1(x_1(s_k)) & f_2(x_2(s_k)) & \cdots & f_n(x_n(s_k)) \end{array} \right]^T.$$

Now denote the *updating instant* of the ZOH as t_k, $k = 1, \ldots, \infty$, and suppose that the updating signal (the successfully transmitted signal from the sampler to the controller and to the ZOH) at the instant t_k has experienced signal transmission delays η_k ($\eta_k = \tau_k + d_k$, where τ_k is the delay from the quantizer to the controller and d_k is the delay from the controller to the ZOH. It is assumed that there is no delay between the sensor and quantizer). Therefore, the state-feedback controller takes the following form:

$$u'(t_k) = Kf(x(t_k - \eta_k)), \tag{6}$$

where K is the state-feedback control gain. Thus, considering the behavior of the ZOH, we have

$$u(t) = Kf(x(t_k - \eta_k)) + w(t), \quad t_k \le t < t_{k+1}, \tag{7}$$

with t_{k+1} being the next updating instant of the ZOH after t_k, and $w(t) \in \mathbb{R}^p$ being the external input.

A natural assumption on the network-induced delays η_k can be made as follows:

$$\eta_m \le \eta_k \le \eta_M, \tag{8}$$

where η_m and η_M denote the minimum and the maximum delays, respectively. In addition, at the updating instant t_k the number of accumulated data packet dropouts since the last updating instant t_{k-1} is denoted as δ_k. We assume that the maximum number of data packet dropouts is $\bar{\delta}$, that is,

$$\delta_k \le \bar{\delta}. \tag{9}$$

Then, it can be seen from (8) and (9) that

$$t_{k+1} - t_k = (\delta_k + 1)h + \eta_{k+1} - \eta_k. \tag{10}$$

Remark 1. It is worth noting that the assumption on the network-induced delays η_k made in (8) is more general than those assumptions in [43, 44, 45]. The main difference lies in the lower bound we introduced. By assuming $\eta_m = 0$, we see that (8) is the same as those assumptions in [43, 44, 45]. The introduction of the lower bound η_m will be shown later, via a numerical example, and will be advantageous for reducing conservativeness.

Therefore, from (1)–(7) we obtain the following closed-loop system:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + BKf(x(t_k - \eta_k)) + Bw(t), \\ z(t) &= Cx(t) + DKf(x(t_k - \eta_k)) + Dw(t), \\ & t_k \le t < t_{k+1}. \end{aligned} \tag{11}$$

Remark 2. It is important to note that in (7), t_k refers to the *updating instant* of the ZOH. While in [43], the controller is expressed as

$$u(t) = F\bar{x}(t_k), \quad t_k \le t < t_{k+1}, \tag{12}$$

with t_k standing for the *sampling instant*. It should be noted that when the controller and actuator are event-driven, we cannot use the sampling instant to model the behavior of the ZOH. The reason is that the signal transmission delays may not

necessarily be integer multiples of the sampling period, and thus the ZOH may be updated between sampling instants. By using the updating instant in this paper, we do not need to synchronize the ZOH and the sampler, and thus the networked control model formulated here is essentially different from that in [43] and is more general, though they appear to be similar.

Before proceeding further, we introduce the following definition [26].

DEFINITION 1. *The closed-loop NCS in* (11) *is said to be passive if there exists a scalar* $\gamma > 0$ *such that*

$$2\int_0^T w^T(t)z(t)dt \geq -\gamma \int_0^T w^T(t)w(t)dt \tag{13}$$

for all $T > 0$ *under zero initial conditions.*

Then, the problems to be addressed in this paper are expressed as follows.

PROBLEM 1 (passivity analysis). *Consider the NCS in Figure* 1. *Given the system matrices* A, B, C, D *in* (1) *and the controller gain matrix* K *in* (6), *determine under what condition the closed-loop networked control system in* (11) *is passive in the sense of Definition* 1.

PROBLEM 2 (passification). *Consider the NCS in Figure* 1. *Given the system matrices* A, B, C, D *in* (1), *determine the controller gain matrix* K *in* (6) *such that the closed-loop NCS in* (11) *is passive in the sense of Definition* 1.

3. Passivity analysis. This section is concerned with the problem of passivity analysis. More specifically, assuming that the matrices A, B, C, D in (1) and the controller gain matrix K in (6) are known, we shall study the conditions under which the closed-loop NCS in (11) is passive in the sense of Definition 1. The following theorem shows that the closed-loop passivity can be guaranteed if there exist some matrices satisfying certain LMIs. This theorem will play an instrumental role in the problem of passification for NCSs (the proof is given in the appendix).

THEOREM 1. *Consider the NCS in Figure* 1. *Given the matrices* A, B, C, D, *the controller gain matrix* K, *and a positive constant* γ, *the closed-loop system in* (11) *is passive in the sense of Definition* 1 *if there exist matrices* $P > 0$; $Q > 0$; $M_i > 0$, $i = 1, 2$; U_i, V_i, $i = 1, \ldots, 4$; *and a diagonal matrix* $R > 0$ *satisfying*

$$\begin{bmatrix}
\Gamma_{11} & \Gamma_{12} & \Gamma_{13} & \Gamma_{14} & U_1 & V_1 & A^TM_1 & A^TM_2 & PBK \\
* & \Gamma_{22} & \Gamma_{23} & \Gamma_{24} & U_2 & V_2 & 0 & 0 & 0 \\
* & * & \Gamma_{33}+\Lambda^2R & \Gamma_{34} & U_3 & V_3 & K^TB^TM_1 & K^TB^TM_2 & 0 \\
* & * & * & \Gamma_{44} & U_4 & V_4 & B^TM_1 & B^TM_2 & -DK \\
* & * & * & * & -\eta_m^{-1}M_1 & 0 & 0 & 0 & 0 \\
* & * & * & * & * & -\kappa^{-1}M_2 & 0 & 0 & 0 \\
* & * & * & * & * & * & -\eta_m^{-1}M_1 & 0 & M_1BK \\
* & * & * & * & * & * & * & -\kappa^{-1}M_2 & M_2BK \\
* & * & * & * & * & * & * & * & -R
\end{bmatrix} < 0, \tag{14}$$

where

$$\begin{aligned}
&\Gamma_{11} = PA + A^TP + Q + U_1 + U_1^T, && \Gamma_{12} = -U_1 + U_2^T + V_1, \\
&\Gamma_{22} = -Q - U_2 - U_2^T + V_2 + V_2^T, && \Gamma_{13} = U_3^T - V_1 + PBK, \\
&\Gamma_{23} = -U_3^T - V_2 + V_3^T, && \Gamma_{33} = -V_3 - V_3^T, \\
&\Gamma_{14} = PB + U_4^T - C^T, && \Gamma_{24} = -U_4^T + V_4^T, \\
&\Gamma_{34} = -V_4^T - K^TD^T, && \Gamma_{44} = -\gamma I - D - D^T, \\
&\kappa = \eta_M - \eta_m + \left(\bar{\delta} + 1\right)h, && \Lambda = \operatorname{diag}\{\sigma_1, \ldots, \sigma_n\}.
\end{aligned} \tag{15}$$

Theorem 1 deserves some remarks.

Remark 3. The basic idea behind Theorem 1 is to transform the sampled-data system in (11) into the state-delay system in (40). The passivity of the transformed time-delay system is then analyzed by defining a new Lyapunov–Krasovskii functional plus free weighting matrix techniques. The most significant feature is that no model transformation has been performed to the delay system in (40), which is essentially different from the results obtained in [43] based on a descriptor model transformation. This helps us avoid using a bounding technique for seeking upper bounds of the inner product between two vectors. Similar ideas appear in [19, 20], whose techniques have been shown to be potentially less conservative than those using model transformation methods.

Remark 4. It is worth noting that if the closed-loop NCS in (11) is passive according to Theorem 1, the asymptotic stability of (11) with $w(t) = 0$ is also guaranteed. This is shown as follows: First, define the Lyapunov–Krasovskii functional in (44). Then, by following along lines similar to the proof Theorem 1, we see that the time derivative of $V(t)$ along the solution of (41) with $w(t) = 0$ is given by

$$\begin{aligned}\dot{V}(t) &\le \bar{\zeta}^T(t)\left(\bar{\Gamma} + \eta_m \bar{U} M_1^{-1} \bar{U}^T + \kappa \bar{V} M_2^{-1} \bar{V}^T\right)\bar{\zeta}(t) \\ &- \int_{t-\eta_m}^{t} \left[\bar{\zeta}^T(t)\bar{U} + \dot{x}^T(\alpha) M_1\right] M_1^{-1} \left[\bar{U}^T \bar{\zeta}(t) + M_1 \dot{x}(\alpha)\right] d\alpha \\ &- \int_{t-\eta_m-\eta(t)}^{t-\eta_m} \left[\bar{\zeta}^T(t)\bar{V} + \dot{x}^T(\alpha) M_2\right] M_2^{-1} \left[\bar{V}^T \bar{\zeta}(t) + M_2 \dot{x}(\alpha)\right] d\alpha,\end{aligned}$$

where

$$\bar{\zeta}(t) = \begin{bmatrix} x(t) \\ x(t-\eta_m) \\ x(t-\eta_m-\eta(t)) \end{bmatrix}, \quad \bar{U} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix},$$
$$\bar{V} = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix}, \quad \bar{\Gamma} = \begin{bmatrix} \bar{\Gamma}_{11} & \Gamma_{12} & \bar{\Gamma}_{13} \\ * & \Gamma_{22} & \Gamma_{23} \\ * & * & \bar{\Gamma}_{33} \end{bmatrix}.$$

By following along lines similar to the proof of Theorem 1, we see that (14) guarantees $\bar{\Gamma} + \eta_m \bar{U} M_1^{-1} \bar{U}^T + \kappa \bar{V} M_2^{-1} \bar{V}^T < 0$, and the asymptotic stability is established.

Remark 5. If there is no quantizer in the NCS shown in Figure 1, then (40) in the above proof reads

$$\begin{aligned}\dot{x}(t) &= Ax(t) + BKx(t-\eta_m-\eta(t)) + Bw(t), \\ z(t) &= Cx(t) + DKx(t-\eta_m-\eta(t)) + Dw(t).\end{aligned} \tag{16}$$

Then, we have the following corollary, which can be proved by following arguments similar to the proof of Theorem 1.

COROLLARY 1. *Consider the NCS in Figure* 1*, but without the quantizer. Given the matrices* A, B, C, D*, the controller gain matrix* K *and a positive constant* γ*, the closed-loop system in* (16) *is passive in the sense of Definition* 1 *if there exist matrices*

$P > 0$; $Q > 0$; $M_i > 0$, $i = 1, 2$; U_i, V_i, $i = 1, \ldots, 4$, *satisfying*

$$\begin{bmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{13} & \Gamma_{14} & U_1 & V_1 & A^T M_1 & A^T M_2 \\ * & \Gamma_{22} & \Gamma_{23} & \Gamma_{24} & U_2 & V_2 & 0 & 0 \\ * & * & \Gamma_{33} & \Gamma_{34} & U_3 & V_3 & K^T B^T M_1 & K^T B^T M_2 \\ * & * & * & \Gamma_{44} & U_4 & V_4 & B^T M_1 & B^T M_2 \\ * & * & * & * & -\eta_m^{-1} M_1 & 0 & 0 & 0 \\ * & * & * & * & * & -\kappa^{-1} M_2 & 0 & 0 \\ * & * & * & * & * & * & -\eta_m^{-1} M_1 & 0 \\ * & * & * & * & * & * & * & -\kappa^{-1} M_2 \end{bmatrix} < 0,$$

where Γ_{ij} *is given in* (15).

Remark 6. It is worth noting that in the proof of Theorem 1, the transformed system in (40) contains two successive delay components η_m and $\eta(t)$, where η_m is a constant delay, and $\eta(t)$ is a nondifferentiable time-varying delay with bound κ. If the lower bound of the network-induced delays is assumed to be zero, that is, $\eta_m = 0$, (40) takes the following form:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + BKf(x(t - \eta(t))) + Bw(t), \\ z(t) &= Cx(t) + DKf(x(t - \eta(t))) + Bw(t), \end{aligned} \tag{17}$$

with

$$0 \le \eta(t) \le \bar{\kappa}, \tag{18}$$

where

$$\bar{\kappa} = \eta_M + \left(\bar{\delta} + 1\right) h. \tag{19}$$

Compared with (39), the upper bound of $\eta(t)$ in (18) is increased by η_m. In other words, without taking the lower bound of the transmission delays into consideration, η_m will be treated as a nondifferentiable time-varying delay instead of a constant one when it is nonzero. Therefore, the introduction of the lower bound η_m will naturally reduce conservativeness, which will be showed via a numerical example later. However, the existing results on NCSs, such as [43, 44, 45], did not offer to take the lower bound η_m into consideration. The following corollary gives a passivity analysis result for the case when $\eta_m = 0$.

COROLLARY 2. *Consider the NCS in Figure* 1*, and suppose the network-induced delays satisfy* $0 \le \eta_k \le \eta_M$*. Given the matrices* A, B, C, D*, the controller gain matrix* K *and a positive constant* γ*, the closed-loop system in* (17) *is passive in the sense of Definition* 1 *if there exist matrices* $P > 0$; $M > 0$; U_i, $i = 1, 2, 3$; *and a diagonal matrix* $R > 0$ *satisfying*

$$\begin{bmatrix} \Pi_{11} & \Pi_{12} & \Pi_{13} & U_1 & A^T M & PBK \\ * & \Pi_{22} + \Lambda^2 R & \Pi_{23} & U_2 & K^T B^T M & 0 \\ * & * & \Pi_{33} & U_3 & B^T M & -DK \\ * & * & * & -\bar{\kappa}^{-1} M & 0 & 0 \\ * & * & * & * & -\bar{\kappa}^{-1} M & MBK \\ * & * & * & * & * & -R \end{bmatrix} < 0, \tag{20}$$

where $\bar{\kappa}$ *is as given in* (19)*,* Λ *is as given in* (15)*, and*

$$\begin{aligned} \Pi_{11} &= PA + A^T P + U_1 + U_1^T, & \Pi_{12} &= PBK - U_1 + U_2^T, \\ \Pi_{22} &= -U_2 - U_2^T, & \Pi_{13} &= PB - C^T + U_3^T, \\ \Pi_{23} &= -K^T D^T - U_3^T, & \Pi_{33} &= -D^T - D - \gamma I. \end{aligned}$$

Proof. The proof follows along lines similar to the proof of Theorem 1 and thus is outlined briefly. First, by considering the quantization behavior, the closed-loop NCS in (17) can be transformed into

$$\begin{aligned} \dot{x}(t) &= Ax(t) + BK(I + \Lambda(t))(x(t - \eta(t))) + Bw(t), \\ z(t) &= Cx(t) + DK(I + \Lambda(t))(x(t - \eta(t))) + Dw(t) \end{aligned} \tag{21}$$

with $\Lambda(t)$ as given in (42). Define the following Lyapunov–Krasovskii functional:

$$V(t) = x^T(t)Px(t) + \int_{-\bar{\kappa}}^{0} \int_{t+\beta}^{t} \dot{x}^T(\alpha) M \dot{x}(\alpha) d\alpha d\beta, \tag{22}$$

where $P > 0$ and $M > 0$ are matrices to be determined. Then, the corollary can be proved by following along lines similar to the proof of Theorem 1. □

4. Passification. This section is devoted to solving the problem of passification for NCSs.

PROPOSITION 1. *Consider the NCS in Figure* 1. *Given a positive constant* γ, *there exists a state-feedback controller in the form of* (6) *such that the closed-loop system in* (11) *is passive in the sense of Definition* 1 *if there exist matrices* $\bar{P} > 0$; $\bar{Q} > 0$; $\bar{M}_i > 0$, $i = 1, 2$; $\bar{U}_i$, $\bar{V}_i$, $i = 1, \ldots, 4$; $\bar{K}$; *and a diagonal matrix* $\bar{R} > 0$ *satisfying*

(23)

$$\begin{bmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} & \Omega_{14} & \bar{U}_1 & \bar{V}_1 & \bar{P}A^T & \bar{P}A^T & B\bar{K} & 0 \\ * & \Omega_{22} & \Omega_{23} & \Omega_{24} & \bar{U}_2 & \bar{V}_2 & 0 & 0 & 0 & 0 \\ * & * & \Omega_{33} & \Omega_{34} & \bar{U}_3 & \bar{V}_3 & \bar{K}^T B^T & \bar{K}^T B^T & 0 & \bar{P} \\ * & * & * & \Omega_{44} & \bar{U}_4 & \bar{V}_4 & B^T & B^T & -D\bar{K} & 0 \\ * & * & * & * & -\eta_m^{-1} \bar{P} \bar{M}_1^{-1} \bar{P} & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & -\kappa^{-1} \bar{P} \bar{M}_2^{-1} \bar{P} & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & -\eta_m^{-1} \bar{M}_1 & 0 & B\bar{K} & 0 \\ * & * & * & * & * & * & * & -\kappa^{-1} \bar{M}_2 & B\bar{K} & 0 \\ * & * & * & * & * & * & * & * & -\bar{P}\bar{R}^{-1}\bar{P} & 0 \\ * & * & * & * & * & * & * & * & * & -\Lambda^{-2}\bar{R} \end{bmatrix} < 0,$$

where κ *and* Λ *are given in* (15) *and*

$$\begin{aligned} \Omega_{11} &= A\bar{P} + \bar{P}A^T + \bar{Q} + \bar{U}_1 + \bar{U}_1^T, & \Omega_{12} &= -\bar{U}_1 + \bar{U}_2^T + \bar{V}_1, \\ \Omega_{22} &= -\bar{Q} - \bar{U}_2 - \bar{U}_2^T + \bar{V}_2 + \bar{V}_2^T, & \Omega_{13} &= \bar{U}_3^T - \bar{V}_1 + B\bar{K}, \\ \Omega_{23} &= -\bar{U}_3^T - \bar{V}_2 + \bar{V}_3^T, & \Omega_{33} &= -\bar{V}_3 - \bar{V}_3^T, \\ \Omega_{14} &= B + \bar{U}_4^T - \bar{P}C^T, & \Omega_{24} &= -\bar{U}_4^T + \bar{V}_4^T, \\ \Omega_{34} &= -\bar{V}_4^T - \bar{K}^T D^T, & \Omega_{44} &= -\gamma I - D - D^T. \end{aligned} \tag{24}$$

Moreover, if the above condition is feasible, the gain matrix of a desired controller in the form of (6) *is given by*

$$K = \bar{K}\bar{P}^{-1}. \tag{25}$$

Proof. From Theorem 1, we know that there exists a state-feedback controller in the form of (6) such that the closed-loop NCS in (11) is passive in the sense of Definition 1 if there exist matrices $P > 0$; $Q > 0$; $M_i > 0$, $i = 1, 2$; U_i, V_i, $i = 1, \ldots, 4$; and a diagonal matrix $R > 0$ satisfying (14). Performing a congruence

transformation to (14) by diag $\{P^{-1}, P^{-1}, P^{-1}, I, P^{-1}, P^{-1}, M_1^{-1}, M_2^{-1}, P^{-1}\}$, and a Schur complement operation to the term $\Lambda^2 P^{-1} R P^{-1}$ in the (3,3) block, together with the change of matrix variables defined by

$$\bar{P} = P^{-1}, \quad \bar{M}_i = M_i^{-1}, \quad \bar{R} = R^{-1}, \quad \bar{K} = KP^{-1}, \quad \bar{Q} = P^{-1}QP^{-1},$$

$$\bar{U} = \begin{bmatrix} \bar{U}_1 \\ \bar{U}_2 \\ \bar{U}_3 \\ \bar{U}_4 \end{bmatrix} = \begin{bmatrix} P^{-1} & 0 & 0 & 0 \\ * & P^{-1} & 0 & 0 \\ * & * & P^{-1} & 0 \\ * & * & * & I \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{bmatrix} P^{-1},$$

$$\bar{V} = \begin{bmatrix} \bar{V}_1 \\ \bar{V}_2 \\ \bar{V}_3 \\ \bar{V}_4 \end{bmatrix} = \begin{bmatrix} P^{-1} & 0 & 0 & 0 \\ * & P^{-1} & 0 & 0 \\ * & * & P^{-1} & 0 \\ * & * & * & I \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} P^{-1},$$

we obtain (23), and the proposition is proved. □

The condition in Proposition 1 still cannot be implemented by using standard numerical software due to the existence of the terms $\bar{P}\bar{M}_i^{-1}\bar{P}$ and $\bar{P}\bar{R}^{-1}\bar{P}$. By noticing $\bar{M}_i > 0$ and $\bar{R} > 0$, we have

$$\left(\bar{M}_i - \bar{P}\right) \bar{M}_i^{-1} \left(\bar{M}_i - \bar{P}\right) \geq 0, \quad \left(\bar{R} - \bar{P}\right) \bar{R}^{-1} \left(\bar{R} - \bar{P}\right) \geq 0,$$

which is equivalent to

$$-\bar{P}\bar{M}_i^{-1}\bar{P} \leq \bar{M}_i - 2\bar{P}, \quad -\bar{P}\bar{R}^{-1}\bar{P} \leq \bar{R} - 2\bar{P}. \tag{26}$$

By combining (23) and (26), we readily obtain the following theorem.

THEOREM 2. *Consider the NCS in Figure* 1. *Given a positive constant* γ, *there exists a state-feedback controller in the form of* (6) *such that the closed-loop system in* (11) *is passive in the sense of Definition* 1 *if there exist matrices* $\bar{P} > 0$; $\bar{Q} > 0$; $\bar{M}_i > 0$, $i = 1, 2$; $\bar{U}_i$, $\bar{V}_i$, $i = 1, \ldots, 4$; $\bar{K}$; *and a diagonal matrix* $\bar{R} > 0$ *satisfying*

(27)

$$\begin{bmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} & \Omega_{14} & \bar{U}_1 & \bar{V}_1 & \bar{P}A^T & \bar{P}A^T & B\bar{K} & 0 \\ * & \Omega_{22} & \Omega_{23} & \Omega_{24} & \bar{U}_2 & \bar{V}_2 & 0 & 0 & 0 & 0 \\ * & * & \Omega_{33} & \Omega_{34} & \bar{U}_3 & \bar{V}_3 & \bar{K}^T B^T & \bar{K}^T B^T & 0 & \bar{P} \\ * & * & * & \Omega_{44} & \bar{U}_4 & \bar{V}_4 & B^T & B^T & -D\bar{K} & 0 \\ * & * & * & * & \eta_m^{-1}\left(\bar{M}_1 - 2\bar{P}\right) & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & \kappa^{-1}\left(\bar{M}_2 - 2\bar{P}\right) & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & -\eta_m^{-1}\bar{M}_1 & 0 & B\bar{K} & 0 \\ * & * & * & * & * & * & * & -\kappa^{-1}\bar{M}_2 & B\bar{K} & 0 \\ * & * & * & * & * & * & * & * & \bar{R} - 2\bar{P} & 0 \\ * & * & * & * & * & * & * & * & * & -\Lambda^{-2}\bar{R} \end{bmatrix} < 0,$$

where Ω_{ii} *is as given in* (24). *Moreover, if the above condition is feasible, the gain matrix of a desired controller in the form of* (6) *is given by* (25).

Remark 7. Note that (27) is an LMI not only over the matrix variables, but also over the scalar γ. This implies that the scalar γ can be included as an optimization variable to obtain a reduction of the passivity performance bound. Then, the minimum (in terms of the feasibility of (27)) passivity performance bound with admissible controllers can be readily found by solving the following convex optimization problem using the LMI toolbox in MATLAB:

$$\begin{aligned} &\text{Minimize } \gamma \text{ subject to (27) over } \bar{P} > 0;\ \bar{Q} > 0;\ \bar{M}_i > 0,\ i = 1, 2; \\ &\bar{U}_i,\ \bar{V}_i,\ i = 1, \ldots, 4;\ \bar{K};\ \text{and diagonal } \bar{R} > 0. \end{aligned}$$

Theorem 2 presents an LMI condition for the existence of desired state-feedback controllers based on the inequalities in (26). In the following, we present another approach to solving the condition in Proposition 1.

Now introduce additional matrix variables $\bar{N}_i > 0$ and $\bar{S} > 0$, and replace (23) with

$$\begin{bmatrix} \Omega_{11} & \Omega_{12}\Omega_{13} & \Omega_{14} & \bar{U}_1 & \bar{V}_1 & \bar{P}A^T & \bar{P}A^T & B\bar{K} & 0 \\ * & \Omega_{22}\Omega_{23} & \Omega_{24} & \bar{U}_2 & \bar{V}_2 & 0 & 0 & 0 & 0 \\ * & * \ \Omega_{33} & \Omega_{34} & \bar{U}_3 & \bar{V}_3 & \bar{K}^T B^T & \bar{K}^T B^T & 0 & \bar{P} \\ * & * & * & \Omega_{44} & \bar{U}_4 & \bar{V}_4 & B^T & B^T & -D\bar{K} & 0 \\ * & * & * & * & -\eta_m^{-1}\bar{N}_1 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & -\kappa^{-1}\bar{N}_2 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & -\eta_m^{-1}\bar{M}_1 & 0 & B\bar{K} & 0 \\ * & * & * & * & * & * & * & -\kappa^{-1}\bar{M}_2 & B\bar{K} & 0 \\ * & * & * & * & * & * & * & * & -\bar{S} & 0 \\ * & * & * & * & * & * & * & * & * & -\Lambda^{-2}\bar{R} \end{bmatrix} < 0, \tag{28}$$

$$\bar{N}_i - \bar{P}\bar{M}_i^{-1}\bar{P} \le 0, \quad i = 1, 2, \tag{29}$$

$$\bar{S} - \bar{P}\bar{R}^{-1}\bar{P} \le 0. \tag{30}$$

By the Schur complement, (29) and (30) are equivalent to

$$\begin{bmatrix} -\bar{N}_i^{-1} & \bar{P}^{-1} \\ * & -\bar{M}_i^{-1} \end{bmatrix} \le 0, \quad i = 1, 2, \tag{31}$$

$$\begin{bmatrix} -\bar{S}^{-1} & \bar{P}^{-1} \\ * & -\bar{R}^{-1} \end{bmatrix} \le 0. \tag{32}$$

Then, we readily obtain the following theorem.

THEOREM 3. *Consider the NCS in Figure* 1. *Given a positive constant* γ, *there exists a state-feedback controller in the form of* (6) *such that the closed-loop system in* (11) *is passive in the sense of Definition* 1 *if there exist matrices* $\bar{P} > 0$; $P > 0$; $\bar{S} > 0$; $S > 0$; $\bar{Q} > 0$; $\bar{N}_i > 0$; $N_i > 0$; $\bar{M}_i > 0$; $M_i > 0$, $i = 1, 2$; $\bar{U}_i$, $\bar{V}_i$, $i = 1, \ldots, 4$; $\bar{K}$; *and diagonal matrices* $\bar{R} > 0$, $R > 0$ *satisfying* (28) *and*

$$\begin{bmatrix} -S & P \\ * & -R \end{bmatrix} \le 0, \quad \begin{bmatrix} -N_i & P \\ * & -M_i \end{bmatrix} \le 0, \quad i = 1, 2, \tag{33}$$

$$\bar{P}P = I, \quad \bar{R}R = I, \quad \bar{M}_i M_i = I, \quad \bar{N}_i N_i = I, \quad i = 1, 2. \tag{34}$$

Moreover, if the above condition is feasible, the gain matrix of a desired controller in the form of (6) *is given by* (25).

The condition presented in Theorem 3 is equivalent to that in Proposition 1. It is noted that this condition is not a convex set due to the matrix equality constraints in (34). Several approaches have been proposed to solve such nonconvex feasibility problems, among which the cone complementarity linearization (CCL) method [9] is the most commonly used (for instance, the CCL algorithm has been used for solving the controller design problems as well as model reduction problems [15, 16, 46]). The basic idea in CCL algorithm is that if the LMI $\left[\begin{smallmatrix} P & I \\ I & L \end{smallmatrix}\right] \ge 0$ is feasible in the $n \times n$ matrix variables $L > 0$ and $P > 0$, then $\mathrm{tr}(PL) \ge n$, and $\mathrm{tr}(PL) = n$ if and only if $PL = I$.

Now using a cone complementarity approach [9], we suggest the following nonlinear minimization problem involving LMI conditions instead of the original nonconvex feasibility problem formulated in Theorem 3.

PROBLEM PCD (PASSIFICATION CONTROLLER DESIGN).

$$\min \ \operatorname{tr}\left(\bar{P}P + \bar{R}R + \sum_{i=1}^{2}\left(\bar{M}_i M_i + \bar{N}_i N_i\right)\right) \ \textit{subject to (28), (33) and}$$

$$\begin{bmatrix} \bar{P} & I \\ I & P \end{bmatrix} \geq 0, \quad \begin{bmatrix} \bar{R} & I \\ I & R \end{bmatrix} \geq 0, \quad \begin{bmatrix} \bar{M}_i & I \\ I & M_i \end{bmatrix} \geq 0, \quad \begin{bmatrix} \bar{N}_i & I \\ I & N_i \end{bmatrix} \geq 0, \quad i = 1, 2.$$

According to [9], if the solution of the above minimization problem is $6n$, that is,

$$\min \ \operatorname{tr}\left(\bar{P}P + \bar{R}R + \sum_{i=1}^{2}\left(\bar{M}_i M_i + \bar{N}_i N_i\right)\right) = 6n,$$

then the conditions in Theorem 3 are solvable. Algorithm 1 in [9] can be easily adapted to solve Problem PCD.

5. Illustrative example. In this section, two examples are provided to illustrate the results developed above. We first use a numerical example to show the advantage of introducing the lower bound of transmission delays, and to show how much the quantization behavior affects the passivity performance. The second example shows the applicability of the passification results.

Example 1. Suppose the system of matrices A, B, C, D in (1) and the controller gain K in (6) are given by

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad D = 0.3, \quad K = \begin{bmatrix} -1 & 1 \end{bmatrix}.$$

The parameters for the quantizer $f(\cdot)$ are given by $\rho_1 = 0.8$ and $\rho_2 = 0.9$; thus according to (4) we have $\sigma_1 = 0.1111$ and $\sigma_2 = 0.0526$. It is assumed that the network-induced delays η_k satisfy $\eta_m \leq \eta_k \leq 0.4$ s, the maximum number of data packet dropouts is 2, and the sampling period is 10 ms. Our purpose is to determine the minimum guaranteed passivity performances for different values of the lower delay bound η_m.

When we do not consider the lower bound of the network-induced delays, that is, $\eta_m = 0$, by using Corollary 2 and Theorem 1 (assuming η_m is sufficiently small), the minimum guaranteed passivity performance obtained is $\gamma_{\min} = 5.4425$. However, if we assume $\eta_m = 0.1$ s, the minimum guaranteed passivity performance obtained is $\gamma_{\min} = 0.9578$. A more detailed comparison for different values of η_m is provided in Table 1, which shows that considering the lower bound of the signal transmission delay gives rise to less conservative results.

TABLE 1
Comparison for different values of η_m.

η_m (s)	0	0.05	0.1	0.15	0.2
Guaranteed passivity performance $\gamma_{\min}$	5.4425	1.7528	0.9578	0.6019	0.3958

The second task in this example is to show how much the quantization behavior affects the guaranteed passivity performance. Now we assume $\eta_m = 0.15$, and other parameters, except that related to the quantizer $f(\cdot)$, are the same as above. When $\rho_1 = 0.8$ and $\rho_2 = 0.9$ (corresponding to $\sigma_1 = 0.1111$ and $\sigma_2 = 0.0526$), by Theorem 1, the minimum guaranteed passivity performance obtained is $\gamma_{\min} = 0.6019$. When

$\rho_1 = 0.7$ and $\rho_2 = 0.8$ (corresponding to $\sigma_1 = 0.1765$ and $\sigma_2 = 0.1111$), the minimum guaranteed passivity performance obtained is $\gamma_{\min} = 2.8398$. This shows that for a coarser quantizer (corresponding to smaller ρ_j or larger σ_j), the obtained minimum guaranteed passivity performance is usually larger. A more detailed comparison for different values of ρ_j is provided in Table 2. To facilitate the presentation, we assume $\rho_1 = \rho_2 = \rho$ (corresponding to $\sigma_1 = \sigma_2 = \sigma$) in the comparison.

TABLE 2
Comparison for different quantizer parameters.

ρ	0.9	0.85	0.8	0.75
σ	0.0526	0.0811	0.1111	0.1429
Guaranteed passivity performance $\gamma_{\min}$	0.3704	0.6652	1.2815	3.3814

Example 2. Suppose the physical plant in Figure 1 is a satellite system, which appears in [4, 14]. The satellite system consists of two rigid bodies joined by a flexible link. This link is modeled as a spring with torque constant k and viscous damping f. Denoting the yaw angles for the two bodies (the main body and the instrumentation module) by θ_1 and θ_2, the control torque by $u(t)$, and the moments of inertia of the two bodies by J_1 and J_2, we see that the dynamic equations are given by

$$J_1\ddot{\theta}_1(t) + f(\dot{\theta}_1(t) - \dot{\theta}_2(t)) + k((\theta_1(t) - \theta_2(t))) = u(t),$$
$$J_2\ddot{\theta}_2(t) + f(\dot{\theta}_1(t) - \dot{\theta}_2(t)) + k((\theta_1(t) - \theta_2(t))) = 0.$$

Assume the output is the angular positions $\theta_2(t)$. Thus, the state-space representation of the above equation is given by

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & J_1 & 0 \\ 0 & 0 & 0 & J_2 \end{bmatrix} \begin{bmatrix} \dot{\theta}_1(t) \\ \dot{\theta}_2(t) \\ \ddot{\theta}_1(t) \\ \ddot{\theta}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -k & k & -f & f \\ k & -k & f & -f \end{bmatrix} \begin{bmatrix} \theta_1(t) \\ \theta_2(t) \\ \dot{\theta}_1(t) \\ \dot{\theta}_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} u(t),$$

$$y(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \theta_1(t) \\ \theta_2(t) \\ \dot{\theta}_1(t) \\ \dot{\theta}_2(t) \end{bmatrix}. \tag{35}$$

Here we choose $J_1 = J_2 = 1$, $k = 0.09$, and $f = 0.04$ (the values of k and f are chosen within their respective ranges). Then, the corresponding matrices described in section 2 are given by

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -0.3 & 0.3 & -0.004 & 0.004 \\ 0.3 & -0.3 & 0.004 & -0.004 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix},$$
$$C = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}, \quad D = 0.$$

It is assumed that the sampling period $h = 10$ ms; the network-induced delay bound in (8) are given by $\eta_m = 10$ ms and $\eta_M = 20$ ms; and the maximum number of data packet dropouts $\bar{\delta} = 2$. In addition, the parameters for the quantizer $f(\cdot)$ are assumed to be $\rho_1 = \rho_2 = \rho_3 = \rho_4 = 0.9$.

The eigenvalues of A are $-0.04 + 0.4224j$, $-0.0400 - 0.4224j$, 0, 0, and thus the above system is not stable. Our purpose is to design a state-feedback controller in

the form of (6) such that the closed-loop system is passive in the sense of Definition 1. By using Theorem 2 (minimizing γ in (27)), we obtain the following matrices (for space consideration we do not list all the obtained matrices here):

$$\bar{P} = \begin{bmatrix} 3.8994 & 0.4268 & -0.4130 & -0.4647 \\ 0.4268 & 0.7929 & 0.5678 & -0.2679 \\ -0.4130 & 0.5678 & 2.0103 & -0.2784 \\ -0.4647 & -0.2679 & -0.2784 & 0.1820 \end{bmatrix},$$
$$\bar{K} = \begin{bmatrix} -1.4010 & -0.0335 & -1.9367 & 0.0873 \end{bmatrix},$$
$$\bar{R} = \operatorname{diag}\{0.0874,\ 0.2912,\ 0.0598,\ 0.0052\}.$$

Thus, according to (25), the gain matrix for the state-feedback controller in (6) is given by

$$K = \begin{bmatrix} -1.2647 & 0.2377 & -2.0589 & -5.5490 \end{bmatrix},$$

and the obtained minimum guaranteed passivity performance in terms of the feasibility of (27) is $\gamma^* = 1.1736$.

We first show that the closed-loop system is asymptotically stable. The initial condition is assumed to be

$$\begin{bmatrix} -0.5 & 0.2 & 0.3 & -0.3 \end{bmatrix}^T.$$

The state responses are depicted in Figure 2, from which we can see that all four states converge to zero. In the simulation, the network-induced delays and the data packet dropouts are generated randomly (meanly distributed within their ranges) according to the above assumption and are shown in Figures 3 and 4. The computed control inputs arriving at the ZOH are shown in Figure 5 (with zoomed area given in Figure 6), where we can see the discontinuous holding behavior of the control inputs.

Now, we will show the passivity of the closed-loop system. To this end, let us assume zero initial conditions and select a set of input signals as follows:

$$w(t) = \begin{cases} \sin t, & 5 \le t \le 15\text{s}, \\ 0 & \text{otherwise}. \end{cases} \tag{36}$$

Figures 7 and 8 depict the state responses and the control input, respectively. Now denote

$$L(t) \triangleq 2\int_0^t w^T(s)z(s)ds, \quad R(t) \triangleq -\gamma\int_0^t w^T(s)w(s)ds,$$

which correspond to the left-hand and right-hand sides of (13), respectively. $L(t)$ and $R(t)$ are depicted in Figure 9, which shows that $L(t) \ge R(t)$ for all $t \ge 0$, and thus (13) is guaranteed and the effectiveness of the passification design is clear.

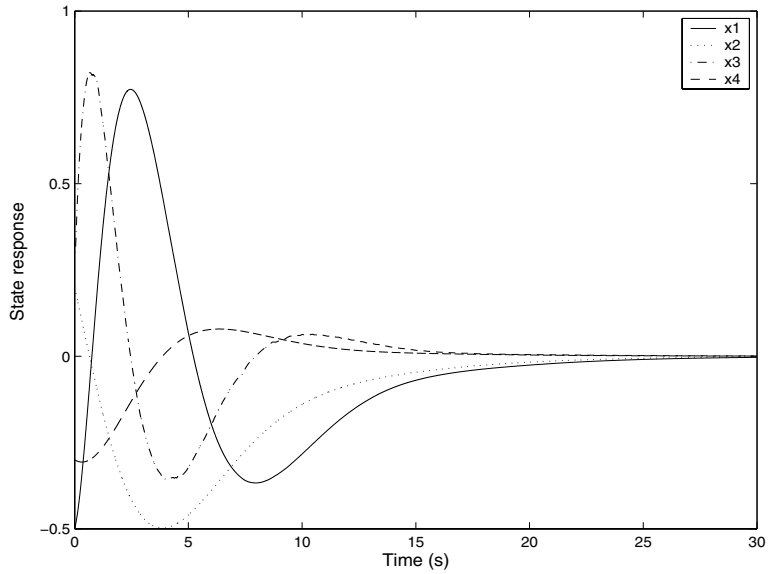


FIG. 2. *State response.*

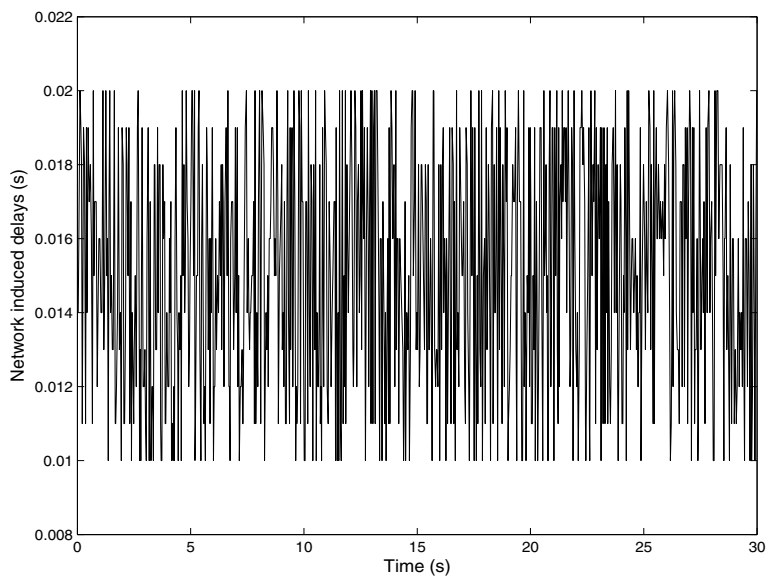


FIG. 3. *Networked-induced delays.*

6. Conclusions. The problems of passivity analysis and passification for network-based linear control systems have been investigated. The physical plant is in continuous time, and the controller is in discrete time. The problems are solved by using a sampled-data approach, which has taken the network-induced delays, data packet dropouts, and measurement quantization into consideration. The measurement quantizer is assumed to be logarithmic, and the network-induced delays are assumed to have both a lower bound and an upper bound, which is more general than those assumptions used in the literature. A new model based on the updating instants of the

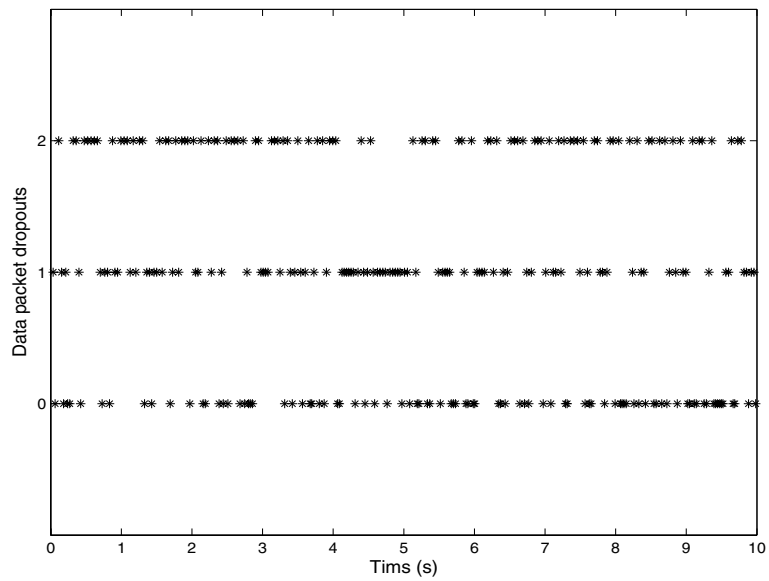


FIG. 4. *Data packet dropouts.*

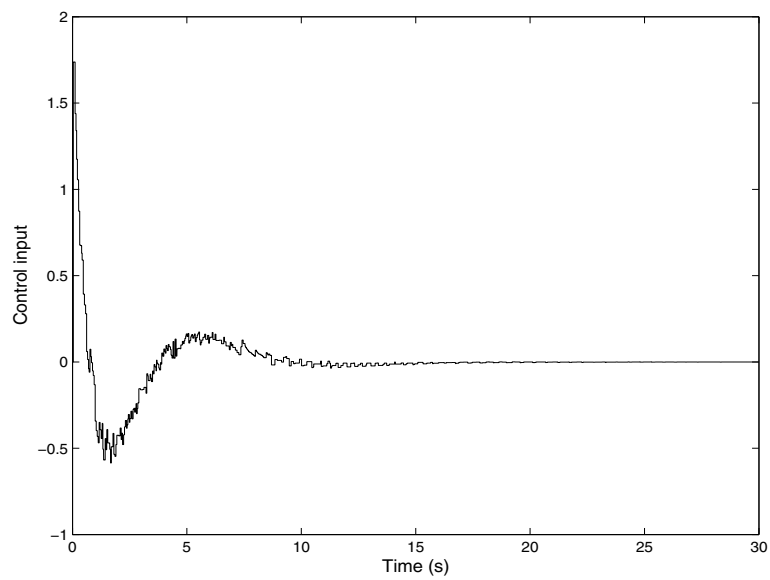


FIG. 5. *Control input.*

ZOH (instead of the sampling instants) has been formulated, and a passivity analysis performance condition has been proposed in the form of an LMI. Based on this condition, a controller design procedure has been developed, which guarantees that the closed-loop NCS is passive. The results developed here can be further extended to linear systems with parameter uncertainties, represented in either norm-bounded or polytopic frameworks. Two illustrative examples have been provided to show the usefulness and effectiveness of the proposed theoretical results. It is worth noting that in this paper, only one side of the quantization effect (from sampler to controller) has

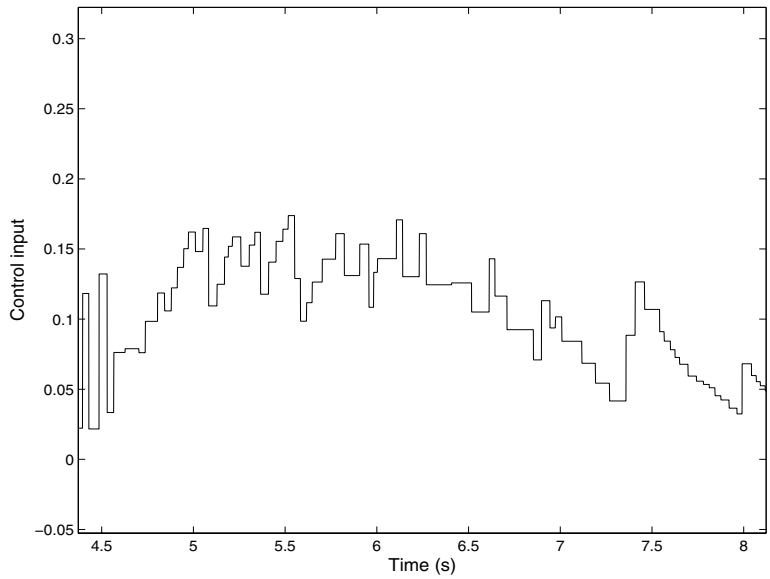


FIG. 6. *Zoomed area of control input.*

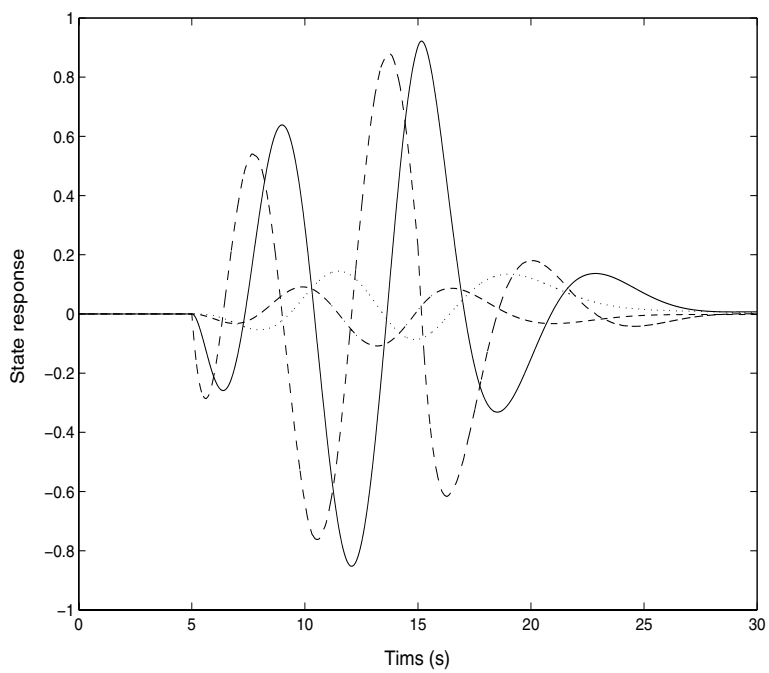


FIG. 7. *State response under* (36).

been taken into consideration. A more challenging problem is to consider the case where quantization effects appear in both sides (from sampler to controller and from controller to ZOH). Moreover, future research effort could be directed towards solving the passification problem via output-feedback controllers, which is useful when the state variables are not measurable.

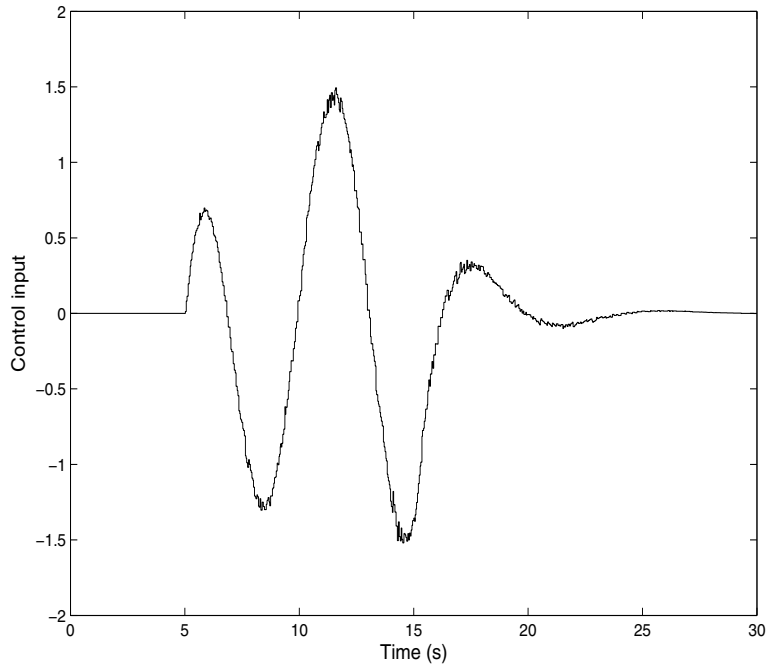


FIG. 8. *Control input under* (36).

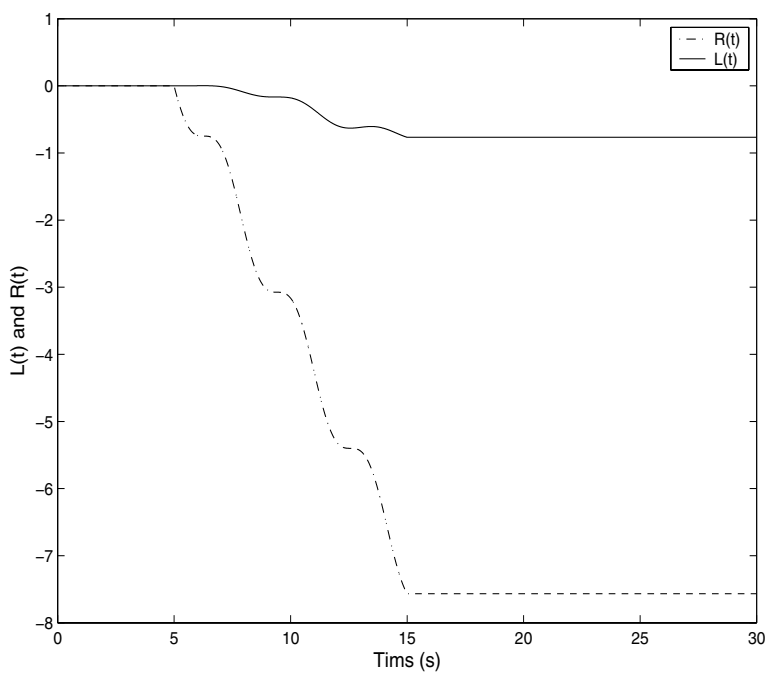


FIG. 9. *Calculation of* $L(t)$ *and* $R(t)$.

Appendix A. Proof of Theorem 1. First, let us represent $t_k - \eta_k$ in (11) as

$$t_k - \eta_k = t - \eta_m - \eta(t), \tag{37}$$

where

$$\eta(t) = t - t_k + (\eta_k - \eta_m). \tag{38}$$

Then, from (10) we have

$$0 \le \eta(t) \le \kappa, \tag{39}$$

where κ is as given in (15). By substituting (37) into (11), we obtain

$$\begin{aligned} \dot{x}(t) &= Ax(t) + BKf(x(t - \eta_m - \eta(t))) + Bw(t), \\ z(t) &= Cx(t) + DKf(x(t - \eta_m - \eta(t))) + Dw(t). \end{aligned} \tag{40}$$

In addition, considering the quantization behavior shown in (2)–(5), and according to [10, 13], (40) can be expressed as

$$\begin{aligned} \dot{x}(t) &= Ax(t) + BK(I + \Lambda(t))(x(t - \eta_m - \eta(t))) + Bw(t), \\ z(t) &= Cx(t) + DK(I + \Lambda(t))(x(t - \eta_m - \eta(t))) + Dw(t), \end{aligned} \tag{41}$$

where

$$\Lambda(t) = \operatorname{diag}\{\Lambda_1(t), \Lambda_2(t), \ldots, \Lambda_n(t)\}, \tag{42}$$

with

$$\Lambda_j(t) \in [-\sigma_j, \sigma_j], \quad j = 1, \ldots, n. \tag{43}$$

Choose the following Lyapunov–Krasovskii functional:

$$\begin{aligned} V(t) &= x^T(t)Px(t) + \int_{t-\eta_m}^{t} x^T(\alpha)Qx(\alpha)\,d\alpha \\ &\quad + \int_{-\eta_m}^{0}\int_{t+\beta}^{t} \dot{x}^T(\alpha)M_1\dot{x}(\alpha)d\alpha d\beta + \int_{-\eta_m-\kappa}^{-\eta_m}\int_{t+\beta}^{t} \dot{x}^T(\alpha)M_2\dot{x}(\alpha)d\alpha d\beta, \end{aligned} \tag{44}$$

where $P > 0$, $Q > 0$, $M_i > 0$ are matrices to be determined. Then, along the solution of system (41), the time derivative of $V(t)$ is given by

$$\begin{aligned} \dot{V}(t) &= 2x^T(t)P\dot{x}(t) + x^T(t)Qx(t) - x^T(t - \eta_m)Qx(t - \eta_m) \\ &\quad + \eta_m\dot{x}^T(t)M_1\dot{x}(t) - \int_{t-\eta_m}^{t} \dot{x}^T(\alpha)M_1\dot{x}(\alpha)d\alpha \\ &\quad + \kappa\dot{x}^T(t)M_2\dot{x}(t) - \int_{t-\eta_m-\kappa}^{t-\eta_m} \dot{x}^T(\alpha)M_2\dot{x}(\alpha)\,d\alpha \\ &\le 2x^T(t)P\dot{x}(t) + x^T(t)Qx(t) - x^T(t - \eta_m)Qx(t - \eta_m) \\ &\quad + \dot{x}^T(t)\Psi\dot{x}(t) - \int_{t-\eta_m}^{t} \dot{x}^T(\alpha)M_1\dot{x}(\alpha)d\alpha - \int_{t-\eta_m-\eta(t)}^{t-\eta_m} \dot{x}^T(\alpha)M_2\dot{x}(\alpha)\,d\alpha, \end{aligned} \tag{45}$$

where $\Psi = \eta_m M_1 + \kappa M_2$. By the Newton–Leibniz formula, we have

$$\int_{t-\eta_m}^{t} \dot{x}(\alpha)\, d\alpha = x(t) - x(t-\eta_m), \tag{46}$$

$$\int_{t-\eta_m-\eta(t)}^{t-\eta_m} \dot{x}(\alpha) d\alpha = x(t-\eta_m) - x(t-\eta_m-\eta(t)). \tag{47}$$

Then, for any matrices

$$U = \begin{bmatrix} U_1^T & U_2^T & U_3^T & U_4^T \end{bmatrix}^T \quad \text{and} \quad V = \begin{bmatrix} V_1^T & V_2^T & V_3^T & V_4^T \end{bmatrix}^T,$$

we have

$$2\zeta^T(t) U \left[x(t) - x(t-\eta_m) - \int_{t-\eta_m}^{t} \dot{x}(\alpha)\, d\alpha \right] = 0, \tag{48}$$

$$2\zeta^T(t) U \left[x(t-\eta_m) - x(t-\eta_m-\eta(t)) - \int_{t-\eta_m-\eta(t)}^{t-\eta_m} \dot{x}(\alpha) d\alpha \right] = 0, \tag{49}$$

where

$$\zeta(t) = \begin{bmatrix} x^T(t) & x^T(t-\eta_m) & x^T(t-\eta_m-\eta(t)) & w^T(t) \end{bmatrix}^T.$$

Then, from (41), (45), (48), (49), we obtain

(50)

$$\begin{aligned}
&\dot{V}(t) - 2w^T(t)z(t) - \gamma w^T(t)w(t) \\
&\quad \le 2x^T(t)P\dot{x}(t) + x^T(t) Q x(t) - x^T(t-\eta_m) Q x(t-\eta_m) \\
&\qquad + \dot{x}^T(t)\Psi\dot{x}(t) - \int_{t-\eta_m}^{t} \dot{x}^T(\alpha) M_1 \dot{x}(\alpha) d\alpha - \int_{t-\eta_m-\eta(t)}^{t-\eta_m} \dot{x}^T(\alpha) M_2 \dot{x}(\alpha)\, d\alpha \\
&\qquad - 2w^T(t)z(t) - \gamma w^T(t)w(t) + 2\zeta^T(t) U \left[x(t) - x(t-\eta_m) - \int_{t-\eta_m}^{t} \dot{x}(\alpha)\, d\alpha \right] \\
&\qquad + 2\zeta^T(t) V \left[x(t-\eta_m) - x(t-\eta_m-\eta(t)) - \int_{t-\eta_m-\eta(t)}^{t-\eta_m} \dot{x}(\alpha) d\alpha \right] \\
&\quad \le \zeta^T(t) \left(\Gamma + \eta_m U M_1^{-1} U^T + \kappa V M_2^{-1} V^T \right) \zeta(t) \\
&\qquad - \int_{t-\eta_m}^{t} \left[\zeta^T(t) U + \dot{x}^T(\alpha) M_1 \right] M_1^{-1} \left[U^T \zeta(t) + M_1 \dot{x}(\alpha) \right] d\alpha \\
&\qquad - \int_{t-\eta_m-\eta(t)}^{t-\eta_m} \left[\zeta^T(t) V + \dot{x}^T(\alpha) M_2 \right] M_2^{-1} \left[V^T \zeta(t) + M_2 \dot{x}(\alpha) \right] d\alpha,
\end{aligned}$$

where

$$\Gamma = \begin{bmatrix} \bar{\Gamma}_{11} & \Gamma_{12} & \bar{\Gamma}_{13} & \bar{\Gamma}_{14} \\ * & \Gamma_{22} & \Gamma_{23} & \Gamma_{24} \\ * & * & \bar{\Gamma}_{33} & \bar{\Gamma}_{34} \\ * & * & * & \bar{\Gamma}_{44} \end{bmatrix},$$

$$\begin{aligned}
&\bar{\Gamma}_{11} = \Gamma_{11} + A^T \Psi A, \quad \bar{\Gamma}_{14} = \Gamma_{14} + A^T \Psi B, \\
&\bar{\Gamma}_{13} = \Gamma_{13} + PBK\Lambda(t) + A^T \Psi BK(I + \Lambda(t)), \\
&\bar{\Gamma}_{33} = \Gamma_{33} + (I + \Lambda(t)) K^T B^T \Psi BK(I + \Lambda(t)), \\
&\bar{\Gamma}_{34} = \Gamma_{34} + (I + \Lambda(t)) K^T B^T \Psi B - \Lambda(t) K^T D^T, \\
&\bar{\Gamma}_{44} = \Gamma_{44} + B^T \Psi B.
\end{aligned}$$

In the following, we will show that (14) guarantees that $\dot{V}(t) - 2w^T(t)z(t) - \gamma w^T(t)w(t) \leq 0$. Notice that $M_1 > 0$ and $M_2 > 0$; then we have

$$\begin{aligned}
&\left[\zeta^T(t) U + \dot{x}^T(\alpha) M_1\right] M_1^{-1} \left[U^T \zeta(t) + M_1 \dot{x}(\alpha)\right] \geq 0, \\
&\left[\zeta^T(t) V + \dot{x}^T(\alpha) M_2\right] M_2^{-1} \left[V^T \zeta(t) + M_2 \dot{x}(\alpha)\right] \geq 0.
\end{aligned}$$

Therefore, from (51) we know that $\dot{V}(t) - 2w^T(t)z(t) - \gamma w^T(t)w(t) \leq 0$ if

$$\Gamma + \eta_m U M_1^{-1} U^T + \kappa V M_2^{-1} V^T < 0,$$

which, by the Schur complement, is equivalent to

(51)

$$\begin{bmatrix}
\Gamma_{11} & \Gamma_{12}\tilde{\Gamma}_{13} & \Gamma_{14} & U_1 & V_1 & A^T M_1 & A^T M_2 \\
* & \Gamma_{22}\Gamma_{23} & \Gamma_{24} & U_2 & V_2 & 0 & 0 \\
* & * \ \Gamma_{33} & \tilde{\Gamma}_{34} & U_3 & V_3 & (I + \Lambda(t)) K^T B^T M_1 & (I + \Lambda(t)) K^T B^T M_2 \\
* & * \ * & \Gamma_{44} & U_4 & V_4 & B^T M_1 & B^T M_2 \\
* & * \ * & * & -\eta_m^{-1} M_1 & 0 & 0 & 0 \\
* & * \ * & * & * & -\kappa^{-1} M_2 & 0 & 0 \\
* & * \ * & * & * & * & -\eta_m^{-1} M_1 & 0 \\
* & * \ * & * & * & * & * & -\kappa^{-1} M_2
\end{bmatrix} < 0,$$

where $\tilde{\Gamma}_{13} = \Gamma_{13} + PBK\Lambda(t)$ and $\tilde{\Gamma}_{34} = \Gamma_{34} - \Lambda(t) K^T D^T$. Rewrite (51) in the form of

$$\Sigma_1 + \Sigma_3 \Sigma_2 + \Sigma_2^T \Sigma_3^T < 0 \tag{52}$$

with

$$\Sigma_1 = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{13} & \Gamma_{14} & U_1 & V_1 & A^T M_1 & A^T M_2 \\ * & \Gamma_{22} & \Gamma_{23} & \Gamma_{24} & U_2 & V_2 & 0 & 0 \\ * & * & \Gamma_{33} & \Gamma_{34} & U_3 & V_3 & K^T B^T M_1 & K^T B^T M_2 \\ * & * & * & \Gamma_{44} & U_4 & V_4 & B^T M_1 & B^T M_2 \\ * & * & * & * & -\eta_m^{-1} M_1 & 0 & 0 & 0 \\ * & * & * & * & * & -\kappa^{-1} M_2 & 0 & 0 \\ * & * & * & * & * & * & -\eta_m^{-1} M_1 & 0 \\ * & * & * & * & * & * & * & -\kappa^{-1} M_2 \end{bmatrix},$$

$$\Sigma_2 = \begin{bmatrix} 0 & 0 & \Lambda(t) & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\Sigma_3 = \begin{bmatrix} K^T B^T P & 0 & 0 & -K^T D^T & 0 & 0 & K^T B^T M_1 & K^T B^T M_2 \end{bmatrix}^T.$$

It is noted that for some matrix $R > 0$ we have

$$\left(\Sigma_3 R^{-\frac{1}{2}} + \Sigma_2^T R^{\frac{1}{2}}\right)\left(\Sigma_3 R^{-\frac{1}{2}} + \Sigma_2^T R^{\frac{1}{2}}\right)^T \geq 0,$$

which gives rise to $\Sigma_3\Sigma_2 + \Sigma_2^T\Sigma_3^T \leq \Sigma_3 R^{-1}\Sigma_3^T + \Sigma_2^T R\Sigma_2$. Therefore, (52) holds if for some matrix $R > 0$,

$$\Sigma_1 + \Sigma_3 R^{-1}\Sigma_3^T + \Sigma_2^T R\Sigma_2 < 0. \tag{53}$$

Note that R is required to be diagonal positive definite. Then, by using a Schur complement operation and by considering (43), (14) guarantees (51), and thus we have $\dot{V}(t) - 2w^T(t)z(t) - \gamma w^T(t)w(t) \leq 0$. Integrating both sides with respect to t over the time period $[0, T]$, we have

$$V(T) - V(0) - 2\int_0^T w^T(t)z(t)dt - \gamma\int_0^T w^T(t)w(t)dt \leq 0. \tag{54}$$

Under the zero initial conditions, we have $V(0) = 0$ and $V(T) \geq 0$; thus (54) guarantees (13), and the proof is completed. □

REFERENCES

[1] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis: A Modern Systems Theory Approach*. Prentice–Hall, Englewood Cliffs, NJ, 1973.

[2] P. Antsaklis and J. Baillieul, *Guest editorial. Special issue on networked control systems*, IEEE Trans. Automat. Control, 49 (2004), pp. 1421–1423.

[3] A. Bicchi, A. Marigo, and B. Piccoli, *On the reachability of quantized control systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 546–563.

[4] R. M. Biernacki, H. Hwang, and S. P. Battacharyya, *Robust stability with structured real parameter perturbations*, IEEE Trans. Automat. Control, 32 (1987), pp. 495–506.

[5] E. K. Boukas, *Stabilization of stochastic nonlinear hybrid systems*, Int. J. Innovative Computing, Information and Control, 1 (2005), pp. 131–141.

[6] R. W. Brockett and D. Liberzon, *Quantized feedback stabilization of linear systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1279–1289.

[7] G. Calcev, R. Gorez, and M. De Neyer, *Passivity approach to fuzzy control systems*, Automatica, 34 (1998), pp. 339–344.

[8] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

[9] L. El Ghaoui, F. Oustry, and M. Ait Rami, *A cone complementarity linearization algorithm for static output-feedback and related problems*, IEEE Trans. Automat. Control, 42 (1997), pp. 1171–1176.

[10] N. Elia and S. K. Mitter, *Stabilization of linear systems with limited information*, IEEE Trans. Automat. Control, 46 (2001), pp. 1384–1400.

[11] F. Fagnani and S. Zampieri, *Stability analysis and synthesis for scalar linear systems with a quantized feedback*, IEEE Trans. Automat. Control, 48 (2003), pp. 1569–1584.

[12] X. Feng and K. A. Loparo, *Active probing for information in control systems with quantized state measurements: A minimum entropy approach*, IEEE Trans. Automat. Control, 42 (1997), pp. 216–238.

[13] M. Fu and L. Xie, *The sector bound approach to quantized feedback control*, IEEE Trans. Automat. Control, 50 (2005), pp. 1698–1711.

[14] P. Gahinet, A. Nemirovskii, A. J. Laub, and M. Chilali, *LMI Control Toolbox User's Guide*, The Math Works, Inc., Natick, MA, 1995.

[15] H. Gao, J. Lam, C. Wang, and S. Xu, *H_∞ model reduction for discrete time-delay systems: Delay independent and dependent approaches*, Int. J. Control, 77 (2004), pp. 321–335.

[16] H. Gao and C. Wang, *Comments and further results on: "A descriptor system approach to H_∞ control of linear time-delay systems"*, IEEE Trans. Automat. Control, 48 (2003), pp. 520–525.

[17] G. C. Goodwin, H. Haimovich, D. E. Quevedo, and J. S. Welsh, *A moving horizon approach to networked control system design*, IEEE Trans. Automat. Control, 49 (2004), pp. 1427–1445.

[18] K. Gu, V. L. Kharitonov, and J. Chen, *Stability of Time-Delay Systems*, Springer-Verlag, Berlin, 2003.

[19] Y. He, M. Wu, J. H. She, and G. P. Liu, *Delay-dependent robust stability criteria for uncertain neutral systems with mixed delays*, Systems Control Lett., 51 (2004), pp. 57–65.

[20] Y. He, M. Wu, J. H. She, and G. P. Liu, *Parameter-dependent Lyapunov functional for stability of time-delay systems with polytopic-type uncertainties*, IEEE Trans. Automat. Control, 49 (2004), pp. 828–832.

[21] H. Ishii and B. A. Francis, *Limited Data Rate in Control Systems with Networks*, Lecture Notes in Control and Inform. Sci. 275, Springer-Verlag, Berlin, 2002.

[22] H. Ishii and B. A. Francis, *Quadratic stabilization of sampled-data systems with quantization*, Automatica, 39 (2003), pp. 1793–1800.

[23] F.-L. Lian, J. Moyne, and D. Tilbury, *Modelling and optimal controller design of networked control systems with multiple delays*, Int. J. Control, 76 (2003), pp. 591–606.

[24] D. Liberzon, *Hybrid feedback stabilization of systems with quantized signals*, Automatica, 39 (2003), pp. 1543–1554.

[25] J. Liu and N. Elia, *Quantized feedback stabilization of non-linear affine systems*, Int. J. Control, 77 (2004), pp. 239–249.

[26] R. Lozano, B. Brogliato, O. Egeland, and B. Maschke, *Dissipative Systems Analysis and Control: Theory and Applications*, Springer-Verlag, London, 2000.

[27] R. Lozano Leal and S. M. Joshi, *Strictly positive real transfer functions revisited*, IEEE Trans. Automat. Control, 35 (1990), pp. 1243–1245.

[28] L. Lu, L. Xie, and W. Cai, *H_2 controller design for networked control systems*, Asian J. Control, 6 (2004), pp. 88–96.

[29] L. A. Montestruque and P. Antsaklis, *Stability of model-based networked control systems with time-varying transmission times*, IEEE Trans. Automat. Control, 49 (2004), pp. 1562–1572.

[30] K. Narendra and J. Taylor, *Frequency Domain Criteria for Absolute Stability*, Academic Press, New York, 1973.

[31] S.-I. Niculescu and R. Lozano, *On the passivity of linear delay systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 460–464.

[32] N. Patel and S. K. Nguang, *Real time digital controller implementation using deterministic uniformly weighted bit streams*, Int. J. Innovative Computing, Information and Control, 2 (2006), pp. 505–517.

[33] P. Seiler and R. Sengupta, *An H_∞ approach to networked control*, IEEE Trans. Automat. Control, 50 (2005), pp. 356–364.

[34] M. Sznaier and A. Sideris, *Feedback control of quantized constrained systems with applications to neuromorphic controllers design*, IEEE Trans. Automat. Control, 39 (1994), pp. 1497–1502.

[35] M. Vidyasagar, *Nonlinear Systems Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1993.

[36] G. C. Walsh, H. Ye, and L. Bushnell, *Stability analysis of networked control systems*, IEEE Trans. Control Systems Technology, 10 (2002), pp. 438–446.
[37] Z. Wang, D. W. C. Ho, and X. Liu, *Variance-constrained control for uncertain stochastic systems with missing measurements*, IEEE Trans. Systems Man Cybernet. Part A: Systems and Humans 35 (2005), pp. 746–753.
[38] Z. Wang, F. Yang, D. Ho, and X. Liu, *Robust finite-horizon filtering for stochastic systems with missing measurements*, IEEE Signal Process. Lett., 12 (2005), pp. 437–440.
[39] L. Xie, M. Fu, and H. Li, *Passivity analysis and passification for uncertain signal processing systems*, IEEE Trans. Signal Process., 46 (1998), pp. 2394–2403.
[40] L. Xie and Y. C. Soh, *Positive real control problem for uncertain linear time-invariant systems*, Systems Control Lett., 24 (1995), pp. 265–271.
[41] S. Xu, J. Lam, Z. Lin, and K. Galkowski, *Positive real control for uncertain two-dimensional systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 49 (2002), pp. 1659–1666.
[42] S. Xu, J. Lam, and C. Yang, *H_∞ and positive-real control for linear neutral delay systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 1321–1326.
[43] M. Yu, L. Wang, and T. Chu, *Sampled-data stabilisation of networked control systems with nonlinearity*, IEE Proc. Control Theory Appl., 152 (2005), pp. 609–614.
[44] M. Yu, L. Wang, T. Chu, and F. Hao, *Stabilization of networked control systems with data packet dropout and transmission delays: Continuous-time case*, European J. Control, 11 (2005), pp. 40–55.
[45] D. Yue, Q.-L. Han, and J. Lam, *Network-based robust H_∞ control of systems with uncertainty*, Automatica, 41 (2005), pp. 999–1007.
[46] L. Zhang, B. Huang, and J. Lam, *H_∞ model reduction of Markovian jump linear systems*, Systems Control Lett., 50 (2003), pp. 103–118.
[47] L. Zhang, Y. Shi, T. Chen, and B. Huang, *A new method for stabilization of networked control systems with random delays*, IEEE Trans. Automat. Control, 50 (2005), pp. 1177–1181.
[48] W. Zhang, M. Branicky, and S. Phillips, *Stability of networked control systems*, IEEE Control Systems Mag., 21 (2001), pp. 84–99.
[49] P. V. Zhivoglyadov and R. H. Middleton, *Networked control design for linear systems*, Automatica, 39 (2003), pp. 743–750.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1323–1367

AN ANALOGUE OF SHANNON INFORMATION THEORY FOR DETECTION AND STABILIZATION VIA NOISY DISCRETE COMMUNICATION CHANNELS*

ALEXEY S. MATVEEV† AND ANDREY V. SAVKIN‡

Abstract. The paper addresses both detection and stabilization problems involving communication errors and capacity constraints. Discrete-time partially observed linear systems are studied. Unlike the classic theory, the sensor signals are transmitted to the estimator/controller over a noisy digital communication link modeled as a stochastic stationary discrete memoryless channel. It is shown that for noise-free plants, the Shannon capacity of the channel constitutes the border separating the cases where stabilization and reliable detection (asymptotic state estimation) with arbitrarily large probability are and are not possible, respectively.

Key words. control over communications channels, communication constraints, Shannon information theory, stabilization, networked control systems

AMS subject classifications. 93B07, 93E10, 93A24

DOI. 10.1137/040621697

1. Introduction.

The standard assumption in the classical control theory is that data transmission required by the algorithm can be performed with infinite precision. However, due to the growth in communication technology, it is becoming more common to employ digital finite capacity networks for the exchange of information between plant components. Examples concern complex dynamical processes such as advanced aircraft, spacecraft, automotive, industrial and defense systems, arrays of microactuators, and power control in mobile communication. Bandwidth communication constraints are often major obstacles to control system design by means of the classical theory. For instance, as was shown in [58], the design of control systems for platoons of underwater vehicles strongly highlights the need for control strategies that explicitly address the bandwidth limitation on communication between vehicles, which is severely restricted underwater. All these emerging applications motivate the development of a new chapter of control theory that deals with networked systems and combines the control and communication issues, taking into account all the limitations on communication between sensors, controllers, and actuators.

Recently there was a good deal of research activity in this field; e.g., see [2, 5, 7, 12, 13, 19, 20, 21, 22, 23, 24, 26, 27, 29, 30, 31, 32, 34, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 47, 50, 51, 52, 53, 57, 59, 60, 61, 62, 63] and the references therein. In particular, optimization problems for perfect (memoryless and noise-free) finite alphabet channels and noisy Gaussian channels with power constraints were studied in [61, 62, 63]; such problems for discrete perfect channels were also examined in [5, 36, 52]. The related problem of the design of optimal sequential quantization schemes for uncontrolled Markov processes was addressed in [6, 62]. Various schemes for stabilization

*Received by the editors December 28, 2004; accepted for publication (in revised form) February 21, 2007; published electronically September 14, 2007. This work was supported by the Australian Research Council and grant 06-08-01386 from the Russian Foundation for Basic Research.

http://www.siam.org/journals/sicon/46-4/62169.html

†Department of Mathematics and Mechanics, Saint Petersburg University, Universitetskii pr.28, Petrodvoretz, St.Petersburg 198504, Russia (almat@am1540.spb.edu).

‡Corresponding author. School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney 2052, Australia (a.savkin@unsw.edu.au).

and observation of unstable linear plants via limited capacity channels were proposed and analyzed in, e.g., [2, 3, 7, 12, 19, 23, 29, 30, 31, 41, 44, 45, 47, 57, 59, 60, 62]. The smallest data rate above which stabilization/observation of a linear plant is possible was derived in [2, 19, 41, 43, 44, 45, 60, 62] in various settings, including both stochastic [41, 43, 44, 45] and deterministic [2, 19, 52, 60, 62] ones. The focus in these works was on the channel quantization effects and perfect finite alphabet channels. This is a natural necessary step in developing the theory. At the same time, this is in contrast with the classic communication theory, where limited capacity channels are modeled in terms of not only quantization effects but also channel errors and time delays. Moreover, many of the major results in this theory are grown on the ground of noisy channel models. So incorporating noisy discrete channels into control problem models seems to be an unavoidable step in the synthesis of the control and communication theories.

To our knowledge, observability/stabilizability of unstable linear plants over noisy discrete channels was addressed in [24, 29, 30, 31, 50, 51, 57, 59]. In [24, 50, 51, 57], the focus was on scalar linear systems with additive disturbances and mth moment observability/stabilizability. It was shown in [50, 51] that, in general, the Shannon concept of capacity cannot serve as the proper figure-of-merit for erroneous channels in control feedback loops. Both sufficient and necessary criteria for observability/stabilizability were given in [50, 51] in terms of a new parametric notion of the channel capacity (called anytime capacity) introduced in [50]. An encoder-decoder pair for estimating the state of a scalar noisy linear system via a noisy binary symmetric channel was proposed in [24]. It was shown by simulation that the estimation error is bounded. Another such pair was constructed in [57] for such a channel with a perfect feedback link. Conditions ensuring that the mathematical expectation of the estimation error is bounded were obtained. In [67], the focus is on stabilizability by means of memoryless controllers in the case where the channels transmitting both observations and controls are noisy and discrete, and the plant is scalar, linear, and stochastic. The works [29, 30, 31] deal with the moment stabilizability of uncertain linear systems with additive disturbances over truncation channels. Such a channel transmits binary code words by dropping a random number of concluding bits. This generalization of the classic erasure channel is motivated by certain wireless communication applications in [29]. Constructive conditions for the robust moment stabilizability are obtained, and a stabilizing controller of a limited computation complexity is explicitly constructed.

It should be remarked that the moment observability/stability formally permits the error at a given time to be large and ensures only that large errors occur with a small probability. Modulo the strong law of large numbers, this guarantees only that along almost any trajectory, the frequency of large errors is small. The natural stronger control objective is to exclude large errors altogether. For stochastic models, this takes the form of almost sure observability/stabilizability. Such an observability/stabilizability via noisy communication links was addressed in [34, 35, 59] for noise-free linear time-invariant (LTI) plants. A necessary condition was established for general erroneous channels in [59]. This condition is that the Shannon capacity $\mathfrak{c}$ of the channel [56] is greater than or equal to the sum η of the logarithms of the absolute values of the system unstable eigenvalues. The sufficiency of this condition (with the strict inequality sign) was justified for only a particular channel model, i.e., the erasure channel with a perfect feedback. Such a channel either transmits the message correctly or loses it. In the latter case, the transmitter becomes aware of the failure via the feedback communication link. It was shown that whenever $\mathfrak{c} > \eta$ and there are no plant disturbances, the system is almost surely asymptotically observable

and stabilizable over the erasure channel. A similar stabilizability result is obtained in [27] under a stronger assumption about the erasure channel.[1]

This paper considers observability/stabilizability almost surely and with as large a probability as desired for general noisy *discrete memoryless channels* (DMC). We study how both quantization effects and channel errors limit the capability for reliable detection or stabilization. In doing so, we look for criteria that are "almost" exhaustive: they are necessary and "almost" sufficient. Furthermore, we examine both cases where the feedback communication link is and is not available, respectively. The objective is to provide a theoretical benchmark by discovering the role of the fundamental concept introduced by Shannon: the capacity of the erroneous channel [54].

We show that for general DMC, this capacity $\mathfrak{c}$ constitutes the border between the cases, where the noiseless LTI partially observed plant is and is not, respectively, almost surely asymptotically observable/stabilizable. More precisely, for such an observability/stabilizability to hold, the inequality $\mathfrak{c} \geq \eta$ is necessary and $\mathfrak{c} > \eta$ is sufficient. (We recall that η is the sum of the logarithms of the absolute values of the system unstable eigenvalues.) In its necessity part, this fact was previously established in [59] for more general (not necessarily discrete and memoryless) channels. We supplement this result from [59] by showing that the inequality $\mathfrak{c} \geq \eta$ remains necessary not only for almost sure but also certain weaker forms of observability/stabilizability. For example, it holds whenever some observer/controller keeps the time-average estimation/stabilization error bounded with a nonzero probability. We also specify the necessity of the above bound $\mathfrak{c}$ by showing that whenever it is trespassed $\eta > \mathfrak{c}$, any estimation/stabilization algorithm almost surely exponentially diverges. The sufficiency part of our result deals with arbitrary DMC and thus extends the corresponding results from [27, 59], which concern only the erasure channels. This part of our work can be viewed as completing the research [59] in the important case of DMC. We also show that the inequality $\mathfrak{c} > \eta$ is sufficient irrespective of availability of the feedback communication link.

More specifically, we show that in the absence of such a link, the estimation error can be made decaying to zero with as large a probability as desired by a proper design of the observer. However, this is achieved at the expense of using code words whose lengths grow as the estimation process progresses. It should be stressed that despite this, the observer produces an asymptotically exact state estimate online. In other words, the estimate of the state at time t is generated at time t, i.e., with no delay. At the same time, the increasing code word lengths require that the memory used by the estimator should increase accordingly.[2] This disadvantage can be discarded if a perfect feedback communication link is available: the result of any transmission across the (feedforward) channel becomes known to the transmitter by the time of the next transmission. This makes it possible to establish a complete synchronization of the encoder and decoder, within which the encoder duplicates the state estimate generated by the decoder. Our main result concerning the detection problem asserts that whenever $\mathfrak{c} > \eta$, the above feedback enables one to design an observer on the basis of fixed length code words for which the estimation error decays to zero almost surely.

A realistic converging observer is explicitly constructed. However, the scheme for the transmission of information across the channel is not described in detail. The point is that the observer employs block codes transmitting data at a given rate below

[1] This is the assumption on p. 736 about the uniform convergence of the average dropout rate to the limit, which makes the model from [27] a particular case of that from [37].

[2] The same feature is characteristic of the anytime coding-decoding schemes considered in [50, 51].

the channel capacity $\mathfrak{c}$ with a given probability of error. Classic information theory guarantees the existence of such a code. Moreover, the invention of such codes is the standard long-standing task in information sciences. It is supposed that a relevant solution should be employed to construct the observer. Thus it is shown that whenever the observability condition $\mathfrak{c} > \eta$ is satisfied and a perfect feedback link is available, almost sure detection can be ensured by a realistic observer with bounded (as time progresses) algebraic complexity and memory consumption per step, which is based on classic block coding-decoding schemes of communication.

Our results on stabilizability are similar to those concerning observability. However, there is a strong distinction. Specifically, we show that unlike detection, stabilization needs far less feedback communication. As a preliminary fact, we first show that to make the stabilization error decaying to zero almost surely by using fixed length code words, a feedback communication of arbitrarily small rate is sufficient. Second, we demonstrate that in fact such a communication requires no special means (such as a special feedback link), since it can be implemented by means of control. This can be arranged thanks to the fact that, on the one hand, the decoder-controller influences the motion of the system and, on the other hand, the sensor observes this motion and feeds the coder by the observation. So the controller is able to encode a message by imparting the motion a certain specific feature. The coder can receive the message by observing the motion and detecting this feature.

Apparently, control should be employed for information transmission with caution, since this potentially contradicts the main control objective [59]. For example, the best result of stabilization would be to keep the state exactly at the required position. However, the information transmission along the above lines requires us to deviate the state from this position. Thus a certain trade-off between the major control objective and communicating information by means of control should be established. Our first stabilization result serves this trade-off by showing that as little information as desired may be transmitted by means of control to achieve stability.

The focus on noise-free plants is motivated by the objective of the paper: to highlight the role of the Shannon (ordinary) capacity. The point is that in the presence of additive bounded disturbances, the border between the cases where the plant can and cannot, respectively, be observed/stabilized with an almost sure bounded error is constituted not by the ordinary $\mathfrak{c}$ but the zero error capacity $\mathfrak{c}_0$ of the channel, another fundamental characteristic introduced by Shannon [55]. In particular, if $\mathfrak{c}_0 < \eta$, the system affected by uniformly and arbitrarily small external disturbances can never be observed/stabilized: the error is unbounded almost surely, irrespective of which causal algorithm of observation/stabilization is employed. These facts were established for erasure infinite alphabet channels in [32] and general DMC in [39]. A result similar in spirit is obtained in [30, 31] for truncation channels and noise-free LTI plants. It is shown that for such a plant to be uniformly stabilizable (i.e., with an error uniformly bounded over the initial states from the unit ball) over such a channel, it is necessary that a certain number $r_{\min}$ of bits is not lost under any circumstances (i.e., with probability 1), and this number $r_{\min}$ exceeds the above quantity η. Since $r_{\min}$ equals the zero-error capacity of the channel at hand, this claim is in harmony with the results of [39]. It should be remarked that $\mathfrak{c}_0 \le \mathfrak{c}$, in general, and $0 = \mathfrak{c}_0 < \mathfrak{c}$ for many particular communication channels [25, 65]. It follows that an asymptotically unstable ($\eta > 0$) plant can never be observed/stabilized with an almost sure bounded error over such channels in the presence of disturbances, whereas almost sure observation/stabilization is possible, provided the disturbance is zero and $\eta < \mathfrak{c}$. A problem of stabilization via noisy channels in probability was considered for linear

plants with bounded disturbances in [38, 51], and necessary and sufficient conditions were presented. Furthermore, detailed discussions of these topics can be found in the research monograph [33].

The observers/controllers considered in this paper employ quantizers with adjusted sensitivity [7, 62] in the multirate fashion [47]. Such a quantizer can be viewed as a cascade of a multiplier by an adjustable factor and an analog-to-digital converter. To be transmitted across the channel, the outputs of this converter are encoded by means of low-error block codes. This is in the vein of the classic source-channel separation principle [4, 17, 18]. (For the state estimation problem, a similar approach was earlier considered in [50] in the form of the following separation of the source and channel. At first, a coder-decoder pair is designed under the assumption that the channel is perfect. Then another coder-decoder pair is constructed to carry the outputs of the first coder reliably across the channel.) For the stabilization problem, the above scheme is considered in connection with a limited communication feedback of arbitrarily small capacity. This feedback is used to put the values of only one scalar variable computed by both the coder and decoder in harmony. This variable is the above adjustable factor. The synchronization is delayed and, as a result, not complete: the values become coherent only if the current feedforward transmission across the channel is errorless. However, we show that this is enough to make the stabilization error almost surely decaying to zero.

The view of the control loop as a link transmitting information is not new. A posteriori, this means that the control loop does transmit information, though its contents may not be clearly specified a priori [11]. The "constructive" part of the same view is the idea that the control signals can be employed as carriers of a priori prespecified information from the decoder-controller to the coder. For various settings, various schemes of such a transmission were considered in, e.g., [34, 38, 51, 59].

The paper is organized as follows. Sections 2 and 3 contain the statements of the detection and stabilization problems. The main results are formulated in section 5, which is prefaced by the list of basic notations and assumptions in section 4. The necessary conditions for observability and stabilizability are justified in sections 6 and 7, respectively. Sections 8 and 9 are focused on the respective sufficient conditions. There also is an appendix containing the proof of a technical result.

2. Detection problem. We consider unstable discrete-time invariant linear plants of the form

$$x(t+1) = Ax(t); \qquad x(0) = x_0, \qquad y(t) = Cx(t). \tag{2.1}$$

Here $x \in \mathbb{R}^n$ is the state and $y \in \mathbb{R}^{n_y}$ is the measured output. The instability means that there is an eigenvalue λ of the matrix A with $|\lambda| \geq 1$. The initial state x_0 is a random vector. The objective is to estimate the current state on the basis of the prior measurements.

We consider the case where this estimate is required at a remote location. The only way to communicate information from the sensor to this location is via a given random noisy discrete channel. So to be transmitted, measurements must first be translated into a sequence of symbols e from the finite *input alphabet* $\mathcal{E}$ of the channel. This is done by a special system's component, referred to as the *coder*. Its outputs e are then transmitted over the channel and transformed by some sort of random disturbance or noise into a sequence of channel's outputs s from a finite *output alphabet* $\mathcal{S}$. By employing the prior outputs s, the *decoder(-estimator)* produces an estimate $\widehat{x}$ of the current state x. In this situation illustrated in Figure 1, the *observer* is

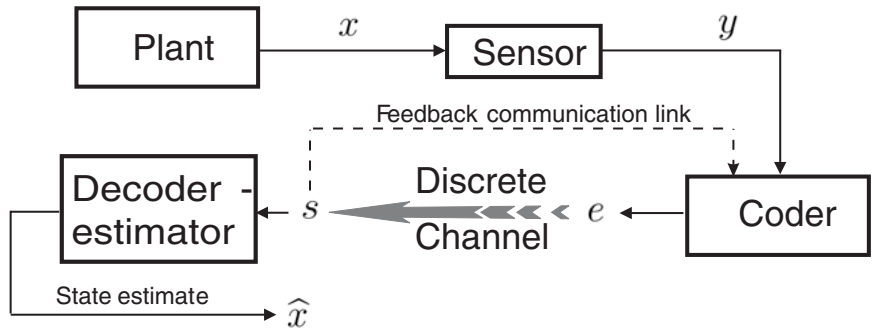


FIG. 1. *Estimation via a limited capacity communication channel.*

constituted by the coder-decoder pair.

The decoder is defined by an equation of the form

$$\widehat{x}(t) = \mathfrak{X}\left[t, s(0), s(1), \ldots, s(t)\right]. \tag{2.2}$$

We consider two classes of coders, each giving rise to a particular problem setup. The first class is related to feedback communication channels [55]: the transmission result $s(t)$ becomes known at the coder site by time $t+1$ of the next transmission. The second class corresponds to the channels with no feedback. The coders from these classes are said to be *with* and *without a feedback* and are given by the following equations, respectively:

$$e(t) = \mathfrak{E}[t, y(0), \ldots, y(t), s(0), \ldots, s(t-1)] \in \mathcal{E}, \tag{2.3}$$

$$e(t) = \mathfrak{E}[t, y(0), \ldots, y(t)] \in \mathcal{E}. \tag{2.4}$$

The communication feedback enables the coder (2.3) to be aware of the actions of the decoder via duplicating the computations in accordance with (2.2). This gives the coder the ground to try to compensate for the previous channel errors. However, it should be noted that this feedback does not increase the rate at which the information can be transmitted across the channel with as small a probability of error as desired [10, 54]. At the same time, it may increase the rate at which information can be transmitted with the zero probability of error [55]. The feedback may also increase the reliability function [68] and simplify coding and decoding operations [65]. For further discussion of this issue and a detailed survey, we refer the reader to [65]. The role of communication feedback in control and state estimation was discussed in [59, 60, 62, 63, 66].

The information received by the decoder is limited to a finite number of bits at any time. So the decoder is hardly able to restore the state with the infinite exactness $\widehat{x}(t) = x(t)$ for a finite time. In this paper, we pursue a more realistic objective of detecting the unstable modes of the system and accept that an observer succeeds if

$$|x(t) - \widehat{x}(t)| \to 0 \quad \text{as} \quad t \to \infty. \tag{2.5}$$

DEFINITION 2.1. *The coder-decoder pair is said to* detect *or* track *the state whenever* (2.5) *is true and to* keep the estimation error (or time-average error) bounded *if the following much weaker properties hold, respectively:*

$$\overline{\lim_{t\to\infty}} |x(t) - \widehat{x}(t)| < \infty, \qquad \overline{\lim_{t\to\infty}} \frac{1}{t} \sum_{\theta=0}^{t-1} |x(\theta) - \widehat{x}(\theta)| < \infty. \tag{2.6}$$

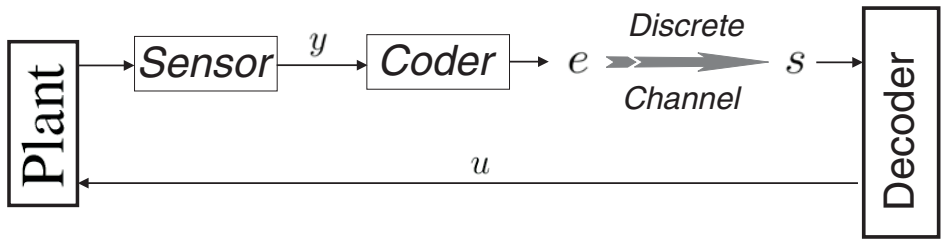


FIG. 2. *Stabilization via a limited capacity communication channel.*

The main question to be discussed is *how low the data rate of the channel can be made before the construction of a coder-decoder pair detecting the state becomes impossible.* In this paper, we focus on the cases where "detecting" means either "detecting with arbitrarily large probability" $p < 1$ or "detecting almost surely."

3. Stabilization problem. Now we consider the controlled version of the unstable plant (2.1):

$$x(t+1) = Ax(t) + Bu(t), \qquad x(0) = x_0, \qquad y(t) = Cx(t), \tag{3.1}$$

where $u \in \mathbb{R}^{n_u}$ is the control. The objective is to construct a controller that asymptotically stabilizes the system: $x(t) \to 0$ and $u(t) \to 0$ as $t \to \infty$.

We examine a remote control setup. Based on the prior observations, the *coder* emits a message $e \in \mathcal{E}$ into the channel. This message may be corrupted during the transmission $e \xrightarrow{\text{channel noise}} s \in \mathcal{S}$. Proceeding from the messages s received over the channel up to the current time t, the *decoder(-controller)* selects a control $u(t)$:

$$u(t) = \mathfrak{U}\left[t, s(0), s(1), \dots, s(t)\right]. \tag{3.2}$$

In this situation, depicted in Figure 2, the controller is assembled of the coder and decoder.

We still consider two classes of coders given by (2.3) and (2.4), respectively. The first of them is associated with the case where there is a perfect feedback in communication between the coder and decoder (see Figure 3). The second class deals with the situation where no such a feedback is available.

DEFINITION 3.1. *A coder-decoder pair is said to* stabilize *the system if*

$$|x(t)| \to 0 \quad \text{and} \quad |u(t)| \to 0 \quad \text{as} \quad t \to \infty \tag{3.3}$$

and to keep the stabilization error (or time-average error) bounded *if the much weaker properties hold, respectively:*

$$\overline{\lim_{t\to\infty}} |x(t)| < \infty, \qquad \overline{\lim_{t\to\infty}} \frac{1}{t} \sum_{\theta=0}^{t-1} |x(\theta)| < \infty. \tag{3.4}$$

What is the bound on the data rate of the channel above which there exists a stabilizing coder-decoder pair? Here "stabilizing" means either "stabilizing almost surely" or "stabilizing with as large a probability as desired."

4. Notations and assumptions. The symbols $\boldsymbol{P}$ and $\boldsymbol{E}$ stand for probability and expectation, respectively. For a random variable $A \in \mathcal{A} = \{a\}$, the conditional

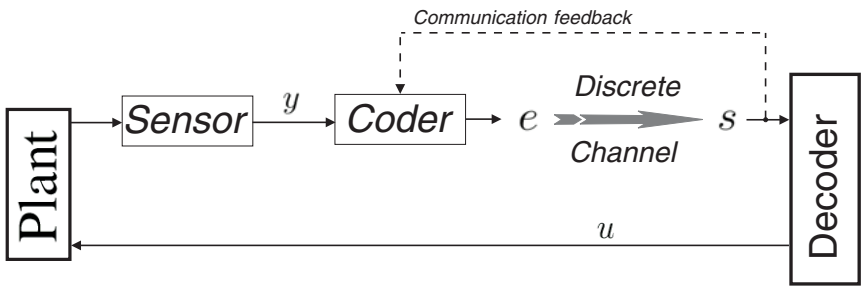


FIG. 3. *Stabilization under a communication feedback.*

probability given $A = a$ is denoted by $\boldsymbol{P}(\cdot|A = a)$ or $\boldsymbol{P}(\cdot|a)$, and $\boldsymbol{P}(a) := \boldsymbol{P}(A = a)$. For two such variables A and $B \in \mathcal{B} = \{b\}$, the symbol $I(A, B)$ stands for the *mutual information*:

$$I(A,B) = \int \boldsymbol{P}(da,db) \log_2 \frac{\boldsymbol{P}(da,db)}{\boldsymbol{P}(da) \otimes \boldsymbol{P}(db)} = \boldsymbol{E} \log_2 \frac{\boldsymbol{P}(da,db)}{\boldsymbol{P}(da) \otimes \boldsymbol{P}(db)} \tag{4.1}$$

if the joint distribution $\boldsymbol{P}(da, db)$ of A and B has the density $\frac{\boldsymbol{P}(da,db)}{\boldsymbol{P}(da)\otimes\boldsymbol{P}(db)}$ with respect to the probability measure $\boldsymbol{P}(da) \otimes \boldsymbol{P}(db)$, and $I(A, B) = \infty$ otherwise [48]. (Here and throughout, $\log_2 0 := -\infty$, $\pm\infty \cdot 0 := 0$.) If the sets $\mathcal{A}, \mathcal{B}$ are finite, then

$$I(A,B) = H(B) - H(B|A) = H(A) - H(A|B) = \sum_{a\in\mathcal{A}, b\in\mathcal{B}} \boldsymbol{P}(a,b) \log_2 \frac{\boldsymbol{P}(a,b)}{\boldsymbol{P}(a)\boldsymbol{P}(b)},$$

where the symbols $H(B), H_a(B) = H_{A=a}(B)$, and $H(B|A)$ denote the *entropy*, the *conditional entropy* given $A = a$, and the *averaged conditional entropy*, respectively, i.e.,

$$H(B) := -\sum_{b\in\mathcal{B}} \boldsymbol{P}(b) \log_2 \boldsymbol{P}(b), \qquad H_a(B) = -\sum_{b\in\mathcal{B}} \boldsymbol{P}(b|a) \log_2 \boldsymbol{P}(b|a), \tag{4.2}$$

$$H(B|A) = \boldsymbol{E} H_A(B) = \sum_{a\in\mathcal{A}} \boldsymbol{P}(a) H_a(B).$$

The probability density of a random vector $V \in \mathbb{R}^s$ is denoted by $p_V(\cdot)$ and that given $A = a$ by $p_V(\cdot|A = a) = p_V(\cdot|a)$. The *differential entropy* of V is the quantity

$$h(V) := -\boldsymbol{E} \log_2 p_V(V) = -\int_{\mathbb{R}^s} p_V(v) \log_2 p_V(v)\, dv. \tag{4.3}$$

This entropy can be viewed as a measure of information required to describe the random vector to a particular accuracy. Approximately $h(V) + sb + \log_2 \mathbf{mes}\, B_0^1$ bits suffice to describe $V \in \mathbb{R}^s$ to b-bit accuracy. (Here B_0^1 is the unit ball in $\mathbb{R}^s$ and **mes** is the Lebesgue measure.) The differential entropy can take negative and infinite values. The symbol $h_{B=b}(V)$ stands for the conditional differential entropy given $B = b$.

The following assumptions are adopted throughout the paper.

Assumption 4.1. The coder sends signals to the decoder over a given stationary discrete noisy memoryless channel [14, 18]. In other words, given a current channel input $e(t)$, the current output $s(t)$ is statistically independent of all other inputs and outputs $e(j), s(j)$, $j \neq t$, and the conditional probability $W(s|e) := \boldsymbol{P}[s(t) = s|e(t) = e]$, $s \in \mathcal{S}$, $e \in \mathcal{E}$, does not depend on time t.

Note that this model incorporates the effect of message loss by including a special "void" symbol $\emptyset$ in the output alphabet $\mathcal{S}$. Then $s(t) = \emptyset$ means that the message $e(t)$ is lost by the channel.

Assumption 4.2. The plant does not affect the operation of the channel: given an input $e(t)$, the output $s(t)$ is statistically independent of the initial state x_0.

Assumption 4.3. The initial state x_0 has a probability density $p_0(x)$.

Assumption 4.4. The pair (A, C) is detectable.

When dealing with the stabilization problem, we impose one more assumption.

Assumption 4.5. The pair (A, B) is stabilizable.

To state the results of the paper, we need the notion of the Shannon *capacity* of the stationary discrete memoryless channel. This is the maximum mutual information between the input and output of the channel [14]:

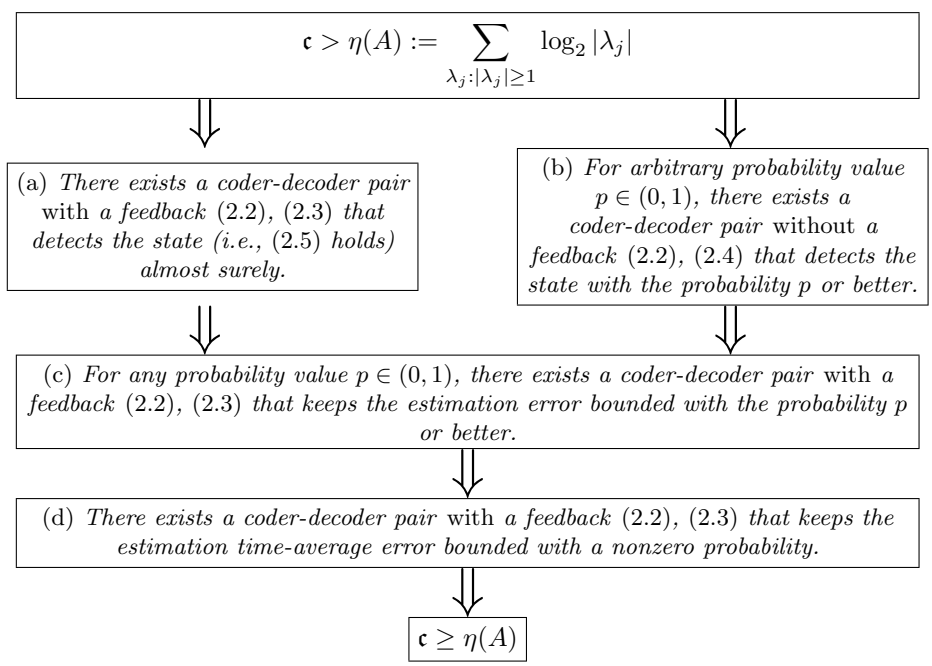
$$\mathfrak{c} = \max_{\boldsymbol{P}_E} I(e, s). \tag{4.4}$$

Here the maximum is over all probability distributions $\boldsymbol{P}_E$ on the input channel alphabet $\mathcal{E} = \{e\}$. Whereas $\boldsymbol{P}_E$ is interpreted as the probability distribution of e, the joint distribution of the channel input e and output s is taken to be that of (e, s) when s results from sending e over the channel: $\boldsymbol{P}_{E,S}(e, s) := W(s|e)\boldsymbol{P}_E(e)$.

5. The domains of observability and stabilizability are determined by the Shannon channel capacity.

5.1. Detection problem.

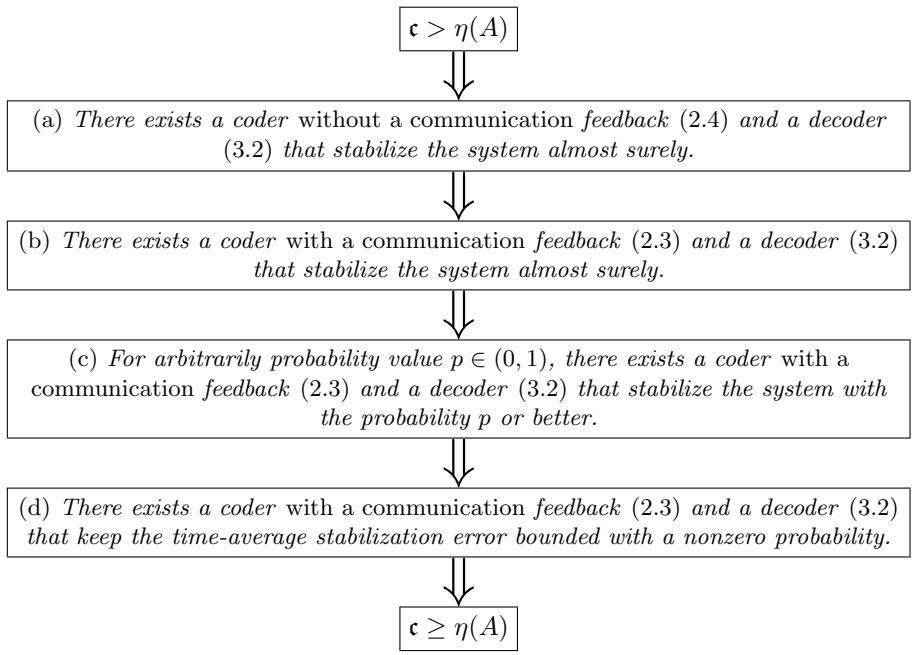
THEOREM 5.1. *Suppose that Assumptions* 4.1–4.4 *hold. Denote by* $\lambda_1, \dots, \lambda_n$ *the eigenvalues of the system* (2.1) *repeating in accordance with their algebraic multiplicities and by* $\mathfrak{c}$ *the capacity* (4.4) *of the communication channel. Then the following implications are true:*



The proof of this theorem will be given in sections 6 and 8. The explicit constructions of tracking coder-decoder pairs will be presented in subsections 8.3 and 8.5.

5.2. Stabilization problem. Similar results are valid for the stabilization problem.

THEOREM 5.2. *Suppose that Assumptions* 4.1–4.5 *hold, and adopt the notations* $\mathfrak{c}$ *and* $\eta(A)$ *from Theorem* 5.1. *Then the following implications are true:*



The proof of this theorem will be given in sections 7 and 9. The explicit construction of a stabilizing coder-decoder pair will be offered in subsections 9.1 and 9.2.

5.3. Comments. The implications (a) $\Rightarrow$ $\mathfrak{c} \geq \eta(A)$ and (b) $\Rightarrow$ $\mathfrak{c} \geq \eta(A)$ contained in Theorems 5.1 and 5.2, respectively, were proved in [59] for general noisy channels (not necessarily discrete and memoryless). The implications $\mathfrak{c} > \eta(A) \Rightarrow$ (a) and $\mathfrak{c} > \eta(A) \Rightarrow$ (b) from Theorems 5.1 and 5.2, respectively, were justified in [59] for a particular DMC: the erasure channel with arbitrary finite alphabet. The implications (a) $\vee$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) from Theorem 5.1 and (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) from Theorem 5.2 are evident. They are mentioned to stress that the claims (a), (b), (c), and (d) are included in the chains of implications with approximately identical extreme terms. Thus these statements are "almost equivalent," and the inequality $\mathfrak{c} > \eta(A)$ is sufficient and "almost necessary" for the systems (2.1) and (3.1) to be detectable and stabilizable, respectively, via the noisy communication channel.

In spirit, Theorem 5.1 resembles the celebrated Shannon channel coding theorem [14, 18, 54]. Indeed, the latter states that whenever the source produces information at the rate $R < \mathfrak{c}$ bits per unit time, the success, i.e., errorless transmission, can be ensured with as large a probability as desired. If, conversely, $R > \mathfrak{c}$, this is impossible. Here the means to ensure success are the rules to encode and decode information before and after transmission, respectively. Theorem 5.1 asserts just the same, provided the "success" is understood as asymptotic tracking (2.5) of the state, the "means" are the coder and decoder-estimator, and R is replaced by $\eta(A)$. This analogy is enhanced by the similarity between the quantities R and $\eta(A)$. Each of them can be interpreted as the unit-time increment of the number of bits required to describe the entity that the receiver wants to know.

Indeed, in the case of the Shannon theorem, this entity is abstract information generated by a source at the rate R, and the interpretation is apparent. In the case considered in this paper, this entity is the "unstable" part x_+ of the state x (since asymptotic tracking (2.5) does not concern the "stable" x_- one).

Explanation 5.1. Here and throughout the paper, $x_\pm \in L_\pm$, where L_+ and L_- are the invariant subspaces of A related to the unstable $\{\lambda_j : |\lambda_j| \geq 1\}$ and stable $\{\lambda_j : |\lambda_j| < 1\}$ parts of its spectrum, and $A_\pm$ denotes the operator A acting on its invariant subspace $L_\pm$.

With regard to the relation $x_+(t+1) = A_+ x_+(t)$, simple calculus (see, e.g., [45] or (6.6)) shows that the entropy (4.3) of $x_+(t)$ evolves as follows:

$$h[x_+(t+1)] = h[x_+(t)] + \log_2 |\det A_+|,$$

where det is the determinant. Thus the number of bits required to describe $x_+(t)$ to any given accuracy b increases by $\log_2 |\det A_+|$ per unit time. It remains to note that $\log_2 |\det A_+| = \eta(A)$.

The above remark on the similarity between the channel coding theorem and Theorem 5.1 clearly extends on Theorem 5.2. Another point of similarity between Theorem 5.2 and the classic information theory concerns the communication feedback. Whereas the classic theory states that this feedback does not increase the rate at which data can be reliably transmitted across the noisy channel [10, 54], Theorem 5.2 shows that the feedback does not visibly extend the class of stabilizable noise-free plants.

Equations (2.2)–(2.4), (3.2) impose no restrictions on the memories of the coder and decoder. At the same time, the observer reliably tracking the state in the presence of the communication feedback and the stabilizing controller that will be explicitly constructed further consume limited (as time progresses) memories.

In the case of the perfect channel ($\mathcal{E} = \mathcal{S}$ and $W(e|e) = 1$), Theorems 5.1 and 5.2 come to results from [41, 42, 43, 44, 45, 47, 53, 60, 62], and the strict inequality $\mathfrak{c} > \eta(A)$ is necessary for the existence of both the tracking observer and the stabilizing controller.

5.4. Complements to the necessary conditions. The last implication (d) $\Rightarrow$ $\mathfrak{c} \geq \eta(A)$ from both Theorems 5.1 and 5.2 can be complemented and enhanced by the following facts.

PROPOSITION 5.3. *Suppose that* $\mathfrak{c} < \eta(A)$. *Then the state can never be observed and the plant can never be stabilized with a bounded error:*

$$\left.\begin{array}{r} \overline{\lim_{t\to\infty}}\, |x(t) - \widehat{x}(t)| = \infty \\ \overline{\lim_{t\to\infty}}\, |x(t)| = \infty \end{array}\right\} \quad \text{a.s. for the plant } \left\{\begin{array}{l}(2.1)\\(3.1)\end{array}\right\}, \text{ any coder,} \tag{5.1}$$

$$\text{and } \left\{\begin{array}{l}\text{decoder-estimator}\\ \text{decoder-controller}\end{array}\right. .$$

This divergence is as fast as exponential. Specifically, pick $\alpha > 1$ *so that* $\log_2 \alpha < \frac{\eta(A)-\mathfrak{c}}{\dim x}$. *Then*

$$\left.\begin{array}{r} \overline{\lim_{t\to\infty}}\, \alpha^{-t}|x(t) - \widehat{x}(t)| = \infty \\ \overline{\lim_{t\to\infty}}\, \alpha^{-t}|x(t)| = \infty \end{array}\right\} \quad \text{a.s. for the plant } \left\{\begin{array}{l}(2.1)\\(3.1)\end{array}\right\}, \text{ any coder,} \tag{5.2}$$

$$\text{and } \left\{\begin{array}{l}\text{decoder-estimator}\\ \text{decoder-controller}\end{array}\right. .$$

REMARK 5.1. *Proposition* 5.3 *entails that* (d) $\Rightarrow \mathfrak{c} \geq \eta(A)$ *in the context of both Theorems* 5.1 *and* 5.2.

Indeed, consider the context of Theorem 5.1 for the definiteness. Suppose that $\mathfrak{c} < \eta(A)$. Then by (5.2), there exist random times $0 < \tau_1 < \tau_2 < \cdots$ such that almost surely $|x(\tau_i) - \widehat{x}(\tau_i)| \geq \alpha^{\tau_i}$ for all i. Then

$$\frac{1}{\tau_i + 1} \sum_{\theta=0}^{\tau_i} |x(\theta) - \widehat{x}(\theta)| \geq \frac{|x(\tau_i) - \widehat{x}(\tau_i)|}{\tau_i + 1} \geq \frac{\alpha^{\tau_i}}{\tau_i + 1} \to \infty \quad \text{as} \quad i \to \infty \quad \text{a.s.}$$

in violation of (d). Hence (d) $\Rightarrow \mathfrak{c} \geq \eta(A)$.

The probability of large observation and stabilization errors is estimated in the following proposition.

PROPOSITION 5.4. *Suppose that the system has no stable modes,* $\eta(A) > \mathfrak{c}$*, and the initial state* x_0 *is almost surely bounded and has a finite differential entropy. Then*

$$\left.\begin{array}{r} \varlimsup\limits_{t\to\infty} \boldsymbol{P}\Big[|x(t) - \widehat{x}(t)| \geq b(t)\Big] \\ \varlimsup\limits_{t\to\infty} \boldsymbol{P}\Big[|x(t)| \geq b(t)\Big] \end{array}\right\} \geq 1 - \frac{\mathfrak{c}}{\eta(A)} \text{ for the plant } \left\{\begin{array}{l} (2.1) \\ (3.1) \end{array}\right\}, \text{ any coder,}$$

$$\text{and } \left\{\begin{array}{l} \text{decoder-estimator} \\ \text{decoder-controller} \end{array}\right. .$$

Here $b(t) > 0$, $t = 0, 1, \ldots$, *is any sequence such that* $\frac{\log_2 b(t)}{t} \to 0$ *as* $t \to \infty$.

REMARK 5.2. *This proposition implies that* $\mathfrak{c} \geq \eta(A)$ *whenever there exist a coder and decoder that keep the mathematical expectation of the error* $\epsilon(t)$ *(or at least the time average* $\frac{1}{t}\sum_{\theta=0}^{t-1} \boldsymbol{E}\epsilon(\theta)$*) bounded.*

The proofs of Propositions 5.3 and 5.4 will be given in sections 6 and 7 (where the detection and stabilization problems will be addressed, respectively).

6. Proofs of Propositions 5.3 and 5.4 for the state estimation problem and the (d) $\Rightarrow \mathfrak{c} \geq \eta(A)$ part of Theorem 5.1. In this section, the focus is on the proof of Proposition 5.3. Proposition 5.4 will be justified within this proof, whereas the above part of Theorem 5.1 follows from Proposition 5.3 by Remark 5.1.

We start with a simple observation. So far as the asymptotic tracking does not concern the stable modes, it seems more or less clear that the proof can be confined to systems with only unstable ones. This is ensured by the following lemma, which employs the notations from Explanation 5.1.

LEMMA 6.1. *Suppose that some coder-decoder pair keeps the estimation error bounded with the probability better than* p *for the primal system* (2.1)*. Then such a pair can also be constructed for the system*

$$x_+(t+1) = A_+ x_+(t), \quad x_+(t) \in L_+, \quad x_+(0) = x_0^+, \quad y(t) = C x_+(t) \tag{6.1}$$

with some initial random vector x_0^+ *that satisfies Assumptions* 4.2 *and* 4.3*, is almost surely bounded, and has a finite entropy.*

It should be remarked that generally speaking, the system (6.1) is considered on a new underlying probability space. However, Assumptions 4.1–4.4 are still true and the channel parameters $W(s|e)$ remain unchanged.

It is easy to see that equations (6.1) describe the processes in the primal system (2.1) starting at $x(0) = x_0^+ \in L_+$. A certain technical nontriviality of Lemma 6.1 comes from the fact that due to Assumption 4.3, the probability to start at $x(0) \in L_+$

is zero (if $L_+ \neq \mathbb{R}^n$). At the same time, the assumptions of the lemma allow the initial coder-decoder pair to produce asymptotically infinite estimation errors not only with zero but a positive probability. To keep the estimation error bounded for the processes (6.1), this pair should be modified in general.

The proof of Lemma 6.1 is placed in the appendix.

To prove Proposition 5.3, we also need the concept of the joint entropy of a vector and a discrete quantity [48]. Let a random vector $V \in \mathbb{R}^s$ have a probability density and a random quantity B take values in a finite set $\mathcal{B}$ with elements b. We denote by $|M|$ the size of the set M, and by db the counting measure, i.e., the function $\mathcal{B}' \mapsto |\mathcal{B}'|$ of the set $\mathcal{B}' \subset \mathcal{B}$. The random quantity B has a probability density $p_B(b) := \boldsymbol{P}(B = b)$ with respect to the measure db, and the definition of the entropy $H(B)$ from (4.2) takes the form similar to (4.3): $H(B) = -\int_{\mathcal{B}} p_B(b) \log_2 p_B(b)\, db$. The joint entropy of $(V, B) \in \mathbb{R}^s \times \mathcal{B}$ is introduced by continuing this analogy. We note that (V, B) has a probability density $p_{V,B}(v, b) := p_V(v|B = b)\boldsymbol{P}(B = b)$ with respect to the measure $dx \otimes db$, and we set

$$H(V, B) := -\boldsymbol{E} \log_2 p_{V,B}(V, B) = -\int_{\mathbb{R}^s \times \mathcal{B}} p_{V,B}(v, b) \log_2 p_{V,B}(v, b)\, dvdb.$$

This entropy inherits many properties of (4.2) and (4.3). Now we list some of them. In doing so, we assume that B_i are random quantities taking finitely many values, either $\widehat{V} := V$ or $\widehat{V} := (V, B_1)$, Ψ is a deterministic function, and $h(V) \in \mathbb{R}$. The conditional entropy $H_{B=b}(\widehat{V}) = H_b(\widehat{V}) =: \mathfrak{H}(b)$ is the entropy of $\widehat{V}$ with respect to the probability given $B = b$, and $H(\widehat{V}|B) = \boldsymbol{E}\mathfrak{H}(B)$ is the averaged conditional entropy. (If $\widehat{V} = V$, we use h instead of H here.) The following properties hold:

$$h(V) \in \mathbb{R} \Rightarrow -\infty < H(\widehat{V}|B) \leq H(\widehat{V}) < +\infty, \qquad I(V, B) = h(V) - h(V|B), \tag{6.2}$$

$$H(V, B_1|B) = h(V|B_1, B) + H(B_1|B) \geq h(V|B), \tag{6.3}$$

$$h(V) \overset{[9]}{\leq} \frac{s}{2} \log_2\Bigl(2\pi e \boldsymbol{E}|V|^2\Bigr), \qquad h(V|B) = h(V - \Psi(B)|B). \tag{6.4}$$

The next preliminary fact is an estimate of uncertainty about the current state $x(t)$ given the output of the decoder. Further, the symbols S and E stand for sequences $\{s(t)\}_{t=0}^{\infty}$ and $\{e(t)\}_{t=0}^{\infty}$. Whenever $0 \leq m_- \leq m_+$, we put $S_{m_-}^{m_+} := \{s(j)\}_{j=m_-}^{m_+}$ and define $E_{m_-}^{m_+}$ similarly.

LEMMA 6.2. *Suppose that* $\det A \neq 0$, $h(x_0) \in \mathbb{R}$, *and* $\mathfrak{c}$ *is the capacity* (4.4) *of the channel. Then for any coder-decoder pair with a feedback* (2.2), (2.3), *the entropy* $h\bigl[x(t)|S_0^t\bigr]$ *is finite and*

$$h\bigl[x(t)|S_0^t\bigr] \geq h\bigl[x_0\bigr] + t\bigl(\log_2 |\det A| - \mathfrak{c}\bigr) - \mathfrak{c}. \tag{6.5}$$

Proof. Note first that by (2.1), the probability densities of $x(t)$ given S_0^θ evolve as follows: $p_j(x) = |\det A|^{-j} \times p_0(A^{-j}x)$, where $p_j(\cdot) := p_{x(j)}(\cdot|S_0^\theta)$ for all j and θ is a fixed time. By (4.3), this implies

$$h\bigl[x(t)\bigl|S_0^\theta\bigr] = h\bigl[x_0\bigl|S_0^\theta\bigr] + t \log_2 |\det A|. \tag{6.6}$$

By the standard arguments [10, 17, 48, 54], $I\bigl[x_0, S_0^t\bigr] \leq \mathfrak{c}(t + 1)$. It remains to note that $I\bigl[x_0, S_0^t\bigr] = h[x_0] - h\bigl[x_0\bigl|S_0^t\bigr]$ and to employ (6.6) with $\theta := t$.

LEMMA 6.3. *Suppose that* $|\det A| > 1$, $h(x_0) \in \mathbb{R}$, *and* $|x_0| < b_0$ *almost surely. Then, for any coder-decoder pair,*

$$\boldsymbol{P}\Big[|x(t) - \widehat{x}(t)| \le b\Big] \le \frac{\mathfrak{c}}{\log_2 |\det A|} + \frac{1}{t} \times \frac{1 - h(x_0) + \mathfrak{c} + \frac{n}{2}\log_2\big(2\pi e \max\{b^2, b_0^2\}\big)}{\log_2 |\det A|} \qquad \forall b > 0, t \ge 1. \tag{6.7}$$

A similar inequality can be obtained from Lemma 3.2 in [59].

REMARK 6.1. *Lemma* 6.3 *evidently justifies Proposition* 5.4 *for the detection problem.*

Proof of Lemma 6.3. Pick t and denote by $\mathfrak{B}$ the random event $\{|x(t) - \widehat{x}(t)| \le b\}$ and by $\mathcal{I}$ its indicator: $\mathcal{I} = 1$ if $\mathfrak{B}$ holds and $\mathcal{I} = 0$ otherwise. We also put $\eta := \log_2 |\det A|$ and $p := \boldsymbol{P}\big[\mathfrak{B}\big]$. Then

$$H\big[x(t), \mathcal{I}\big|S_0^t\big] \overset{(6.3)}{\ge} h\big[x(t)\big|S_0^t\big] \overset{(6.5)}{\ge} h[x_0] - \mathfrak{c} + t\big[\eta - \mathfrak{c}\big]. \tag{6.8}$$

The random variable $\mathcal{I}$ takes only two values. So its entropy (given any event) does not exceed 1. Hence

$$H\big[x(t), \mathcal{I}\big|S_0^t\big] \overset{(6.3)}{=\!=} h\big[x(t)\big|\mathcal{I}, S_0^t\big] + H\big[\mathcal{I}\big|S_0^t\big] \le 1 + \sum_{\sigma=0,1} \boldsymbol{P}(\mathcal{I} = \sigma) h_{\mathcal{I}=\sigma}\big[x(t)\big|S_0^t\big].$$

Repeating the arguments underlying (6.6) shows that $h_{\mathcal{I}=0}\big[x(t)\big|S_0^t\big] = h_{\mathcal{I}=0}\big[x_0\big|S_0^t\big] + t\eta$. Hence

$$\begin{aligned} H\big[x(t), \mathcal{I}\big|S_0^t\big] &\le 1 + (1-p) h_{\mathcal{I}=0}\big[x_0\big|S_0^t\big] + (1-p)t\eta + p h_{\mathcal{I}=1}\big[x(t)\big|S_0^t\big] \\ &\overset{(6.2)}{\le} 1 + (1-p) h_{\mathcal{I}=0}\big[x_0\big] + (1-p)t\eta + p h_{\mathcal{I}=1}\big[x(t)\big|S_0^t\big]. \end{aligned}$$

Here $|x_0| \le b_0$ almost surely, and so $h_{\mathcal{I}=0}[x_0] \overset{(6.4)}{\le} \frac{n}{2}\log_2[2\pi e \boldsymbol{E}(|x_0|^2|\mathcal{J} = 0)] \le \frac{n}{2}\log_2[2\pi e b_0^2]$. Furthermore,

$$\begin{aligned} h_{\mathcal{I}=1}\big[x(t)\big|S_0^t\big] &\overset{(2.2),(6.4)}{=\!=} h_{\mathcal{I}=1}\big[x(t) - \widehat{x}(t)\big|S_0^t\big] \overset{(6.2)}{\le} h_{\mathcal{I}=1}\big[x(t) - \widehat{x}(t)\big] \\ &\overset{(6.4)}{\le} \frac{n}{2}\log_2\Big[2\pi e \boldsymbol{E}\big(\underbrace{|x(t) - \widehat{x}(t)|^2}_{\le b^2 \text{ whenever } \mathfrak{B} \text{ holds}} \big|\mathfrak{B}\big)\Big] \le \frac{n}{2}\log_2\Big[2\pi e b^2\Big]. \end{aligned}$$

Thus we see that

$$\begin{aligned} H\big[x(t), \mathcal{I}\big|S_0^t\big] &\le 1 + (1-p)t\eta + \frac{n}{2}\big[(1-p)\log_2\big(2\pi e b_0^2\big) + p\log_2\big(2\pi e b^2\big)\big] \\ &\le 1 + (1-p)t\eta + \frac{n}{2}\log_2\big(2\pi e \max\{b_0^2, b^2\}\big). \end{aligned}$$

By combining this with (6.8), we get the following formula, which clearly implies (6.7):

$$t\Big\{\big[1 - (1-p)\big]\eta - \mathfrak{c}\Big\} \le 1 + \frac{n}{2}\log_2\big(2\pi e \max\{b_0^2, b^2\}\big) - h(x_0) + \mathfrak{c}\,. \qquad \square$$

Proof of Proposition 5.3 *for the state estimation problem.* Evidently, it suffices to consider the system with the full observation $y = x$, $C = I$ in (2.1). Suppose that

(5.1) fails to be true for the detection problem; i.e., there exists a coder-decoder pair that keeps the estimation error bounded with a positive probability. By Lemma 6.1, such a pair also exists for the auxiliary system (6.1). Since $\eta(A) = \eta(A_+)$, this system can be put in place of (2.1) in the proof. In other words, one can assume in the proof that $h(x_0) \in \mathbb{R}$, $|x_0| \le b_0 < \infty$ almost surely, and the system (2.1) at hand has no stable modes, and so $\eta(A) = \log_2 |\det A|$.

By sacrificing a small probability, the error boundedness can be made uniform: there exist $b > 0$ such that

$$\boldsymbol{P}\Big[\big|x(t) - \widehat{x}(t)\big| \le b \ \forall t\Big] > 0. \tag{6.9}$$

At the same time, Lemma 6.3 ensures that for any $\rho > \frac{\mathfrak{c}}{\eta(A)}$, there exists a nonrandom time $\tau_1 > 0$ such that

$$\boldsymbol{P}\Big[\big|x(t) - \widehat{x}(t)\big| \le b\Big] \le \rho \qquad \forall t \ge \tau_1.$$

Since $\eta(A) > \mathfrak{c}$ by the hypotheses of Proposition 5.3, one may pick $\rho < 1$ here.

Now we consider the tail of the process $x(t), \widehat{x}(t), e(t), s(t)$, $t \ge \tau_1 + 1$, in the conditional probability space given that $|x(\tau_1) - \widehat{x}(\tau_1)| \le b$ and $S_0^{\tau_1} = \boldsymbol{S}$. Here we employ an $\boldsymbol{S} \in \mathcal{S}^{\tau_1+1}$ such that

$$\boldsymbol{P}\big[\mathfrak{B}^1_{\boldsymbol{S}}\big] > 0, \qquad \text{where} \quad \mathfrak{B}^1_{\boldsymbol{S}} := \big\{|x(\tau_1) - \widehat{x}(\tau_1)| \le b \wedge S_0^{\tau_1} = \boldsymbol{S}\big\}.$$

The initial state $x(\tau_1+1) = A^{\tau_1+1}x_0$ of this tail is almost surely bounded and $h(x_0) \in \mathbb{R} \overset{(6.2),(6.6)}{\Longrightarrow} h\big[x(\tau_1+1)|\mathfrak{B}^1_{\boldsymbol{S}}\big] \in \mathbb{R}$. At the same time, the above conditioning does not alter the channel (considered for $t > \tau_1$) due to Assumptions 4.1 and 4.2. The signals $\widehat{x}(t), e(t), s(t)$, $t \ge \tau_1 + 1$, are still generated by (2.2) and (2.3) (or (2.4)), where $\boldsymbol{S}$ and $A^{-\tau_1-1}x(\tau_1+1), A^{-\tau_1}x(\tau_1+1), \dots, A^{-1}x(\tau_1+1)$ are substituted for $s(0), \dots, s(\tau_1)$ and $y(0), \dots, y(\tau_1)$, respectively. Thus Lemma 6.3 can be applied once more. It follows that $\boldsymbol{P}[|x(t) - \widehat{x}(t)|\mathfrak{B}^1_{\boldsymbol{S}}| \le b] \le \rho$ for all $t \ge \tau_2(\boldsymbol{S})$. For $\tau_2 := \max_{\boldsymbol{S}} \tau_2(\boldsymbol{S})$, we have

$$\begin{aligned}\boldsymbol{P}\Big[\big|x(\tau_2) - \widehat{x}(\tau_2)\big| \le b\Big|\big|x(\tau_1) - \widehat{x}(\tau_1)\big| \le b\Big] &= \sum_{\boldsymbol{S}} \boldsymbol{P}\Big[S^{\tau_1} = \boldsymbol{S}\Big|\big|x(\tau_1) - \widehat{x}(\tau_1)\big| \le b\Big] \\ &\times \boldsymbol{P}\Big[\big|x(\tau_2) - \widehat{x}(\tau_2)\big| \le b\Big|\mathfrak{B}^1_{\boldsymbol{S}}\Big] \le \rho \sum_{\boldsymbol{S}} \boldsymbol{P}\Big[S^{\tau_1} = \boldsymbol{S}\Big|\big|x(\tau_1) - \widehat{x}(\tau_1)\big| \le b\Big] = \rho.\end{aligned}$$

Now we repeat the above arguments with respect to the tail on $t > \tau_2$ and conditioning given that $|x(\tau_1) - \widehat{x}(\tau_1)| \le b, |x(\tau_2) - \widehat{x}(\tau_2)| \le b$, $S_0^{\tau_2} = \boldsymbol{S}$. By continuing likewise, we get a sequence $0 < \tau_1 < \tau_2 < \cdots$ such that

$$\begin{aligned}p_{i+1|1,\dots,i} &:= \boldsymbol{P}\Big[\big|x(\tau_{i+1}) - \widehat{x}(\tau_{i+1})\big| \le b\Big|\big|x(\tau_1) - \widehat{x}(\tau_1)\big| \le b, \dots, \big|x(\tau_i) - \widehat{x}(\tau_i)\big| \le b\Big] \\ &\le \rho \qquad \forall i.\end{aligned}$$

Hence

$$\begin{aligned}\boldsymbol{P}\Big[\big|x(t) - \widehat{x}(t)\big| \le b \ \forall t\Big] &\le \boldsymbol{P}\Big[\big|x(\tau_i) - \widehat{x}(\tau_i)\big| \le b \ \forall i\Big] \\ &= \lim_{k\to\infty} \boldsymbol{P}\Big[\big|x(\tau_i) - \widehat{x}(\tau_i)\big| \le b \ \forall i = 1, \dots, k\Big] \\ &= \lim_{k\to\infty} \boldsymbol{P}\Big[\big|x(\tau_1) - \widehat{x}(\tau_1)\big| \le b\Big] \times \prod_{i=2}^{k} p_{i|1,\dots,i-1} \le \lim_{k\to\infty} \prod_{i=1}^{k} \rho \overset{\rho<1}{=\!=} 0\end{aligned}$$

in violation of (6.9). The contradiction obtained proves (5.1) for the detection problem.

To prove (5.2), we apply (5.1) to the process $x_*(t) := \alpha^{-t}x(t)$, $\widehat{x}_*(t) := \alpha^{-t}\widehat{x}(t)$, $e(t)$, $s(t)$. This is possible, since it is generated by (2.1), (2.2), and (2.3) (or (2.4)), where $A := \alpha^{-1}A$, $\mathfrak{X}_*[t,\cdot] := \alpha^{-t}\mathfrak{X}[t,\cdot]$, and $[y(0),\dots,y(t)]$ is replaced by $x_*(0)$, $\alpha x_*(1),\dots,\alpha^t x_*(t)$. The condition $\eta(\alpha^{-1}A) > \mathfrak{c}$ holds, since

$$\eta(\alpha^{-1}A) = \sum_{\lambda_j} \max\{\log_2(\alpha^{-1}|\lambda_j|),0\} = \sum_{\lambda_j}\big[\max\{\log_2|\lambda_j|,\log_2\alpha\} - \log_2\alpha\big]$$
$$\geq \sum_{\lambda_j} \max\{\log_2|\lambda_j|,0\} - n\log_2\alpha = \eta(A) - n\log_2\alpha > \mathfrak{c},$$

where $n = \dim x$ and the last inequality follows from the assumption $\log_2\alpha < \frac{\eta(A)-\mathfrak{c}}{n}$ of Proposition 5.3. □

Proof of the (d) $\Rightarrow \mathfrak{c} \geq \eta(A)$ *part of Theorem* 5.1. As was pointed out, this part of Theorem 5.1 follows from Proposition 5.3 by Remark 5.1.

7. Proofs of Propositions 5.3 and 5.4 for the stabilization problem and the (d) $\Rightarrow \mathfrak{c} \geq \eta(A)$ part of Theorem 5.2. These proofs result from the arguments from the previous section, along with the following simple observation, which is close to similar facts from [50, 59, 60, 62].

LEMMA 7.1. *Consider a coder* (2.3) *and decoder* (3.2). *Then there exist other coders and decoder-estimator*

$$e(t) = \mathfrak{E}_{un}[t, y_{un}(0),\dots,y_{un}(t), s(0),\dots,s(t-1)], \qquad \widehat{x}_{un}(t) = \mathfrak{X}\,[t, s(0), s(1),\dots,s(t)]$$

that generate an estimate $\widehat{x}_{un}(t)$ *of the state of the uncontrolled system* (2.1)

$$x_{un}(t+1) = Ax_{un}(t), \qquad x_{un}(0) = x_0, \qquad y_{un}(t) = Cx_{un}(t)$$

and produce the estimation error identical to the stabilization error of the original coder-decoder pair:

$$|x_{un}(t) - \widehat{x}_{un}(t)| = |x(t)|. \tag{7.1}$$

Proof. Let the new decoder generate the estimate via the recursion

$$\widehat{x}_{\rm un}(t+1) = A\widehat{x}_{\rm un}(t) - Bu(t), \quad u(t) := \mathfrak{U}\,[t, s(0), s(1),\dots,s(t)]\,, \quad \widehat{x}_{\rm un}(0) = 0,$$

where $\mathfrak{U}(\cdot)$ is taken from (3.2), and let the new coder be defined by the formula

$$e(t) := \mathfrak{E}[t, y_{\rm un}(0) - C\widehat{x}_{\rm un}(0),\dots,y_{\rm un}(t) - C\widehat{x}_{\rm un}(t), s(0),\dots,s(t-1)],$$

where $\mathfrak{E}(\cdot)$ is taken from (2.3). This formula presupposes that the coder also computes the estimate $\widehat{x}_{\rm un}(t)$.

Now we consider the process $\{x(t),u(t)\}_{t=0}^{\infty}$ generated in the system (3.1) by the original coder and decoder. Arguing by induction on t, it is easy to see that, first, both coder-decoder pairs give rise to common sequences $\{e(t)\},\{s(t)\},\{u(t)\}$, and, second, $y(t) = y_{\rm un}(t) - C\widehat{x}_{\rm un}(t)$ and (7.1) does hold. □

Modulo Lemma 7.1, Propositions 5.3 and 5.4 for the stabilization problem are immediate from the same propositions concerning the detection problem, whereas the (d) $\Rightarrow \mathfrak{c} \geq \eta(A)$ part of Theorem 5.2 follows from the matching part of Theorem 5.1.

8. Proof of Theorem 5.1. The necessity part of this theorem was justified in section 6. In this section, the focus is on proving that the inequality $\mathfrak{c} > \eta(A)$ is sufficient for almost sure observability.

From now until subsection 8.6, we consider the plants (2.1) with no stable modes. In this case, $\eta(A) = \log_2 |\det A|$. We start with some preliminaries.

8.1. Error exponents for discrete memoryless channels. We shall use the convenient notations $\lesssim$ and $\approx$ for inequality and equality up to a polynomial factor. In other words, $\varphi(m) \lesssim \psi(m) \Leftrightarrow \varphi(m) \le \psi(m)g(m)$ for all $m = 1, 2, \dots$, where $g(m)$ is a polynomial in m, and $\varphi(m) \approx \psi(m) \Leftrightarrow \varphi(m) \lesssim \psi(m)$ & $\psi(m) \lesssim \varphi(m)$. When $\varphi(m)$ and $\psi(m)$ depend on some other variables, the polynomial is assumed to be independent of them. The symbols $\mathcal{E}^m = \{\boldsymbol{e}\}$ and $\mathcal{S}^m = \{\boldsymbol{s}\}$ stand for the sets of all m-words over the input and output channel alphabets, respectively: $\boldsymbol{e} = (e_0, \dots, e_{m-1})$, $\boldsymbol{s} = (s_0, \dots, s_{m-1})$. The following result is straightforward from Lemma IV.1 and Theorem IV.1 in [8] (see also [14, 16]). We recall that $\mathfrak{c}$ is the capacity (4.4) of the channel.

THEOREM 8.1. *For any $0 < R < \mathfrak{c}$ and $m = 1, 2, \dots$, there exist $N \approx 2^{mR}$ input code words $\boldsymbol{e}^{[1]}, \dots, \boldsymbol{e}^{[N]} \in \mathcal{E}^m$ and a decoding rule $\mathfrak{D}_m : \mathcal{S}^m \to \{1, 2, \dots, N\}$ such that the maximum probability of error obeys the bound*

$$\max_{i=1,\dots,N} f_i \lesssim 2^{-mF(R,W)}. \tag{8.1}$$

Here $F(R, W) > 0$ does not depend on m (but depends on the rate R and the channel W) and

$$f_i = \boldsymbol{P}\Big[\mathfrak{D}_m(\boldsymbol{s}) \ne i \Big| \boldsymbol{e}^{[i]}\Big] := \sum_{\boldsymbol{s}:\mathfrak{D}_m(\boldsymbol{s})\ne i} \prod_{j=0}^{m-1} W(s_j|e_j^{[i]}), \qquad \boldsymbol{e}^{[i]} = (e_0^{[i]}, \dots, e_{m-1}^{[i]}),$$

is the probability of incorrect decoding, provided that the code word $\boldsymbol{e}^{[i]}$ was sent over the channel.

8.2. Contracted quantizers and detection in the case of errorless transmission. The facts presented in this subsection are mainly based on the ideas and results from [7, 40, 42, 53, 60, 62].

An N-level *quantizer* $\mathfrak{Q}$ in $\mathbb{R}^n$ is a partition of the unit ball B_0^1 with respect to some norm $|\cdot|$ in $\mathbb{R}^n$ into N disjoint sets $Q_1, \dots, Q_N \subset \mathbb{R}^n$, each equipped with a point $q_i \in Q_i$ called the *centroid* of Q_i. Such a quantizer associates any vector $x \in Q_i$ with its *quantized value* q_i and any vector outside B_0^1 with an alarm symbol ✠.

DEFINITION 8.2. *The quantizer $\mathfrak{Q}$ is said to be* m-contracted *($m = 1, 2, \dots$) for the system* (2.1) *if*

$$A^m (Q_i - q_i) \subset \rho_{\mathfrak{Q}} B_0^1 \qquad \forall i = 1, \dots, N, \tag{8.2}$$

where $\rho_{\mathfrak{Q}} \in (0, 1)$ is called the contraction rate.

The role of such quantizers in the solution of the estimation problem is revealed by the following lemma.

LEMMA 8.3. *Suppose that the following statements hold:*

(i) *at a time instant $t = t_*$, an estimate $\widehat{x}(t)$ and its exactness $\delta = \delta(t) \ge |\widehat{x}(t) - x(t)|$ are known at both the coder and decoder sites;*

(ii) *an N-level m-contracted quantizer is given;*

(iii) *there exist ways to both encode N distinct messages for transmission over the channel within the subsequent time interval of duration m and decode the received data by time $t_* + m$ so that the overall transmission is errorless.*

Then the estimate $\widehat{x}(t)$ for $t = t_*, \dots, t_* + m$ can be constructed so that

$$\begin{aligned} &|\widehat{x}(t) - x(t)| \le \rho_{\mathfrak{Q}}\delta \quad \text{for } t = t_* + m \\ \text{and} \quad &|\widehat{x}(t) - x(t)| \le \|A\|^{t-t_*}\delta \quad \text{for} \quad t = t_*, \dots, t_* + m - 1. \end{aligned} \tag{8.3}$$

Proof. We apply the quantizer to the scaled error $\delta^{-1}[x(t) - \widehat{x}(t)]$ and, by employing (iii), make the decoder aware of the corresponding quantized value q_i by time $t_* + m$. The estimate is defined by

$$\widehat{x}(t) := A^{t-t_*}\widehat{x}(t_*) \text{ for } t = t_*, \dots, t_* + m - 1 \quad \text{and} \quad \widehat{x}(t_* + m) := A^m[\widehat{x}(t_*) + \delta q_i].$$

Then for $t = t_*, \dots, t_* + m - 1$, it follows from (2.1) that

$$|\widehat{x}(t) - x(t)| = |A^{t-t_*}\widehat{x}(t_*) - A^{t-t_*}x(t_*)| \le \|A\|^{t-t_*}|\widehat{x}(t_*) - x(t_*)| \le \|A\|^{t-t_*}\delta;$$

i.e. the second relation from (8.3) holds. Furthermore,

$$\begin{aligned} |\widehat{x}(t_* + m) - x(t_* + m)| &= \left|A^m[\widehat{x}(t_*) + \delta q_i] - A^m x(t_*)\right| \\ &= \delta\Big|A^m\Big\{\underbrace{\delta^{-1}[x(t_*) - \widehat{x}(t_*)]}_{\in Q_i} - q_i\Big\}\Big| \overset{(8.2)}{\le} \rho_{\mathfrak{Q}}\delta. \end{aligned}$$

□

A coder-decoder pair tracking the state in the case of errorless transmission. Let (iii) be true at any time $t = t_*$. Then the scheme from the above proof can be successively repeated. In doing so, we put $\delta(t_* + m) := \rho_{\mathfrak{Q}}\delta(t_*)$ in accordance with (8.3). (We assume that computations of $\widehat{x}(t)$ and $\delta(t)$ are performed by both the decoder and the coder. So (i) holds for $t = t_* + m$ whenever it is true for $t = t_*$.) This ensures tracking, since (8.3) $\Rightarrow$ (2.5).

This conclusion tacitly assumes that (i) does hold for some t_*. To ensure this, suppose that (iii) is true not only at any time but also with an increased number of messages $N := N + 1$. (One message is reserved to carry the alarm signal ✠.) Let the coder and decoder be given common initial estimate $\widehat{x}(0)$ and its exactness $\delta(0)$. Whenever $\Delta(t_*) := |x(t_*) - \widehat{x}(t_*)| \le \delta(t_*)$, the above algorithm is applied. Otherwise, the signal ✠ is sent over the channel. On its decoding at time $t_* + m$, the value of δ is increased, $\delta(t_* + m) := \gamma\delta(t_*)$, where $\gamma > 1$ is a given coefficient. In this case, the estimate is defined by $\widehat{x}(t+1) := A\widehat{x}(t)$ for $t = t_* + m - 1$. Then we put $t_* := t_* + m$ and repeat the described operations. Note that

$$\begin{aligned} \Delta(t_*) > \delta(t_*) \Rightarrow \Delta(t_* + m) &= |A^m[\widehat{x}(t_*) - x(t_*)]| \\ &\le \|A\|^m\Delta(t_*) \quad \text{and} \quad \delta(\mathrm{t}_* + \mathrm{m}) = \gamma\delta(\mathrm{t}_*). \end{aligned}$$

So until (i) holds, the error $\Delta(jm)$ increases no faster than $\|A\|^{jm}$, whereas $\delta(jm) = \gamma^j\delta(0)$. Now pick $\gamma > \|A\|^m$. Then necessarily $\Delta(jm) \le \delta(jm)$ for large j; i.e., (i) does hold sooner or later.

Thus a contracted quantizer is a gate to an observer successfully tracking the state.

A problem with the construction of such an observer along the above lines is that assumptions (ii) and (iii) imply converse requirements to the parameters N and m.

Indeed, (iii) presupposes that the number N cannot be large given m. (In the case of the perfect channel, $N \le |\mathcal{E}|^m$.) On contrary, (ii) means that the number N of quantizer levels must be large enough: $N > 2^{m\eta(A)}$. (Indeed, let **mes** denote the Lebesgue measure. Then (8.2) implies $|\det A|^m \,\mathbf{mes}\,(Q_i) \le \rho_{\mathfrak{Q}}^n \,\mathbf{mes}\, B_0^1$. Summing over i gives $|\det A|^m \,\mathbf{mes}\, B_0^1 \le N\rho_{\mathfrak{Q}}^n \,\mathbf{mes}\, B_0^1 \Rightarrow N > 2^{m\eta(A)}$.)

To look for a trade-off between the above converse requirements, it is important to know whether the bound $N > 2^{m\eta(A)}$ is tight. In other words, is it sufficient for the existence of a contracted quantizer? The next theorem shows that the answer is in a sense affirmative.

THEOREM 8.4. *For any square matrix A with no stable modes and $m = 1, 2, \dots$, there exists an m-contracted quantizer with $N \approx 2^{m\eta(A)}$ levels.*

Sketch of proof. We restrict ourselves to only a sketch, since the claim can be derived from the arguments scattered over [7, 40, 42, 53, 60, 62]. Note first that whenever the statement is true for two matrices A_1 and A_2, it is also true for the block matrix $\left(\begin{smallmatrix} A_1 & 0 \\ 0 & A_2 \end{smallmatrix}\right)$. By employing the canonical Jordan form of A as in [60, 62], this reduces the proof to the case where the matrix is a real Jordan block. Let s denote its size, λ its eigenvalue, and $\omega := |\lambda|$. As follows from, e.g., [64, Lemma 3.1, p. 64], $\Xi(m) := \omega^{-m} g(m)^{-1} A^m \to 0$ as $m \to \infty$ for some polynomial $g(\cdot)$. So $\|\Xi(m)\| < \rho < 1$ for $m \approx \infty$. Here $\|\cdot\|$ is the operator norm associated with the norm $|z| := \max_i |z_i|$ in $\mathbb{R}^s = \{z = (z_1, \dots, z_s)\}$. Multiplying the polynomial $g(m)$ by a sufficiently large scalar factor makes the inequality $\|\Xi(m)\| < \rho$ true for all m. Now consider the uniform quantizer $\mathfrak{Q}$ partitioning the unit ball B_0^1 into $N := k^s$ balls Q_i of radius $\frac{1}{k}$. Here $k := \lceil \omega^m g(m) \rceil$ and $\lceil d \rceil$ is the minimum integer exceeding d. The centroid q_i is the center of the ball Q_i. Then $\|\Xi(m)\| < \rho \Rightarrow \Xi(m)[Q_i - q_i] \subset \rho[Q_i - q_i] = \frac{\rho}{k} B_0^1 \Rightarrow A^m[Q_i - q_i] \subset \rho \frac{\omega^m g(m)}{k} B_0^1 \subset \rho B_0^1$. Thus the quantizer is m-contracted. It remains to note that $N = k^s \approx \omega^{sm} = |\det A|^m = 2^{m\eta(A)}$. □

The contraction rate ρ of the proposed quantizer does not depend on m. However, this rate can be made geometrically decreasing in m, provided the number of levels N is slightly increased.

LEMMA 8.5. *Suppose that a square matrix A with no stable modes and $\eta > \eta(A)$ are given. Then for any $m = 1, 2, \dots$, there exists an m-contracted quantizer with the contraction rate $\varkappa^{2m}$ and $N \lesssim 2^{m\eta}$ levels, where $\varkappa = \varkappa_{\eta,A} \in (0, 1)$ does not depend on m.*

Sketch of proof. In the above sketched proof of Theorem 8.4, one must alter the choice of k by $k := \lceil \alpha^m \omega^m g(m) \rceil$. Here $\alpha > 1$ is a parameter to be adjusted. This evidently provides the rate of contraction $\rho_{\mathfrak{Q}} \le \rho\alpha^{-m} \le \alpha^{-m}$ and gives rise to a quantizer with $N \approx \alpha^{sm}\omega^{sm} = 2^{m[\eta(A) + s \log_2 \alpha]}$ levels. Then the statement of the lemma results from properly adjusting the value of $\alpha > 1$. □

8.3. A coder-decoder pair without a feedback for a noisy channel. In this subsection, we assume that $\mathfrak{c} > \eta(A)$. The observer to be constructed resembles that from the previous subsection. The difference is that now employed are code words and quantizers with increasing length m_i and number of levels, respectively. So the observer operation is composed of time cycles $[\tau_i : \tau_{i+1}]$ of increasing durations $\tau_{i+1} - \tau_i = m_i$.

To construct an observer, we pick

1. two numbers η and R such that $\eta(A) < \eta < R < \mathfrak{c}$; and then for any $m = 1, 2, \dots$, we choose
2. a set $\mathfrak{E}^{[m]} \subset \mathcal{E}^m$ of $N = N'_m \approx 2^{mR}$ input code words, each of length m, and a decoding rule $\mathfrak{D}_m$ with the properties described in Theorem 8.1; and

3. an m-contracted quantizer $\mathfrak{Q}_m$ described in Lemma 8.5.

REMARK 8.1. *The quantizer outputs including the alarm signal* ✠ *can be encoded by the code words from the set* $\boldsymbol{\mathcal{E}}^{[m]}$*, provided that* m *is large enough.*

Indeed, let N''_m denote the number of the quantizer levels. Then

$$N'_m \approx 2^{mR} \quad \text{and} \quad \mathrm{N}''_{\mathrm{m}} \lesssim 2^{\mathrm{m}\eta} \quad \text{and} \quad \eta < \mathrm{R} \Rightarrow \mathrm{N}''_{\mathrm{m}} + 1 \le \mathrm{N}''_{\mathrm{m}} \qquad \forall \mathrm{m} \ge \mathrm{m}_*,$$

where m_* is large enough.

Finally, we pick $\gamma > \|A\|$ and consider the sequence of integers

$$m_i := i + m_0, \qquad i = 0, 1, \dots, \tag{8.4}$$

where $m_0 \ge n, m_*$ is an integer parameter of the observer.

Operation of the observer. Both the coder and the decoder compute their own estimates $\widehat{x}_c(t)$, $\widehat{x}_d(t)$ and bounds for the estimate exactness $\delta_c(t)$, $\delta_d(t)$, respectively. Initially, they are given common and arbitrarily chosen values of $\widehat{x}_c(0) = \widehat{x}_d(0) = \widehat{x}_0$ and $\delta_c(0) = \delta_d(0) = \delta_0 > 0$. (The inequality $\delta_0 \ge |\widehat{x}_0 - x(0)|$ may be violated.)

At any time t, both the coder and the decoder compute the next estimates and the bounds by the formulas

(8.5)
$$\widehat{x}_c(t{+}1) := A\widehat{x}_c(t), \qquad \widehat{x}_d(t{+}1) := A\widehat{x}_d(t), \qquad \delta_c(t{+}1) := \delta_c(t), \qquad \delta_d(t{+}1) := \delta_d(t).$$

However, at times $t = \tau_i$, where

$$\tau_i := m_0 + \dots + m_{i-1} = i \cdot m_0 + \frac{i(i-1)}{2}, \tag{8.6}$$

they preface these computations by the following operations.

The coder (at times $t = \tau_i$, $i = 1, 2, \dots$) does the following:

c.1. proceeding from the previous measurements calculates the current state $x(\tau_i)$;

c.2. employs the quantizer $\mathfrak{Q}_{m_i}$ and computes the quantized value $q_c(\tau_i)$ of the current scaled estimation error

$$\varepsilon(\tau_i) := \big[\delta_c(\tau_i)\big]^{-1}\big[x(\tau_i) - \widehat{x}_c(\tau_i)\big] \tag{8.7}$$

(we recall that the quantized value of any vector outside the unit ball is the alarm symbol $q_c(\tau_i) = ✠$);

c.3. encodes the quantized value $q_c(\tau_i)$ by means of the code book $\boldsymbol{\mathcal{E}}^{[m_i]}$, and thus obtained code word of length m_i is transmitted over the channel during the next operation cycle $[\tau_i : \tau_{i+1})$;

c.4. corrects the estimate and then the exactness bound

(8.8)
$$\widehat{x}_c(\tau_i) := \widehat{x}_c(\tau_i) + \delta_c(\tau_i)\, \overset{\star}{q}_c\,(\tau_i), \quad \delta_c(\tau_i) := \delta_c(\tau_i) \times \Big(\langle q_c(\tau_i)\rangle_{\varkappa,\gamma}\Big)^{m_i}, \quad \text{where}$$
$$\overset{\star}{q} := \begin{cases} q & \text{if } q \ne ✠, \\ 0 & \text{otherwise,} \end{cases} \qquad \langle q\rangle_{\varkappa,\gamma} := \begin{cases} \varkappa & \text{if } q \ne ✠, \\ \gamma & \text{otherwise,} \end{cases}$$

and $\varkappa \in (0, 1)$ is the parameter from Lemma 8.5.

Only after this does the coder perform the computations in accordance with (8.5). Note that step c.1 is possible, since the system (2.1) (with no stable modes) is observable thanks to Assumption 4.4.

The decoder (at times $t = \tau_i, i = 2, 3, \dots$) does the following:

d.1. applies the decoding rule $\mathfrak{D}_{m_{i-1}}$ to the data received within the previous operation cycle $[\tau_{i-1} : \tau_i)$ and thus computes the decoded value $q_d(\tau_i)$ of the quantized and scaled estimation error $q_c(\tau_{i-1})$ (this value may be incorrect due to transmission errors);

d.2. corrects successively the estimate and the exactness bound (8.9)

$$\widehat{x}_d(\tau_i) := \widehat{x}_d(\tau_i) + \delta_d(\tau_i) A^{m_{i-1}} \overset{\star}{q}_d(\tau_i), \quad \delta_d(\tau_i) := \delta_d(\tau_i) \times \Big(\langle q_d(\tau_i) \rangle_{\varkappa,\gamma} \Big)^{m_{i-1}}.$$

Only after this does it perform the computations from (8.5).

The coder can alter instantaneous multiplication of $\delta_c(\tau_i)$ by $\varkappa^{m_i}$ or γ^{m_i} at the time τ_i with keeping $\delta_c(t)$ constant during the next operation cycle $[\tau_i : \tau_{i+1})$ by multiplying by $\varkappa$ or γ, respectively, at each step of this cycle. Likewise, computing the large power $A^{m_{i-1}}$ employed in (8.9) can be distributed over the cycle $[\tau_{i-1} : \tau_i]$. This hint cannot be directly applied to computing $\delta_d(t)$, since the decoder becomes aware of the multiplier ($\varkappa$ or γ) only at the end of the current cycle $[\tau_{i-1} : \tau_i]$. However, one can perform both the computations and then choose the correct quantity ($\delta_d(\tau_{i-1})\varkappa^{m_{i-1}}$ or $\delta_d(\tau_{i-1})\gamma^{m_{i-1}}$) at the end τ_i of the cycle.

8.4. Tracking with arbitrarily large probability. Now we show that the coder-decoder pair constructed in the previous subsection tracks the state (2.5) with as large a probability as desired. More precisely, it does so, provided the parameter m_0 from (8.4) is chosen properly. (This choice depends on the desired probability.)

The values of the quantities $\widehat{x}_d, \widehat{x}_c, \delta_c, \delta_d$ before and after the update at time τ_i are marked by $^-$ and $^+$, respectively . We start with the following key fact.

LEMMA 8.6. *In any event where the decoder always decodes the data correctly* $q_d(\tau_i) = q_c(\tau_{i-1})$ *for all* $i \geq 2$*, the coder-decoder pair ensures asymptotic tracking* (2.5). *Furthermore,*

$$\big| x(\tau_i) - \widehat{x}_c^-(\tau_i) \big| \leq k \varkappa^{\tau_i}, \qquad i = 1, 2, \ldots, \tag{8.10}$$

where the constant k *does not depend on* i *(but may depend on the event).*

Proof. We start with showing that there exists an index $i = 1, 2, \ldots$ for which

$$|\varepsilon(\tau_i)| \leq 1, \tag{8.11}$$

where the scaled error $\varepsilon(\tau_i)$ is defined in (8.7). Indeed, otherwise, $q_c(\tau_i) = \maltese$, $\delta_c^-(\tau_{i+1}) = \delta_c^-(\tau_i)\gamma^{m_i}$ for all $i \geq 1$ and $\widehat{x}_c(t+1) = A\widehat{x}_c(t)$ for all t. So for $i \geq 2$, we have

$$\begin{aligned} |\varepsilon(\tau_i)| &= \Big[\gamma^{\sum_{j=1}^{i-1} m_j} \delta_0 \Big]^{-1} \Big| A^{\tau_i} \big[x_0 - \widehat{x}_0 \big] \Big| \overset{(8.6)}{=\!=} \frac{\gamma^{m_0}}{\delta_0} \gamma^{-\tau_i} \Big| A^{\tau_i} \big[x_0 - \widehat{x}_0 \big] \Big| \\ &\leq \left(\frac{\|A\|}{\gamma} \right)^{\tau_i} \gamma^{m_0} \frac{|x_0 - \widehat{x}_0|}{\delta_0} \xrightarrow{\gamma > \|A\|} 0 \end{aligned}$$

as $i \to \infty$ in violation of the hypothesis $\varepsilon(\tau_i) > 1$ for all i. Thus (8.11) does hold for some i.

Now consider an index i such that (8.11) holds. Then (8.11) is still true for

$i := i+1$. Indeed,

$$|\varepsilon(\tau_{i+1})| \overset{(8.7)}{=\!=} \big[\delta_c^-(\tau_{i+1})\big]^{-1}\big|x(\tau_{i+1}) - \widehat{x}_c^-(\tau_{i+1})\big| \overset{(2.1),(8.6),(8.8)}{=\!=\!=} \varkappa^{-m_i}\big[\delta_c^-(\tau_i)\big]^{-1} \times \Big|A^{m_i}x(\tau_i) - A^{m_i}\big[\widehat{x}_c^-(\tau_i) + \delta_c^-(\tau_i)q_c(\tau_i)\big]\Big| = \varkappa^{-m_i}\Big|A^{m_i}\Big\{\underbrace{\delta_c^-(\tau_i)\big]^{-1}\big[x(\tau_i) - \widehat{x}_c^-(\tau_i)\big]}_{v} - q_c(\tau_i)\Big\}\Big|.$$

Here $q_c(\tau_i)$ is the quantized value of the vector v, and an m_i-contracted quantizer with the contraction rate $\rho_{\mathfrak{Q}} = \varkappa^{2m_i}$ is applied. So (8.2) yields

$$|\varepsilon(\tau_{i+1})| \le \varkappa^{m_i} < 1; \tag{8.12}$$

i.e., (8.11) does hold for $i := i+1$.

It follows that (8.11) is true for all $i \ge \overline{i}$, where $\overline{i}$ is large enough. Hence (8.8) yields

$$\delta_c^-(\tau_i) = \delta_c^-(\tau_{\overline{i}})\varkappa^{\sum_{j=\overline{i}}^{i-1} m_j}.$$

We proceed by taking into account (8.7) and (8.11):

$$\big|x(\tau_i) - \widehat{x}_c^-(\tau_i)\big| \le \delta_c^-(\tau_i) = \overline{\delta}\varkappa^{\sum_{j=0}^{i-1} m_j} \overset{(8.6)}{=\!=} \overline{\delta}\varkappa^{\tau_i}, \qquad \text{where} \quad \overline{\delta} := \delta_c(\tau_{\overline{i}})\varkappa^{-\sum_{j=0}^{\overline{i}-1} m_j}.$$

This evidently implies (8.10) and shows that the coder tracks the state. As for the decoder, note that

$$\widehat{x}_d^+(\tau_i) = \widehat{x}_c^-(\tau_i), \qquad i = 1, 2, \dots . \tag{8.13}$$

Indeed, for $i = 1$, this relation is evident. Suppose that this relation is true for some $i \ge 1$. Due to the absence of transmission errors, $\delta_d^{\pm}(\tau_j) = \delta_c^{\pm}(\tau_{j-1})$, $j = 2, 3, \dots$. So

$$\widehat{x}_d^+(\tau_{i+1}) \overset{(8.9)}{=\!=} \widehat{x}_d^-(\tau_{i+1}) + \delta_d^-(\tau_{i+1})A^{m_i} \overset{\star}{q}_d(\tau_{i+1}) \overset{(8.5)}{=\!=} A^{m_i}\widehat{x}_d^+(\tau_i) + \delta_c^-(\tau_i)A^{m_i} \overset{\star}{q}_c(\tau_i) \overset{(8.13)}{=\!=} A^{m_i}\Big[\widehat{x}_c^-(\tau_i) + \delta_c^-(\tau_i) \overset{\star}{q}_c(\tau_i)\Big] \overset{(8.8)}{=\!=} A^{m_i}\widehat{x}_c^+(\tau_i) \overset{(8.5)}{=\!=} \widehat{x}_c^-(\tau_{i+1});$$

i.e., (8.13) holds for $i := i+1$. Thus this relation is true for all $i \ge 1$.

Whenever $\tau_i < t \le \tau_{i+1}$, we have by (8.5)

$$\big|x(t) - \widehat{x}_d(t)\big| = \Big|A^{t-\tau_i}\big[x(\tau_i) - \widehat{x}_d^+(\tau_i)\big]\Big| \overset{(8.13)}{\le} \|A\|^{t-\tau_i}\big|x(\tau_i) - \widehat{x}_c^-(\tau_i)\big|, \tag{8.14}$$

$$\max_{\tau_i < t \le \tau_{i+1}} \big|x(t) - \widehat{x}_d(t)\big| \overset{(8.10)}{\le} k\|A\|^{m_i}\varkappa^{\tau_i} = k2^{m_i \log_2 \|A\| + \log_2 \varkappa \tau_i} \overset{(8.4),(8.6)}{=\!=\!=} k2^{(i+m_0)\log_2 \|A\| + \left[i\cdot m_0 + \frac{i(i-1)}{2}\right]\log_2 \varkappa}.$$

So far as $\log_2 \varkappa < 0$, this maximum converges to 0 as $i \to \infty$; i.e., (2.5) does hold with $\widehat{x}(t) := \widehat{x}_d(t)$. □

Now we show that the assumption of Lemma 8.6 holds with large probability, provided that the parameter m_0 in (8.4) is chosen large.

LEMMA 8.7. *The probability* $\boldsymbol{p}_{err}$ *that the decoder decodes at least one message incorrectly does not exceed*

$$\boldsymbol{p}_{err} \le K_{R,W,F} 2^{-m_0 F}.$$

Here the constant $K_{R,W,F}$ *does not depend on* m_0*, and the inequality holds with any* $F \in (0, F(R,W))$*, where* $F(R,W)$ *is taken from* (8.1).

Proof. Denote by $\boldsymbol{e}(i)$ and $\boldsymbol{s}(i)$ the messages of length m_{i-1} formed by the coder at time τ_{i-1} and received by the decoder at time τ_i, respectively. For simplicity of notation, we assume that the map $\mathfrak{D}_m$ from Theorem 8.1 takes values in the set $\mathcal{E}^{[m]}$ of input code words. The symbol $\boldsymbol{p}_{\rm err}(i)$ stands for the probability that the decoding of $\boldsymbol{s}(i)$ is wrong: $\boldsymbol{p}_{\rm err}(i) = \boldsymbol{P}\{\mathfrak{D}_{m_j}[\boldsymbol{s}(j)] \ne \boldsymbol{e}(j)\}$. Since the estimate (8.1) implies $\max_i f_i \le c_{R,W,F} q^{-mF}$, we have

$$\begin{aligned}
\boldsymbol{p}_{\rm err}(i) &= \sum_{\boldsymbol{e}\in\mathcal{E}^{[m_{i-1}]}} \boldsymbol{P}\big[\boldsymbol{e}(i)=\boldsymbol{e}\big]\boldsymbol{P}\Big\{\mathfrak{D}_{m_{i-1}}[\boldsymbol{s}(i)] \ne \boldsymbol{e}(i)\Big|\boldsymbol{e}(i)=\boldsymbol{e}\Big\} \\
&\le c_{R,W,F} \sum_{\boldsymbol{e}\in\mathcal{E}^{[m_{i-1}]}} \boldsymbol{P}\big[\boldsymbol{e}(i)=\boldsymbol{e}\big] 2^{-m_{i-1}F} \overset{(8.4)}{=\!=} c_{R,W,F} 2^{-(i-1+m_0)F}; \\
\boldsymbol{p}_{\rm err} &\le \sum_{i=1}^{\infty} \boldsymbol{p}_{\rm err}(i+1) \le c_{R,W,F} \sum_{i=1}^{\infty} 2^{-(i+m_0)F} = \frac{c_{R,W,F}}{2^F-1} 2^{-m_0 F}. \qquad \square
\end{aligned}$$

As was shown in [50], the probability of error cannot be made small when stationary fixed length block coding-decoding schemes are employed.

By combining Lemmas 8.6 and 8.7, we arrive at the following conclusion.

COROLLARY 8.8. *Suppose that the system* (2.1) *has no stable modes and* $\eta(A) < \mathfrak{c}$*. Then statement* (b) *from Theorem* 5.1 *holds.*

8.5. Tracking almost surely by means of fixed length code words. The observer from the previous subsection employs code words whose lengths increase as the estimation process progresses. So the memories of the coder and decoder should increase accordingly.[3] In this subsection, we show that almost sure asymptotic state tracking can be achieved on the basis of fixed length code words whenever a communication feedback is available. In doing so, we still consider the system (2.1) with no stable modes. Extensions on systems with both stable and unstable modes will be given in subsection 8.6.

The almost sure tracking coder-decoder pair is that from subsection 8.3 modified as follows:

(i) The operation cycles are of equal and fixed duration m_0, i.e., (8.4) and (8.6) are replaced by $m_i = m_0$ and $\tau_i := i m_0$, respectively.

(ii) Instead of forming its own sequences of state estimates $\{\widehat{x}_c(t)\}$ and exactness bounds $\{\delta_c(t)\}$, the coder duplicates those generated by the decoder.

To accomplish (ii), the coder must be aware of the results $\boldsymbol{s}(i)$ of the transmissions across the channel. This becomes possible thanks to the communication feedback. Specifically, now the coder operates as follows.

At times $t = \tau_i$, $i = 1, 2, \dots$, the *coder* prefaces (8.5) by the following actions:

- It carries out step c.1 of the previous coder (see subsection 8.3). Then it in fact duplicates steps d.1 and d.2 of the decoder, i.e., for $t = \tau_i$, $i = 2, 3, \dots$.

[3] The same feature is characteristic of the anytime coding-decoding schemes considered in [50].

- The coder applies the decoding rule $\mathfrak{D}_{m_0}$ to the data received by the decoder within the previous operation cycle $[\tau_{i-1} : \tau_i)$ and thus gets $q_d(\tau_i)$.
- It corrects $\widehat{x}_c(\tau_i)$ and $\delta_c(\tau_i)$ in accordance with the formulas (8.15)

$$\widehat{x}_c(\tau_i) := \widehat{x}_c(\tau_i) + \delta_c(\tau_i) A^{m_0} \overset{\star}{q}_d(\tau_i), \qquad \delta_c(\tau_i) := \delta_c(\tau_i) \left(\langle q_d(\tau_i) \rangle_{\varkappa,\gamma} \right)^{m_0}.$$

- After this steps c.2 and c.3 of the previous coder are performed if $i = 1, 2, \ldots$.

For technical convenience, we put $q_c(\tau_0) := q_d(\tau_1) := ✠$ and suppose that at times $t = \tau_0, \tau_1$ the coder and decoder act accordingly. By comparing (8.9) and (8.15), we see that $\widehat{x}_c^{\pm}(\tau_i) = \widehat{x}_d^{\pm}(\tau_i)$ and $\delta_c^{\pm}(\tau_i) = \delta_d^{\pm}(\tau_i)$.

The main result of this subsection is as follows.

PROPOSITION 8.9. *Suppose that Assumptions* 4.1–4.4 *hold, the system* (2.1) *has no stable modes, and* $\mathfrak{c} > \eta(A)$*, where* $\mathfrak{c}$ *and* $\eta(A)$ *are taken from Theorem* 5.1*. Then the modified coder-decoder pair detects the state (i.e.,* (2.5) *holds) almost surely, provided the duration of the operation cycle is large enough:* $m_0 \geq M(A, B, \varkappa, \gamma, W, R)$.

An explicit formula for $M(A, B, \varkappa, \gamma, W, R)$ can be derived from the proof of this proposition.

COROLLARY 8.10. *Suppose that the system* (2.1) *has no stable modes and* $\eta(A) < \mathfrak{c}$*. Then statement* (a) *from Theorem* 5.1 *holds.*

In the remainder of the subsection, we prove Proposition 8.9; so its hypotheses are assumed to hold. At first, we informally discuss the idea of the proof. The operation cycle $[\tau_{i-1} : \tau_i)$ is said to be *regular* if during it a message different from the alarm one is correctly transmitted from the coder to the decoder $q_d(\tau_i) = q_c(\tau_{i-1}) \neq ✠$ and $\delta_d^+(\tau_{i-1})$ is a true bound for the estimation error: $\delta_d^+(\tau_{i-1}) \geq |x(\tau_{i-1}) - \widehat{x}_d^+(\tau_{i-1})|$. Then the update (8.9) at time $t = \tau_i$ improves the error bound via multiplying by $\varkappa^{m_0} < 1$, while keeping it correct for the updated estimate, which can be proved similarly to (8.12). However, a cycle is not necessarily regular. First, the initial bound δ_0 may be incorrect. This is a weak reason, since the algorithm would make the bound correct for a finite time in the absence of decoding errors at step d.1 (see the proof of Lemma 8.6). Second, the cycle may be irregular due to such errors. Any of them may make not only the current cycle irregular but also launch a whole "tail" of irregular cycles even if the messages transmitted across the channel during the subsequent cycles were decoded correctly. This holds if the transmission error makes the upper bound δ incorrect. During this tail, the error bound would increase via multiplying by $\gamma^{m_0} > 1$ in order to become correct once more. So any error has an aftereffect, which evidently remains true in the real circumstances where the subsequent cycles are not necessarily "errorless." A priori, it is not even clear that the chain of consecutive irregular cycles will be terminated and a regular one will occur.

The proof is based on the fact that the probability of the decoding errors can be made as small as desired by properly picking m_0. By the strong law of large numbers, this entails that the average frequency of the decoding errors is small almost surely. In other words, the errors are rarely encountered. The next step is to evaluate the duration of the aftereffect of each such error and to show that the average frequency ω_{irr} of the irregular cycles does not exceed the average frequency of the above errors multiplied by a fixed factor. So ω_{irr} is also small. Hence not only do regular cycles follow any irregular one but also the average frequency of regular cycles $\omega_{reg} \gg \omega_{irr}$. By taking into account that at any irregular cycle the bound δ_d is increased at most by multiplying by γ^{m_0}, we conclude that (approximately) $\delta_d^-(\tau_i) \leq \delta_0 \varkappa^{im_0\omega_{reg}} \gamma^{im_0\omega_{irr}} \to 0$ as $i \to \infty$. This convergence is extended on the estimation error on the ground that

$\delta_d^-(\tau_i)$ is the correct bound for this error for most of the i's.

To carry out the first step of this plan, we need the following variant of the strong law of large numbers [28, sect. 32, p. 53] (see also [46, 49]).

THEOREM 8.11. *Suppose that $\mathcal{F}_i$ is a flow of nondecreasing σ-algebras in a probability space, the random variable ξ_i is $\mathcal{F}_i$-measurable, and $b_i \uparrow \infty$, $b_i > 0$, $i = 1, 2, \ldots$. Suppose also that $\boldsymbol{E}\big|\xi_i - \boldsymbol{E}(\xi_i|\mathcal{F}_{i-1})\big| < \infty$ and*

$$\sum_{i=1}^{\infty} \frac{1}{b_i^2} \boldsymbol{E}\Big\{ \big[\xi_i - \boldsymbol{E}(\xi_i|\mathcal{F}_{i-1})\big]^2 \Big\} < \infty. \tag{8.16}$$

Then with probability 1,

$$\frac{1}{b_r} \sum_{i=1}^{r} \big[\xi_i - \boldsymbol{E}(\xi_i|\mathcal{F}_{i-1})\big] \to 0 \quad \text{as} \quad r \to \infty. \tag{8.17}$$

Now we consider the stochastic process generated by the coder-decoder pair from Proposition 8.9. The symbols $\boldsymbol{e}(i)$ and $\boldsymbol{s}(i)$ stand for the messages formed by the coder at time τ_{i-1} and received by the decoder at time τ_i, respectively. We also introduce the error indicator function:

$$I^{\mathrm{err}}(i) := 1 \quad \text{if} \quad \mathfrak{D}_{m_0}[\boldsymbol{s}(i)] \neq \boldsymbol{e}(i),\ i \geq 2, \qquad \text{and} \quad I^{\mathrm{err}}(i) := 0 \quad \text{otherwise.} \tag{8.18}$$

LEMMA 8.12. *We pick $0 < F < F(R, W)$, where $F(R, W)$ is taken from* (8.1). *Then the following relation holds almost surely, provided that m_0 is sufficiently large:*

$$\overline{\lim_{r \to \infty}} \frac{1}{r} \sum_{i=1}^{r} I^{err}(i) \leq 2^{-Fm_0}. \tag{8.19}$$

Proof. We are going to apply Theorem 8.11 to $\xi_i := I^{\mathrm{err}}(i)$ and $b(i) := i$. The σ-algebra $\mathcal{F}_i$ is taken to be that generated by the random quantities $x_0, \boldsymbol{s}(0), \ldots, \boldsymbol{s}(i)$. Due to the construction of the coder, $\boldsymbol{e}(i) = \mathfrak{E}_*[i, x_0, \boldsymbol{s}(0), \ldots, \boldsymbol{s}(i-1)]$, where $\mathfrak{E}_*(\cdot)$ is a deterministic function. It follows that $I^{\mathrm{err}}(i)$ is $\mathcal{F}_i$-measurable. Furthermore, $0 \leq I^{\mathrm{err}}(i) \leq 1 \Rightarrow 0 \leq \boldsymbol{E}[I^{\mathrm{err}}(i)|\mathcal{F}_{i-1}] \leq 1$ almost surely, which implies (8.16). So by Theorem 8.11,

$$\frac{1}{r} \sum_{i=1}^{r} \Big\{ I^{\mathrm{err}}(i) - \boldsymbol{E}[I^{\mathrm{err}}(i)|\mathcal{F}_{i-1}] \Big\} = 0 \quad \text{a.s.} \tag{8.20}$$

Now we are going to estimate $\boldsymbol{E}[I^{\mathrm{err}}(i)|\mathcal{F}_{i-1}]$. By invoking Assumptions 4.1, 4.2, and that $\boldsymbol{e}(i) = \mathfrak{E}_*[i, x_0, \boldsymbol{s}(0), \ldots, \boldsymbol{s}(i-1)]$, we get

$$\begin{aligned}
\boldsymbol{E}[I^{\mathrm{err}}(i)|\mathcal{F}_{i-1}] &= \boldsymbol{E}\Big\{ I^{\mathrm{err}}(i) \big| x_0, \boldsymbol{s}(0), \ldots, \boldsymbol{s}(i-1), \boldsymbol{e}(i) \Big\} \\
&= \boldsymbol{P}\big[\mathfrak{D}_{m_0}[\boldsymbol{s}(i)] \neq \boldsymbol{e}(i) \big| x_0, \boldsymbol{s}(0), \ldots, \boldsymbol{s}(i-1), \boldsymbol{e}(i) \big] \\
&= \sum_{\boldsymbol{\epsilon}} \boldsymbol{P}\big[\mathfrak{D}_{m_0}[\boldsymbol{s}(i)] \neq \boldsymbol{\epsilon} \big| \boldsymbol{e}(i) = \boldsymbol{\epsilon} \big] I_{\boldsymbol{e}(i) = \boldsymbol{\epsilon}} \overset{(8.1)}{\lesssim} 2^{-m_0 F(R,W)}.
\end{aligned}$$

Since $F < F(R, W)$, this implies

$$\boldsymbol{E}[I^{\mathrm{err}}(i)|\mathcal{F}_{i-1}] \leq 2^{-Fm_0} \qquad \forall i \qquad \text{for} \quad m_0 \approx \infty. \tag{8.21}$$

So by invoking (8.20), we see that almost surely

$$\begin{aligned}\overline{\lim_{r\to\infty}}\,\frac{1}{r}\sum_{i=1}^{r} I^{\rm err}(i) \;&=\; \overline{\lim_{r\to\infty}}\,\frac{1}{r}\sum_{i=1}^{r} \boldsymbol{E}[I^{\rm err}(i)|\mathcal{F}_{i-1}]\\ &\quad+\lim_{r\to\infty}\frac{1}{r}\sum_{i=1}^{r}\Big\{I^{\rm err}(i)-\boldsymbol{E}[I^{\rm err}(i)|\mathcal{F}_{i-1}]\Big\}\le 2^{-Fm_0}. \qquad \square\end{aligned}$$

Now we start to analyze the influence of the channel noise on the estimation errors. To this end, introduce the indicator functions of the following events:

$$\begin{array}{ccc} I_0(i) & \longleftrightarrow & q_d(\tau_i)=q_c(\tau_{i-1})\neq \text{✠},\\ I_{\text{✠}}(i) & \longleftrightarrow & q_d(\tau_i)=q_c(\tau_{i-1})=\text{✠},\\ I^{\rm err}_{\rm c\text{✠}}(i) & \longleftrightarrow & q_d(\tau_i)\neq q_c(\tau_{i-1})=\text{✠},\\ I^{\rm err}_{\rm d\text{✠}}(i) & \longleftrightarrow & \text{✠}=q_d(\tau_i)\neq q_c(\tau_{i-1}),\\ I^{\rm err}_{0}(i) & \longleftrightarrow & \text{✠}\neq q_d(\tau_i)\neq q_c(\tau_{i-1})\neq\text{✠}.\end{array} \tag{8.22}$$

Note that $I^{\rm err}_{\rm c\text{✠}}(i)+I^{\rm err}_{\rm d\text{✠}}(i)+I^{\rm err}_{0}(i)=I^{\rm err}(i)$ and $I_0(i)+I_{\text{✠}}(i)=1-I^{\rm err}(i)$. We first study the evolution of

$$\delta_i:=\delta_c^+(\tau_i)\quad\text{and}\quad z_i:=|\widehat{x}_c^+(\tau_i)-x(\tau_i)|. \tag{8.23}$$

Lemma 8.13. *The following relations hold for any $i\ge 1$:*

$$\delta_i=\delta_{i-1}\Big\{\varkappa^{m_0}\big[I_0(i)+I_0^{err}(i)+I_{c\text{✠}}^{err}(i)\big]+\gamma^{m_0}\big[I_{\text{✠}}(i)+I_{d\text{✠}}^{err}(i)\big]\Big\}, \tag{8.24}$$

$$\begin{aligned} z_i &\le z_{i-1}\|A\|^{m_0}\big[I_{\text{✠}}(i)+I_{d\text{✠}}^{err}(i)\big]\\ &\quad+\delta_{i-1}\varkappa^{2m_0}I_0(i)+\|A\|^{m_0}(z_{i-1}+\delta_{i-1})\big[I_0^{err}(i)+I_{c\text{✠}}^{err}(i)\big].\end{aligned} \tag{8.25}$$

Here $\varkappa\in(0,1)$ is taken from Lemma 8.5, and $\gamma>\|A\|$ is the parameter of the estimator.

Proof. To prove (8.24), we note that

$$\delta_i\overset{(8.23)}{=\!=}\delta_c^+(\tau_i)\overset{(8.15)}{=\!=}\delta_c^-(\tau_i)\left(\langle q_d(\tau_i)\rangle_{\varkappa,\gamma}\right)^{m_0}\overset{(8.23)}{=\!=}\delta_{i-1}\left(\langle q_d(\tau_i)\rangle_{\varkappa,\gamma}\right)^{m_0}.$$

So (8.24) is immediate from (8.22) and the definition of $\langle\cdot\rangle_{\varkappa,\gamma}$ from (8.8). To justify (8.25), we observe that

$$\begin{aligned} z_i &\overset{(8.23)}{=\!=} |\widehat{x}_c^+(\tau_i)-x(\tau_i)|\overset{(8.15)}{=\!=}|\widehat{x}_c^-(\tau_i)+\delta_c^-(\tau_i)A^{m_0}\overset{\star}{q}_d(\tau_i)-x(\tau_i)|\\ &\overset{(2.1),(8.5)}{=\!=\!=} \Big|A^{m_0}\Big[\widehat{x}_c^+(\tau_{i-1})+\delta_c^+(\tau_{i-1})\overset{\star}{q}_d(\tau_i)-x(\tau_{i-1})\Big]\Big|\\ &\overset{(8.23)}{=\!=} \Big|A^{m_0}\Big[\widehat{x}_c^+(\tau_{i-1})-x(\tau_{i-1})+\delta_{i-1}\overset{\star}{q}_d(\tau_i)\Big]\Big|.\end{aligned}$$

If $I_{\text{✠}}(i)+I^{\rm err}_{\rm d\text{✠}}(i)=1$, then $q_d(\tau_i)=\text{✠}$ and $\overset{\star}{q}_d(\tau_i)=0$ by (8.8). So $z_i\le\|A\|^{m_0}|\widehat{x}_c^+(\tau_{i-1})-x(\tau_{i-1})|\overset{(8.23)}{=\!=}\|A\|^{m_0}z_{i-1}$. If $I_0(i)=1$, then $\text{✠}\neq q_c(\tau_{i-1})=q_d(\tau_i)=\overset{\star}{q}_d(\tau_i)$. So $z_i=\delta_{i-1}\,|A^{m_0}\,[\varepsilon(\tau_{i-1})-q_c(\tau_{i-1})]|$ due to (8.7), where $q_c(\tau_{i-1})$ is the quantized value of $\varepsilon(\tau_{i-1})$. Hence by invoking (8.2), where $\rho_{\mathfrak{Q}}=\varkappa^{2m_0}$ thanks to Lemma 8.5, we get $z_i\le\varkappa^{2m_0}\delta_{i-1}$. Finally, suppose that $I_0^{\rm err}(i)+I^{\rm err}_{\rm c\text{✠}}(i)=1$. Then $|\overset{\star}{q}_d(\tau_i)|\le 1$, and

so $z_i \le \|A\|^{m_0}\,(|\widehat{x}_c^+(\tau_{i-1}) - x(\tau_{i-1})| + \delta_{i-1}) \overset{(8.23)}{=\!=} \|A\|^{m_0}\,(z_{i-1} + \delta_{i-1})$. Summarizing, we arrive at (8.25). □

Lemma 8.13 entails an important conclusion about the evolution of the ratio $\xi_i := z_i/\delta_i$, which determines whether the alarm symbol ✠ is sent over the channel:

$$q_c(\tau_i) = \text{✠} \Leftrightarrow \xi_i = z_i/\delta_i > 1. \tag{8.26}$$

Corollary 8.14. *For $i \ge 1$, the following inequality holds:*

$$\xi_i \le \left\{ \begin{array}{ll} \rho\xi_{i-1} & \text{if} \quad \xi_{i-1} > 1 \\ \varkappa^{m_0} & \text{if} \quad \xi_{i-1} \le 1 \end{array} \right\} \big[1 - I^{err}(i)\big] + b/2\big[\xi_{i-1} + 1\big] I^{err}(i), \tag{8.27}$$

where $I^{err}(i)$ is the error indicator function (8.18)*, and*

$$\rho := \left(\frac{\|A\|}{\gamma}\right)^{m_0}, \qquad b := 2\left(\frac{\|A\|}{\varkappa}\right)^{m_0}. \tag{8.28}$$

The proof of this claim is by merely checking (8.27) on the basis of (8.24) and (8.25).

Now we are going to study how often the alarm symbol ✠ is sent or, in other words, the frequency of the event $\xi_i > 1$. The following lemma reveals a relationship between this event and the channel errors.

Lemma 8.15. *Whenever $\xi_i > 1$ for $i = \bar{i} + 1, \dots, \bar{i} + r$, the number l of channel errors within the interval $[\bar{i} + 1 : \bar{i} + r]$ obeys the lower bound*

$$l := \big|j = \bar{i} + 1, \dots, \bar{i} + r : I^{err}(j) = 1\big| \ge r\frac{\log_2[\rho^{-1}]}{\log_2 b + \log_2[\rho^{-1}]} - \frac{\log_2 \max\{\xi_{\bar{i}}, \frac{\xi_{\bar{i}}+1}{2}\}}{\log_2 b + \log_2[\rho^{-1}]}. \tag{8.29}$$

Proof. If $\xi_{\bar{i}} \le 1$ and $I^{\rm err}(\bar{i}+1) = 0$, then (8.27) implies $\xi_{\bar{i}+1} = \varkappa^{m_0} < 1$ in violation of the hypotheses of the lemma. Thus $\xi_{\bar{i}} \le 1 \Rightarrow I^{\rm err}(\bar{i}+1) = 1$. By invoking (8.27) once more, we get, for $i = \bar{i} + 1, \dots, \bar{i} + r$,

$$\begin{aligned} \xi_i &\le \rho\xi_{i-1}[1 - I^{\rm err}(i)] + b\xi_{i-1}I^{\rm err}(i) + \frac{b}{2}[1 - \xi_{i-1}]I^{\rm err}(i) \\ &\le \xi_{i-1}\big\{\rho[1 - I^{\rm err}(i)] + bI^{\rm err}(i)\big\} + \frac{b}{2}\max\{1 - \xi_{i-1}, 0\}I^{\rm err}(i). \end{aligned}$$

The last summand does not vanish only if $i = \bar{i}+1$ and $I^{\rm err}(\bar{i}+1) = 1$. It follows that

$$\begin{aligned} 1 < \xi_{\bar{i}+r} \le \xi_{\bar{i}}\rho^{r-l}b^l + \frac{b}{2}\max\{1 - \xi_{\bar{i}}, 0\}\rho^{r-l}b^{l-1} = \rho^{r-l}b^l \max\left\{\xi_{\bar{i}}, \frac{\xi_{\bar{i}}+1}{2}\right\}, \\ 0 < (l-r)\log_2[\rho^{-1}] + l\log_2 b + \log_2 \max\left\{\xi_{\bar{i}}, \frac{\xi_{\bar{i}}+1}{2}\right\} \;\Big|\; \Rightarrow \text{(8.29)}. \end{aligned}$$ □

Now we are in a position to estimate the frequency of sending the alarm signal ✠.

Corollary 8.16. *For the indicator function $I_{\xi>1}(i) \longleftrightarrow \xi_i > 1$, the following relation holds almost surely:*

$$\varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=1}^{r} I_{\xi>1}(i-1) \le \beta := \mu^{-1} \varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=1}^{r} I^{err}(i), \qquad \text{where} \quad \mu := \frac{\log_2[\rho^{-1}]}{\log_2 b + \log_2[\rho^{-1}]}. \tag{8.30}$$

Proof. If $\beta \ge 1$, the claim is obvious. Suppose that $\beta < 1$. Then $\xi_i \le 1$ for some $i = i^*$. Indeed, otherwise, Lemma 8.15 with $\bar{i} := 0$ and arbitrary r yields

$$\varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=1}^{r} I^{\mathrm{err}}(i) \ge \mu + \varlimsup_{r\to\infty} -\frac{1}{r}\frac{\log_2 \max\{\xi_0, \frac{\xi_0+1}{2}\}}{\log_2 b + \log_2[\rho^{-1}]} = \mu,$$

which implies $\beta \ge 1$ in violation of the hypothesis. For $r > i^*$, the set $\{i^* \le i \le r : I_{\xi>1}(i) = 1\}$ disintegrates into several intervals of durations $r_1, \dots, r_s$, respectively, not containing i^* and separated by intervals where $\xi_i \le 1$. Now we apply Lemma 8.15 to the jth interval, picking $\bar{i}$ to be the integer preceding its left end. Then $\xi_{\bar{i}} \le 1$, the second ratio in (8.29) is nonpositive, and so the number l_j of errors contained by the interval at hand is no less than $r_j\mu$. Hence

$$\begin{aligned}\varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=1}^{r} I^{\mathrm{err}}(i) &\ge \varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=j}^{s} l_j \ge \mu \varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=j}^{s} r_j \\ &= \mu \varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=i^*}^{r} I_{\xi>1}(i) = \mu \varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=1}^{r} I_{\xi>1}(i-1) \;\Big|\; \Rightarrow \text{ (8.30).} \qquad \square\end{aligned}$$

Corollary 8.17. *The indicator function $I(i) \longleftrightarrow I^{err}(i) = 1 \vee I_{\xi>1}(i-1) = 1$ almost surely obeys the inequality*

$$\begin{aligned}(8.31)\quad \varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=1}^{r} I(i) &\le \left\{2 + \frac{\log_2 b}{\log_2[\rho^{-1}]}\right\} \\ &\times \varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=1}^{r} I^{err}(i) \overset{(8.19)}{\le} \overline{p} := 2^{-Fm_0}\left\{2 + \frac{\log_2 b}{\log_2[\rho^{-1}]}\right\}.\end{aligned}$$

Indeed, this is immediate from Corollary 8.16 and the apparent inequality $I(i) \le I^{\mathrm{err}}(i) + I_{\xi>1}(i-1)$.

Observation 8.1. It is easy to see that the first inequality in (8.31) is a direct consequence of (8.27). In other words, it holds for any nonnegative solution ξ_i of the recursive inequalities (8.27) with $i = 1, 2, \dots$, where $\{I^{\mathrm{err}}(i)\}$ is an arbitrary sequence of reals $I^{\mathrm{err}}(i) = 0, 1$ and $\rho, \varkappa \in (0,1)$, $b > 1$ are arbitrary numbers.

Lemma 8.18. *The coder-decoder pair considered in Proposition* 8.9 *tracks the state almost surely if*

(8.32)

$$\omega := \log_2[\varkappa^{-1}] - \overline{p}\big\{\log_2\gamma + \log_2[\varkappa^{-1}]\big\} > 0 \quad \text{and} \quad \chi := \omega(1-\overline{p}) - \overline{p}\log_2\|A\| > 0.$$

Proof. The symbol c (with a possible index) will be used to denote random constants independent of i and r. For any $\alpha > 0$, (8.31) implies $S(r) := \sum_{i=1}^{r} I(i) \le r(\overline{p} + \alpha)$ for $r \approx \infty$. Since $\varkappa < 1 < \gamma$, (8.22) and (8.24) yield

$$\begin{aligned}\delta_i &\le \delta_{i-1}\Big\{\varkappa^{m_0}[1 - I(i)] + \gamma^{m_0} I(i)\Big\} \qquad \forall i \ge 1 \;\Rightarrow\; \delta_r \\ &\le \delta_0 \varkappa^{rm_0} \prod_{i=1}^{r}\left(\frac{\gamma}{\varkappa}\right)^{m_0 I(i)} = \delta_0 \varkappa^{rm_0}\left(\frac{\gamma}{\varkappa}\right)^{m_0 S(r)} \\ &\overset{r\approx\infty}{\le} \delta_0 \varkappa^{rm_0}\left(\frac{\gamma}{\varkappa}\right)^{rm_0(\overline{p}+\alpha)} = \delta_0 2^{-rm_0\omega_\alpha}, \\ &\qquad \text{where } \omega_\alpha := \log_2[\varkappa^{-1}] - [\overline{p}+\alpha]\big\{\log_2\gamma + \log_2[\varkappa^{-1}]\big\} \xrightarrow{\alpha\to 0} \omega > 0.\end{aligned}$$

Thus for $\alpha \approx 0$, we have $\omega_\alpha > 0$ and

$$\delta_i \le c' 2^{-im_0\omega_\alpha} \to 0 \quad \text{as} \quad i \to \infty. \tag{8.33}$$

Now we note that due to (8.25),

$$z_i \le \delta_{i-1}\varkappa^{2m_0}\big[1 - I(i)\big] + \|A\|^{m_0}(z_{i-1} + \delta_{i-1})I(i) \le \|A\|^{m_0} z_{i-1} I(i) + c2^{-im_0\omega_\alpha},$$

$$z_r \le z_0 \prod_{i=1}^{r}\Big[\|A\|^{m_0} I(i)\Big] + c\sum_{i=1}^{r} 2^{-im_0\omega_\alpha} \prod_{j=i+1}^{r}\Big[\|A\|^{m_0} I(j)\Big].$$

The first relation from (8.32) implies $\overline{p} < 1$. So $\{i \ge 1 : I(i) = 1\} \ne \{i = 1, 2, \dots\}$ due to (8.31). It follows that for $r \approx \infty$, the first summand vanishes and

$$z_r \le c \sum_{i=r-l}^{r} 2^{-im_0\omega_\alpha} \|A\|^{m_0(r-i)},$$

where $\{r-l+1, \dots, r\}$ is the largest subinterval of the set $\mathfrak{I}_r := \{1 \le i \le r : I(i) = 1\}$ containing r. (If $r \notin \mathfrak{I}_r$, then $l := 0$.) We proceed by taking into account the inequality $l \le \sum_{i=1}^{r} I(i) = S(r) \le r(\overline{p} + \alpha)$ for all $r \approx \infty$:

$$\begin{aligned} z_r &\le c2^{-rm_0\omega_\alpha} \sum_{i=r-l}^{r} 2^{(r-i)m_0\omega_\alpha} \|A\|^{m_0(r-i)} = c2^{-rm_0\omega_\alpha} \sum_{i=0}^{l} \big(2^{\omega_\alpha}\|A\|\big)^{m_0 i} \\ &\le c\frac{2^{-rm_0\omega_\alpha}\big(2^{\omega_\alpha}\|A\|\big)^{m_0 l}}{1 - \big(2^{\omega_\alpha}\|A\|\big)^{-m_0}} \\ &\overset{r\approx\infty}{\le} c\frac{2^{-rm_0\omega_\alpha}\big(2^{\omega_\alpha}\|A\|\big)^{m_0 r(\overline{p}+\alpha)}}{1 - \big(2^{\omega_\alpha}\|A\|\big)^{-m_0}} = \frac{c}{1 - \big(2^{\omega_\alpha}\|A\|\big)^{-m_0}} 2^{-rm_0\chi_\alpha}, \end{aligned}$$

where $\chi_\alpha := \omega_\alpha[1 - (\overline{p} + \alpha)] - (\overline{p} + \alpha)\log_2\|A\| \xrightarrow[\alpha\to 0]{(8.32)} \chi > 0$.

Thus $\chi_\alpha > 0$ for $\alpha \approx 0$. So $z_r \overset{(8.23)}{=\!=} |\widehat{x}_c^+(\tau_r) - x(\tau_r)| \to 0$ as $r \to \infty$. Here $\widehat{x}_c^+(\tau_r) = \widehat{x}_d^+(\tau_r)$ thanks to (8.9) and (8.15). It remains to note that for $\tau_r < t \le \tau_{r+1} = \tau_i + m_0$,

$$|\widehat{x}_d(t) - x(t)| \overset{(2.1),(8.5)}{=\!=} \big|A^{t-\tau_r}[\widehat{x}_d^+(\tau_r) - x(\tau_r)]\big| \le \|A\|^{m_0}|\widehat{x}_d^+(\tau_r) - x(\tau_r)| \xrightarrow{t\to\infty} 0,$$

since $r \to \infty$ as $t \to \infty$. □

Proof of Proposition 8.9. By Lemma 8.18, it suffices to show that (8.32) does hold whenever m_0 is large enough. In turn, this is true if $\overline{p} \to 0$ as $m_0 \to \infty$. The required property is established as follows:

$$\overline{p} \overset{(8.31)}{=\!=} 2^{-Fm_0}\left\{2 + \frac{\log_2 b}{\log_2[\rho^{-1}]}\right\} \overset{(8.28)}{=\!=} 2^{-Fm_0}\left\{2 + \frac{1 + m_0[\log_2\|A\| + \log_2\varkappa^{-1}]}{m_0\log_2\gamma - \log_2\|A\|}\right\} \to 0$$

as $m_0 \to \infty$. □

8.6. Completing the proof of Theorem 5.1. The implication $\mathfrak{c} > \eta(A) \Rightarrow$ (a) and (b) has already been justified for systems with no stable modes. Now we consider the general case. Suppose that $\mathfrak{c} > \eta(A)$. Since $\eta(A) = \eta(A_+)$, claims (a) and (b)

are true for the system (6.1) with $x_0^+ := \pi_+ x_0$ by Corollary 8.8 and Proposition 8.9. We apply the corresponding coder-decoder pair to the primal system (2.1). In doing so, we also alter the coder's step c.1, where it identifies the current state $x_+(\tau_i)$ of (6.1). Formerly this was done on the basis of the past measurements from (6.1). Now we employ the observations from (2.1). Then thanks to Assumption 4.4, it is possible to compute $\pi_+ x(\tau_i)$, provided $m_0 \geq n$ in (8.4). Here π_+ is the projector onto L_+ parallel to L_-, and the notations $L_\pm$ were introduced in Explanation 5.1. Since evidently $x_+(t) := \pi_+ x(t)$, this does not alter the operation of the observer. It remains to note that so far as $x_-(t) := x(t) - x_+(t) \to 0$ as $t \to \infty$, this observer tracks that state of (2.1)

$$|x(t) - \widehat{x}_+(t)| = |x_-(t) + x_+(t) - \widehat{x}_+(t)| \leq |x_-(t)| + |x_+(t) - \widehat{x}_+(t)| \to 0$$

whenever it detects the state of (6.1): $|x_+(t) - \widehat{x}_+(t)| \to 0$.

The implications (a) $\vee$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) are apparent, whereas (d) $\Rightarrow \mathfrak{c} \geq \eta(A)$ was justified in section 6.

9. Proof of Theorem 5.2. The necessity part of this theorem was justified in section 7. In this section, the focus is on proving that the inequality $\mathfrak{c} > \eta(A)$ is sufficient for almost sure stabilizability.

The section is organized as follows. Subsection 9.1 describes a stabilizing coder-decoder pair. Its convergence is proved in subsection 9.3. In subsection 9.2, we show in which way communication feedback required for the almost sure stabilization can be arranged by means of control. In these subsections, we consider the plant (3.1) with no stable modes. In subsection 9.4, all arguments are consolidated, and the proof of Theorem 5.2 is completed.

9.1. Coder-decoder pair stabilizing the plant almost surely by means of fixed length code words. We present it assuming that the stabilizability condition $\mathfrak{c} > \eta(A)$ holds and the plant (3.1) has no stable modes. Then

$$\eta(A) = \log_2 |\det A|, \tag{9.1}$$

and the system is controllable and observable thanks to Assumptions 4.4 and 4.5. In the general case, a stabilizing controller can be obtained by applying that presented below to the unstable part of the system (see subsection 9.4).

Though a stabilizing controller can be constructed along the lines of subsection 8.3, we omit this and focus on stabilization by means of fixed length code words. In doing so, we show that much less communication feedback is required for stabilization than for detection. In fact, a feedback with arbitrarily small rate is sufficient.

As in subsection 8.3, we start with introducing basic components of which the coder and decoder will be assembled. To this end, we first pick two numbers η and R such that

$$\eta(A) < \eta < R < \mathfrak{c}. \tag{9.2}$$

Then for any $m = 1, 2, \ldots$, we choose

- a set $\boldsymbol{\mathcal{E}}^{[m]} \subset \mathcal{E}^m$ of $N = N'_m \approx 2^{mR}$ input code words $\mathfrak{D}_m$ with the properties described in Theorem 8.1,
- an m-contracted quantizer $\mathfrak{Q}_m$ described in Lemma 8.5.

We recall that such a quantizer partitions the unit ball B_0^1 with respect to some norm $|\cdot|$ in $\mathbb{R}^n$ into a collection of disjoint sets $Q_1, \ldots, Q_N$, each equipped with a

centroid $q_i \in Q_i$, and associates any vector $x \in Q_i$ with its quantized value q_i and any vector outside the ball B_0^1 with an alarm symbol ✠. Furthermore,

$$A^m (Q_i - q_i) \subset \rho_m B_0^1 \qquad \forall i = 1, \dots, N, \qquad \text{where} \quad \rho_m := \varkappa^{2m}. \tag{9.3}$$

Deadbeat stabilizer. We also pick a *deadbeat stabilizer*, i.e., a linear transformation of an initial state

$$x(0) = x \xrightarrow{\mathfrak{S}} [u(0), u(1), \dots, u(n-1), 0, 0, \dots] \tag{9.4}$$

into a sequence of controls driving the state to zero: $x(n) = 0$. Since the system (3.1) is controllable, such a stabilizer exists [1, p. 253]. We also define $\mathfrak{S}(✠) := \overline{\boldsymbol{u}}_✠$ by picking an *alarm control sequence* $\overline{\boldsymbol{u}}_✠ = [u_0, \dots, u_{s-1}, 0, 0, \dots]$, which drives the system from $x(0) = 0$ to $x(s) = 0$. This sequence will be specified further, and its role will be explained in subsection 9.2. The number $L(\mathfrak{S}) := \max\{n, s\}$ is called the *length* of the stabilizer.

Operation cycles and parameters. The controller operation consists of time cycles $[\tau_i : \tau_{i+1}]$,

$$\tau_i := i m_0, \tag{9.5}$$

of equal duration m_0. A fixed and independent of the cycle sequence of operations is executed within any cycle. The integer parameter m_0 of the controller, along with one more parameter γ, is chosen so that

$$m_0 \geq n + L(\mathfrak{S}), \qquad \gamma > \|A\|, \tag{9.6}$$

and for all $m \geq m_0$, the outputs of the quantizer $\mathfrak{Q}_m$ including the alarm signal ✠, can be encoded by the code words from the set $\boldsymbol{\mathcal{E}}^{[m]}$. (This is possible by Remark 8.1.) Encoding will be carried out at the beginning τ_i of each operation cycle, and then the code word thus obtained will be transmitted to the decoder during this cycle.

We also suppose that a limited feedback communication is available.

Assumption 9.1. By the end τ_{i+1} of the current operation cycle, the coder almost surely becomes aware whether or not the message received by the decoder at the beginning τ_i of this cycle was the alarm one ✠.

Thus the required feedback concerns only one message ✠ and has the size of one bit per operation cycle. By increasing the cycle duration m_0, the average amount of information transmitted across the feedback link can be made arbitrarily small. Assumption 9.1 may also be true due not to the feedback link but the fact that the alarm signal is transmitted over an especially reliable subchannel.

Remark 9.1. *In subsection* 9.2*, we shall show that Assumption* 9.1 *can always be ensured by means of control via a special choice of the alarm control sequence.*

A stabilizing *coder-decoder pair* operates as follows. Both coder and decoder compute controls $u_c(t)$, $u_d(t)$ and upper bounds for the state norm $\delta_c(t)$, $\delta_d(t)$, respectively. Actually acting upon the plant is the control $u_d(t)$. The initial bound is common: $\delta_c(0) = \delta_d(0) = \delta_0$. (The inequality $\delta_0 \geq |x(0)|$ may be violated.) Within any operation cycle $[\tau_i : \tau_{i+1})$, the coder successively sends over the channel the symbols of the code word of length m_0 formed at time τ_i, and the decoder carries out the control program $u_d(\tau_i), u_d(\tau_i + 1), \dots, u_d(\tau_{i+1} - 1)$ generated at time τ_i. These actions are prefaced at times $t = \tau_i$ by the following operations.

The coder (at times $t = \tau_i$, $i = 1, 2, \dots$) does the following:

c.1. proceeding from the previous measurements calculates the current state $x(\tau_i)$;
c.2. computes the prognosis of the state at time $t = \tau_{i+1}$:

$$\widehat{x}_c(t) := A^{m_0} x(\tau_i) + \sum_{j=\tau_i}^{t-1} A^{t-1-j} B u_c(j); \tag{9.7}$$

c.3. if $i = 3, 4, \dots$ corrects the state norm upper bound:

$$\delta_c(\tau_i) := \delta_c(\tau_i) \frac{\langle q_d(\tau_{i-1})\rangle_{\varkappa,\gamma}^{m_0}}{\langle q_c(\tau_{i-2})\rangle_{\varkappa,\gamma}^{m_0}} \tag{9.8}$$

$$= \delta_c(\tau_i) \times \begin{cases} \left(\frac{\gamma}{\varkappa}\right)^{m_0} & \text{if } q_d(\tau_{i-1}) = \maltese \text{ and } q_c(\tau_{i-2}) \neq \maltese, \\ \left(\frac{\varkappa}{\gamma}\right)^{m_0} & \text{if } q_d(\tau_{i-1}) \neq \maltese \text{ and } q_c(\tau_{i-2}) = \maltese, \\ 1 & \text{if } \begin{array}{c} q_d(\tau_{i-1}) = \maltese \text{ and } q_c(\tau_{i-2}) = \maltese \\ \text{or} \\ q_d(\tau_{i-1}) \neq \maltese \text{ and } q_c(\tau_{i-2}) \neq \maltese; \end{array} \end{cases}$$

here $\varkappa$ and γ are taken from (9.3) and (9.6), respectively, and

$$\langle q\rangle_{\varkappa,\gamma} := \begin{cases} \varkappa & \text{if } q \neq \maltese, \\ \gamma & \text{otherwise;} \end{cases} \tag{9.9}$$

c.4. employs the quantizer $\mathfrak{Q}_{m_0}$ and computes the quantized value $q_c(\tau_i)$ of the scaled state at time τ_{i+1}:

$$\varepsilon(\tau_i) := \big[\delta_c(\tau_i)\big]^{-1} \widehat{x}_c(\tau_{i+1}), \qquad q_c(\tau_i) := \mathfrak{Q}_{m_0}\big[\varepsilon(\tau_i)\big]; \tag{9.10}$$

c.5. encodes this quantized value $q_c(\tau_i)$ by means of the code book $\mathfrak{E}^{[m_0]}$ and thus obtains the code word to be transmitted over the channel during the next operation cycle $[\tau_i : \tau_{i+1})$;
c.6. computes the control program $\overline{\boldsymbol{u}}_{i+1}^c = \big[u_c(\tau_{i+1}), \dots, u_c(\tau_{i+2} - 1)\big]$ for not the next but the overtaking operation cycle $[\tau_{i+1}, \tau_{i+2} - 1)$ and then corrects the state upper bound:

$$\overline{\boldsymbol{u}}_{i+1}^c := \delta_c(\tau_i) \mathfrak{S}\big[q_c(\tau_i)\big], \qquad \delta_c(\tau_i) := \delta_c(\tau_i) \times \langle q_c(\tau_i)\rangle_{\varkappa,\gamma}^{m_0}, \tag{9.11}$$

where $\langle q\rangle_{\varkappa,\gamma}$ is given by (9.9) and $\mathfrak{S}$ is the deadbeat stabilizer.

The decoder (at the times $t = \tau_i$, $i = 2, 3, \dots$) does the following:
d.1. applies the decoding rule $\mathfrak{D}_{m_0}$ to the data received within the previous operation cycle $[\tau_{i-1} : \tau_i]$ and thus acquires the decoded value $q_d(\tau_i)$ of $q_c(\tau_{i-1})$ (which may be incorrect due to the channel errors);
d.2. computes the control program $\overline{\boldsymbol{u}}_i^d = \big[u_d(\tau_i), \dots, u_d(\tau_{i+1} - 1)\big]$ for the next operation cycle $[\tau_i : \tau_{i+1})$ and corrects the state upper bound:

$$\overline{\boldsymbol{u}}_i^d := \delta_d(\tau_i) \mathfrak{S}\big[q_d(\tau_i)\big], \qquad \delta_d(\tau_i) := \delta_d(\tau_i) \times \langle q_d(\tau_i)\rangle_{\varkappa,\gamma}^{m_0}. \tag{9.12}$$

For the definiteness, the initial control programs $\overline{\boldsymbol{u}}_0^c, \overline{\boldsymbol{u}}_0^d, \overline{\boldsymbol{u}}_1^c, \overline{\boldsymbol{u}}_1^d$ are taken to be the alarm ones.

Note that step c.1 is possible. Indeed, due to (9.4), (9.5), (9.12), and the first relation from (9.6), the dynamics of the closed-loop system (3.1) is free, $u(t) = 0$, for at least n time steps before τ_i. Since the system (3.1) is observable, this makes

it possible to identify the state $x(\tau_i)$ proceeding from the measurements even if the coder is unaware of the entire sequence of controls u_d actually acting upon the plant.

Step c.3 is possible by Assumption 9.1. The role of this step is to make the bounds δ_c and δ_d identical whenever the transmission across the channel is errorless. To specify this claim, we mark the values of δ_c and δ_d after and just before the updates in accordance with (9.11), (9.12) with the $^+$ and $^-$ indices, respectively. So the value $\delta_c^-(\tau_i)$ is taken after the correction (9.8). For consistency, we also assume that $q_c(\tau_0) := q_d(\tau_1) := \maltese$.

LEMMA 9.1. *Step* c.3 *ensures that whenever the current transmission is errorless, the next state norm upper bounds used by the coder and decoder, respectively, are identical:*

$$q_c(\tau_{i-1}) = q_d(\tau_i) \quad \Longrightarrow \quad \delta_c^-(\tau_i) = \delta_d^-(\tau_{i+1}), \qquad i = 1, 2, \dots. \tag{9.13}$$

Proof. It suffices to show that for $i = 1, 2, \dots$

$$\delta_c^-(\tau_i) = \delta_d^-(\tau_{i+1}) \left[\frac{\langle q_c(\tau_{i-1}) \rangle_{\varkappa,\gamma}}{\langle q_d(\tau_i) \rangle_{\varkappa,\gamma}} \right]^{m_0}. \tag{9.14}$$

The proof will be by induction on i. For $i = 1$, the claim is evident. Suppose that (9.14) holds for some $i \geq 1$. Then

$$\begin{aligned} \delta_c^-(\tau_{i+1}) \overset{(9.8)}{=\!=} \delta_c^+(\tau_i) \frac{\langle q_d(\tau_i) \rangle_{\varkappa,\gamma}^{m_0}}{\langle q_c(\tau_{i-1}) \rangle_{\varkappa,\gamma}^{m_0}} &\overset{(9.11)}{=\!=} \delta_c^-(\tau_i) \langle q_c(\tau_i) \rangle_{\varkappa,\gamma}^{m_0} \frac{\langle q_d(\tau_i) \rangle_{\varkappa,\gamma}^{m_0}}{\langle q_c(\tau_{i-1}) \rangle_{\varkappa,\gamma}^{m_0}} \\ &\overset{(9.14)}{=\!=} \delta_d^-(\tau_{i+1}) \langle q_c(\tau_i) \rangle_{\varkappa,\gamma}^{m_0} \\ &\overset{(9.12)}{=\!=} \delta_d^-(\tau_{i+2}) \frac{\langle q_c(\tau_i) \rangle_{\varkappa,\gamma}^{m_0}}{\langle q_d(\tau_{i+1}) \rangle_{\varkappa,\gamma}^{m_0}}; \end{aligned}$$

i.e., (9.14) with $i := i + 1$ does hold. □

The main property of the proposed coder-decoder pair is given by the following proposition.

PROPOSITION 9.2. *Suppose that Assumptions* 4.1–4.5*, Assumption* 9.1*, and relations* (9.6) *hold,* $\mathfrak{c} < \eta(A)$*, and the system* (3.1) *has no stable modes. Then the coder-decoder pair introduced in this subsection stabilizes the system almost surely if* m_0 *is large enough:* $m_0 \geq M(A, B, \varkappa, \gamma, W, R)$.

An explicit expression for the bound $M(A, B, \varkappa, \gamma, W, R)$ can be derived from the proof of this proposition, which is given in subsection 9.3. The proof resembles that of Proposition 8.9. However, there are important differences. They mainly proceed from the fact that now the coder and decoder are not completely synchronized via the communication feedback, contrary to the situation from Proposition 8.9. More precisely, the coder and decoder from Proposition 8.9 produce common error upper bounds δ_c, δ_d and state estimates $\widehat{x}_c, \widehat{x}_d$. Now only the bounds δ_c, δ_d are synchronized in a weaker sense: they are not always common but only when the previous transmission across the "feedforward" channel is errorless (see Lemma 9.1). At the same time, the feedback link is not used to put the controls produced by the coder in harmony with those generated by the decoder. In fact, the goal of the proof is to demonstrate that being properly adjusted to the current circumstances, the arguments from the proof of Proposition 8.9 (see subsection 8.5) are not destroyed by this difference.

We also stress that Proposition 9.2 holds for any choice of the alarm control sequence.

9.2. Communication feedback by means of control. Now we show that no special means are required to ensure Assumption 9.1. The point is that the decoder may notify the coder about receiving the alarm signal by means of the alarm control sequence. The idea is roughly as follows. Whenever $q_d(\tau_i) \neq \maltese$, the control program $\overline{\boldsymbol{u}}_i^d = [u_d(\tau_i), \dots, u_d(\tau_{i+1}-1)]$ acting upon the plant is a linear function of the quantized state thanks to (9.12). Since the latter is n-dimensional, this program lies in a certain n-dimensional linear space. During the time interval $[\tau_i : \tau_{i+1}-1]$, the coder observes the sequence of measurements $\overline{\boldsymbol{y}}_i := [y(\tau_i), \dots, y(\tau_{i+1}-1)]$, which is a linear transformation of both this program and the n-dimensional initial state $x(\tau_i)$. So the sequence lies in a $2n$-dimensional linear subspace $\mathfrak{L}$. However, the space of all observation sequences $\overline{\boldsymbol{y}}_i$ may be of larger dimension for large $m_0 = \tau_{i+1} - \tau_i$. This makes it possible to pick the alarm control sequence so that it generates the sequence of observations not in $\mathfrak{L}$. Then the coder may recognize the event $q_d(\tau_i) = \maltese$ by checking the relation $\overline{\boldsymbol{y}}_i \notin \mathfrak{L}$.

To be specific, now we consider a particular example of this scheme.

Alarm control sequence $\overline{\boldsymbol{u}}_{\maltese}$. We first pick a control u_* and a sequence $\overline{\boldsymbol{u}}_- := [u_0^-, \dots, u_{n-1}^-]$ such that $Bu_* \neq 0$ and the sequence $\overline{\boldsymbol{u}}_-$ drives the system from the state $x(0) = A^n B u_*$ to $x(n) = 0$. Then we put

$$\overline{\boldsymbol{u}}_{\maltese} := \big[\underbrace{0, \dots, 0}_{2n}, u_*, \underbrace{0, \dots, 0}_{n}, u_0^-, \dots, u_{n-1}^-\big]. \tag{9.15}$$

It is easy to see that this sequence drives the system from $x(0) = 0$ to $x(4n+1) = 0$, as required. To recognize the event $q_d(\tau_i) = \maltese$ by the end of the current operation cycle $[\tau_i : \tau_{i+1}]$, the coder prefaces step c.3 with the following two steps and then proceeds by executing steps c.3–c.6:

c.3$^{\maltese}$1 proceeding from the previous measurements, the coder computes the states $x(\tau_i + 2n)$ and $x(\tau_i + 3n + 1)$;

c.3$^{\maltese}$2 the coder decides that $q_d(\tau_i) = \maltese$ if and only if $x(\tau_i{+}3n{+}1) \neq A^{n+1} x(\tau_i{+}2n)$.

Step c.3$^{\maltese}$1 is possible, since the dynamics of the system is free, $u(t) = 0$, for at least n time steps before times $\tau_i + 2n$ and $\tau_i + 3n + 1$. So the coder can identify the required states on the basis of the measurements.

Steps c.3$^{\maltese}$1 and c.3$^{\maltese}$2 ensure correct recognition of the event $q_d(\tau_i) = \maltese$. Indeed, it suffices to note that $x(\tau_i + 3n + 1) - A^{n+1} x(\tau_i + 2n)$ amounts to $\delta_d(\tau_i) A^n B u_* \neq 0$ if $q_d(\tau_i) = \maltese$, and 0 otherwise.

9.3. Proof of Proposition 9.2. The assumptions of this proposition are supposed to hold throughout the subsection. To start with, we rewrite the state prognosis (9.7) in a more convenient form.

LEMMA 9.3. *The state prognosis* (9.7) *is given by the formula*

$$\widehat{x}_c(\tau_{i+1}) = \delta_c^-(\tau_{i-1}) A^{m_0} \big[\varepsilon(\tau_{i-1}) - \overset{\star}{q}_c(\tau_{i-1})\big] + A^{m_0} \sum_{j=\tau_{i-1}}^{\tau_i - 1} A^{\tau_i - 1 - j} B \big[u_d(j) - u_c(j)\big]. \tag{9.16}$$

Here $\overset{\star}{q} := 0$ *if* $q = \maltese$ *and* $\overset{\star}{q} := q$ *otherwise,* $q_c(\tau_0) := \maltese$*, and* $\varepsilon(\tau_{i-1})$ *is defined by* (9.10)*, where* $\delta_c(\tau_i) = \delta_c^-(\tau_i)$.

Proof. Suppose first that $q_c(\tau_{i-1}) \neq \maltese$. Due to the first formula from (9.11) with $i := i - 1$ and the definition of the deadbeat stabilizer, the sequence of controls $u_c(\tau_i), \dots, u_c(\tau_{i+1} - 1)$ drives the system from the state $\delta_c^-(\tau_{i-1}) q_c(\tau_{i-1})$ at time τ_i

to the state 0 at time $\tau_i + n$. Since $u_c(t) = 0$ for $t = \tau_i + n, \dots, \tau_{i+1} - 1$, the state 0 is kept until time $\tau_{i+1} = \tau_i + m_0$. Hence

$$\delta_c^-(\tau_{i-1})A^{m_0}\overset{\star}{q}_c(\tau_{i-1}) + \sum_{j=\tau_i}^{\tau_{i+1}-1} A^{\tau_{i+1}-1-j}Bu_c(j) = 0. \tag{9.17}$$

This is still true if $q_c(\tau_{i-1}) = \maltese$. Indeed, then $\overset{\star}{q}_c(\tau_{i-1}) = 0$ and $u_c(j), \tau_i \le j \le \tau_{i+1}-1$ is the alarm control sequence, which drives the system from the state $x(\tau_i) = 0$ to $x(\tau_{i+1}) = 0$. Subtracting (9.7) and (9.17) yields

$$\begin{aligned}\widehat{x}_c(\tau_{i+1}) &= A^{m_0}\big[x(\tau_i) - \delta_c^-(\tau_{i-1})\overset{\star}{q}_c(\tau_{i-1})\big] \\ &= \delta_c^-(\tau_{i-1})A^{m_0}\big[\delta_c^-(\tau_{i-1})^{-1}\widehat{x}_c(\tau_i) - \overset{\star}{q}_c(\tau_{i-1})\big] + A^{m_0}\big[x(\tau_i) - \widehat{x}_c(\tau_i)\big].\end{aligned}$$

Here by (9.7) with $i := i - 1$ and (3.1),

$$\widehat{x}_c(\tau_i) = A^{m_0}x(\tau_{i-1}) + \sum_{j=\tau_{i-1}}^{\tau_i-1} A^{\tau_i-1-j}Bu_c(j),$$

$$x(\tau_i) = A^{m_0}x(\tau_{i-1}) + \sum_{j=\tau_{i-1}}^{\tau_i-1} A^{\tau_i-1-j}Bu_d(j).$$

As a result, we arrive at (9.16) by taking into account (9.10). □

Now we consider the stochastic process generated by the coder and decoder. The symbols $\boldsymbol{e}(i)$ and $\boldsymbol{s}(i)$ stand for the messages formed by the coder at time τ_{i-1} and received by the decoder at time τ_i, respectively. We also introduce the error indicator function

$$I^{\mathrm{err}}(i) := 1 \quad \text{if} \quad \mathfrak{D}_{m_0}[\boldsymbol{s}(i)] \ne \boldsymbol{e}(i), \quad i \ge 2, \qquad \text{and} \quad I^{\mathrm{err}}(i) := 0 \quad \text{otherwise},$$

and pick $0 < F < F(R, W)$, where $F(R, W)$ is taken from (8.1). By retracing the arguments from the proof of Lemma 8.12, it is easy to see that the following relation holds almost surely for all sufficiently large m_0:

$$\varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=1}^{r} I^{\mathrm{err}}(i) \le 2^{-Fm_0}. \tag{9.18}$$

Corollary 9.4. *For the indicator function* $\widehat{I}^{err}(i) \leftrightarrow I^{err}(i) = 1 \vee I^{err}(i-1) = 1 \vee I^{err}(i-2) = 1$, *we have*

$$\varlimsup_{r\to\infty} \frac{1}{r}\sum_{i=1}^{r} \widehat{I}^{err}(i) \le 3 \cdot 2^{-Fm_0}. \tag{9.19}$$

Indeed, this is immediate from (9.18) and the inequality $\widehat{I}^{\mathrm{err}}(i) \le I^{\mathrm{err}}(i) + I^{\mathrm{err}}(i-1) + I^{\mathrm{err}}(i-2)$.

Now we start to analyze the influence of the channel errors on the evolution of the closed-loop system. To this end, we introduce the following linear operators by employing the deadbeat stabilizer (9.4) and its length $L(\mathfrak{S})$:

$$\mathfrak{C}[u_0, \dots, u_{L(\mathfrak{S})-1}] := \sum_{j=0}^{L(\mathfrak{S})-1} A^{-1-j}Bu_j, \tag{9.20}$$

$$x \in \mathbb{R}^n \xrightarrow{\mathfrak{S}} \big[u(0), \dots, u(n-1), 0, 0, \dots\big] \xrightarrow{\mathfrak{C}} \mathfrak{B}(x).$$

We also put $\tau_{-1} := -1$, $q_c(\tau_l) := q_d(\tau_j) := \maltese$, $l = -1, 0$, $j = 0, 1$, and for $i \geq 0$, we consider the indicator functions (8.22). We first study the evolution of

$$\delta_i := \delta_c^-(\tau_i) \quad \text{and} \quad z_i := |\widehat{x}_c(\tau_{i+1})|. \tag{9.21}$$

LEMMA 9.5. *The following relations hold for $j \geq 1$ and $i \geq 2$:*

$$\begin{aligned}
\delta_j &= \delta_{j-1}\Big\{\varkappa^{m_0}\big[I_0(j) + I_0^{err}(j) + I_{d\maltese}^{err}(j)\big] + \gamma^{m_0}\big[I_\maltese(j) + I_{c\maltese}^{err}(j)\big]\Big\} \\
&\quad \times \left\{\left(\frac{\gamma}{\varkappa}\right)^{m_0} I_{d\maltese}^{err}(j-1) + \left(\frac{\varkappa}{\gamma}\right)^{m_0} I_{c\maltese}^{err}(j-1) + [1 - I_{d\maltese}^{err}(j-1) \right. \\
&\quad \left. - I_{c\maltese}^{err}(j-1)]\right\}, \qquad (9.22) \\
z_i &\leq z_{i-1}\|A\|^{m_0}\big[I_\maltese(i) + I_{c\maltese}^{err}(i)\big] \\
&\quad + \delta_{i-1}\varkappa^{2m_0}\big[I_0(i) + I_0^{err}(i) + I_{d\maltese}^{err}(i)\big] + \delta_{i-2}d(m_0)\widehat{I}^{err}(i). \qquad (9.23)
\end{aligned}$$

Here $\varkappa \in (0,1)$, $\gamma > \|A\|$, and $\widehat{I}^{err}(i)$ are taken from (9.3), (9.6), *and Corollary* 9.4, *respectively, and*

$$d(m) := \|A\|^{2m}\left[1 + \left(\frac{\gamma}{\varkappa}\right)^m\right]\max\big\{\|\mathfrak{B}\|, |\mathfrak{C}\overline{\boldsymbol{u}}_\maltese|\big\}, \tag{9.24}$$

where $\overline{\boldsymbol{u}}_\maltese$ is the alarm control sequence.

Proof. We start with proving (9.22):

$$\delta_j \overset{(9.21)}{=\!=} \delta_c^-(\tau_j) \overset{(9.8)}{=\!=} \delta_c^+(\tau_{j-1})\frac{\langle q_d(\tau_{j-1})\rangle_{\varkappa,\gamma}^{m_0}}{\langle q_c(\tau_{j-2})\rangle_{\varkappa,\gamma}^{m_0}} \overset{(9.11),(9.21)}{=\!=\!=} \delta_{j-1}\langle q_c(\tau_{j-1})\rangle_{\varkappa,\gamma}^{m_0}\frac{\langle q_d(\tau_{j-1})\rangle_{\varkappa,\gamma}^{m_0}}{\langle q_c(\tau_{j-2})\rangle_{\varkappa,\gamma}^{m_0}}.$$

It remains to note that due to (9.9) and (8.22), the second multiplier and the ratio in the last expression equal the first and second expressions in the curly brackets $\{\ \}$ from (9.22), respectively. To justify (9.23), we denote by s' and s'' the first and second summands from (9.16), respectively. Since in (9.16), $q_c(\tau_{i-1})$ is the quantized value of $\varepsilon(\tau_{i-1})$ by means of the quantizer $\mathfrak{Q}_{m_0}$, relation (9.3) yields

$$\begin{aligned}
|s'| &\leq \delta_{i-1} \times \left\{\begin{array}{lll} \varkappa^{2m_0} & \text{if} & q_c(\tau_{i-1}) \neq \maltese \\ \|A\|^{m_0}|\varepsilon(\tau_{i-1})| & \text{if} & q_c(\tau_{i-1}) = \maltese \end{array}\right\} \\
&\overset{(8.22)}{=\!=} \delta_{i-1}\varkappa^{2m_0}\big[I_0(i) + I_0^{\mathrm{err}}(i) + I_{\mathrm{d}\maltese}^{\mathrm{err}}(i)\big] + \delta_c^-(\tau_{i-1})|\varepsilon(\tau_{i-1})|\|A\|^{m_0}\big[I_\maltese(i) + I_{\mathrm{c}\maltese}^{\mathrm{err}}(i)\big].
\end{aligned}$$

Here $\delta_c^-(\tau_{i-1})|\varepsilon(\tau_{i-1})| = |\widehat{x}_c(\tau_i)| = z_{i-1}$ by (9.10) and (9.21). As a result, we see that $|s'|$ does not exceed the sum of the first two summands from (9.23).

The second summand s'' from (9.16) can be rewritten in the following form due to (9.11), (9.12), and (9.20):

$$\begin{aligned}
s'' &= A^{2m_0}\Big\{\delta_d^-(\tau_{i-1})\beta\big[q_d(\tau_{i-1})\big] - \delta_c^-(\tau_{i-2})\beta\big[q_c(\tau_{i-2})\big]\Big\}, \quad \text{where} \\
\beta(q) &:= \left\{\begin{array}{ll} \mathfrak{B}(q) & \text{if } q \neq \maltese, \\ \mathfrak{C}\overline{\boldsymbol{u}}_\maltese & \text{otherwise,} \end{array}\right. \\
s'' &\overset{(9.14),(9.21)}{=\!=\!=} \delta_{i-2}A^{2m_0}\left\{\frac{\langle q_d(\tau_{i-2})\rangle_{\varkappa,\gamma}^{m_0}}{\langle q_c(\tau_{i-3})\rangle_{\varkappa,\gamma}^{m_0}}\beta\big[q_d(\tau_{i-1})\big] - \beta\big[q_c(\tau_{i-2})\big]\right\}.
\end{aligned}$$

Whenever $\widehat{I}^{\rm err}(i)=0$, we have $q_d(\tau_{i-1})=q_c(\tau_{i-2})$, $q_d(\tau_{i-2})=q_c(\tau_{i-3})$, and so the last expression in the brackets $\{\}$ vanishes. In any case, $|\beta(q)|\le\max\{\|\mathfrak{B}\|,|\mathfrak{C}\overline{\boldsymbol{u}}_{\maltese}|\}$ for $q:=q_d(\tau_{i-1}),q_c(\tau_{i-2})$, since $q\neq\maltese\Rightarrow|q|\le1\Rightarrow|\beta(q)|\le\|\mathfrak{B}\|$. At the same time, $\frac{\langle q_d(\tau_{i-2})\rangle_{\varkappa,\gamma}}{\langle q_c(\tau_{i-3})\rangle_{\varkappa,\gamma}}\le\gamma/\varkappa$ due to (9.8). As a result, we see that $|s''|$ does not exceed the last summand from (9.23), which completes the proof. □

Now we focus on the evolution of the ratio $\xi_i:=z_i/\delta_i$ determining whether $\maltese$ is sent over the channel.

LEMMA 9.6. *For $i\ge2$, inequality* (8.27) *holds with $I^{err}(i):=\widehat{I}^{err}(i)$, where the indicator function $\widehat{I}^{err}(i)$ was introduced in Corollary* 9.4*, and*

$$\rho:=\left(\frac{\|A\|}{\gamma}\right)^{m_0},\qquad b:=2\left(\frac{\gamma}{\varkappa^2}\right)^{2m_0}\bigl[1+d(m_0)\bigr]. \tag{9.25}$$

Proof. Thanks to (9.22), (9.23), and (9.25)

$$\begin{aligned}\xi_i\ \le\ &\Bigl\{\xi_{i-1}\rho\bigl[I_{\maltese}(i)+I^{\rm err}_{{\rm c}\maltese}(i)\bigr]+\varkappa^{m_0}\bigl[I_0(i)+I^{\rm err}_0(i)+I^{\rm err}_{{\rm d}\maltese}(i)\bigr]\Bigr\}\\ &\times\ \underbrace{\left\{\left(\frac{\varkappa}{\gamma}\right)^{m_0}I^{\rm err}_{{\rm d}\maltese}(i-1)+\left(\frac{\gamma}{\varkappa}\right)^{m_0}I^{\rm err}_{{\rm c}\maltese}(i-1)+[1-I^{\rm err}_{{\rm d}\maltese}(i-1)-I^{\rm err}_{{\rm c}\maltese}(i-1)]\right\}}_{\lambda}\\ &+\frac{\delta_{i-2}}{\delta_i}d(m_0)\widehat{I}^{\rm err}(i).\end{aligned}$$

By (9.22), $\frac{\delta_{j-1}}{\delta_j}\le\gamma^{m_0}\varkappa^{-2m_0}\Rightarrow\frac{\delta_{i-2}}{\delta_i}\le\gamma^{2m_0}\varkappa^{-4m_0}$. Due to the definition of $\widehat{I}^{\rm err}(i)$ from Corollary 9.4,

$$\lambda\le\left(\frac{\gamma}{\varkappa}\right)^{m_0}\widehat{I}^{\rm err}(i)+1-\widehat{I}^{\rm err}(i),\qquad I(i)[1-\widehat{I}^{\rm err}(i)]=0,$$

for $I:=I^{\rm err}_0,I^{\rm err}_{{\rm d}\maltese},I^{\rm err}_{{\rm c}\maltese}$. Hence

$$\begin{aligned}\xi_i&\le\Bigl[\xi_{i-1}\rho I_{\maltese}(i)+\varkappa^{m_0}I_0(i)\Bigr]\Bigl[1-\widehat{I}^{\rm err}(i)\Bigr]\\ &\quad+\Bigl[\xi_{i-1}\rho\bigl\{I_{\maltese}(i)+I^{\rm err}_{{\rm c}\maltese}(i)\bigr\}+\varkappa^{m_0}\bigl\{I^{\rm err}_0(i)+I^{\rm err}_{{\rm d}\maltese}(i)+I_0(i)\bigr\}\Bigr]\\ &\quad\times\left(\frac{\gamma}{\varkappa}\right)^{m_0}\widehat{I}^{\rm err}(i)+\gamma^{2m_0}\varkappa^{-4m_0}d(m_0)\widehat{I}^{\rm err}(i)\\ &\le\Bigl[\xi_{i-1}\rho I_{\maltese}(i)+\varkappa^{m_0}I_0(i)\Bigr]\Bigl[1-\widehat{I}^{\rm err}(i)\Bigr]\\ &\quad+\Bigl\{\bigl[\xi_{i-1}\rho+\varkappa^{m_0}\bigr]\times\left(\frac{\gamma}{\varkappa}\right)^{m_0}+\gamma^{2m_0}\varkappa^{-4m_0}d(m_0)\Bigr\}\widehat{I}^{\rm err}(i).\end{aligned}$$

Here $\rho<1$ by (9.6) and (9.25), and $\varkappa<1$. So the factor multiplying $\widehat{I}^{\rm err}(i)$ in the last summand does not exceed

$$(\xi_{i-1}+1)\left(\frac{\gamma}{\varkappa}\right)^{m_0}+\left(\frac{\gamma}{\varkappa^2}\right)^{2m_0}d(m_0)\le\left(\frac{\gamma}{\varkappa^2}\right)^{2m_0}[1+d(m_0)](\xi_{i-1}+1)\overset{(9.25)}{=\!=}b/2(\xi_{i-1}+1).$$

Summarizing, we arrive at (8.27) with $I^{\rm err}(i):=\widehat{I}^{\rm err}(i)$. □

COROLLARY 9.7. *The indicator function $I(i)\longleftrightarrow\widehat{I}^{err}(i)=1\vee I_{\xi>1}(i-1)=1$ almost surely obeys the inequality*

$$\varlimsup_{r\to\infty}\frac{1}{r}\sum_{i=1}^{r}I(i)\le\overline{p}:=3\cdot2^{-Fm_0}\left\{2+\frac{\log b}{\log[\rho^{-1}]}\right\}. \tag{9.26}$$

Indeed, thanks to Lemma 9.6 and Observation 8.1,

$$\varlimsup_{r\to\infty}\frac{1}{r}\sum_{i=1}^{r}I(i)\le\left\{2+\frac{\log b}{\log[\rho^{-1}]}\right\}\times\varlimsup_{r\to\infty}\frac{1}{r}\sum_{i=1}^{r}\widehat{I}^{\rm err}(i)\overset{(9.19)}{\Longrightarrow}(9.26).$$

LEMMA 9.8. *Under the assumptions of Proposition* 9.2*, the coder and decoder stabilize the system almost surely if*

(9.27)
$$\omega:=\log_2[\varkappa^{-1}]-2\overline{p}\big\{\log_2\gamma+\log_2[\varkappa^{-1}]\big\}>0\quad\text{and}\quad\chi:=\omega(1-\overline{p})-\overline{p}\log_2\|A\|>0.$$

Proof. The symbol c (with a possible index) will be used to denote random constants independent of i and r. For any $\alpha>0$, (9.26) implies $S(r):=\sum_{i=1}^{r}I(i)\le r(\overline{p}+\alpha)$ for $r\approx\infty$. Since $\varkappa<1<\gamma$, (8.22) and (9.22) yield

$$\begin{aligned}\delta_i&\le\delta_{i-1}\left\{\varkappa^{m_0}[1-I(i)]+\left(\frac{\gamma^2}{\varkappa}\right)^{m_0}I(i)\right\}\quad\forall i\ge1\Rightarrow\delta_r\\&\le\delta_0\varkappa^{rm_0}\prod_{i=1}^{r}\left(\frac{\gamma}{\varkappa}\right)^{2m_0I(i)}=\delta_0\varkappa^{rm_0}\left(\frac{\gamma}{\varkappa}\right)^{2m_0S(r)}\\&\overset{r\approx\infty}{\le}\delta_0\varkappa^{rm_0}\Big(\frac{\gamma}{\varkappa}\Big)^{2rm_0(\overline{p}+\alpha)}=\delta_02^{-rm_0\omega_\alpha},\\&\qquad\text{where }\omega_\alpha:=\log_2[\varkappa^{-1}]-2[\overline{p}+\alpha]\big\{\log_2\gamma+\log_2[\varkappa^{-1}]\big\}\xrightarrow{\alpha\to0}\omega>0.\end{aligned}$$

Thus for $\alpha\approx0$, we have $\omega_\alpha>0$ and

$$\delta_i\le c'2^{-im_0\omega_\alpha}\to0\quad\text{as}\quad i\to\infty.\tag{9.28}$$

This, along with (9.11), (9.12), (9.14), and (9.21), implies

$$u_c(t)\to0,\quad u(t)=u_d(t)\to0\qquad\text{as}\quad t\to\infty.\tag{9.29}$$

In particular, the second relation from (3.3) holds. To prove the first one, we note that due to (9.23) and (9.28)

$$\begin{aligned}z_i\le z_{i-1}\|A\|^{m_0}I(i)+c''[\delta_{i-1}+\delta_{i-2}]\le z_{i-1}\|A\|^{m_0}I(i)+c2^{-im_0\omega_\alpha}\qquad\forall i\ge2;\\z_r\le z_1\prod_{i=2}^{r}\Big[\|A\|^{m_0}I(i)\Big]+c\sum_{i=2}^{r}2^{-im_0\omega_\alpha}\prod_{j=i+1}^{r}\Big[\|A\|^{m_0}I(j)\Big].\end{aligned}$$

The first relation from (9.27) implies $\overline{p}<1$. So $\{i\ge2:I(i)=1\}\ne\{i=2,3,\dots\}$ due to (9.26). It follows that for $r\approx\infty$, the first summand vanishes and

$$z_r\le c\sum_{i=r-l}^{r}2^{-im_0\omega_\alpha}\|A\|^{m_0(r-i)},$$

where $\{r-l+1,\dots,r\}$ is the largest subinterval of the set $\mathfrak{I}_r:=\{2\le i\le r:I(i)=1\}$ containing r. (If $r\notin\mathfrak{I}_r$, then $l:=0$.) We proceed by taking into account the inequality

$l \le \sum_{i=1}^{r} I(i) = S(r) \le r(\overline{p}+\alpha)$ for all $r \approx \infty$:

$$\begin{aligned} z_r &\le c2^{-rm_0\omega_\alpha} \sum_{i=r-l}^{r} 2^{(r-i)m_0\omega_\alpha} \|A\|^{m_0(r-i)} = c2^{-rm_0\omega_\alpha} \sum_{i=0}^{l} \bigl(2^{\omega_\alpha}\|A\|\bigr)^{m_0 i} \\ &\le c\frac{2^{-rm_0\omega_\alpha}\bigl(2^{\omega_\alpha}\|A\|\bigr)^{m_0 l}}{1-\bigl(2^{\omega_\alpha}\|A\|\bigr)^{-m_0}} \\ &\overset{r\approx\infty}{\le} c\frac{2^{-rm_0\omega_\alpha}\bigl(2^{\omega_\alpha}\|A\|\bigr)^{m_0 r(\overline{p}+\alpha)}}{1-\bigl(2^{\omega_\alpha}\|A\|\bigr)^{-m_0}} = \frac{c}{1-\bigl(2^{\omega_\alpha}\|A\|\bigr)^{-m_0}} 2^{-rm_0\chi_\alpha}, \\ &\qquad \text{where} \quad \chi_\alpha := \omega_\alpha[1-(\overline{p}+\alpha)] - (\overline{p}+\alpha)\log_2\|A\| \xrightarrow[\alpha\to 0]{(9.27)} \chi > 0. \end{aligned}$$

Thus $\chi_\alpha > 0$ for $\alpha \approx 0$. So $z_r \overset{(9.21)}{=\!=} |\widehat{x}_c(\tau_{r+1})| \to 0$ as $r \to \infty$. This and (9.7), (9.29) yield $A^{m_0}x(\tau_r) \to 0$ as $r \to \infty$. Since the matrix A has no stable modes, the matrix A^{-m_0} is well defined, and so $x(\tau_r) \to 0$ as $r \to \infty$. To obtain the first relation from (3.3), it remains to note that for $\tau_r \le t < \tau_{r+1} = \tau_i + m_0$,

$$|x(t)| = \left| A^{t-\tau_r}x(\tau_r) + \sum_{j=\tau_r}^{t-j} A^{t-1-j}Bu_d(j) \right| \le \|A\|^{m_0} \left(|x(\tau_r)| + \|B\| \sum_{j=\tau_i}^{\tau_{i+1}-1} |u_d(j)| \right)$$

and to invoke (9.29). □

Completing the proof of Proposition 9.2. By Lemma 9.8, it suffices to show that (9.27) holds whenever m_0 is large enough. Owing to (9.24) and (9.25),

$$\begin{aligned} \frac{1}{m_0}\log_2[1+d(m_0)] &= \frac{1}{m_0}\Bigl(\log_2 d(m_0) + \log_2[1+d(m_0)^{-1}]\Bigr) \\ &= 2\log_2\|A\| + \log_2\gamma + \log_2\varkappa^{-1} + \frac{1}{m_0}\log_2\left[1+\left(\frac{\varkappa}{\gamma}\right)^{m_0}\right] \\ &\quad + \frac{1}{m_0}\log_2\max\{\|\mathfrak{B}\|, |\mathfrak{C}\overline{u}_{\mathfrak{A}}|\} + \frac{1}{m_0}\log_2[1+d(m_0)^{-1}] \\ &\xrightarrow{m_0\to\infty} \Delta_\infty := 2\log_2\|A\| + \log_2\gamma + \log_2\varkappa^{-1}, \\ \frac{\log_2 b}{\log_2[\rho^{-1}]} &= \frac{1+2m_0[\log_2\gamma + 2\log_2\varkappa^{-1}] + \log_2[1+d(m_0)]}{m_0[\log_2\gamma - \log_2\|A\|]} \\ &\xrightarrow{m_0\to\infty} \frac{2[\log_2\gamma + 2\log_2\varkappa^{-1}] + \Delta_\infty}{\log_2\gamma - \log_2\|A\|}. \end{aligned}$$

This and (9.26) yield $\overline{p} \to 0$ as $m_0 \to \infty$, and we see that (9.27) does hold for $m_0 \approx \infty$. □

9.4. Completing the proof of Theorem 5.2. The arguments from subsection 9.2 and Proposition 9.2 ensure that $\mathfrak{c} > \eta(A) \Rightarrow$ (a) for systems with no stable modes. Now we consider the general case. Suppose that $\mathfrak{c} > \eta(A)$ and invoke the notations from Explanation 5.1. Since $\eta(A) = \eta(A_+)$, claim (a) is true for the system

$$x_+(t+1) = A_+x_+(t) + \pi_+Bu(t), \qquad x_+(0) := \pi_+x_0, \qquad y_+(t) = Cx_+(t), \tag{9.30}$$

where π_+ and π_- are the projectors onto L_+ parallel to L_- and vice versa, respectively. While picking the parameter m_0 in (9.6) and the alarm control sequence (9.15) for

the coder and decoder stabilizing the system (9.30), we employ the dimension n of the state of the original system. Now we apply this coder and this decoder to the primal system (3.1). In doing so, we alter the coder's steps c.1 and c.3✠1, where it identifies the state $x_+(\tau)$ for $\tau = \tau_i, \tau_i + 2n, \tau_i + 3n + 1$. Formerly this was done on the basis of the past measurements from (9.30). Now we employ the observations from (3.1). Then thanks to Assumption 4.4, it is possible to compute $\pi_+ x(\tau) = x_+(\tau)$ because the dynamics of the system (3.1) is free, $u(t) = 0$, at least n time steps before τ. It follows that

$$|\pi_+ x(t)| \to 0 \quad \text{and} \quad |u(t)| \to 0 \qquad \text{as} \quad t \to \infty \qquad \text{a.s.} \tag{9.31}$$

So to complete the proof, it suffices to show that $x_-(t) := \pi_- x(t) \to 0$ whenever (9.31) holds. To this end, we note that

$$x_-(t+1) = A_- x_-(t) + \pi_- B u(t), \qquad \|A_-^m\| \le c\rho^m, \qquad m = 0, 1, 2, \dots,$$

for some $\rho \in (0, 1)$. Hence for any given t_* and $t > t_*$, we have

$$\begin{aligned}
|x_-(t)| &= \Big| A_-^t x_-(0) + \sum_{j=0}^{t-1} A_-^{t-1-j} \pi_- B u(j) \Big| \\
&\le c\rho^t |x_-(0)| + c\|B\|\|\pi_-\| \Big[\sum_{j=0}^{t_*} \rho^{t-1-j} |u(j)| + \sum_{j=t_*}^{t-1} \rho^{t-1-j} |u(j)| \Big], \\
\varlimsup_{t\to\infty} |x_-(t)| &= c\|B\|\|\pi_-\| \varlimsup_{t\to\infty} \sum_{j=t_*}^{t-1} \rho^{t-1-j} |u(j)| \\
&\le \frac{c\|B\|\|\pi_-\|}{1-\rho} \sup_{t \ge t_*} |u(t)| \to 0 \quad \text{as} \quad t_* \to \infty,
\end{aligned}$$

where the last relation follows from (9.31). Thus (a) does hold.

The implications (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) are apparent, whereas (d) $\Rightarrow \mathfrak{c} \ge \eta(A)$ holds due to Remark 5.1.

Appendix: Proof of Lemma 6.1. The proof is prefaced by two technical facts.

LEMMA A.1. *Suppose that Assumptions* 4.1 *and* 4.2 *hold and a decoder* (2.2) *and a feedback coder* (2.3) *are taken. Then the joint distribution of the variables* x_0, $E_0^t = \{e_j\}_{j=0}^t$, $S_0^t = \{s_j\}_{j=0}^t$ *is given by*

$$\boldsymbol{P}[dx, dS_0^t, dE_0^t] = \prod_{j=0}^{t} W(s_j|e_j) \delta\Big[e_j, \mathfrak{E}_x\big(j, x, S_0^{j-1}\big)\Big] ds_j de_j \boldsymbol{P}_0(dx), \quad \text{where} \tag{A.1}$$

$$\mathfrak{E}_x[t, x_0, S_0^{t-1}] := \mathfrak{E}[t, Cx_0, \dots, CA^t x_0, S_0^{t-1}] \overset{(2.1),(2.3)}{=\!=} e(t). \tag{A.2}$$

Here $\boldsymbol{P}_0(dx)$ *is the probability distribution of* x_0, *and* $\delta(e, e') := 1$ *if* $e = e'$ *and* $\delta(e, e') := 0$ *otherwise.*

Proof. The proof will be by induction on t. For $t = 0$,

$$\begin{aligned}
\boldsymbol{P}[dx, dS_0^t, dE_0^t] &= \boldsymbol{P}[dx, ds_0, de_0] \\
&= \boldsymbol{P}[dx, ds_0|e_0] \boldsymbol{P}(de_0) \overset{\text{Assumptions 4.1, 4.2}}{=\!=\!=} \boldsymbol{P}[dx|e_0] \boldsymbol{P}[ds_0|e_0] \boldsymbol{P}(de_0) \\
\overset{\text{Assumption 4.1,}}{=\!=\!=} W(s_0|e_0) \boldsymbol{P}[dx, de_0] ds_0 &\overset{(A.2)}{=\!=} W(s_0|e_0) \delta\big[e_0, \mathfrak{E}_x(0, x)\big] ds_0 de_0 \boldsymbol{P}_0(dx);
\end{aligned}$$

i.e., (A.1) does hold for $t = 0$. Suppose that it holds for some $t = 0, 1, \dots$. Then

$$\begin{aligned}\boldsymbol{P}\big[dx, dS_0^{t+1}, dE_0^{t+1}\big] &= \boldsymbol{P}\big[dx, dS_0^t, ds_{t+1}, dE_0^t, de_{t+1}\big]\\ &= \boldsymbol{P}\big[dx, dS_0^t, ds_{t+1}, dE_0^t|e_{t+1}\big]\boldsymbol{P}\big[de_{t+1}\big]\\ \overset{\text{Assumptions 4.1, 4.2}}{=\!=\!=} &\ \boldsymbol{P}\big[dx, dS_0^t, dE_0^t|e_{t+1}\big]\boldsymbol{P}\big[ds_{t+1}|e_{t+1}\big]\boldsymbol{P}\big[de_{t+1}\big]\\ = &\ W(s_{t+1}|e_{t+1})\boldsymbol{P}\big[dx, dS_0^t, dE_0^t, de_{t+1}\big]ds_{t+1}\\ \overset{\text{(A.2)}}{=\!=} &\ W(s_{t+1}|e_{t+1})\delta\big[e_{t+1}, \mathfrak{E}_x(t+1, x, S_0^t)\big]\boldsymbol{P}\big[dx, dS_0^t, dE_0^t,\big]ds_{t+1}de_{t+1}.\end{aligned}$$

This and the induction hypothesis show that (A.1) does hold for $t = t + 1$. □

Corollary A.2. *Given a coder and decoder-estimator, we denote by* $\mathfrak{B}$ *the random event of keeping the error bounded* (2.6)*. The conditional probability of this event given* $x(0) = x$ *can be chosen so that it does not depend on the distribution of the initial state* $x(0)$*, provided Assumption* 4.2 *holds. This is true irrespective of whether or not this distribution has a probability density.*

Indeed, thanks to Lemma A.1, the conditional distribution

$$\begin{aligned}\boldsymbol{P}\big[dS_0^t|x(0) = x\big] &= \sum_{E_0^t}\prod_{j=0}^{t} ds_j W(s_j|e_j)\delta\big[e_j, \mathfrak{E}_x(j, x, S_0^{j-1})\big]\\ &= \prod_{j=0}^{t} ds_j W\big[s_j|\mathfrak{E}_x(j, x, S_0^{j-1})\big]\end{aligned}$$

does not depend on the distribution of the initial state. It remains to note that

$$\boldsymbol{P}(\mathfrak{B}|x) = \lim_{k\to\infty}\lim_{l\to\infty}\int_{\left\{S_0^l:|A^t x - \mathfrak{X}(t,S_0^t)|<k\ \forall t=0,\dots,l\right\}}\boldsymbol{P}\big[dS_0^l|x(0) = x\big].$$

Proof of Lemma 6.1. Consider a coder (2.3) and decoder (2.2) that keep the estimation error bounded with the probability better than p for the primal system (2.1). By invoking the notation $p_0(\cdot)$ from Assumption 4.3 and putting $Q_{>p} := \{x \in \mathbb{R}^n : \boldsymbol{P}(\mathfrak{B}|x) > p\}$, we get

$$\begin{aligned}p < \boldsymbol{P}(\mathfrak{B}) &= \int_{\mathbb{R}^n}\boldsymbol{P}(\mathfrak{B}|x)p_0(x)\,dx \Rightarrow \int_{Q_{>p}} p_0(x)\,dx\\ &= \boldsymbol{P}(x_0 \in Q_{>p}) > 0 \Rightarrow \exists c > 0 : \boldsymbol{P}(x_0 \in Q) > 0,\end{aligned}$$

where $Q := \{x \in Q_{>p} : p_0(x) \le c\}$. Then there exists a compact subset $\overset{\circ}{Q} \subset Q$ such that $\boldsymbol{P}[x_0 \in \overset{\circ}{Q}] = \int_{\overset{\circ}{Q}} p_0(x)\,dx > 0$ [15, sect. 134Fb]. Now we pass to the probability space related to the probability given $x_0 \in \overset{\circ}{Q}$. This evidently keeps Assumptions 4.1–4.3 true and the channel parameters $W(s|e)$ unchanged. We assume that all random variables inherit their initial notations. Note also that in the new probability space, the initial vector x_0 is almost surely bounded and has a bounded density. Hence $h(x_0) \in \mathbb{R}$.

Now we introduce the projector π_+ onto L_+ parallel to L_- and the compact set $Q_+ := \pi_+\overset{\circ}{Q} \subset L_+$, and we define the initial vector in (6.1) to be $x_0^+ := \pi_+ x_0$. This vector evidently has a bounded probability density,

$$x_0^+ \in Q_+ \quad \text{a.s.}, \tag{A.3}$$

and so the second moment of x_0 is finite. It follows that $h(x_0^+) \in \mathbb{R}$.

The multivalued function $x_+ \in Q_+ \mapsto \mathfrak{V}(x_+) := \{x_- \in L_- : x_+ + x_- \in \overset{\circ}{Q}\}$ has a closed graph $\overset{\circ}{Q}$ and so is upper-hemicontinuous. Thus there exists a single-valued measurable selector $x_+ \in Q_+ \mapsto \chi_-(x_+) \in \mathfrak{V}(x_+)$. By extending it as a measurable function on L_+ and putting $\chi(x_+) := x_+ + \chi_-(x_+)$, we get

$$x_+ \in Q_+ \Rightarrow \chi(x_+) \in \overset{\circ}{Q} \subset Q \Rightarrow \boldsymbol{P}\big[\mathfrak{B}|\chi(x_+)\big] > p. \tag{A.4}$$

Now we are in a position to transform the original coder-decoder pair (2.3), (2.2) serving the primal system into that keeping the estimation error bounded for the auxiliary system (6.1). We note first that the system (6.1) is observable thanks to Assumption 4.4. So for any $t \geq n-1$, there exists a *dead-beat observer*, i.e., a linear transformation $y(0), \dots, y(t) \overset{\mathfrak{S}_t}{\mapsto} x_+(0)$, where $y(i)$ are taken from (6.1). We define a new coder and decoder as follows. For $t = 0, \dots, n-1$, they in fact do nothing. However, for the sake of definiteness, we pick $e_* \in \mathcal{E}$ and put $\mathfrak{E}_+[t, y(0), \dots, y(t), S_0^{t-1}] := e_*$, $\mathfrak{X}_+[t, S_0^t] := 0$. For $t \geq n$, the new coder and decoder act as follows:

$$\begin{aligned} \omega &= \big[y(0), \dots, y(t), S_0^{t-1}\big] \overset{\mathfrak{S}_t}{\mapsto} x_+(0), S_0^{t-1} \mapsto \mathfrak{E}_+[t, \omega] \\ &:= \mathfrak{E}\big\{t-n, C\chi[x_+(0)], \dots, CA^{t-n}\chi[x_+(0)], S_n^{t-1}\big\}, \\ \widehat{x}_+(t) &:= \mathfrak{X}_+\big[t, S_0^t\big] := \pi_+ A^n \mathfrak{X}\big[t-n, S_n^t\big]. \end{aligned}$$

Now consider the process $\xi(t) = [x(t), y(t), e(t), s(t), \widehat{x}(t)]$, $t = 0, 1, \dots$, generated by the original coder-decoder pair in the system (2.1) when started with the initial random state $\chi[x_0^+]$. It is easy to see that $\pi_+ x(t)$, $C\pi_+ x(t)$, $e(t-n)$, $s(t-n)$, $\pi_+ A^n \widehat{x}(t-n)$ is a process generated by the new coder and decoder in the auxiliary system (6.1). Here $\widehat{x}(t) := 0$, $e(t) := e_*$ for $t < 0$, and $s(-n), \dots, s(-1)$ are mutually independent and independent of $\xi(t)$, $t = 0, 1, \dots$ random quantities, each with the distribution $W(s|e_*)$. Hence for $t \geq n$,

$$\begin{aligned} |x_+(t) - \widehat{x}_+(t)| &= |\pi_+ x(t) - \pi_+ A^n \widehat{x}(t-n)| \leq \|\pi_+\| |x(t) - A^n \widehat{x}(t-n)| \\ &= \|\pi_+\| |A^n x(t-n) - A^n \widehat{x}(t-n)| \leq \|\pi_+\| \|A^n\| |x(t-n) - \widehat{x}(t-n)|. \end{aligned}$$

So for the new coder-decoder pair and the system (6.1), the probability of keeping the estimation error bounded is no less than that for the process $\xi(t)$, $t = 0, 1, \dots$. It remains to note that the latter amounts to $\boldsymbol{EP}[\mathfrak{B}|\chi(x_0^+)]$ and is greater than p by (A.3) and (A.4).

REFERENCES

[1] K. Aström and B. Wittenmark, *Computer-Controlled Systems: Theory and Design*, Prentice–Hall, New York, 1997.

[2] J. Baillieul, *Feedback designs for controlling device arrays with communication channel bandwidth constraints*, in Proceedings of the 4th ARO Workshop on Smart Structures, 1999.

[3] J. Baillieul, *Feedback designs in information-based control*, in Stochastic Theory and Control, Lecture Notes in Control and Inform. Sci., B. Pasik-Duncan, ed., Springer, New York, 2002, pp. 35–57.

[4] T. Berger, *Rate Distortion Theory*, Prentice–Hall, Englewood Cliffs, NJ, 1971.

[5] V. Borkar and S. Mitter, *LQG control with communication constraints*, in Communications, Computation, Control, and Signal Processing: A Tribute to Thomas Kailath, Kluwer Academic Publishers, Norwell, MA, 1997, pp. 365–373.

[6] V. S. Borkar, S. K. Mitter, and S. Tatikonda, *Optimal sequential vector quantization of Markov sources*, SIAM J. Control Optim., 40 (2001), pp. 135–148.

[7] R. W. Brockett and D. Liberzon, *Quantized feedback stabilization of linear systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1279–1289.
[8] I. Csisár, *The method of types*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2505–2523.
[9] A. Dembo, Th. M. Cover, and J. A. Thomas, *Information theoretic inequalities*, IEEE Trans. Inform. Theory, 37 (1991), pp. 1501–1518.
[10] R. L. Dobrushin, *Information transmission in a channel with feedback*, Theory Probab. Appl., 34 (1958), pp. 367–383.
[11] N. Elia, *When Bode meets Shannon: Control-oriented feedback communication schemes*, IEEE Trans. Automat. Control, 49 (2004), pp. 1477–1492.
[12] N. Elia and S. Mitter, *Stabilization of linear systems with limited information*, IEEE Trans. Automat. Control, 46 (2001), pp. 1384–1400.
[13] F. Fagnani and S. Zampieri, *Stability analysis and synthesis for scalar linear systems with a quantized feedback*, IEEE Trans. Automat. Control, 48 (2003), pp. 1569–1584.
[14] R. M. Fano, *Transmission of Information, a Statistical Theory of Communication*, Wiley, New York, 1961.
[15] D. H. Fremlin, *Measure Theory*, Torres Fremlin, Colchester, UK, 2000.
[16] R. G. Gallager, *A simple derivation of the coding theorem and some applications*, IEEE Trans. Inform. Theory, 11 (1965), pp. 3–18.
[17] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.
[18] S. Guiasu, *Information Theory with Applications*, McGraw–Hill, New York, 1977.
[19] J. Hespanha, A. Ortega, and L. Vasudevan, *Towards the control of linear systems with minimum bit-rate*, in Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems, 2002, http://www.ece.ucsb.edu/~hespanha/published.html.
[20] H. Ishii and T. Başar, *Remote control of LTI systems over networks with state quantization*, Systems Control Lett., 54 (2005), pp. 15–31.
[21] H. Ishii, T. Başar, and R. Tempo, *Randomized algorithms for quadratic stability quantized sampled-data systems*, Automatica, 40 (2004), pp. 839–646.
[22] H. Ishii and T. Basar, *Feedback designs in information-based control*, in Proceedings of the 15th IFAC World Congress on Automatic Control, Barcelona, Spain, 2002.
[23] H. Ishii and B. Francis, *Limited Data Rate in Control Systems with Networks*, Lecture Notes in Control and Inform. Sci. 275, M. Thoma and M. Morari, eds., Springer, Berlin, 2002.
[24] R. Jain, T. Simsek, and P. Varaiya, *Control under communication constraints*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 3209–3216.
[25] J. Körner and A. Orlitsky, *Zero-error information theory*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2207–2229.
[26] K. Li and J. Baillieul, *Robust quantization for digital finite communication bandwidth control*, IEEE Trans. Automat. Control, 49 (2004), pp. 1573–1584.
[27] Q. Ling and M. D. Lemmon, *Stability of quantized control systems under dynamic bit assignment*, IEEE Trans. Automat. Control, 50 (2005), pp. 734–740.
[28] M. Loéve, *Probability Theory*, Vol. 2, 4th ed., Springer, New York, 1978.
[29] N. C. Martins, *Information Theoretic Aspects of the Control and Mode Estimation of Stochastic Systems*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2004.
[30] N. C. Martins, M. A. Dahleh, and N. Elia, *Feedback stabilization of uncertain systems using a stochastic digital link*, in Proceedings of the 43rd IEEE Conference on Decision and Control, Atlantis, Bahamas, 2004, pp. 1889–1895.
[31] N. C. Martins, M. A. Dahleh, and N. Elia, *Feedback stabilization of uncertain systems in the presence of a direct link*, IEEE Trans. Automat. Control, 51 (2006), pp. 438–447.
[32] A. Matveev and A. Savkin, *Comments on "Control over noisy channels" and relevant negative results*, IEEE Trans. Automat. Control, 50 (2005), pp. 2105–2110.
[33] A. S. Matveev and A. V. Savkin, *Estimation and Control over Communication Networks*, Birkhäuser, Boston, to appear.
[34] A. S. Matveev and A. V. Savkin, *An analogue of Shannon information theory for networked control systems: Stabilization via a noisy discrete channel*, in Proceedings of the 43rd IEEE Conference on Decision and Control, Atlantis, Bahamas, 2004, pp. 4491–4496.
[35] A. S. Matveev and A. V. Savkin, *An analogue of Shannon information theory for networked control systems: State estimation via a noisy discrete channel*, in Proceedings of the 43rd IEEE Conference on Decision and Control, Atlantis, Bahamas, 2004, pp. 4485–4490.
[36] A. S. Matveev and A. V. Savkin, *The problem of LQG optimal control via a limited capacity communication channel*, Systems Control Lett., 53 (2004), pp. 51–64.
[37] A. S. Matveev and A. V. Savkin, *Multirate stabilization of linear multiple sensor systems via limited capacity communication channels*, SIAM J. Control Optim., 44 (2005), pp. 584–617.

[38] A. S. Matveev and A. V. Savkin, *Stabilization of stochastic linear plants via limited capacity stochastic communication channels*, in Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, CA, 2006, pp. 484–489.

[39] A. S. Matveev and A. V. Savkin, *Shannon zero error capacity in the problems of state estimation and stabilization via noisy communication channels*, Internat. J. Control, 80 (2007), pp. 241–255.

[40] G. N. Nair and R. J. Evans, *State estimation under bit-rate constraints*, in Proceedings of the 37th IEEE Conference on Decision and Control, 1998.

[41] G. N. Nair and R. J. Evans, *Stabilization with data-rate-limited feedback: Tightest attainable bounds*, Systems Control Lett., 41 (2000), pp. 49–56.

[42] G. N. Nair and R. J. Evans, *Exponential stabilisability of multidimensional linear systems with limited data rates*, in Proceedings of the 15th IFAC World Congress on Automatic Control, Barcelona, Spain, 2002.

[43] G. N. Nair and R. J. Evans, *Mean square stabilisability of stochastic linear systems with data rate constraints*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 1632–1637.

[44] G. N. Nair and R. J. Evans, *Exponential stabilizability of finite-dimensional linear systems with limited data rate*, Automatica, 39 (2003), pp. 585–593.

[45] G. N. Nair and R. J. Evans, *Stabilizability of stochastic linear systems with finite feedback data rates*, SIAM J. Control Optim., 43 (2004), pp. 413–436.

[46] J. Neveu, *Mathematical Foundations of the Calculus of Probabilities*, Holden–Day, San Francisco, 1965.

[47] I. R. Petersen and A. V. Savkin, *Multi-rate stabilization of multivariable discrete-time linear systems via a limited capacity communication channel*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 304–309.

[48] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*, Holden–Day Series in Time Series Analysis, G. M. Jenkins and E. Parzen, eds., Holden–Day, San Francisco, 1964.

[49] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, in Optimization Methods in Statistics, Academic Press, New York, 1971, pp. 233–257.

[50] A. Sahai, *Anytime Information Theory*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2001.

[51] A. Sahai and S. Mitter, *The necessity and sufficiency of anytime capacity for stabilization of a linear system over noisy communication links—Part* I: *Scalar systems*, IEEE Trans. Inform. Theory, 52 (2006), pp. 3369–3395.

[52] A. V. Savkin, *Analysis and synthesis of networked control systems: Topological entropy, observability, robustness, and optimal control*, Automatica, 42 (2006), pp. 51–62.

[53] A. V. Savkin and I. R. Petersen, *Set-valued state estimation via a limited capacity communication channel*, IEEE Trans. Automat. Control, 48 (2003), pp. 676–681.

[54] C. E. Shannon, *A mathematical theory of communication*, Bell System Tech. J., 27 (1948), pp. 379–423, 623–656.

[55] C. E. Shannon, *The zero error capacity of a noisy channel*, IEEE Trans. Inform. Theory, 2 (1956), pp. 8–19.

[56] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL, 1949.

[57] T. Simsek, R. Jain, and P. Varaiya, *Scalar estimation and control with noisy binary observation*, IEEE Trans. Automat. Control, 49 (2004), pp. 1598–1603.

[58] D. J. Stiwell and B. E. Bishop, *Platoons of underwater vehicles*, IEEE Control Syst. Mag., 20 (2000), pp. 45–52.

[59] S. Tatikonda and S. Mitter, *Control over noisy channels*, IEEE Trans. Automat. Control, 49 (2004), pp. 1196–1201.

[60] S. Tatikonda and S. Mitter, *Control under communication constraints*, IEEE Trans. Automat. Control, 49 (2004), pp. 1056–1068.

[61] S. Tatikonda, A. Sahai, and S. Mitter, *Stochastic linear control over a communication channel*, IEEE Trans. Automat. Control, 49 (2004), pp. 1549–1561.

[62] S. C. Tatikonda, *Control under Communication Constraints*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2000.

[63] S. C. Tatikonda, A. Sahai, and S. K. Mitter, *Control of LQG systems under communication constraints*, in Proceedings of the 37th IEEE Conference on Decision and Control, 1998.

[64] R. Varga, *Iterative Matrix Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1965.

[65] S. Verdú, *Fifty years of Shannon theory*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2057–

2078.

[66] H. S. Witsenhausen, *Separation of estimation and control for discrete time systems*, Proc. IEEE, 59 (1971), pp. 1557–1566.

[67] S. Yüksel and T. Başar, *Coding and control over discrete noisy forward and feedback channels*, in Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain, 2005, pp. 2517–2522.

[68] K. S. Zigangirov, *Upper bounds on the error probability for channels with feedback*, Problems Inform. Transmission, 6 (1970), pp. 87–92.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1368–1397

OPTIMAL VORTEX REDUCTION FOR INSTATIONARY FLOWS BASED ON TRANSLATION INVARIANT COST FUNCTIONALS*

K. KUNISCH† AND B. VEXLER‡

Abstract. We consider the problem of an appropriate choice of a cost functional for vortex reduction for unsteady flows described by the Navier–Stokes equations. This choice is directly related to a physically correct definition of a vortex. Therefore, we discuss different possibilities for the cost functional and analyze the resulting optimal control problems. Moreover, we present an efficient numerical realization of this concept based on space-time finite element discretization and demonstrate its behavior in some numerical experiments. It is demonstrated that the choice of cost functionals has a significant effect on the reduction of vortices.

Key words. optimal control, vortex reduction, Navier–Stokes equations

AMS subject classifications. 35Q30, 76D05, 76D55, 48J20

DOI. 10.1137/050632774

1. Introduction.

This work focuses on the choice of proper cost functionals in optimal control formulations for vortex reduction in incompressible fluids. The formalization of vorticity is still a major challenge and a subject of intense research within fluid mechanics research itself. In the context of optimal control the quantification must satisfy the additional requirement that it allows the description of vorticity as a scalar-valued functional in terms of observables of the fluid. Moreover, the mathematical properties of the functional have significant consequences for mathematical programming considerations and for the numerical realization of the resulting optimization problems.

Let us first summarize some of the cost functionals that were already used in the optimal control literature to formulate vortex reduction problems. We denote by $y(t,x)$ the velocity vector and by $p(t,x)$ the pressure of an incompressible fluid which extends over the time horizon $[0,T]$ and the spatial domain Ω. Further, let $\tilde{\Omega}$ be the subset of Ω over which vortex reduction is desired.

An intensively studied cost functional in the context of optimal control of vortex reduction is given by

$$\int_0^T \int_{\tilde{\Omega}} |\operatorname{curl} y(t,x)|^2 \, dx \, dt; \tag{1.1}$$

see, e.g., [AT, G]. One of the objections against this functional is that it is not Galilean invariant; i.e., it is not invariant under transformations of the form $\mathcal{Q}x + d\,t$ of the flow field y, where $\mathcal{Q}$ is a time-independent matrix and d is a constant vector. Another frequently used functional is of the form

$$\int_0^T \int_{\tilde{\Omega}} |y(t,x) - y_{des}(t,x)|^2 \, dx \, dt, \tag{1.2}$$

*Received by the editors May 31, 2005; accepted for publication (in revised form) February 26, 2007; published electronically September 14, 2007.
http://www.siam.org/journals/sicon/46-4/63277.html

†Institute for Mathematics, University of Graz, Heinrichstraße 36, A-8010 Graz, Austria (karl.kunisch@uni-graz.at).

‡Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenberger Straße 69, 4040 Linz, Austria (boris.vexler@oeaw.ac.at).

where y_{des} stands for a given desired flow field which contains some of the expected features of the controlled flow field without the undesired vortices. Typically y_{des} is chosen as the solution to the Stokes problem on the same flow geometry and with the same boundary conditions as those which are involved for the characterization of y. This functional is referred to as a tracking-type functional. Just like the functional in (1.1), the tracking-type functional is not Galilean invariant. From the mathematical programming point of view the functionals in (1.1), (1.2) behave quite differently, however. To explain this fact let us consider the following prototype boundary optimal control problem:

$$\begin{cases} \min J(y) + G(u) \\ \text{subject to} \\ y_t - \nu\Delta y + y \cdot \nabla y + \nabla p = f \text{ in } (0,T] \times \Omega, \\ -\text{div } y = 0 \text{ in } (0,T] \times \Omega, \\ y(0,\cdot) = y_0 \text{ on } \Omega, \\ y = u \text{ on } (0,T] \times \Gamma_c,\ y = 0 \text{ on } (0,T] \times (\partial\Omega \setminus \Gamma_c), \end{cases} \tag{1.3}$$

where $\nu > 0$, f and y_0 are given, and u denotes the control variable acting on $(0,T]\times\Gamma_c$ and satisfying $\int_{\Gamma_c} u(t)\, n\, dx = 0$, with n denoting the outer normal to $\partial\Omega$. Further J and G are real-valued functionals penalizing vorticity and control-action, respectively, with J as in (1.1) or (1.2), and $G(u) = \frac{1}{2}|u|^2$, where $|\cdot|$ denotes an appropriate norm on the control space. If u is an optimal solution to (1.3), then u, together with the associated velocity y and pressure p, satisfies the primal equations, which are the equations in (1.3), the adjoint equation

$$\begin{cases} -\lambda_t - \nu\Delta\lambda + (\nabla y)^t\lambda - (y \cdot \nabla)\lambda + \nabla\pi = J'(y) \text{ in } (0,T] \times \Omega, \\ -\text{div } \lambda = 0 \text{ in } (0,T] \times \Omega, \\ \lambda(T,\cdot) = 0 \text{ on } \Omega, \\ \lambda = 0 \text{ on } (0,T] \times \partial\Omega, \end{cases} \tag{1.4}$$

with adjoint velocity λ and adjoint pressure π, and satisfies as well the so-called optimality condition formally given by

$$\nu\frac{\partial\lambda}{\partial n} + G'(u) - \pi n = 0 \text{ on } \Gamma_c. \tag{1.5}$$

We refer to [FGH] and [HK], for example, for rigorous frameworks for boundary control of the Navier–Stokes equations. Note that the adjoint equations related to (1.1) and (1.2) differ significantly with respect to the regularity of the right-hand sides: in the former case the right-hand side involves second order derivatives of the velocity field, whereas for the tracking-type cost functional only y without derivatives appears in (1.4). Moreover, in the case when the residue between y_{des} and y at the optimal control is sufficiently small, a second order optimality condition for (1.3) with J as in (1.2) holds [HK]. It appears to be difficult to obtain conditions which

lend themselves to an intuitive interpretation and which imply second order sufficient optimality for optimal control problems involving (1.1). For second order sufficient optimality conditions related to optimal control of the Navier–Stokes equations, we also refer to [TW].

To discuss candidates for Galilean invariant measures we decompose the velocity gradient tensor ∇y as

$$\nabla y = S + \Omega,$$

where $S = \frac{1}{2}(\nabla y + (\nabla y)^t)$ is the rate of strain tensor and $\Omega = \frac{1}{2}(\nabla y - (\nabla y)^t)$ is the vorticity tensor. The fact that Ω is used for both the spatial domain and the antisymmetric part of ∇y should not create confusion. We use this notation for both since they are quite standard in the literature.

The Δ-criterion (see [CPC] and [BMC]) is based on a local phase plane analysis of

$$\dot{\xi} = A\,\xi \tag{1.6}$$

with $A = \nabla y(t, x)$. For two-dimensional systems the geometry of the trajectories in terms of the eigenvalues of A can be found in many textbooks. For the case when A is a 3×3 matrix, a detailed analysis is given in [CPC], for example. In particular, if

$$\Delta = \frac{1}{2}\left(\frac{Q}{3}\right)^3 + \left(\frac{\det \nabla y}{2}\right)^2 > 0, \tag{1.7}$$

where

$$Q = \frac{1}{2}(|\Omega|^2 - |S|^2), \tag{1.8}$$

then the characteristic equation associated with A has one real and two complex eigenvalues. Thus, the regions in $(0, T) \times \Omega$, where Δ is positive, are candidates for local instantaneous stirring. In (1.8) we denote $|\Omega|^2 = \sum_{i,j} \Omega_{ij}^2$ and similarly for $|S|$. For incompressible fluids we have

$$Q = -\frac{1}{2}\mathrm{trace}(A^2).$$

The research in [JH] contains an interesting discussion of some of the shortcomings of earlier characterizations of vortices, including (1.7), and it proposes to define vortices as regions where the second eigenvalue of the symmetric matrix $S^2 + \Omega^2$ satisfies

$$\lambda_2(S^2 + \Omega^2) < 0.$$

Under appropriate conditions this criterion guarantees an instantaneous local pressure minimum in a two-dimensional plane in a three-dimensional flow.

In the case when spatial domain Ω is two-dimensional, it can be easily verified by direct computation that the following criteria are equivalent:

(i) The smaller eigenvalue of $S^2 - \Omega^2$ is negative;
(ii) ∇y has complex eigenvalues;
(iii) $Q > 0$;
(iv) $\det \nabla y > 0$.

These considerations suggest the use of

$$\int_0^T \int_{\tilde{\Omega}} \max(0, \det \nabla y(t,x))\, dx\, dt \tag{1.9}$$

as a cost functional in vortex-reduction formulations. Note that due to the max-operation the cost functional in (1.9) is nondifferentiable and therefore, for numerical optimization routines, regularization of the max-operator may be necessary. This cost functional was used for optimal vortex reduction in a driven-cavity problem in [HKSV].

Let us return to the cost functional (1.1) involving the curl-operator and note that it can be equivalently expressed as

$$\int_0^T \int_{\tilde{\Omega}} |\Omega(t,x)|^2\, dx\, dt.$$

Vorticity, together with thresholding, has been widely used for representing vortices; see [JH] and the references given there. However, it is now well accepted as an inadequate vorticity measure, for example, in the context of boundary layers. In particular, it was shown in [Lu] that maxima and minima of $|\Omega|$ in planar wall-bounded flows occur only at the wall.

A well-known Galilean invariant measure defines the vorticity region as the domain where the vorticity tensor dominates the rate of strain tensor, i.e.,

$$Q = \frac{1}{2}(|\Omega|^2 - |S|^2) = -\frac{1}{2}(\lambda_1 + \lambda_2 + \lambda_3) > 0,$$

where λ_i are the eigenvalues of $S^2 + \Omega^2$.

This criterion was originally proposed in [O] and [W] for two-dimensional domains and investigated in [HWM] for three-dimesional domains. It is referred to as the Okubo–Weiss criterion, or Q-criterion, and readily lends itself to being used in a cost functional of the form

$$\int_0^T \int_{\tilde{\Omega}} \max(0, |\Omega(t,x)|^2 - |S(t,x)|^2)\, dx\, dt.$$

As mentioned above, the Okubo–Weiss criterion coincides with the $\det \nabla y > 0$ criterion in two dimensions.

Galilean invariant vortex criteria allow a classification which is invariant under frame changes that move at a constant speed relative to each other. In a variety of different theoretical and example-driven approaches (see, e.g., [H1], [H2], [LHK], [LKH], [TK]), it was established that Galilean invariance is not sufficient for reliable vortex identification. Rather, criteria must be invariant also under coordinate transformations of the form $\mathcal{Q}(t)x + d(t)$, where $\mathcal{Q}$ is a time-dependent orthogonal matrix and d a time-dependent velocity vector. Such transformations are called *objective* in continuum mechanics and, in particular, they allow time-dependent rotations. An alternative way to point out the deficiencies of Galilean invariant criteria is based on taking the point of view of tracer dynamics. The gradient of the tracer q satisfies

$$\frac{D\nabla q}{Dt} = -(\nabla y)^t \nabla q,$$

where $\frac{D}{Dt}$ is the material derivative and $(\nabla y)^t$ denotes the transpose of the velocity gradient tensor. Studies have shown that the acceleration gradient tensor or second

derivatives of the pressure must be considered as well; see [LKH, LHK] and the references given there. In [LKH] an objective criterion is obtained in two dimensions which defines a rotation dominated region by means of

$$|r(y,p)| > 1, \tag{1.10}$$

where

$$r(y,p) = \frac{\omega}{\sigma} - \frac{\sigma_s(p_{x_1x_1} - p_{x_2x_2}) - 2\sigma_n p_{x_1x_2}}{\sigma^{\frac{3}{2}}}, \tag{1.11}$$

and $\omega = (y_2)_{x_1} - (y_1)_{x_2}$, $\sigma_s = (y_2)_{x_1} + (y_1)_{x_2}$, $\sigma_n = (y_1)_{x_1} - (y_2)_{x_2}$, $\sigma = (\sigma_s^2 + \sigma_n^2)^{1/2}$. It can readily be used for vortex-reduction by introducing the cost functional

$$\int_0^T \int_{\tilde{\Omega}} \max(r(t,x)^2 - 1, 0)\, dx\, dt, \tag{1.12}$$

for example.

In [TK] the Okubo–Weiss criterion is reconsidered through study of the stability of fluid particles in the eigenbasis of the rate of strain tensor S. This results in the modified criterion

$$Q_s = \frac{1}{2}(|\Omega - \Omega_S|^2 - |S|^2) > 0, \tag{1.13}$$

where Ω_S is the matrix containing the time derivatives of the unit eigenvectors of S in the Lagrangian frame. This criterion is well defined and objective regardless of the spatial dimension; however, as noted in [TK], the physical principles used in deriving (1.13) are restricted to two dimensions. In the appendix of [TK] it is verified that in two dimensions the criteria (1.10) and (1.13) coincide.

An interesting new vorticity criterion [H1, H2] again departs from a stability consideration of $\xi = 0$ in (1.7). It utilizes the Lyapunov functional

$$V(t,\xi) = \frac{1}{2}\frac{d}{dt}|\xi|^2 = \xi^t S(t, x(t))\xi.$$

By incompressibility of the velocity field, it can be argued that $Z = Z(t) = \{\xi : V(t,\xi) = 0\}$ defines a cone in $\mathbb{R}^3(\mathbb{R}^2)$ separating regions with different qualitative properties of the flow of (1.6) depending on the signs of the real eigenvalues of $\nabla y(t, x(t))$. To analyze the qualitative behavior of the flow, the strain acceleration tensor $M = S_t + (\nabla S)y + S\nabla y + (\nabla y)^t S$ is defined and its restriction M_Z to Z is considered. The elliptic region is defined as the set in (t,x) space, where M_Z is indefinite or $S(t,x)$ vanishes on Z. A vortex is a bounded connected set of fluid trajectories that remain in the elliptic region. This is an objective criterion valid for two and three dimensions. It appears that further considerations are necessary regarding how to design a practical cost functional based on this vortex definition which can be used in optimal control formulations.

In this paper we shall show the practical efficiency of the objective functional (1.12) for vortex reduction. We shall further conduct a comparison among the four functionals (1.1), (1.2), (1.9), and (1.12). Optimal control problems based on these four functionals can give surprisingly different results. A comparison among different cost functionals is, at first, impeded by the following difficulty: As indicated in the prototype problem (1.3), usually a term $G(u)$ representing control costs is utilized.

Mathematically it guarantees a priori bounds on minimizing sequences for (1.3) and, subsequently, existence of a minimizer for the optimal control problem (1.3). The optimal solution depends on G and therefore, if J is taken as one of the four functionals (1.1), (1.2), (1.9), (1.12), the question must be addressed of how to eliminate the effect of the control-cost term on the solutions of these optimal control problems. Here we take the approach of eliminating G altogether. As a consequence, we have to consider atypical existence problems for optimal control problems with the Navier–Stokes equations as constraints. In fact, there are no obvious a priori bounds for the control. We can only hope for a priori bounds for y due to J. Assuming that such bounds can be obtained it is, however, unfeasible to assume that boundedness of y implies boundedness of u for most practical norms for y and u, where y and u are linked through the Navier–Stokes equations. For this reason we consider finite dimensional control spaces only. This still leaves us with interesting existence problems for optimal control problems without control costs in the functional to be minimized.

We should also note the fact that some arbitrariness remains due to the fact that vorticity criteria of the type $c(t,x) > 0$ pointwise in the space-time cylinder must be converted to scalar-valued functionals; compare (1.10) and (1.12), for example.

Let us briefly outline the following sections. Section 2 is devoted to existence results for optimal control problems with the Navier–Stokes equations as constraints. Specifically we also consider the situation without control costs, where a priori bounds on the controls can result only from the differential equation which appears as a constraint. In section 3 we discuss optimality systems for the optimal control problems under consideration. Section 4 is devoted to algorithmic aspects concerning the optimization algorithm and the space-time finite element discretization. Numerical examples for a channel flow with an obstacle are given in section 5. In section 6 (appendix) the proofs for the theorems and propositions of sections 2 and 3 are provided.

2. Optimal control problem. In this section we formulate optimal control problems for vortex reduction and discuss the existence of solutions in some prototypical cases. In order to compare different vortex descriptions leading to different cost functionals, we choose a formulation where the control variable does not explicitly enter the cost functional; i.e., we consider optimal control problems without control costs. In general, existence of a solution for such problems cannot be guaranteed. Therefore we restrict ourselves to the consideration of finite dimensional control spaces. Even then, due to nonlinearity of the state equation and possible nonconvexity of the cost functional, existence of optimal controls does not follow from standard arguments. These arguments can be employed if the cost functional is radially unbounded with respect to the control, but this is not the case in our work; see, e.g., [AT, Li]. In practice, as well, the control variables are often restricted to a finite dimensional setting.

Throughout this section we consider the optimal control problem of vortex reduction on the spacial domain $\Omega \subset \mathbb{R}^2$, with boundary $\partial\Omega$ of C^1-class, in the time interval $I = (0,T)$. The space-time cylinder is denoted by $Q = (0,T) \times \Omega$.

In order to formulate the optimal control problem we introduce the following spaces:

$$\mathcal{V} = \left\{ v \in H^1(\Omega)^2 \,:\, \operatorname{div} v = 0 \right\}, \quad \mathcal{V}_0 = \left\{ v \in H_0^1(\Omega)^2 \,:\, \operatorname{div} v = 0 \right\},$$

$$\mathcal{H} = \left\{ v \in L^2(\Omega)^2 \,:\, \operatorname{div} v = 0 \right\}, \quad H = \left\{ v \in H_0^1(\Omega)^2 \,:\, \operatorname{div} v = 0 \right\}^{-_{L^2(\Omega)^2}},$$

where $-_{L^2(\Omega)^2}$ denotes the closure in $L^2(\Omega)^2$, and $\mathcal{V}^*$ is the dual space to $\mathcal{V}$. These

spaces build a Gelfand triple $\mathcal{V} \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{V}^*$; i.e., the imbeddings $\mathcal{V} \hookrightarrow \mathcal{H}$ and $\mathcal{H} \hookrightarrow \mathcal{V}^*$ are continuous and $\mathcal{V}$ is dense in $\mathcal{H}$.

For an arbitrary space Y we use the abbreviations $L^p(Y) = L^p(0,T;Y)$, for $1 \le p < \infty$, and $C(Y) = C([0,T];Y)$. We further set

$$W = \left\{ w \in L^2(\mathcal{V}) \,:\, w_t \in L^2(\mathcal{V}_0^*) \right\}, \quad W_0 = W \cap L^2(\mathcal{V}_0), \tag{2.1}$$

$$L^2(\Omega)/\mathbb{R} = \left\{ v \in L^2(\Omega) \,:\, \int_\Omega v \, dx = 0 \right\}.$$

The space W_Σ of admissible functions appearing in the Dirichlet boundary conditions is chosen as

$$W_\Sigma = \{ \hat{g} = \tau g \,:\, g \in W \},$$

where $\tau : W \to L^2(H^{1/2}(\partial\Omega)^2)$ is the trace operator onto the lateral boundary $(0,T) \times \partial\Omega$ of the cylinder Q; see [HK].

As motivated in the introduction, we choose a finite dimensional control space $U \cong \mathbb{R}^n$ $(n \in \mathbb{N})$ and consider a control operator $\hat{B} \in \mathcal{L}(U, W_\Sigma)$. Then $\hat{B}$ can be expressed as

$$\hat{B}u = \sum_{i=1}^{n} u_i \hat{\psi}_i, \quad \hat{\psi}_i = \tau\psi_i, \quad \text{with } \psi_i \in W .$$

Throughout we assume the operator $\hat{B}$ to be injective, i.e., the functions $\{\hat{\psi}_i\}$ are linearly independent, and that

$$\psi_i(0) \in H, \quad i = 1, 2, \dots, n.$$

The latter condition implies that $\psi_i(0) \cdot n = 0$ on $\partial\Omega$ and also will be required for the initial condition y_0 below. Our results can easily be generalized to just requiring the compatibility condition

$$\left((\hat{B}u)(0) - y_0|_{\partial\Omega} \right) \cdot n = 0 .$$

For later use, we introduce a prolongation operator $B \in \mathcal{L}(U, W)$ of the control operator $\hat{B}$ with the property

$$\tau(Bu) = \hat{B}u \quad \text{for all } u \in U .$$

This prolongation may be defined by $Bu = \sum_i u_i \psi_i$ or as in Lemma 6.2, but each prolongation satisfying the above condition is admissible.

The state equation for the velocity field $y = y(t,x)$ and pressure $p = p(t,x)$ is formulated as follows:

$$\begin{cases} y_t - \nu\Delta y + y \cdot \nabla y + \nabla p = f \text{ in } (0,T] \times \Omega, \\ -\operatorname{div} y = 0 \text{ in } (0,T] \times \Omega, \\ y(0,\cdot) = y_0 \text{ on } \Omega, \\ y = \hat{B}u \text{ on } (0,T] \times \partial\Omega. \end{cases} \tag{2.2}$$

The data are assumed to satisfy $\nu > 0$, $f \in L^2(H^{-1}(\Omega)^2)$, and $y_0 \in H$, where $H^{-1}(\Omega)$ is the dual space of $H_0^1(\Omega)$. The state equation (2.2) is understood in the distributional sense, allowing for a variational formulation for the velocity component y. The introduction of the pressure component and its regularity is discussed below in Proposition 2.1.

To introduce the weak formulation for the velocity component we define a semilinear form $\bar{a}\colon W \times W_0 \to \mathbb{R}$ by

$$\bar{a}(y,\psi) = \int_0^T \{(y_t,\psi) + \nu(\nabla y, \nabla\psi) + (y\nabla y,\psi) - (f,\psi)\}\, dt + (y(0) - y_0, \psi(0)).$$

The velocity component y is called the variational solution of (2.2) if $y \in W$ satisfies

$$y \in Bu + W_0 \ :\ \bar{a}(y)(\psi) = 0 \quad \text{for all } \psi \in L^2(\mathcal{V}_0). \tag{2.3}$$

For the state equation formulated in this setting we have the following existence result.

PROPOSITION 2.1. *For every $u \in U$ there exists a unique variational solution $y \in W$ of the state equation* (2.3) *defining a continuous solution operator $S\colon U \to W$. Moreover, there exists a distribution p fulfilling* (2.2) *such that $p = \partial_t P$ with some $P \in C(L^2(\Omega)/\mathbb{R})$. The mapping $u \mapsto P$ is continuous from U to $C(L^2(\Omega)/\mathbb{R})$.*

The proof of this proposition is given in section 6 (appendix), where it is shown that the pair (y, P) satisfies

$$y(t) - y(0) - \nu\int_0^t \Delta y(s)\, ds + \int_0^t (y(s)\cdot\nabla)y(s)\, ds + \nabla P(t) = \int_0^t f(s)\, ds$$

in $C(H^{-1}(\Omega)^2)$.

The space of all pairs $x = (y,p)$ satisfying $y \in W$ and $p = \partial_t P$ (as distribution) with some $P \in C(L^2(\Omega)/\mathbb{R})$ is denoted by X, i.e.,

$$X = \left\{(y,p) \ :\ y \in W \text{ and } p = \partial_t P,\ P \in C(L^2(\Omega)/\mathbb{R})\right\}.$$

In the following proposition a regularity result for the solution $x = (y,p)$ of (2.2) is given. It will be shown that x lies in the space

$$\mathcal{X} = L^2(H^2(\Omega)^2) \cap H^1(L^2(\Omega)^2) \times L^2(H^1(\Omega)),$$

provided that additional assumptions are satisfied. This regularity in particular allows us to interpret the pressure component p as an almost everywhere defined function rather than as only a distribution.

PROPOSITION 2.2. *If $\partial\Omega$ is of C^2-class, $f \in L^2(L^2(\Omega)^2)$, $y_0 \in \mathcal{V}_0$, $\psi_i \in W \cap L^2(H^2(\Omega)^2) \cap H^1(L^2(\Omega)^2)$, and $\psi_i(0) \in \mathcal{V}_0$, then $x \in \mathcal{X}$.*

The proof of this proposition is given in section 6 (appendix).

Under the assumptions of Proposition 2.2 a variational formulation incorporating the pressure component can be stated. To this end we introduce the space $\mathcal{X}_0$ by

$$\mathcal{X}_0 = \{(y,p) \in \mathcal{X} \ :\ \tau y = 0\}$$

and define for $x = (y, p) \in \mathcal{X}$ and $\zeta = (\psi, \xi) \in \mathcal{X}_0$ the semilinear form $a\colon \mathcal{X} \times \mathcal{X}_0 \to \mathbb{R}$ by

$$a(x)(\zeta) = \int_0^T \{(y_t, \psi) + \nu(\nabla y, \nabla \psi) + (y \nabla y, \psi) - (p, \operatorname{div} \psi) - (f, \psi) + (\operatorname{div} y, \xi)\}\, dt + (y(0) - y_0, \psi(0)).$$

We introduce a prolongation $\mathcal{B} \in \mathcal{L}(U, \mathcal{X})$ by $\mathcal{B}u = (Bu, 0)$ and state the corresponding variational formulation as follows:

$$x \in \mathcal{B}u + \mathcal{X}_0 \ : \ a(x, \zeta) = 0 \quad \text{for all } \zeta \in \mathcal{X}_0\,. \tag{2.4}$$

Under the assumptions of Proposition 2.2 the solution x satisfies the variational problem (2.4).

We are now prepared to introduce the optimization problems which will further be investigated. In the following section we shall derive optimality systems under the mild regularity requirements of Proposition 2.1 as well as under the stronger ones of Proposition 2.2. The corresponding variational formulations are given in (2.3) and (2.4), respectively.

The optimization problems are of the form

$$\text{minimize } J(x) \ \text{ subject to (2.2)}, \quad x \in X,\ u \in U, \tag{2.5}$$

where $J\colon X \to \mathbb{R}$ and the solutions to (2.2) are understood in the sense of Proposition 2.1.

We stress that due to the absence of a control cost term, one cannot use standard techniques to ensure the existence of a solution of (2.5). In the following, we first provide the existence of solutions for two choices of the functional J:

$$J_1(y) = \int_0^T \int_\Omega |y(t, x) - y_{des}(t, x)|^2\, dx\, dt, \tag{2.6}$$

$$J_2(y) = \int_0^T \int_\Omega |\operatorname{curl} y(t, x)|^2\, dx\, dt, \tag{2.7}$$

where $y_{des} \in L^2(L^2(\Omega))$ is a given desired velocity field.

THEOREM 2.3. *There exists a solution for the optimal control problem* (2.5) *for both choices of the cost functional $J = J_1$ and $J = J_2$.*

The proof of this theorem is given in section 6 (appendix).

As discussed in the previous section, the cost functionals J_1 and J_2 defined in (2.6) and (2.7) are not based on Galilean invariant or objective vortex definitions. Therefore we additionally consider the functional obtained from the Q-criterion:

$$J_3(y) = \int_0^T \int_\Omega g_3(\det \nabla y)\, dx\, dt, \tag{2.8}$$

which is Galilean invariant, and the functional based on the vortex criterion from [LKH]:

$$J_4(y, p) = \int_0^T \int_\Omega g_4(r(y, p))\, dx\, dt, \tag{2.9}$$

which is even objective. Here, $r(y,p)$ is defined as in (1.11) and the functions $g_3, g_4 \in C^2(\mathbb{R})$ are chosen as follows:

$$g_3(t) = \begin{cases} 0, & t \le 0, \\ l(t), & t > 0, \end{cases} \qquad g_4(t) = \begin{cases} l(-t-1), & t < -1, \\ 0, & -1 \le t \le 1\,, \\ l(t-1), & t > 1, \end{cases} \qquad l(t) = \frac{t^3}{t^2+1}.$$

The techniques presented in section 6 for ensuring the existence of optimal solutions for optimal control problem (2.5) with $J = J_1$ and $J = J_2$ (without the control cost term) cannot be directly applied for the cost functionals J_3 and J_4. For J_3 we obtain the following result.

THEOREM 2.4. *The optimal control problem* (2.5) *with* $J = J_3$ *and additional control constraints* $u_a \le u \le u_b$ *(*$u_a, u_b \in U$*) possesses an optimal solution.*

The proof of this theorem is given in section 6 (appendix).

Remark 2.1. The discussion of the case $J = J_4$ requires more regularity of the state variable for this cost functional to be well defined. For the required regularity, including $p \in L^2(0,T;H^2(\Omega))$, strong compatibility assumptions on the data are necessary; see, e.g., [T]. A detailed analysis for existence and optimality conditions is not within the scope of this paper.

3. Optimality system. In this section we discuss necessary optimality conditions for (2.5). The derivation is rigorous for J_1, J_2, J_3, but only formal for J_4.

In order to set up the optimality system, we introduce the adjoint equation for $z = (\lambda, \pi)$:

$$\begin{cases} -\lambda_t - \nu\Delta\lambda + (\nabla y)^t\lambda - (y\cdot\nabla)\lambda + \nabla\pi = J_y'(y,p) \text{ in } (0,T]\times\Omega, \\ -\text{div } \lambda = J_p'(y,p) \text{ in } (0,T]\times\Omega, \\ \lambda(T,\cdot) = 0 \text{ on } \Omega, \\ \lambda = 0 \text{ on } (0,T]\times\partial\Omega\,. \end{cases} \tag{3.1}$$

This equation is understood in the distributional sense, allowing for a variational formulation for the velocity component λ. The adjoint pressure π is introduced in Theorem 3.1 similarly to how the primal pressure p was introduced in Proposition 2.1.

We note that for $J = J_1, J_2, J_3$ the term J_p' vanishes and the adjoint velocity field λ is divergence-free. This is not the case for the choice $J = J_4$. For J_4, moreover, regularity beyond $x \in X$ is required to make $J_{4,y}'$ rigorous; see Remark 2.1 above. In fact, the derivatives $J_{4,y}'$ and $J_{4,p}'$ are given by

$$J_{4,y}'(x)(\delta y) = \int_0^T \int_\Omega g_4'(r(y,p))r_y'(y,p)(\delta y)\,dx\,dt,$$

$$J_{4,p}'(x)(\delta y) = \int_0^T \int_\Omega g_4'(r(y,p))r_p'(y,p)(\delta p)\,dx\,dt,$$

where $r_y'(y,p)(\delta y)$ and $r_p'(y,p)(\delta p)$ are directional derivatives of $r(y,p)$ defined in (1.11).

The following theorem ensures the existence of the solution for this adjoint equation for the choices $J = J_1, J_2, J_3$, where the velocity component λ of the adjoint state z is given in the following variational sense:

$$\lambda \in L^2(\mathcal{V}_0) \,:\, \bar{a}'(y)(\psi,\lambda) = J'(y)(\psi) \quad \text{for all } \psi \in W_0\,. \tag{3.2}$$

THEOREM 3.1. *The functionals $J = J_1, J_2, J_3$ are Gateaux differentiable on $L^2(H^1(\Omega)^2)$, and for every $x = (y, p) \in X$ there exists a unique distributional solution $z = (\lambda, \pi)$ of the adjoint equation* (3.1) *with $\lambda \in L^2(\mathcal{V}_0)$, $\lambda_t \in L^{4/3}(\mathcal{V}_0^*)$, and $\pi = \partial_t \Pi$ with $\Pi \in C(L^2(\Omega)/\mathbb{R})$. If, in addition, the assumptions of Proposition* 2.2 *are fulfilled, then $z \in \mathcal{X}_0$.*

The proof is given in section 6 (appendix). It implies that the solution of Theorem 3.1 satisfies

$$\lambda(t) - \nu \int_t^T \Delta\lambda \, ds + \int_t^T \big((\nabla y)^t \lambda - (y \cdot \nabla)\lambda\big) \, ds + \nabla\Pi(t) = \int_t^T J_y'(y) \, ds.$$

The existence of the adjoint state allows the formulation of first order optimality conditions for the problem (2.5). Due to the fact that the functionals J_i $(i = 1, 2, 3)$ do not depend on p, we can formulate the optimality system using only the velocity components y of x and λ of z, respectively:

$$y \in Bu + W_0 : \bar{a}(y)(\psi) = 0 \quad \text{for all } \psi \in L^2(\mathcal{V}_0), \tag{3.3}$$

$$\lambda \in L^2(\mathcal{V}_0) : \bar{a}'(y)(\psi, \lambda) = J'(y)(\psi) \quad \text{for all } \psi \in W_0, \tag{3.4}$$

$$u \in U_{ad} : J'(y)(B(v-u)) - \bar{a}'(y)(B(v-u), \lambda) \geq 0 \quad \text{for all } v \in U_{ad}, \tag{3.5}$$

where $U_{ad} = U$ in the case when $J = J_1$ or $J = J_2$, and $U_{ad} = \{u \in U : u_a \leq u \leq u_b\}$ in the case when $J = J_3$.

THEOREM 3.2. *Let $(u, x) \in U \times X$ be a local solution of the optimal control problem* (2.5) *for the choices $J = J_1, J_2, J_3$. Then the triple (u, x, z) fulfills the optimality system* (3.3)–(3.5)*, where $z = (\lambda, \pi)$ is the adjoint state. In the case when $U = U_{ad}$ the inequality* (3.5) *can be replaced by an equality.*

The proof is given in section 6 (appendix).

If the assumptions of Proposition 2.2 are fulfilled, then we have $x, z \in \mathcal{X}$. In this case the optimality system can be equivalently rewritten using the semilinear form $a(\cdot)(\cdot)$ involving pressure components:

$$x \in \mathcal{B}u + \mathcal{X}_0 : a(x)(\zeta) = 0 \quad \text{for all } \zeta \in \mathcal{X}_0, \tag{3.6}$$

$$z = (\lambda, \pi) \in \mathcal{X}_0 : a'(x)(\zeta, z) = J'(x)(\zeta) \quad \text{for all } \zeta \in \mathcal{X}_0, \tag{3.7}$$

$$u \in U_{ad} : J'(x)(\mathcal{B}(v-u)) - a'(x)(\mathcal{B}(v-u), z) \geq 0 \quad \text{for all } v \in U_{ad}\,. \tag{3.8}$$

Moreover, integration by parts, Green's formula, and the fact that $Bv \in W$ for all $v \in U$ imply that

$$J'(x)(\mathcal{B}(v-u)) - a'(x)(\mathcal{B}(v-u), z) = \int_0^T \int_{\partial\Omega} (\nu \nabla\lambda \cdot n) \cdot \hat{B}(v-u) \, ds \, dt\,. \tag{3.9}$$

Here, the trace $\nu\nabla\lambda \cdot n$ in (3.9) can be understood in a usual $L^2(L^2(\partial\Omega))$ sense.

Remark 3.1. On the discrete level the equality (3.9) does not hold anymore due to the lack of the appropriate formulas for integration by parts of the discretized solutions. As suggested in [V], we use the integrated residual (3.8) of the adjoint equation on the discrete level, which allows a higher order of convergence, with respect to the maximal cell-size h, than the discretization based on the boundary integral of the flux $\nu\nabla\lambda\cdot n$. A similar technique for the approximation of boundary integrals using a representation as volume integrals in the context of finite element discretization is discussed, e.g., in [GLLS].

4. Algorithmic aspects. In this section we describe a solution algorithm for optimal control problems of vortex reduction. The problem is reformulated as an unconstrained optimization problem by eliminating the state equation. Based on this formulation we describe the Newton method for solving this problem on the continuous level. Subsequently the optimization problem is discretized by space-time finite element methods. This allows a natural translation of the optimality conditions from the continuous to the discrete level due to the fact that the approaches *optimize-then-discretize* and *discretize-then-optimize* coincide for Galerkin-type discretizations. For more details of the finite element discretization of nonstationary optimal control problems we refer to [BMV].

4.1. Optimization algorithm. Before describing the discretization, we discuss the solution algorithm based on Newton's method on the continuous level. Since a finite element discretization is used, the continuous algorithm can then be simply translated into a discrete one by projection. Throughout this section, we require the assumptions of Proposition 2.2, which ensures the existence of a solution operator $S\colon U \to \mathcal{X}$ for the state equation in formulation (2.4) such that

$$S(u) \in \mathcal{B}u + \mathcal{X}_0 \,:\, a(S(u))(\zeta) = 0 \quad \text{for all } \zeta \in \mathcal{X}_0 \quad \text{for all } u \in U.$$

This gives rise to the introduction of a reduced cost functional $j\colon U \to \mathbb{R}$ by

$$j(u) = J(S(u)), \tag{4.1}$$

and allows us to reformulate the optimization problem (2.5) as an unconstrained problem

$$\text{minimize } j(u), \quad u \in U. \tag{4.2}$$

For J_3 and J_4 we should, in principle, replace U by U_{ad}. Since, for the numerical examples we consider, the inequality constraints did not become active, we do not consider U_{ad} here.

For the application of Newton's method to this optimization problem, we have to compute the derivatives of the reduced cost functional j. This is addressed in the following proposition.

PROPOSITION 4.1. *Let j be the reduced cost functional defined in* (4.1). *Its derivatives can be expressed as follows:*

(a) *For an arbitrary direction $\delta u \in U$ we have*

$$j'(u)(\delta u) = J'(x)(\mathcal{B}\delta u) - a'(x)(\mathcal{B}\delta u, z),$$

where $x = S(u)$ is the solution of the state equation (2.4) *and $z \in \mathcal{X}_0$ is the solution of the adjoint equation* (3.7).

(b) *For arbitrary directions $\delta u, \tau u \in U$ we have*

$$j''(u)(\delta u, \tau u) = J''(x)(\delta x, \mathcal{B}\tau u) - a''(x)(\delta x, \mathcal{B}\tau u, z) - a'(x)(\mathcal{B}\tau u, \delta z),$$

where $z \in X_0$ is the solution of the adjoint equation (3.7), *$\delta x \in \mathcal{X}$ is determined by the tangent equation*

$$\delta x \in \mathcal{B}\delta u + \mathcal{X}_0 \,:\, a'(x)(\delta x, \zeta) = 0 \quad \textit{for all } \zeta \in \mathcal{X}_0, \tag{4.3}$$

and $\delta z \in \mathcal{X}_0$ is the solution of the dual Hessian equation

$$\delta z \in \mathcal{X}_0 \,:\, a'(x)(\zeta, \delta z) = J''(x)(\delta x, \zeta) - a''(x)(\delta x, \zeta, z) \quad \textit{for all } \zeta \in \mathcal{X}_0. \tag{4.4}$$

The proof is similar to [HK] and [BMV].

In the following we describe the solution of the optimization problem (4.2) by Newton's method on the continuous level. Starting with an initial guess $u^0 \in U$, the next iterate u^{n+1} is computed by an update step

$$u^{n+1} = u^n + \delta u^n,$$

where δu^n solves

$$j''(u^n)(\delta u^n, v) = -j'(u^n)(v) \quad \text{for all } v \in U. \tag{4.5}$$

To solve (4.5) we use the conjugate gradient (cg) method, which requires only the evaluation of the right-hand side and of matrix-vector products. Thus we have to evaluate $j'(u^n)(v)$ and $j''(u^n)(\delta u^n, v)$ for fixed v. This can be done efficiently based on Proposition 4.1. Note that the second derivative $a''(x)$ involved in the representation of $j''(u)$ does not depend on the state x due to the quadratic structure of the Navier–Stokes equations.

Remark 4.1. For one step of the cg method, we have to solve one tangent equation (4.3) and one dual-Hessian equation (4.4). In some cases, if the dimension of U is small, it might be more efficient to build up the Hessian $\nabla^2 j(u^n)$; see [BMV] for a detailed discussion and a comparison.

4.2. Finite element discretization. In order to apply Newton's method described before, we consider a space-time finite element discretization of the optimal control problem under consideration. For the time discretization we use the dG (discontinuous Galerkin) or the cG (continuous Galerkin) method; see, e.g., [EJT].

For the time grid

$$0 = t_0 < \cdots < t_l < \cdots < t_M = T, \quad k_l = t_l - t_{l-1},$$

and a space mesh $\mathcal{T}_h$ consisting of quadrilaterals, we consider a space of spatially continuous and cellwise bilinear (biquadratic) and discontinuous in time piecewise polynomial functions of order r, X_{hk}^r. A similar space with continuous and piecewise polynomial functions in time of order s is denoted by Y_{hk}^s. The Galerkin method using X_{kh}^r as the trial and the test spaces leads to dG discretization. If the continuous in time space Y_{hk}^s is used as a trial space, this results in a cG discretization. For the detailed description of discrete equations to be solved within one step of the Newton method we refer to [BMV]. In our practical realization we use the dG(0) method, which results in a variant of the backward Euler and the cG(1) method, which is very similar to the Crank–Nicolson method. We emphasize that the space-time finite element discretization leads to the exact representation of the first and second derivatives of the discrete reduced cost functional, which is important for the convergence of the optimization algorithms. The derivation of these representations follows along the same lines as in the continuous case; cf. Proposition 4.1. The first directional derivatives of the reduced cost functional are given by

$$j'_{kh}(u)(\delta u) = J'(x_{kh})(B_h \delta u) - a'_k(x)(B_h \delta u, z_{kh}),$$

where x_{kh} and z_{kh} are the solutions of the discretized state and adjoint equations, respectively, and $a_k(\cdot,\cdot)(\cdot)$ is the discrete analogue of the semilinear form $a(\cdot,\cdot)(\cdot)$. The operator B_h is the extension of the control operator $\hat{B}$ in the discrete state space, with the property that $(B_h \delta u)(t, x_i) = 0$ for all interior nodes x_i of the mesh

$\mathcal{T}_h$. This choice leads to the fact that the integration in the above representation is done only over the cells adjacent to the boundary. We refer to [KV] for a more detailed discussion of this construction.

Remark 4.2. The use of (at least) quadratic elements for the pressure is essential for the practical realization for $J = J_4$. This is due to the fact that the second derivatives of the pressure are involved in the definition of J_4 and they must be accurately approximated by the numerical scheme in order to reliably compare the results of the four cost functionals.

Remark 4.3. The solution of the underlying state equation is required in the whole time interval for the computation of the dual, tangent, and dual-Hessian equations. If all data are stored, the storage grows linearly with respect to the number of time intervals in the time grid and also linearly with respect to the number of degrees of freedom in the space discretization. This makes the optimization procedure prohibitive for fine discretizations. This difficulty can be overcome by using storage reduction techniques known as "check-pointing" or "windowing"; see, e.g., [Gr], [BGL], and [BMV] for an application to optimization problems governed by parabolic equations.

Remark 4.4. For the examples that will be presented in the following section we use isoparametric biquadratic finite elements for the space discretization of both pressure and velocities. We add further terms to the semilinear form a in order to obtain a stable formulation with respect to both the pressure-velocity coupling and convection dominated flows. This type of stabilization technique is based on local projections of the pressure gradients (LPS method) first introduced in [BB]. In the context of optimal control problems this type of stabilization is analyzed in [RV, BV].

5. Numerical examples. In this section we discuss some numerical examples illustrating the effect of different choices of the cost functional in the context of optimal vortex reduction. For these examples we chose the dG(0) method for time and biquadratic elements for space discretization, as described in the previous section. In the context of the optimization, we use trust region techniques for globalization of the convergence; see, e.g., [NW]. The use of such techniques in the examples described below is necessary, particularly for the optimization of the cost functional J_4.

We use two different configurations, both based on the computational domain Ω; see Figure 5.1.

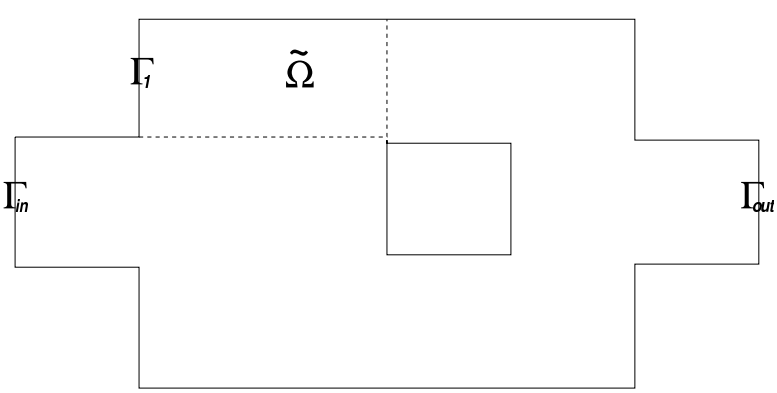


FIG. 5.1. *Computational domain.*

In both configurations we start with the following uncontrolled situation: We have constant parabolic inflow on Γ_{in}, "no-slip" boundary conditions on $\partial\Omega \setminus (\Gamma_{in} \cup \Gamma_{out})$, and "do nothing" boundary conditions on Γ_{out} (see [HRT]), i.e.,

$$\nu \nabla y \cdot n - p \cdot n = 0 \text{ on } \Gamma_{out}.$$

The flow with Reynolds number $Re \approx 10^3$ is considered on the time horizon $(0, T)$ with $T = 3$. The initial velocity field y_0 is chosen as the solution of the nonstationary Stokes equation on the same configuration after several time steps.

The solution of the uncontrolled state equation for $t = 2.4$ is shown in Figure 5.2. In this figure we observe two primary "vortex regions."

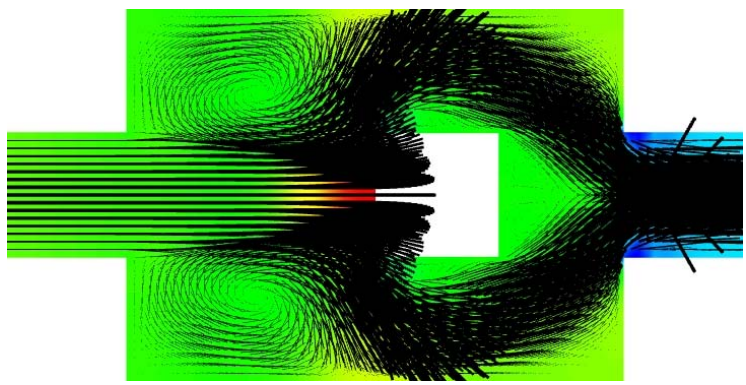


FIG. 5.2. *Uncontrolled flow, $t = 2.4$.*

In our first test we consider Dirichlet control on the part of boundary Γ_1 given as follows: $y = u\hat{y}_1$ on Γ_1, with a parabolic profile $\hat{y}_1$. The control space U is one-dimensional here, i.e., $U = \mathbb{R}$. In the following, we study the dependency of four different cost functionals J_1, J_2, J_3, and J_4 on $u \in [-8; 8]$ with observation region $\widetilde{\Omega}$ (see Figure 5.1) and the whole Ω (see Figures 5.3–5.6).

Figure 5.9 We conclude that for the present situation the vortex reduction with the help of these four cost functions leads to very different results. The reduced cost functional seems to be convex for J_1 and J_3 and to have several local extrema for the functional J_4. In our second configuration we compare the optimal solutions in more detail.

For the second configuration we set the following Dirichlet boundary conditions:

$$\begin{aligned} y &= 0 \qquad \text{on } \partial\Omega \setminus (\Gamma_{in} \cup \Gamma_{out}), \\ y &= g(u)\, \hat{y}_{in} \text{ on } \Gamma_{in}, \end{aligned}$$

where $\hat{y}_{in}$ is a fixed parabolic profile and

$$g(u)(t) = (c_0/T) + \sum_{k=1}^{n} \big(u_{2k-1} \sin(2\pi kt/T) + u_{2k} \cos(2\pi kt/T)\big).$$

The control variable u is searched for in the space $U = \mathbb{R}^{2n}$. For this setting we have for all $u \in U$

$$\int_0^T \int_{\Gamma_{in}} y \cdot n \, ds \, dt = c_0 \int_{\Gamma_{in}} \hat{y}_{in} \cdot n \, ds$$

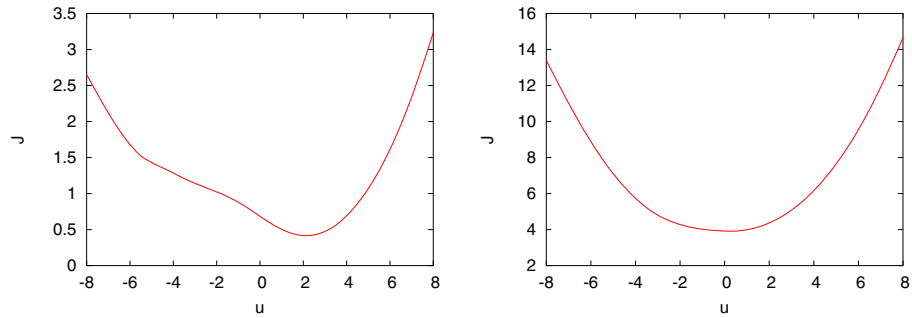


FIG. 5.3. *Cost functional* J_1 *(tracking) for* $u \in [-8, 8]$*, observation domain* $\widetilde{\Omega}$ *(left) and* Ω *(right).*

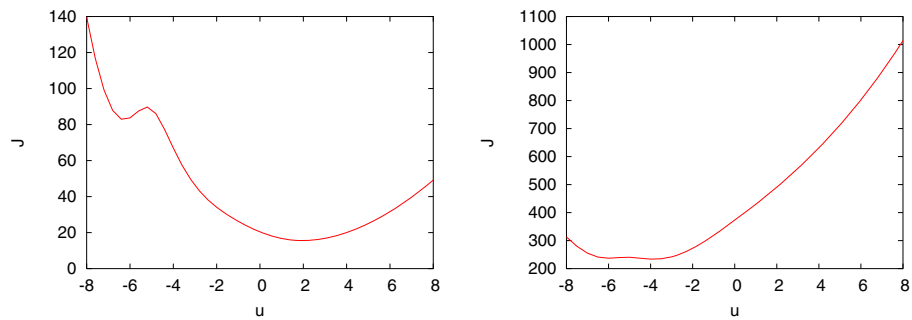


FIG. 5.4. *Cost functional* J_2 (curl) *for* $u \in [-8, 8]$*, observation domain* $\widetilde{\Omega}$ *(left) and* Ω *(right).*

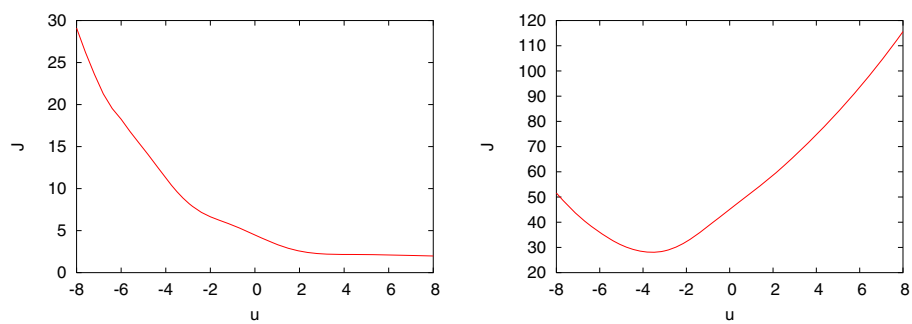


FIG. 5.5. *Cost functional* J_3 (det) *for* $u \in [-8, 8]$*, observation domain* $\widetilde{\Omega}$ *(left) and* Ω *(right).*

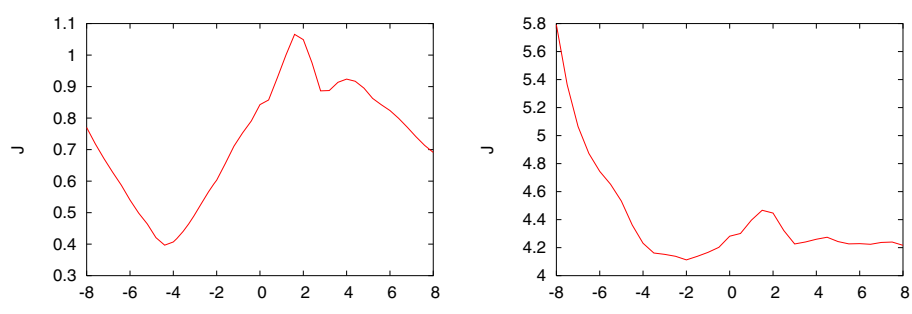


FIG. 5.6. *Cost functional* J_4 *(LKH) for* $u \in [-8, 8]$*, observation domain* $\widetilde{\Omega}$ *(left) and* Ω *(right).*

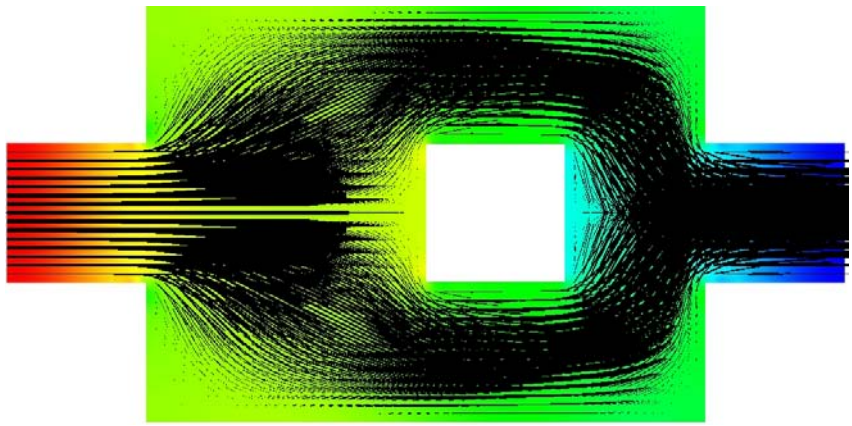


FIG. 5.7. *Stokes flow, used as the desired state for the tracking-type functional.*

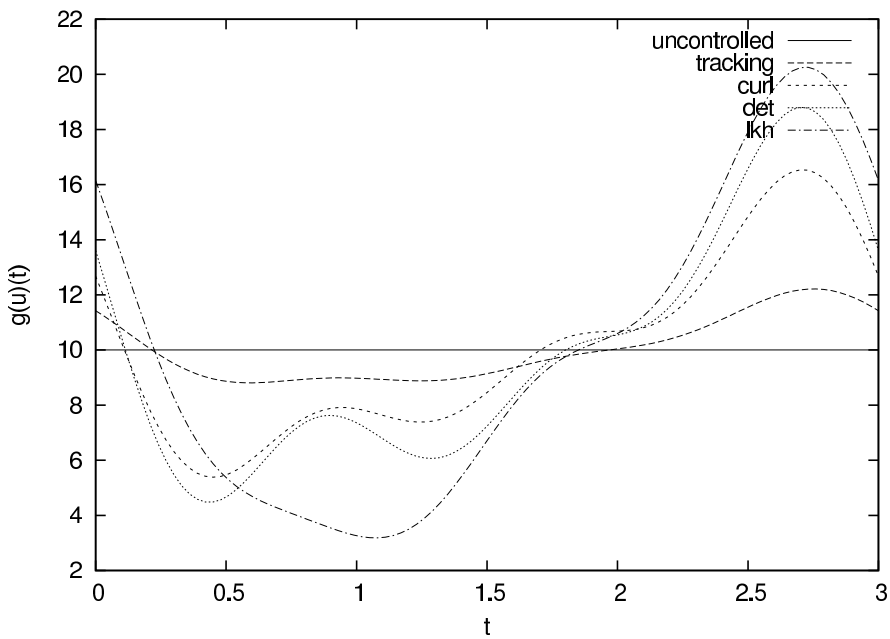


FIG. 5.8. *Optimal controls $g(u)(t)$ for four different cost functionals.*

independently of u. This condition has the following physical interpretation: The total flux through the inflow boundary in the time horizon $(0, T)$ does not depend on the control action. Thus we aim for the vortex reduction under the constraint that the total flux remains unchanged.

In Figures 5.8 and 5.9 we collect the results for the four cost functionals in the specified configuration. For the tracking-type functional we use the solution of the Stokes equation (see Figure 5.7) as the desired state y_{des}. In Figure 5.8 we show the optimal trajectories $g(u)(t)$ of the controls for the four cost functionals under consideration. In Figure 5.9 we collect the solutions of the state equation for the optimal control u with respect to the four different cost functionals J_1, J_2, J_3, J_4. It can be noted that from the point of view of graphical vortex representation there is a significant reduction of "vorticity" as we move from J_1 to J_4.

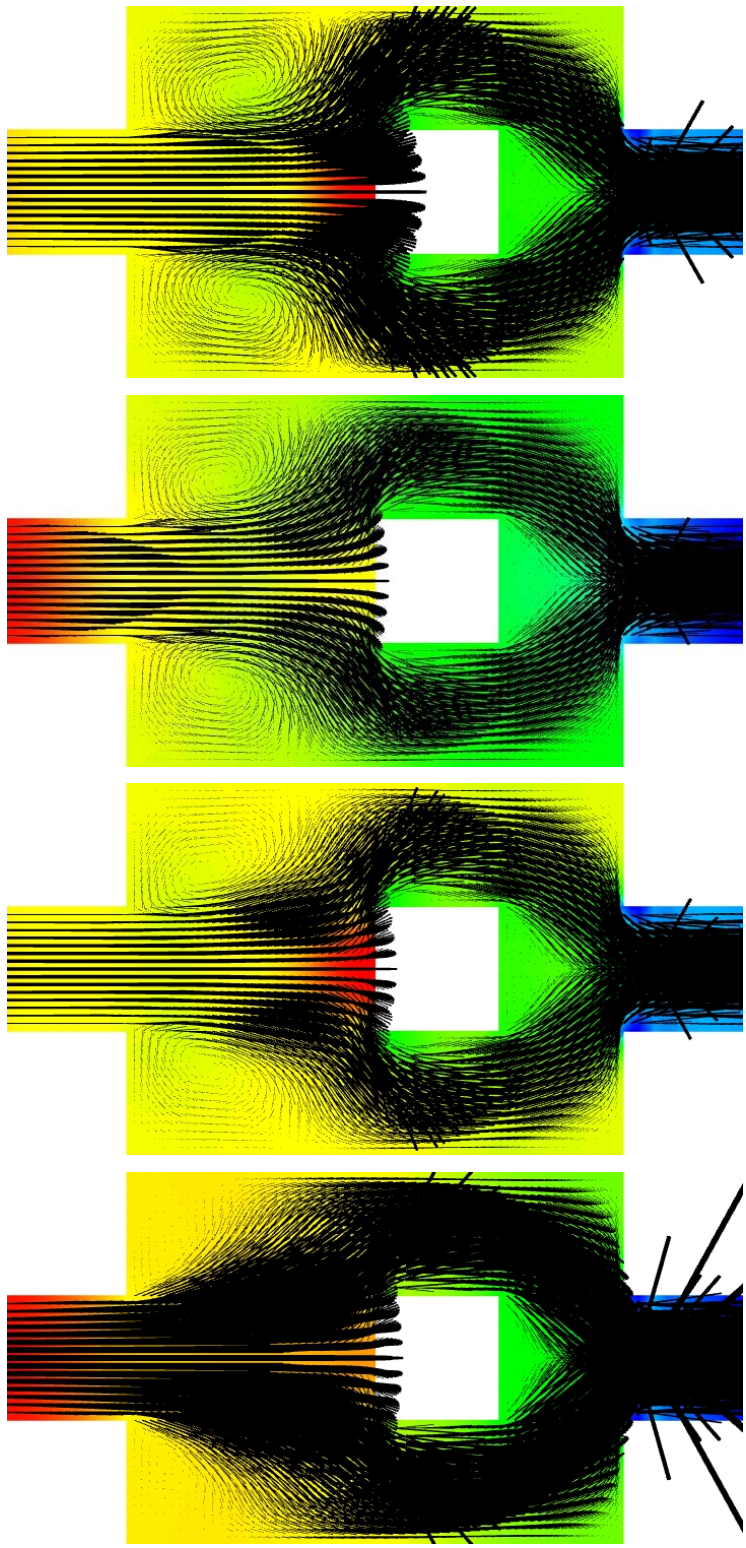


FIG. 5.9. *Optimal flow with respect to four different cost functionals (from top to bottom):* J_1 *(tracking),* J_2 (curl), J_3 (det), J_4 *(LKH).*

TABLE 5.1
Values of the four cost functionals for no control (u^0) and optimal control u^I, u^{II}, u^{III}, and u^{IV}.

	J_1	J_2	J_3	J_4
u^0	0.29608	5.60368	1.76442	0.525471
u^I	**0.14346**	3.72201	0.77298	0.350834
u^{II}	0.17125	**3.52171**	0.65593	0.390835
u^{III}	0.35245	3.79186	**0.46048**	0.422569
u^{IV}	0.18587	4.41063	0.93945	**0.181102**

In the first configuration, we have observed that the cost functionals J_2, J_3, J_4 may have local minima. Although it is impossible to check numerically, if the computed local minimum is a global one, we can provide the following results confirming our belief that we have found global minima. We denote by u^I, u^{II}, u^{III}, and u^{IV} the optimal solutions for the optimal control problems with the functionals J_1, J_2, J_3, and J_4, respectively. Moreover, we denote by $u^0 = 0$ the control which corresponds to the uncontrolled situation. In Table 5.1 we present the values of the four cost functionals J_1, J_2, J_3, and J_4 for these controls. As expected the smallest value in the first column corresponds to the optimal solution with the cost functional J_1, i.e., u^I, etc.

6. Appendix.

Proof of Proposition 2.1. We abbreviate $\|\cdot\| = \|\cdot\|_{L^2(\Omega)^2}$, $(\cdot,\cdot) = (\cdot,\cdot)_{L^2(L^2(\Omega))}$, and $\langle\cdot,\cdot\rangle = \langle\cdot,\cdot\rangle_{L^2(\mathcal{V}_0^*),L^2(\mathcal{V}_0)}$. We shall use the trilinear form

$$c(u,v,w) = \sum_{i,j=1}^{2} \int_\Omega u_j \frac{\partial u_i}{\partial x_j} w_i dx \quad \text{for } u,v,w \in H^1(\Omega)^2,$$

and the following properties of c:

(i) $c(u,v,w) = -c(u,w,v)$ for all $u \in \mathcal{V}_0$, $v,w \in H^1(\Omega)^2$;
(ii) $c(u,v,w) = -c(u,w,v)$ for all $u \in \mathcal{V}$, $v \in H^1(\Omega)^2$, $w \in H_0^1(\Omega)^2$;
(iii) $|c(u,v,w)| \le \sqrt{2}\|u\|^{1/2}\|\nabla u\|^{1/2}\,\|\nabla v\|\,\|w\|^{1/2}\,\|\nabla w\|^{1/2}$ for all $u,v,w \in H^1(\Omega)$;

see, e.g., [T].

First, we recall that $H_0^1(\Omega)^2$ can be expressed as

$$H_0^1(\Omega)^2 = \mathcal{V}_0 \oplus \mathcal{V}_0^\perp\,,$$

where

$$\mathcal{V}_0^\perp = \left\{(-\Delta_0)^{-1}\nabla q \;:\; q \in L^2(\Omega)/\mathbb{R}\right\},$$

and Δ_0 denotes the Laplacian with homogeneous Dirichlet boundary conditions; see [GR]. The forcing function $f \in L^2(H^{-1}(\Omega)^2)$ can therefore be decomposed as $f = (f_1, f_2) \in L^2(\mathcal{V}_0^*) \times L^2((\mathcal{V}_0^\perp)^*)$, and there exists $q_f \in L^2(L^2(\Omega)/\mathbb{R})$ such that

$$f_2(v) = \langle \nabla q_f, v\rangle_{L^2(H^{-1}(\Omega)^2),L^2(H_0^1(\Omega)^2)} \quad \text{for all } v \in L^2(\mathcal{V}_0^\perp)\,.$$

We abbreviate $\hat{g} = \hat{B}u$ and note that $\hat{g}(0) \in H$ since $\psi_i(0) \in H$ for $i = 1,\dots,n$. The results in [HK], specifically Theorems 1.1 and 2.1, imply the existence of a unique velocity component $y \in W$ of (2.2) in the form

$$y = y_L + y_N,$$

where

$$\text{(6.1)}\qquad \begin{cases} \langle y_{L,t}, v\rangle + \nu\,(\nabla y_L, \nabla v) = \langle f_1, v\rangle \quad \text{for all } v \in L^2(\mathcal{V}_0)\,, \\ \tau y_L = \hat{g} \text{ in } W_\Sigma\,, \\ y_L(0) = y_0 \text{ in } H, \end{cases}$$

and

$$\text{(6.2)}\qquad \begin{cases} \langle y_{N,t}, v\rangle + \nu\,(\nabla y_N, \nabla v) + \displaystyle\int_0^T c(y_N + y_L, y_N + y_L, v)dt = 0 \\ \qquad\qquad \text{for all } v \in L^2(\mathcal{V}_0), \\ \tau y_N = 0 \text{ in } W_\Sigma\,, \\ y_N(0,\cdot) = 0 \text{ in } H. \end{cases}$$

Moreover, there exists a constant K_1, and a continuous function $K_2 : \mathbb{R}^3 \to \mathbb{R}^+$, both independent of y_0, f_1, and $\hat{g}$, such that

$$\text{(6.3)}\qquad \|y_L\|_{L^2(\mathcal{V})} + \|y_{L,t}\|_{L^2(\mathcal{V}_0^*)} + \|y_L\|_{C(\mathcal{H})} \le K_1 \left(\|y_0\|_H + \|f_1\|_{L^2(\mathcal{V}_0^*)} + \|\hat{g}\|_{W_\Sigma}\right)$$

and

$$\text{(6.4)}\qquad \|y\|_{L^2(\mathcal{V})} + \|y_t\|_{L^2(\mathcal{V}_0^*)} + \|y\|_{C(\mathcal{H})} \le K_2 \left(\|y_0\|_H, \|f_1\|_{L^2(\mathcal{V}_0^*)}, \|\hat{g}\|_{W_\Sigma}\right).$$

The results in [HK] are for $f = 0$ but can easily be extended to an arbitrary $f \in L^2(H^{-1}(\Omega)^2)$.

To argue continuity of $\hat{g} \mapsto y(\hat{g})$ from W_Σ to W, let $\hat{g}_n \to \hat{g}$ in W_Σ. Then $\{\hat{g}_n\}_{n=1}^\infty$ is bounded in W_Σ, and by (6.2) there exists a constant K_3 such that

$$\text{(6.5)}\qquad \|y(\hat{g}_n)\|_W + \|y(\hat{g}_n)\|_{C(\mathcal{H})} \le K_3 \quad \text{for all } n = 1, 2, \ldots,$$

and this estimate holds for $y(\hat{g}_n)$ replaced by $y = y(\hat{g})$ as well. Let us henceforth denote $y^n = y(\hat{g}_n)$ with decomposition according to (6.1) and (6.2) as $y^n = y_L^n + y_N^n$. Further we write $y = y_L + y_N$ for $y = y(\hat{g})$. From (6.3) we have

$$\text{(6.6)}\qquad \|y_L^n - y_L\|_{L^2(\mathcal{V})} + \|y_{L,t}^n - y_{L,t}\|_{L^2(\mathcal{V}_0^*)} + \|y_L^n - y_L\|_{C(\mathcal{H})} \le K_1 \|\hat{g}^n - \hat{g}\|_{W_\Sigma}.$$

For $y_N^n - y_N$ we have the equation

$$\text{(6.7)}\qquad \begin{cases} \langle y_{N,t}^n - y_{N,t}, v\rangle + \nu\,(\nabla(y_N^n - y_N), \nabla v) \\ \quad + \displaystyle\int_0^T \left(c(y_N^n + y_L^n, y_N^n + y_L^n, v) - c(y_N + y_L, y_N + y_L, v)\right) dt = 0 \text{ for all } v \in L^2(\mathcal{V}_0)\,, \\ \tau(y_N^n - y_N) = 0\,, \\ y_N^n(0) - y_N(0) = 0. \end{cases}$$

Abbreviating $w = y_N^n - y_N \in W_0$ and setting $v = w\chi_{(0,t)}$, where $\chi_{(0,t)} : (0,T) \to \mathbb{R}$ denotes the characteristic function of $(0,t)$ for $t > 0$, using $w(0) = 0$ we find

$$\int_0^t \Big(\frac{1}{2}\frac{d}{ds}\|w(s)\|^2 + \nu\,\|\nabla w\|^2 + c(y_N^n + y_L^n, y_N^n - y_N, w) + c(y_N^n + y_L^n, y_L^n - y_L, w) \\ + c(y_N^n - y_N, y_N + y_L, w) + c(y_L^n - y_L, y_N + y_L, w) \Big)\, ds = 0\,.$$

By (ii) this implies that

$$\int_0^t \Big(\frac{1}{2}\frac{d}{ds}\|w(s)\|^2 + \nu\,\|\nabla w\|^2 + c(y^n, y_L^n - y_L, w) + c(w, y, w) + c(y_L^n - y_L, y, w) \Big)\, ds = 0,$$

and hence by (i),

$$\frac{1}{2}\|w(t)\|^2 + \nu \int_0^t \|\nabla w\|^2 \le \int_0^t \big(|c(y^n, w, y_L^n - y_L)| + |c(w, y, w)| + |c(y_L^n - y_L, w, y)| \big)\, ds.$$

By (iii), (6.5), and (6.6) there exists a constant K_4 independent of n such that

$$\begin{aligned} &\frac{1}{2}\|w(t)\|^2 + \nu \int_0^t \|\nabla w(s)\|^2\, ds \\ &\le \sqrt{2} \int_0^t \Big(\|\nabla w(s)\|\, \|\nabla(y_L^n(s) - y_L(s))\|^{\frac{1}{2}}\, \|y_L^n(s) - y_L(s)\|^{\frac{1}{2}} \big(\|y_n(s)\|^{\frac{1}{2}}\, \|\nabla y_n(s)\|^{\frac{1}{2}} \\ &\qquad + \|y(s)\|^{\frac{1}{2}}\, \|\nabla y(s)\|^{\frac{1}{2}} \big) + \|\nabla w(s)\|\, \|w(s)\|\, \|\nabla y(s)\| \Big) ds \\ &\le K_4 \int_0^t \Big(\|\nabla w(s)\|\, \|\nabla(y_L^n(s) - y_L(s))\|^{\frac{1}{2}} \big(\|\nabla y_n(s)\|^{\frac{1}{2}} + \|\nabla y(s)\|^{\frac{1}{2}} \big) \\ &\qquad + \|\nabla w(s)\|\, \|w(s)\|\, \|\nabla y(s)\| \Big)\, ds\,. \end{aligned}$$

Using Young's inequality and absorbing terms we obtain

(6.8)
$$\begin{aligned} \|w(t)\|^2 + \nu \int_0^t \|\nabla w(s)\|^2 &\le \frac{4K_4^2}{\nu} \int_0^t \|\nabla(y_L^n(s) - y_L(s))\| \big(\|\nabla y^n(s)\| + \|\nabla y(s)\| \big)\, ds \\ &\quad + \frac{2K_4^2}{\nu} \int_0^t \|w(s)\|^2\, \|\nabla y(s)\|^2 ds \le \frac{8K_4^2 K_3}{\nu}\, \|y_L^n - y_L\|_{L^2(\mathcal{V})} \\ &\quad + \frac{2K_4^2}{\nu} \int_0^t \|w(s)\|^2\, \|\nabla y(s)\|^2 ds\,. \end{aligned}$$

By Gronwall's inequality we have with $K_5 = \frac{8K_4^2 K_3}{\nu}$

$$\|w(t)\|^2 \le K_5\, \exp\left(\frac{2K_4^2}{\nu}\, \|y\|_{L^2(\mathcal{V})}^2 \right) \|y_L^n - y_L\|_{L^2(\mathcal{V})}. \tag{6.9}$$

By (6.6) the right-hand side of (6.9) converges to 0 as $n \to \infty$ and from (6.8) we deduce that $w = y_N^n - y_N \to 0$ in W_0 as $n \to \infty$. Together with (6.6) this implies that $y^n \to y$ in W which establishes the desired continuity of $g \mapsto y(g)$ from W_Σ to W. Consequently $u \to y(\hat{B}u)$ is continuous from U to W as well.

So far we worked in solenoidal spaces. To introduce the pressure component we first set

$$Y(t) = \int_0^t y(s)\,ds, \quad \hat{b}(t) = y_0 - y(t) - \int_0^t \big((y(s)\cdot\nabla)y(s) - f_1(s)\big)\,ds\,.$$

We have, using (iii), that $Y \in H^1(H^1(\Omega)^2) \hookrightarrow C(H^1(\Omega)^2)$ and $\hat{b} \in C(H^{-1}(\Omega)^2)$, and

$$\nu(\nabla Y(t), \nabla v) - \langle \hat{b}(t), v\rangle_{H^{-1}(\Omega)^2, H_0^1(\Omega)^2} = 0 \quad \text{for all } v \in \mathcal{V}_0 \text{ and each } t \in (0,T)\,.$$

Then by de Rham's theorem there exists a unique $\hat{P}(t) \in L^2(\Omega)/\mathbb{R}$ such that

$$\nu(\nabla Y(t), \nabla v) - \langle \nabla \hat{P}(t) - \hat{b}(t), v\rangle_{H^{-1}(\Omega)^2, H_0^1(\Omega)^2} = 0$$

$$\text{for all } v \in H_0^1(\Omega)^2 \text{ and each } t \in (0,T);$$

see, e.g., [DL, T]. Since the gradient operator is an isomorphism from $L^2(\Omega)/\mathbb{R}$ into $H^{-1}(\Omega)^2$ and $\hat{b}(t) + \nu\Delta Y(t) \in H^{-1}(\Omega)^2$ we conclude that $\hat{P} \in C(L^2(\Omega)/\mathbb{R})$.

Now we insert the f_2 component of f and define

$$b(t) = y_0 - y(t) - \int_0^t \big((y(s)\cdot\nabla)y(s) - f(s)\big)\,ds \quad \text{and} \quad P(t) = \hat{P}(t) + q_f(t)\,.$$

Then $P \in C(L^2(\Omega/\mathbb{R}))$ and

$$\nu(\nabla Y(t), \nabla v) - \langle \nabla P(t) - b(t), v\rangle_{H^{-1}(\Omega)^2, H_0^1(\Omega)^2} = 0$$

$$\text{for all } v \in H_0^1(\Omega)^2 \text{ and each } t \in (0,T),$$

which is equivalent to the following equality in $H^{-1}(\Omega)^2$:

$$y(t) - y(0) - \nu\int_0^t \Delta y(s)\,ds + \int_0^t (y(s)\cdot\nabla)y(s)\,ds + \nabla P(t) = \int_0^t f(s)\,ds\,. \tag{6.10}$$

This allows us to introduce pressure p as the (distributional) derivative $p = \partial_t P$.

To argue continuity of $\hat{g} \mapsto P(\hat{g})$ from W_Σ to $C(L^2(\Omega/\mathbb{R}))$, we again consider a sequence $\hat{g}^n \to \hat{g}$, $P^n = P(\hat{g}^n)$, $y^n = y(\hat{g}^n)$, and $P = P(\hat{g})$, $y = y(\hat{g})$. For a constant K_6 independent of t and n we have from (6.10)

$$\|\nabla P^n(t) - \nabla P(t)\| \le K_6\Big(\|y^n - y\|_{C(\mathcal{H})} + \|y^n - y\|_{L^2(\mathcal{V})} + \sup_{\|v\|_{L^2(H_0^1(\Omega)^2)}=1} \int_0^T |c(y^n(t), y^n(t), v(t)) - c(y(t), y(t), v(t))|\,dt\Big)\,.$$

For the last term we have, using (ii) and (iii) and the fact that the sequence $\{\|y^n\|_{L^2(\mathcal{H})}\}$ is bounded, for a constant K_7 independent of n,

$$\begin{aligned}
&\int_0^T |c(y^n(t), y^n(t), v(t)) - c(y(t), y(t), v(t))|\, dt \\
&\qquad \le \int_0^T \left(|c(y(t), v(t), y(t) - y^n(t))| + |c(y(t) - y^n(t), v, y^n(t))| \right) dt \\
&\qquad \le K_7 \int_0^T \|\nabla(y(t) - y^n(t))\|^{\frac{1}{2}} \|\nabla v(t)\| \left(\|\nabla y(t)\|^{\frac{1}{2}} + \|\nabla y^n(t)\|^{\frac{1}{2}} \right) dt \\
&\qquad \le \sqrt{2}\, K_7 \left(\int_0^T \|\nabla(y(t) - y^n(t))\| \left(\|\nabla y(t)\| + \|\nabla y^n(t)\| \right) dt \right)^{\frac{1}{2}} \\
&\qquad \le \sqrt{2}\, K_7 \|y - y^n\|^{\frac{1}{2}}_{L^2(\mathcal{V})} \left(\|y\|^{\frac{1}{2}}_{L^2(\mathcal{V})} + \|y^n\|^{\frac{1}{2}}_{L^2(\mathcal{V})} \right) \to 0 \text{ for } n \to \infty .
\end{aligned}$$

This proves that $\nabla P^n(t) \to \nabla P(t)$ in $H^{-1}(\Omega)^2$ uniformly in $t \in [0, T]$, and therefore $P^n \to P$ in $C(L^2(\Omega)/\mathbb{R})$. □

Proof of Proposition 2.2. We recall the following two additional properties of $c(\cdot, \cdot, \cdot)$.

(iv) For all $u \in H^1(\Omega)^2, v \in H^2(\Omega)^2, w \in L^2(\Omega)^2$ there holds

$$|c(u, v, w)| \le c_4 \|u\|^{\frac{1}{2}} \|u\|^{\frac{1}{2}}_{H^1(\Omega)^2} \|v\|^{\frac{1}{2}}_{H^1(\Omega)^2} \|v\|^{\frac{1}{2}}_{H^2(\Omega)^2} \|w\|.$$

(v) For all $u \in H^2(\Omega)^2, v \in H^1(\Omega)^2, w \in L^2(\Omega)^2$,

$$|c(u, v, w)| \le c_5 \|u\|^{\frac{1}{2}} \|u\|^{\frac{1}{2}}_{H^2(\Omega)^2} \|v\|_{H^1(\Omega)^2} \|w\|;$$

see, e.g., [T].

We recall the decomposition

$$L^2(\Omega)^2 = H \oplus H^\perp, \quad H^\perp = \{\nabla q \ : \ q \in H^1(\Omega)\};$$

see, e.g., [GR]. The forcing function $f \in L^2(L^2(\Omega)^2)$ can therefore be decomposed as $f = (f_1, f_2) \in L^2(H) \times L^2(H^\perp)$.

We consider the decomposition $y = y_L + y_N$, where y_L fulfills (6.1) and y_N fulfills (6.2), respectively. Due to the fact that $\psi_i(0) \in V_0$ and $\psi_i \in W \cap L^2(H^2(\Omega)^2) \cap H^1(L^2(\Omega)^2)$ for all i, we have

$$\hat{g} = \tau g, \quad \text{with } g \in W \cap L^2(H^2(\Omega)^2) \cap H^1(L^2(\Omega)^2), \quad g(0) \in \mathcal{V} .$$

We consider $w = y_L - g$ fulfilling

$$\begin{cases}
\langle w_t, v \rangle + \nu\, (\nabla w, \nabla v) = \langle f_1, v \rangle + \langle g_t, v \rangle + \nu\, (\nabla g, \nabla v) \quad \text{for all } v \in L^2(\mathcal{V}_0)\,, \\
\tau w = 0 \text{ in } W_\Sigma\,, \\
w(0) = y_0 - g(0) \text{ in } H\,.
\end{cases}$$

Using the regularity of g we obtain that

$$\langle f_1, v \rangle + \langle g_t, v \rangle + \nu\, (\nabla g, \nabla v)$$

is a linear continuous functional on $L^2(L^2(\Omega)^2)$ and $w(0) \in \mathcal{V}_0$. Therefore using a regularity result for Stokes equations with homogeneous Dirichlet boundary conditions (see, e.g., [DL]), we conclude $w \in L^2(H^2(\Omega)^2) \cap H^1(L^2(\Omega)^2) \cap C(H^1(\Omega)^2)$, $y_L \in L^2(H^2(\Omega)^2) \cap H^1(L^2(\Omega)^2) \cap C(H^1(\Omega)^2)$, and

$$\|y_L\|_{L^2(H^2(\Omega)^2)} + \|y_L\|_{H^1(L^2(\Omega)^2)} + \|y_L\|_{C(H^1(\Omega)^2)} \le C_1\big(\|f_1\|_{L^2(L^2(\Omega)^2)} + \|g\|_{H^1(L^2(\Omega)^2)} + \|g\|_{L^2(H^2(\Omega)^2)}\big), \tag{6.11}$$

with a constant C_1 dependent on f and g. To argue the corresponding result for y_N, we derive an a priori estimate for y_N in $L^2(H^2(\Omega)^2) \cap H^1(L^2(\Omega)^2) \cap C(H^1(\Omega)^2)$ using the fact that y_L satisfies (6.11). Then, the existence of a solution with asserted regularity can be obtained using a standard Galerkin procedure; see, e.g., [T].

We use $v = \chi_{(0,t)}\,\Delta y_N$ as a test function in (6.2), where $\chi_{(0,t)}$ is the characteristic function of $(0,t)$, for $t > 0$, and obtain

$$\begin{aligned}\frac{1}{2}\|\nabla y_N(t)\|^2 + \nu \int_0^t \|\Delta y_N(s)\|^2\,ds \\ \le \int_0^t \big(\,|c(y_N(s), y_N(s), \Delta y_N(s))| + |c(y(s), y_L(s), \Delta y_N(s))| \\ + |c(y_L(s), y_N(s), \Delta y_N(s))|\,\big)\,ds.\end{aligned}$$

For the first term we obtain, using (ii) and (iv),

$$\begin{aligned}\int_0^t |c(y_N(s), y_N(s), \Delta y_N(s))|\,ds &\le c_4 \int_0^t \|y_N(s)\|^{\frac{1}{2}}\,\|\nabla y_N(s)\|\,\|\Delta y_N(s)\|^{\frac{3}{2}}\,ds \\ &\le \frac{27c_4^4}{4\nu^3}\int_0^t \|y_N(s)\|^2\,\|\nabla y_N(s)\|^4\,ds + \frac{\nu}{4}\int_0^t \|\Delta y_N(s)\|^2\,ds \\ &\le C_2 \int_0^t \|\nabla y_N(s)\|^2\,\|\nabla y_N(s)\|^2\,ds + \frac{\nu}{4}\int_0^t \|\Delta y_N(s)\|^2\,ds\end{aligned}$$

for a constant C_2, where we used the fact that $\|y_N\|_{C(H)}$ is bounded according to (6.4). For the second and third terms we have, using (iv), (v), and (6.11),

$$\begin{aligned}\int_0^t \big(|c(y(s), y_L(s), \Delta y_N(s))| + |c(y_L(s), y_N(s), \Delta y_N(s))|\big)\,ds \\ \le C_3 \int_0^t \Big(\|y(s)\|^{\frac{1}{2}}\,\|y(s)\|_{\mathcal{V}}^{\frac{1}{2}}\,\|y_L(s)\|_{\mathcal{V}}^{\frac{1}{2}} + \|y_L(s)\|^{\frac{1}{2}}\|\nabla y_N(s)\|\Big) \\ \times \|y_L(s)\|_{H^2(\Omega)^2}^{\frac{1}{2}}\|\Delta y_N(s)\|\,ds \le C_4 \int_0^t \|y(s)\|_{\mathcal{V}}\|y_L(s)\|_{H^2(\Omega)^2}ds \\ + C_4 \int_0^t \|y_L(s)\|_{H^2(\Omega)^2}\|\nabla y_N(s)\|^2 ds + \frac{\nu}{4}\int_0^t \|\Delta y_N(s)\|^2 ds,\end{aligned}$$

with some constants C_3, C_4. Absorbing terms we obtain

$$\|\nabla y_N(t)\|^2 + \nu \int_0^t \|\Delta y_N(s)\|^2 \, ds \le 2C_4 \|y\|_{L^2(V)} \, \|y_L\|_{L^2(H^2(\Omega)^2)} + \int_0^t \big(2C_2 \|\nabla y_N(s)\|^2 + 2C_4 \|y_L\|_{L^2(H^2(\Omega)^2)}\big) \|\nabla y_N(s)\|^2 \, ds \,.$$

Using Gronwall's inequality we first infer that y_N is bounded in $C(H^1(\Omega)^2) \cap L^2(H^2(\Omega)^2)$. The boundedness of $y_{N,t}$ in $L^2(L^2(\Omega)^2)$ is then obtained using (6.2). Using arguments similar to those for the introduction of the pressure in the proof of Proposition 2.1, we obtain $p \in L^2(H^1(\Omega))$. In fact $\nabla \colon H^1(\Omega)/\mathbb{R} \to H^\perp$ is a homeomorphism,

$$t \mapsto y_t - \nu \Delta y + (y \cdot \nabla) y - f_1 \in L^2(L^2(\Omega)),$$

and

$$\big(y_t(t) - \nu \Delta y(t) + (y\cdot) y(t) - f_1(t), v\big) = 0 \quad \text{for all } v \in H \text{ and a.e. } t \in (0,T).$$

Hence, there is $p_1 \in L^2(H^1(\Omega)/\mathbb{R})$ fulfilling the following equality in $L^2(\Omega)^2$:

$$y_t(t) - \nu \Delta y(t) + (y\cdot) y(t) + \nabla p_1 = f_1 \,.$$

The second component of the pressure is given through the definition of $H^\perp$, i.e., $p = p_1 + p_2$, $\nabla p_2 = f_2$. This completes the proof. ∎

In order to prove Theorem 2.3, we start with a regularity result for the Stokes equation that we need in what follows.

LEMMA 6.1. *Let $(v, s) \in X$ be the solution of the Stokes equation:*

$$\begin{cases} v_t - \nu \Delta v + \nabla s = f \ \textit{in}\ (0,T] \times \Omega, \\ -\operatorname{div} v = 0 \ \textit{in}\ (0,T] \times \Omega, \\ v(0,\cdot) = 0 \ \textit{on}\ \Omega, \\ v = 0 \ \textit{on}\ (0,T] \times \partial\Omega, \end{cases} \tag{6.12}$$

with $f \in L^q(Q)$, $q > d+2$. Then the following estimate holds:

$$\|\nabla v\|_{L^\infty(Q)} + \|v(T)\|_{L^2(\Omega)} \le c \, \|f\|_{L^q(Q)}$$

with a constant c independent of $f \in L^q(Q)$.

Proof. For the proof we introduce the Sobolev space $W_q^{k,l}(Q)$ consisting of functions whose derivatives of order $\le k$ with respect to x and of order $\le l$ with respect to t are in $L^q(Q)$. From [S2] we have the following result:

$$\|v\|_{W_q^{2,1}(Q)} \le c \, \|f\|_{L^q(Q)};$$

see also [DL]. Using an embedding theorem from [S1] we obtain for $q > d+2$,

$$\|\nabla v\|_{L^\infty(Q)} \le c \, \|v\|_{W_q^{2,1}(Q)}.$$

Moreover, the following estimate is well known (see, e.g., [DL]):

$$\|v(T)\|_{L^2(\Omega)} \le c \, \|v\|_W \le c \, \|v\|_{W_2^{2,1}(Q)}.$$

This completes the proof. ∎

In the following, we formulate two core lemmas for functionals J_1 and J_2 that will be used for proving the existence of solutions to (2.5).

LEMMA 6.2. *For a sequence $\{u_k\} \subset U$, let $(y_k, p_k) \in X$ denote the solutions of the state equation (2.2), and assume that $J_1(y_k) \leq C$ for a constant $C > 0$. Then the sequence $\{u_k\}$ is bounded in U.*

Proof. We introduce prolongation $\psi_i \in W$ of the functions $\hat{\psi}_i \in W_\Sigma$, which define $\hat{B}$, by means of the Stokes equations:

$$\begin{cases} (\psi_i)_t - \nu\Delta\psi_i + \nabla\zeta_i = 0 \text{ in } (0,T] \times \Omega, \\ -\text{div } \psi_i = 0 \text{ in } (0,T] \times \Omega, \\ \psi_i(0,\cdot) = 0 \text{ on } \Omega, \\ \psi_i = \hat{\psi}_i \text{ on } (0,T] \times \partial\Omega. \end{cases} \tag{6.13}$$

This allows us to define a prolongation $B\colon U \to W$ of the control operator $\hat{B}$ by means of

$$Bu = \sum_{i=1}^{n} u_i \psi_i$$

and the corresponding operator for the pressure $R\colon U \to L^2(L^2(\Omega)/\mathbb{R})$ by

$$Ru = \sum_{i=1}^{n} u_i \zeta_i.$$

Note that the family $\{\psi_i\}$ is linearly independent in W. Next, we set

$$z_k = y_k - Bu_k, \quad r_k = p_k - Ru_k.$$

These variables satisfy the following equations:

$$\begin{cases} (z_k)_t - \nu\Delta z_k + \nabla r_k = f - y_k \cdot \nabla y_k \text{ in } (0,T] \times \Omega, \\ -\text{div } z_k = 0 \text{ in } (0,T] \times \Omega, \\ z_k(0,\cdot) = y_0 \text{ on } \Omega, \\ z_k = 0 \text{ on } (0,T] \times \partial\Omega. \end{cases} \tag{6.14}$$

We proceed by showing that $\{z_k\}$ is bounded in $L^{q'}(Q)$, where

$$\frac{1}{q'} + \frac{1}{q} = 1, \quad q > d + 2.$$

For this purpose we consider the following "dual" equation for an arbitrary function $\xi \in L^q(Q)$: Find $(v, s) \in X$ such that

$$\begin{cases} -v_t - \nu\Delta v - \nabla s = \xi \text{ in } (0,T] \times \Omega, \\ -\text{div } v = 0 \text{ in } [0,T) \times \Omega, \\ v(T,\cdot) = 0 \text{ on } \Omega, \\ v = 0 \text{ on } [0,T) \times \partial\Omega. \end{cases} \tag{6.15}$$

From Lemma 6.1 we obtain

$$\|\nabla v\|_{L^\infty(Q)} + \|v(0)\|_{L^2(\Omega)} \le c\,\|\xi\|_{L^q(Q)}\,. \tag{6.16}$$

Using (6.14) for z_k and (6.15) for v we obtain

$$\begin{aligned}\int_0^T (z_k,\xi)\,dt &= \int_0^T \{(-v_t,z_k) + \nu(\nabla v,\nabla z_k)\}\,dt \\ &= \int_0^T \{(v,(z_k)_t) + \nu(\nabla v,\nabla z_k)\}\,dt + (v(0),z_k(0)) - (v(T),z_k(T)) \\ &= \int_0^T (f,v)\,dt - \int_0^T (y_k\cdot\nabla y_k,v)\,dt + (v(0),y_0).\end{aligned}$$

For the second term in the last expression we have

$$\Big|-\int_0^T (y_k\cdot\nabla y_k,v)\,dt\Big| = \Big|\int_0^T (y_k\cdot\nabla v,y_k)\Big| \le \|\nabla v\|_{L^\infty(Q)}\,\|y_k\|^2_{L^2(Q)}.$$

Using (6.16) we obtain

$$\Big|\int_0^T (z_k,\xi)\,dt\Big| \le c\left(\|f\|_{L^2(V^*)} + \|y_k\|^2_{L^2(Q)} + \|y_0\|_{L^2(\Omega)}\right)\|\xi\|_{L^q(Q)}\,.$$

Due to the fact that $J_1(y_k)$ is bounded, we have

$$\|z_k\|_{L^{q'}(Q)} = \sup_{\xi\in L^q(Q)} \frac{\int_0^T (z_k,\xi)\,dt}{\|\xi\|_{L^q(Q)}} \le C,$$

with a generic positive constant C. Due to $1 < q' < 2$, this implies that

$$\|Bu_k\|_{L^{q'}(Q)} \le \|z_k\|_{L^{q'}(Q)} + \|y_k\|_{L^{q'}(Q)} \le C + \|y_k\|_{L^2(Q)} \le C.$$

Since B is an injective mapping to $L^{q'}(Q)$, it follows that $\{u_k\}$ is bounded in U. This completes the proof. ∎

LEMMA 6.3. *For a sequence $\{u_k\}\subset U$, let $(y_k,p_k)\in X$ denote the solutions of the state equation* (2.2)*, and assume that $J_2(y_k)\le C$, with a constant $C\in\mathbb{R}_+$. Then the sequence $\{u_k\}$ is bounded in U.*

The proof uses the same techniques as those for Lemma 6.2.

Using Proposition 2.1, Lemma 6.2, and Lemma 6.3 we are now able to prove Theorem 2.3.

Proof of Theorem 2.3. Let $\mathcal{A}\subset X\times U$ be a set of admissible pairs:

$$\mathcal{A} = \{((y,p),u)\in X\times U\,:\,(y,p)\in X \text{ fulfills the state equation (2.2)}\}\,.$$

From Proposition 2.1 we have that $\mathcal{A}$ is not empty, and due to boundedness from below the functionals J_1 and J_2, we obtain in both cases the existence of a nonnegative real number J^* with

$$J^* = \inf_{((y,p),u)\in\mathcal{A}} J(y)$$

and a sequence $\{((y_k, p_k), u_k)\} \subset \mathcal{A}$ with

$$\lim_{k\to\infty} J(y_k) = J^*.$$

Therefore $J(y_k)$ is bounded, and using Lemma 6.2 (respectively, Lemma 6.3) we obtain that $\{u_k\}$ is bounded as well. Choosing a subsequence u_{k_l} we have

$$u_{k_l} \to u^* \in U.$$

We set $(y^*, p^*) = S(u^*)$, and due to Proposition 2.1 we obtain

$$J^* = \lim_{l\to\infty} J(y_{k_l}) = J(y^*).$$

This completes the proof. □

Proof of Theorem 2.4. The functional J_3 is well defined and continuous on $L^2(\mathcal{V})$. The reduced cost functional $j_3 \colon U \to \mathbb{R}$ is defined as $j_3(u) = J_3(S(u))$, where S is the (continuous) solution operator of the state equation (2.2); see Proposition 2.1. Thus, j_3 is continuous as well.

Let $\{u_n\} \subset U$ be a minimizing sequence, i.e.,

$$J^* = \inf_{u\in U,\ u_a\le u\le u_b} j_3(u), \quad j_3(u_n) \to J^*.$$

Due to the facts that $\{u_n\}$ is bounded and U is finite dimensional, there exists a subsequence converging to $u^* \in U$. Continuity of j_3 completes the proof. □

Proof of Theorem 3.1. The functionals $J_i \colon L^2(H^1(\Omega)^2) \to \mathbb{R}$ $(i = 1, 2, 3)$ are continuous. Moreover, for any $y \in L^2(H^1(\Omega)^2)$ and $\psi \in L^2(H^1(\Omega)^2)$, there exist the directional derivatives $J_i'(y)(\psi)$ given by

$$J_1'(y)(\psi) = 2\int_0^T \int_\Omega (y - y_{des})\psi \, dx \, dt,$$

$$J_2'(y)(\psi) = 2\int_0^T \int_\Omega \operatorname{curl} y \cdot \operatorname{curl} \psi \, dx \, dt,$$

$$J_3'(y)(\psi) = \int_0^T \int_\Omega g_3'(\det(\nabla y)) \cdot \left(y^1_{x_1}\psi^2_{x_2} + \psi^1_{x_1}y^2_{x_2} - y^2_{x_1}\psi^1_{x_2} - \psi^2_{x_1}y^1_{x_2}\right) dx \, dt.$$

Due to the fact that $g_3'(t) \in [0, 3]$ for all $t \in \mathbb{R}$, we obtain that $g_3'(\det(\nabla y)) \in L^\infty(Q)$ and that

$$J_i'(y) \in L^2(H^{-1}(\Omega)^2), \qquad i = 1, 2, 3.$$

The existence of the adjoint velocity $\lambda \in L^2(\mathcal{V}_0)$ with $\lambda_t \in L^{4/3}(\mathcal{V}_0^*)$ follows, e.g., from [HK]. The introduction of dual pressure π can be done as in the proof of Proposition 2.1. Under the assumptions from Proposition 2.2 additional regularity $z \in \mathcal{X}_0$ can be argued as in [HK]. □

Proof of Theorem 3.2. The reduced cost functional $j \colon U \to \mathbb{R}$ is defined as $j(u) = J(S(u))$, where $S : U \to W$ is the (continuous) solution operator for the velocity component of the state equation (2.2); see Proposition 2.1. The solution operator S is directionally differentiable, and the directional derivative $\delta y = S'(u)(\delta u)$ fulfills the following linearized equation (see, e.g., [HK]):

$$\delta y \in B\delta u + W_0 \ : \ \bar{a}'(y)(\delta y, \psi) = 0 \quad \text{for all } \psi \in L^2(\mathcal{V}_0).$$

In the proof of Theorem 3.1 it is shown that the functionals $J = J_1, J_2$, and J_3 are directionally differentiable. Therefore, the reduced cost functional j is directionally differentiable too. A necessary optimality condition for the reduced cost functional is given by

$$j'(u)(\delta u - u) \geq 0 \quad \text{for all } \delta u \in U_{ad}.$$

To complete the proof it remains to show the representation of the directional derivative of j. For this purpose we recall the definition of the adjoint equation, and using the fact that $\delta y - B\delta u \in W_0$ we find

$$\begin{aligned} j'(u)(\delta u) &= J'(y)(\delta y) = J'(y)(\delta y - B\delta u) + J'(y)(B\delta u) \\ &= \bar{a}'(y)(\delta y - B\delta u, \lambda) + J'(y)(B\delta u) \\ &= -\bar{a}'(y)(B\delta u, \lambda) + J'(y)(B\delta u). \qquad \square \end{aligned}$$

Acknowledgment. The authors would like to thank Prof. Haller for a very helpful exchange of emails.

REFERENCES

[AT] F. Abergel and R. Temam, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Dynamics, 1 (1990), pp. 303–325.

[BB] R. Becker and M. Braack, *A finite element pressure gradient stabilization for the Stokes equations based on local projections*, Calcolo, 38 (2001), pp. 137–199.

[BMV] R. Becker, D. Meidner, and B. Vexler, *Efficient numerical solution of parabolic optimization problems by finite element methods*, Optim. Methods Softw., to appear.

[BV] R. Becker and B. Vexler, *Optimal control of the convection-diffusion equation using stabilized finite element methods*, Numer. Math., 106 (2007), pp. 349–367.

[BGL] M. Berggren, R. Glowinski, and J.-L. Lions, *A computational approach to controllability issues for flow-related models.* (I): *Pointwise control of the viscous Burgers equation*, Int. J. Comput. Fluid Dyn., 7 (1996), pp. 237–252.

[BMC] H. M. Blackburn, N. N. Mansour, and B. J. Cantwell, *Topology of fine-scale motions in turbulent channel flow*, J. Fluid Mech., 310 (1996), pp. 293–324.

[CPC] M. S. Chong, A. E. Perry, and B. J. Cantwell, *A general classification of three-dimensional flow fields*, Phys. Fluids A, 2 (1990), pp. 765–777.

[C] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.

[DL] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 6, Springer Verlag, Berlin, 1993.

[EJT] K. Eriksson, Cl. Johnson, and V. Thomée, *Time discretization of parabolic problems by the discontinuous Galerkin method*, RAIRO Model. Math. Anal. Numer., 19 (1985), pp. 611–643.

[FGH] A. V. Fursikov, M. D. Gunzburger, and L. S. Hou, *Boundary value problems and optimal boundary control for the Navier–Stokes system: The two-dimensional case*, SIAM J. Control Optim., 36 (1998), pp. 852-894.

[GLLS] M. Giles, M. Larson, M. Levenstam, and E. Suli, *Adaptive Error Control for Finite Element Approximations of the Lift and Drag Coefficients in Viscous Flow*, Technical report NA-76/06, Oxford University Computing Laboratory, Oxford, UK, 1997.

[GR] V. Girault and P.-A. Raviert, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1979.

[Gr] A. Griewank, *Achieving logarithmic growth of temporal and spatial complexity in reverse automatic differentiation*, Optim. Methods Softw., 1 (1992), pp. 35–54.

[G] M. Gunzburger, ed., *Flow Control*, IMA Vol. Math. Appl. 68, Springer, Berlin, 1995.

[H1] G. Haller, *Lagrangian structures and the rate of strain in a partition of two-dimensional turbulence*, Phys. Fluids, 13 (2001), pp. 3365–3385.

[H2] G. Haller, *An objective definition of a vortex*, J. Fluid Mech., 525 (2005), pp. 1–26.

[Ha] A. Haraux, *How to differentiate the projection on a closed convex set in Hilbert space. Some applications to variational inequalities*, J. Math. Soc. Japan, 29 (1977), pp. 615–631.

[HRT] J. G. Heywood, R. Rannacher, and S. Turek, *Artificial boundaries and flux and pressure conditions for the incompressible Navier–Stokes equations*, Internat. J. Numer. Methods Fluids, 22 (1996), pp. 325–352.

[HKSV] M. Hintermüller, K. Kunisch, Y. Spasov, and S. Volkwein, *Dynamical systems based optimal control of incompressible fluids*, Internat. J. Numer. Methods Fluids, 46 (2004), pp. 345–359.

[HK] M. Hinze and K. Kunisch, *Second order methods for boundary control of the instationary Navier–Stokes system*, ZAMM Z. Angew. Math. Mech., 84 (2004), pp. 171–187.

[HWM] J. C. R. Hunt, A. A. Wray. and P. Moin, *Eddies, Stream and Convergence Zones in Turbulent Flows*, Report CTR-S88, Center for Turbulence Research, Stanford University, Stanford, CA, 1988.

[JH] J. Jeong and F. Hussain, *On the identificaton of vortex*, J. Fluid Mech., 285 (1995), pp. 69–94.

[KV] K. Kunisch and B. Vexler, *Constrained Dirichlet boundary control in L^2 for a class of evolution equations*, SIAM J. Control Optim., to appear.

[LHK] G. Lapeyre, B. L. Hua, and P. Klein, *Dynamics of the orientation of active and passive scalars in two-dimensional turbulence*, Phys. Fluids, 13 (2001), pp. 251–264.

[LKH] G. Lapeyre, P. Klein, and B. L. Hua, *Does the tracer gradient vector align with the strain eigenvectors in $2D$ turbulence?*, Phys. Fluids, 11 (1999), pp. 3729–3737.

[Li] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Heidelberg, 1971.

[Lu] H. J. Lugt, *The dilemma of defining a vortex*, in Recent Developments in the Theoretical and Experimental Fluid Mechanics, U. Müller, K. G. Roesner, and B. Schmidt, eds., Springer, Berlin, 1979, pp. 309–321.

[NW] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Ser. Oper. Res., Springer, New York, 1999.

[O] A. Okubo, *Horizontal dispersion of floatable trajectories in the vicinity of velocity singularities such as convergencies*, Deep-Sea Res., 17 (1970), pp. 445–454.

[RV] A. Rösch and B. Vexler, *Optimal control of the Stokes equations: A priori error analysis for finite element discretization with postprocessing*, SIAM J. Numer. Anal., 44 (2006), pp. 1903–1920.

[S2] V. A. Solonnikov, *On the differential properties of the solution of the first boundary value problem for nonstationary systems of Navier–Stokes equations*, Trudy Math. Inst. Steklov, 73 (1964), pp. 222–291.

[S1] V. A. Solonnikov, *Estimates in L^P of solutions of elliptic and parabolic systems*, Proc. Steklov Inst. Math., 102 (1967), pp. 157–185.

[TK] M. Tabor and I. Klapper, *Stretching and alignment in chaotic and turbulent flows*, Chaos Solitons Fractals, 4 (1994), pp. 1031–1055.

[T] R. Temam, *Navier–Stokes Equations. Theory and Numerical Analysis*, North–Holland, Amsterdam, 1984.

[TW] F. Tröltzsch and D. Wachsmuth, *Second-order sufficient optimality conditions for the control of Navier Stokes equations*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 93–119.

[V] B. Vexler, *Finite element approximation of elliptic dirichlet optimal control problems*, Numer. Funct. Anal. Optim., to appear.

[W] J. Weiss, *The dynamics of enstrophy transfer in two-dimensional hydrodynamics*, Phys. D, 48 (1991), pp. 273–294.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1398–1430

WELL-POSEDNESS OF THE SHOOTING ALGORITHM FOR STATE CONSTRAINED OPTIMAL CONTROL PROBLEMS WITH A SINGLE CONSTRAINT AND CONTROL*

J. FRÉDÉRIC BONNANS† AND AUDREY HERMANT†

Abstract. This paper deals with the shooting algorithm for optimal control problems with a scalar control and a regular scalar state constraint. Additional conditions are displayed, under which the so-called alternative formulation is equivalent to Pontryagin's minimum principle. The shooting algorithm appears to be well-posed (invertible Jacobian) iff (i) the no-gap second-order sufficient optimality condition holds, and (ii) when the constraint is of order $q \geq 3$, there is no boundary arc. Stability and sensitivity results without strict complementarity at touch points are derived using Robinson's strong regularity theory, under a minimal second-order sufficient condition. The directional derivatives of the control and state are obtained as solutions of a linear quadratic problem.

Key words. optimal control, Pontryagin's principle, state constraints, junction conditions, shooting algorithm, no-gap second-order optimality conditions, strong regularity, sensitivity analysis, directional derivatives

AMS subject classifications. 49M05, 49K40, 34B15, 34E10

DOI. 10.1137/06065756X

1. Introduction.

For optimal control problems satisfying the strengthened Legendre–Clebsch condition, Pontryagin's principle allows us to express the control as a function of the state and the costate. For unconstrained problems, the resulting two-point boundary value problem reduces to a finite-dimensional "shooting" equation whose unknown is the initial costate (see, e.g., [29]). The extension to control constrained problems is relatively easy, assuming nontangentiality conditions when a constraint becomes active or inactive. This approach allows us to compute accurate solutions at low cost, once the *structure* of active constraints is known, and reasonable initial values of unknowns can be guessed. For state constrained optimal control problems, a reformulation of the optimality conditions is needed, and the shooting equations take into account only some of the optimality conditions. Therefore, checking that the shooting equations are well-posed under minimal hypotheses becomes challenging.

An alternative formulation, suitable for the shooting algorithm in the presence of state constraints, was first introduced by Bryson, Denham, and Dreyfus [8] (see also [9]) in a heuristic manner. Some additional conditions (necessary for optimality) were missing, as shown in Jacobson, Lele, and Speyer [18], where the first results on the regularity of the multiplier and on junction conditions are stated. A significant clarification of their work can be found in the unpublished paper by Maurer [25], where the link between the results of [18] and the alternative formulation of [8, 9] is established. Numerous different versions of Pontryagin's principle with state constraints were given in the literature; see the survey by Hartl, Sethi, and Vickson [16].

Stability results for *first*-order state constraints and directional differentiability of

*Received by the editors April 19, 2006; accepted for publication (in revised form) February 19, 2007; published electronically September 19, 2007.
http://www.siam.org/journals/sicon/46-4/65756.html
†CMAP, Ecole Polytechnique, INRIA Futurs, Route de Saclay, 91128 Palaiseau, France (bonnans@ cmap.polytechnique.fr, hermant@cmap.polytechnique.fr).

solutions in L^2 were first obtained by Malanowski [21] using an infinite-dimensional implicit function theorem and differentiation of the projection on a convex set [15]. The (strong) second-order sufficient condition used in the analysis was later weakened by Malanowski [22], taking into account the strictly active constraints. These results require *no assumptions on the structure* of the trajectory. However, no extensions of this method for higher-order state constraints are known. Dontchev and Hager [12] derived, still for first-order constraints, L^∞ stability results under an additional assumption on the structure of the contact set. Malanowski and Maurer obtain sensitivity results in [23] (first order) and [24] (higher order), when there are finitely many nontangential junction points and strict complementarity holds, by application of the implicit function theorem to the shooting mapping. They obtain derivatives as the solution of an equality constrained linear quadratic problem, but when the order of the constraint is $q \geq 2$, the data of the latter depend on the (precomputed) variation of entry times. Numerical applications of the shooting algorithm to state constrained problems in the aerospace field are presented, e.g., in [10, 3] and in [26], where the role of additional conditions appears crucial to eliminate nonoptimal solutions; numerical examples of sensitivity analysis are given in [2]. Discretization errors are studied in, e.g., [13].

This paper handles the case of a scalar control and a regular scalar state constraint, for which regularity and junction conditions results are known. We assume that the Hamiltonian is uniformly strongly convex w.r.t. the control variable, that there are finitely many nontangential junction times, and that strict complementarity on boundary arcs holds.

We express the additional conditions under which the alternative formulation is equivalent to Pontryagin's principle. When strict complementarity holds at touch points as well, we prove that the shooting algorithm is well-posed (invertible Jacobian) iff (i) the no-gap second-order sufficient condition in [5] holds, and (ii) when the constraint is of order $q \geq 3$, there is no boundary arc. Then stability and sensitivity results, removing the strict complementarity hypothesis at touch points, are derived, applying Robinson's strong regularity theory [28] to the shooting mapping. We give a necessary and sufficient second-order condition characterizing the strong regularity property. The directional derivatives of the control and state are obtained as solutions of an inequality constrained linear quadratic problem, independent of the variations of junction times.

The paper is organized as follow. In section 2, we give the characterization of Pontryagin extremals as solutions of the shooting equations under some minimal additional conditions. Then, in section 3, we give the characterization of the well-posedness of the shooting algorithm and its relation to the no-gap second-order optimality conditions obtained in [5, 6]. Finally, in section 4, we give stability and sensitivity analysis results.

The results of sections 2 and 3 of this paper are extended to the case of vector-valued state constraints and control in the report [4]. The main difficulty is the extension of the junction conditions result of Jacobson, Lele, and Speyer [18] (Proposition 2.5 below). The latter plays a crucial role in the proof of the *necessity* of the condition claimed in this paper as necessary and sufficient for the well-posedness of the shooting algorithm (see Theorem 3.3).

2. Junction conditions. The section is organized as follows. After introducing notation, definitions, assumptions, and basic results needed in the paper, we recall in subsection 2.1 an alternative formulation for optimality conditions (Definition 2.7),

which is useful for the shooting algorithm. This is one of the various formulations existing in the literature (see, e.g., the survey [16]). Therefore, one of the main concerns of this paper is to investigate, in subsection 2.2, the equivalence with Pontryagin's minimum principle (Proposition 2.10). Finally, in subsection 2.3, we formulate the shooting algorithm and show that some of the additional conditions are automatically satisfied by a solution of the shooting equations (Proposition 2.15).

Denote by $L^\infty(0,T)$ the Banach space of measurable and essentially bounded functions and by $W^{1,\infty}(0,T)$ the Sobolev space of functions having a weak derivative in $L^\infty(0,T)$. Let the control and state spaces be, respectively, $\mathcal{U} := L^\infty(0,T)$ and $\mathcal{Y} := W^{1,\infty}(0,T;\mathbb{R}^n)$. We consider the following optimal control problem with a scalar state constraint and a scalar control:

$$(\mathcal{P}) \qquad \min_{(u,y)\in\mathcal{U}\times\mathcal{Y}} \int_0^T \ell(u(t),y(t))\mathrm{d}t + \phi(y(T)) \tag{2.1}$$

$$\text{subject to} \qquad \dot{y}(t) = f(u(t),y(t)) \quad \text{a.e. } t\in[0,T]; \quad y(0)=y_0, \tag{2.2}$$

$$g(y(t)) \le 0 \qquad \text{for all } t\in[0,T]. \tag{2.3}$$

The data of the problem are the distributed cost $\ell : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}$, final cost $\phi : \mathbb{R}^n \to \mathbb{R}$, dynamics $f : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$, state constraint $g : \mathbb{R}^n \to \mathbb{R}$, final time $T > 0$, and initial condition $y_0 \in \mathbb{R}^n$.

We assume throughout the paper that the following hold:

(A0) The mappings ℓ, ϕ, f, and g are k-times continuously differentiable (C^k) with $k \ge 2$ and have locally Lipschitz continuous second-order derivatives when $k = 2$. The dynamics f is Lipschitz continuous.

(A1) The initial condition satisfies $g(y_0) < 0$.

The space of row vectors is denoted by $\mathbb{R}^{n*}$. The space of continuous functions over $[0,T]$ is denoted by $C[0,T]$. The dual space of Radon measures, denoted by $\mathcal{M}[0,T]$, is identified with the space of functions of bounded variation $BV(0,T)$ vanishing at zero. The transposition operator in $\mathbb{R}^n$ is denoted by a star *. Fréchet derivatives of f, ℓ, etc., w.r.t. arguments $u\in\mathbb{R}$, $y\in\mathbb{R}^n$, are denoted by a subscript, for instance $f_u(u,y) = D_u f(u,y)$, $f_{uu}(u,y) = D^2_{uu} f(u,y)$. One exception to this rule, which should not be a source of confusion, is that we denote by y_u the (unique) solution in $\mathcal{W}$ of the state equation (2.2) associated with the control $u\in\mathcal{U}$. Total derivation w.r.t. time is denoted by a dot, i.e., $\dot{y}(t) = \frac{\mathrm{d}y(t)}{\mathrm{d}t}$.

A *trajectory* is an element (u,y) of $\mathcal{U}\times\mathcal{Y}$ satisfying the state equation (2.2). A trajectory (u,y) is said to be *feasible* if it satisfies the state constraint (2.3). Define the classical (resp., generalized) *Hamiltonian* functions of $(\mathcal{P})$, $H : \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^{n*} \to \mathbb{R}$ (resp., $\mathcal{H} : \mathbb{R}\times\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^{n*} \to \mathbb{R}$) by

$$H(u,y,p) := \ell(u,y) + pf(u,y); \quad \mathcal{H}(p_0,u,y,p) := p_0\ell(u,y) + pf(u,y). \tag{2.4}$$

First-order necessary optimality conditions for $(\mathcal{P})$ are given by *Pontryagin's minimum principle*.

Definition 2.1. *A trajectory (u,y) is a* Pontryagin extremal *if there exists $p_0\in\mathbb{R}^+$, $p\in BV([0,T];\mathbb{R}^{n*})$, and $\eta\in\mathcal{M}[0,T]$, with $(p_0,\mathrm{d}\eta)\neq 0$, such that*

$$\dot{y}(t) = \mathcal{H}_p(p_0,u(t),y(t),p(t)) \quad \text{a.e. } t\in[0,T]; \quad y(0)=y_0, \tag{2.5}$$

$$-\mathrm{d}p(t) = \mathcal{H}_y(p_0,u(t),y(t),p(t))\mathrm{d}t + g_y(y(t))\mathrm{d}\eta(t) \quad \text{in } \mathcal{M}([0,T];\mathbb{R}^{n*}), \tag{2.6}$$

$$p(T) = p_0\phi_y(y(T)), \tag{2.7}$$

$$u(t) \in \operatorname{argmin}_{w\in\mathbb{R}} \mathcal{H}(p_0,w,y(t),p(t)) \quad \text{a.e. } t\in[0,T], \tag{2.8}$$

$$g(y(t)) \le 0 \text{ for all } t \in [0,T]; \quad \mathrm{d}\eta \ge 0; \quad \int_0^T g(y(t))\mathrm{d}\eta(t) = 0. \tag{2.9}$$

By $\mathrm{d}\eta \ge 0$, we mean that $\int_0^T \varphi(t)\mathrm{d}\eta(t) \ge 0$ for all nonnegative continuous functions $\varphi \in C[0,T]$, or equivalently, that η is nondecreasing. The costate equation (2.6) with final condition (2.7) is equivalent to

$$p(t) = \int_t^T \mathcal{H}_y(p_0, u(s), y(s), p(s))\mathrm{d}s + \int_t^T g_y(y(s))\mathrm{d}\eta(s) + p_0\phi_y(y(T)).$$

The next theorem is well known (see [11, 14] for nondifferentiable versions).

THEOREM 2.2. *A trajectory (u,y) solution of $(\mathcal{P})$ is a Pontryagin extremal.*

A trajectory $(\bar u, \bar y)$ is a *local solution* of $(\mathcal{P})$ if it minimizes (2.1) subject to (2.2)–(2.3) and $\|u - \bar u\|_\infty \le \rho$ for some $\rho > 0$. We say that $(u,y) \in \mathcal{U} \times \mathcal{Y}$ is a *stationary point* of $(\mathcal{P})$ if there exists a nonzero $(p_0, p, \eta) \in \mathbb{R}^+ \times BV(0,T;\mathbb{R}^{n*}) \times \mathcal{M}(0,T)$ such that (2.5)–(2.7), (2.9) are satisfied and

$$\mathcal{H}_u(p_0, u(t), y(t), p(t)) = 0 \qquad \text{for a.a. } t \in [0,T].$$

It is well known that a local solution of $(\mathcal{P})$ is a stationary point. Obviously a Pontryagin extremal is a stationary point, but the converse is in general false. An exception is when the (generalized) Hamiltonian is convex w.r.t. the control variable along the trajectory (see also our assumption (A2) below). Whenever this holds, definitions of both Pontryagin extremals and stationary points are equivalent.

Definitions. A *boundary* (resp., *interior*) *arc* is a maximal interval of positive measure $\mathcal{I} \subset [0,T]$ such that $g(y(t)) = 0$ (resp., $g(y(t)) < 0$) for all $t \in \mathcal{I}$. If $[\tau_{en}, \tau_{ex}]$ is a boundary arc, τ_{en} and τ_{ex} are called an *entry* and an *exit* point, respectively. Entry and exit points are said to be *regular* if they are endpoints of an interior arc. A *touch* point τ in $(0,T)$ is an isolated contact point (an endpoint of two interior arcs). Entry, exit, and touch points are called *junction points* (or *times*). We say that the junctions are regular when the entry/exit points are regular.

The first-order time derivative of the state constraint along a trajectory (u,y), i.e., $g^{(1)}(u,y) = \frac{\mathrm{d}}{\mathrm{d}t} g(y(t)) = g_y(y) f(u,y)$, is denoted by $g^{(1)}(y)$ if the function $\mathbb{R}\times\mathbb{R}^n \to \mathbb{R}$, $(u,y) \mapsto g_y(y)f(u,y)$ does not depend on u (that is, the function $(u,y) \mapsto g_u^{(1)}(u,y)$ is identically zero). If f and g are C^q, we may define similarly $g^{(2)}, \dots, g^{(q)}$ if $g_u^{(j)} \equiv 0$ for all $j = 1, \dots, q-1$, and we have $g^{(j)}(u,y) = g_y^{(j-1)}(y) f(u,y)$ for $j = 1, \dots, q$.

Let $q \ge 1$ be the smallest number of time derivations of the state constraint so that a dependence w.r.t. u appears, i.e., $g_u^{(q)} \not\equiv 0$. If q is finite, we say that q is the *order* of the state constraint (see, e.g., [8]). A state constraint of order q is said to be *regular* along the trajectory (u,y) if the condition below holds:

$$\exists\, \gamma > 0, \quad |g_u^{(q)}(\hat u, y(t))| \ge \gamma \quad \text{for all } t \in [0,T] \text{ and all } \hat u \in \mathbb{R}. \tag{2.10}$$

Note that the set of generalized multipliers (p_0, p, η) is a cone. When $p_0 = 0$, we say that the multiplier is singular; otherwise it is regular. Dividing then (p,η) by p_0, we obtain the qualified version of Pontryagin's principle, substituting the generalized Hamiltonian with the classical Hamiltonian. It is easily seen that a Pontryagin extremal satisfying (2.10) (and (A1)) has no singular multiplier, and that the multiplier (p,η) in the qualified version of Pontryagin's principle ($p_0 = 1$) is unique. The same is true for a stationary solution.

Being of bounded variation, p has at most countably many discontinuity times and has everywhere on $[0,T]$ left and right limits, denoted by $p(t^\pm) = \lim_{t'\to t^\pm} p(t')$. The jump at $\tau \in (0,T)$ is denoted by $[p(\tau)] = p(\tau^+) - p(\tau^-)$. Similar observations hold for η.

Assumptions. We say that (u,y) is a *regular Pontryagin extremal* if it satisfies Definition 2.1 with $p_0 = 1$, with costate p and multiplier η, and if assumptions (A2)–(A4) below are satisfied.

(A2) The Hamiltonian is strongly convex w.r.t. the control variable, uniformly w.r.t. $t \in [0,T]$:

$$\exists\, \alpha > 0, \quad H_{uu}(\hat u, y(t), p(t^\pm)) \geq \alpha \quad \text{for all } t \in [0,T] \text{ and all } \hat u \in \mathbb{R}. \tag{2.11}$$

(A3) The data of the problem are C^{2q}, i.e., $k \geq 2q$ in (A0), and the state constraint is of order q and regular, i.e., (2.10) holds.

(A4) The trajectory (u,y) has a *finite set of junction times* that will be denoted by $\mathcal{T} =: \mathcal{T}_{en} \cup \mathcal{T}_{ex} \cup \mathcal{T}_{to}$, with $\mathcal{T}_{en}$, $\mathcal{T}_{ex}$, and $\mathcal{T}_{to}$ the *disjoint* (and possibly empty) subsets of, respectively, entry, exit, and touch points, and we assume that $g(y(T)) < 0$.

Hypothesis (A4) implies that all entry and exit points are regular. In what follows, we denote by $\mathcal{I}_b$ the union of boundary arcs, i.e., $\mathcal{I}_b := \cup_{i=1}^{N_b} [\tau_{en}^i, \tau_{ex}^i]$ for $\mathcal{T}_{en} := \{\tau_{en}^1 < \cdots < \tau_{en}^{N_b}\}$ and $\mathcal{T}_{ex} := \{\tau_{ex}^1 < \cdots < \tau_{ex}^{N_b}\}$.

Remark 2.3. Throughout the paper, (A3) can be weakened, replacing (2.10) by

$$\exists\, \gamma, \varepsilon > 0, \quad |g_u^{(q)}(\hat u, y(t))| \geq \gamma \quad \text{for all } t,\ \operatorname{dist}(t, \mathcal{I}_b \cup \mathcal{T}_{to}) < \varepsilon, \text{ and all } \hat u \in \mathbb{R}. \tag{2.12}$$

Notation. Given a finite subset $\mathcal{S}$ of $(0,T)$, we denote by $PC^k_{\mathcal{S}}[0,T]$ the set of functions over $[0,T]$ that are of class C^k outside $\mathcal{S}$ (PC stands for piecewise continuous) and have, as well as their first k derivatives, a left and right limit over $\mathcal{S}$ and a right (resp., left) limit at 0 (resp., T).

Let φ be a real-valued function over $[0,T]$. Assuming w.l.o.g. the elements of $\mathcal{S}$ in increasing order, we may define $\varphi(\mathcal{S}) := (\varphi(\tau))_{\tau\in\mathcal{S}} \in \mathbb{R}^{\operatorname{Card}\mathcal{S}}$. We adopt a similar convention for vectors, $\nu_{\mathcal{S}} := (\nu_\tau)_{\tau\in\mathcal{S}} \in \mathbb{R}^{\operatorname{Card}\mathcal{S}}$, and will also use the following notation:

$$\nu_{\mathcal{S}}^{1:q} := \begin{pmatrix} \nu_{\mathcal{S}}^1 \\ \vdots \\ \nu_{\mathcal{S}}^q \end{pmatrix} \in \mathbb{R}^{q\operatorname{Card}\mathcal{S}};\ g^{(0:q-1)}(y(\mathcal{S})) := \begin{pmatrix} g(y(\mathcal{S})) \\ \vdots \\ g^{(q-1)}(y(\mathcal{S})) \end{pmatrix} \in \mathbb{R}^{q\operatorname{Card}\mathcal{S}}.$$

2.1. Alternative formulation of optimality conditions. Under assumption (A4) we have a finite number of arcs and we can show, with regularity assumptions (A2)–(A3), that the multiplier η is differentiable on the interior of each arc [18, 25]. An analysis of the optimality system on interiors of arcs shows then that a regular Pontryagin extremal satisfies the conditions stated in Proposition 2.4 below. An analysis at junction times leads afterwards to the junction conditions given in Proposition 2.5.

Proposition 2.4. *Let (u,y) be a regular Pontryagin extremal satisfying* (A2)–(A4). *Then we have $u \in PC^q_{\mathcal{T}}[0,T]$, $y \in PC^{q+1}_{\mathcal{T}}([0,T];\mathbb{R}^n)$ and there exists $p \in PC^1_{\mathcal{T}}([0,T];\mathbb{R}^{n*})$, $\eta_0 \in PC^0_{\mathcal{T}}[0,T]$, and jump parameters $\nu_{\mathcal{T}}$, such that the following optimality system is satisfied:*

$$\dot y(t) = H_p(u(t),y(t),p(t)) = f(u(t),y(t)) \text{ on } [0,T]; \qquad y(0) = y_0, \tag{2.13}$$

$$-\dot p(t) = H_y(u(t),y(t),p(t)) + g_y(y(t))\eta_0(t) \text{ on } [0,T]\setminus\mathcal{T}, \tag{2.14}$$

$$p(T) = \phi_y(y(T)), \tag{2.15}$$

$$0 = H_u(u(t), y(t), p(t)) \quad \text{on } [0,T] \setminus \mathcal{T}, \tag{2.16}$$

$$g(y(t)) = 0 \ \text{ on } \mathcal{I}_b; \quad \eta_0(t) = 0 \ \text{ on } [0,T] \setminus \mathcal{I}_b, \tag{2.17}$$

$$g(y(t)) < 0 \ \text{ on } [0,T] \setminus (\mathcal{I}_b \cup \mathcal{T}_{to}); \quad \eta_0(t) \geq 0 \ \text{ on } \operatorname{int} \mathcal{I}_b, \tag{2.18}$$

$$g(y(\tau)) = 0 \quad \text{for all } \tau \in \mathcal{T}_{to}, \tag{2.19}$$

$$[p(\tau)] = -\nu_\tau g_y(y(\tau)); \ \nu_\tau \geq 0 \text{ for all } \tau \in \mathcal{T}. \tag{2.20}$$

We denote by $\operatorname{int} \mathcal{I}_b$ the interior of $\mathcal{I}_b$. A touch point $\tau \in \mathcal{T}_{to}$ is said to be *essential* if $\nu_\tau > 0$ in (2.20); otherwise it is nonessential. We denote by $\mathcal{T}_{to}^{ess}$ the set of essential touch points. Hypotheses (A2)–(A4) also imply the continuity of the control variable and of some of its time derivatives at junction points. The next proposition is due to Jacobson, Lele, and Speyer [18].

PROPOSITION 2.5. *Let (u,y) be a regular Pontryagin extremal satisfying* (A2)–(A4). *Then*

(i) *for all entry or exit points $\tau \in \mathcal{T}_{en} \cup \mathcal{T}_{ex}$,*

(a) *if q is* odd, *u and its $q-1$ first derivatives are continuous at τ, $\nu_\tau = 0$ and p is continuous at τ;*

(b) *if q is* even, *u and its $q-2$ first derivatives are continuous at τ.*

(ii) *for all touch points $\tau \in \mathcal{T}_{to}$,*

(a) *u and its $q-2$ first derivatives are continuous at τ;*

(b) *if τ is nonessential (i.e., $\nu_\tau = 0$), u and its q first derivatives and p are continuous at τ;*

(c) *if $q = 1$, then τ is a nonessential touch point.*

REMARK 2.6. If (u,y) satisfies (A2)–(A4) and (2.13)–(2.20), the multiplier $\eta \in \mathcal{M}[0,T]$ such that (u,y) satisfies Definition 2.1 is given by

$$\mathrm{d}\eta(t) \ = \ \sum_{\tau \in \mathcal{T}} \nu_\tau \delta_\tau + \eta_0(t)\mathrm{d}t, \tag{2.21}$$

where δ_τ denotes the Dirac measure at time τ, $\nu_\tau = [\eta(\tau)]$ is the nonnegative jump at $\tau \in \mathcal{T}$, and the density $\eta_0 \in PC^0_{\mathcal{T}}[0,T]$ equals $\frac{\mathrm{d}\eta}{\mathrm{d}t}$ on $[0,T] \setminus \mathcal{T}$.

We now present the alternative formulation that will be used in the shooting algorithm. First introduced heuristically in [8], it is based on the use of the mixed explicit constraint $g^{(q)}(u(t), y(t)) = 0$ on boundary arcs. Let the *augmented Hamiltonian* $\tilde{H} : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^{n*} \times \mathbb{R} \to \mathbb{R}$ be defined by

$$\tilde{H}(u, y, p_q, \eta_q) = H(u, y, p_q) + \eta_q g^{(q)}(u, y), \tag{2.22}$$

where q denotes the order of the state constraint and H is the classical Hamiltonian (2.4).

DEFINITION 2.7. *We say that a trajectory (u,y) in $PC^q_{\mathcal{T}}[0,T] \times PC^{q+1}_{\mathcal{T}}([0,T];\mathbb{R}^n)$ satisfying* (A3)–(A4) *is a solution of the* alternative formulation *if there exist $p_q \in PC^{q+1}_{\mathcal{T}}([0,T];\mathbb{R}^{n*})$, $\eta_q \in PC^q_{\mathcal{T}}[0,T]$, alternative jump parameters $\nu^j_{\mathcal{T}_{en}}$, $j = 1, \ldots, q$, and $\nu_{\mathcal{T}_{to}}$ such that the following relations are satisfied (we omit dependence in time):*

$$\dot{y} = \tilde{H}_p(u, y, p_q, \eta_q) = f(u,y) \quad \text{on } [0,T]; \quad y(0) = y_0, \tag{2.23}$$

$$-\dot{p}_q = \tilde{H}_y(u, y, p_q, \eta_q) = H_y(u, y, p_q) + \eta_q g^{(q)}_y(u,y) \quad \text{on } [0,T] \setminus \mathcal{T}, \tag{2.24}$$

$$p_q(T) = \phi_y(y(T)), \tag{2.25}$$

$$0 = \tilde{H}_u(u, y, p_q, \eta_q) = H_u(u, y, p_q) + \eta_q g^{(q)}_u(u,y) \quad \text{on } [0,T] \setminus \mathcal{T}, \tag{2.26}$$

$$g^{(j)}(y(\tau)) = 0 \qquad \text{for } j = 0, 1, \ldots, q-1; \quad \tau \in \mathcal{T}_{en}, \tag{2.27}$$

$$g^{(q)}(u, y) = 0 \qquad \text{on } \mathcal{I}_b, \tag{2.28}$$

$$g(y(\tau)) = 0 \qquad \text{for all } \tau \in \mathcal{T}_{to}, \tag{2.29}$$

$$\eta_q(t) = 0 \qquad \text{on } [0,T] \setminus \mathcal{I}_b, \tag{2.30}$$

$$[p_q(\tau)] = -\sum_{j=1}^{q} \nu_\tau^j g_y^{(j-1)}(y(\tau)) \qquad \text{for all } \tau \in \mathcal{T}_{en}, \tag{2.31}$$

$$[p_q(\tau)] = 0 \qquad \text{for all } \tau \in \mathcal{T}_{ex}, \tag{2.32}$$

$$[p_q(\tau)] = -\nu_\tau g_y(y(\tau)) \qquad \text{for all } \tau \in \mathcal{T}_{to}. \tag{2.33}$$

In the heuristic formulation of [8], (2.23)–(2.33) are interpreted as necessary optimality conditions for the problem of minimizing (2.1) subject to (2.2) and *equality constraints* (2.27)–(2.29) for a *fixed* set of junction times $\mathcal{T}$. Alternative jump parameters $\nu_{\tau_{en}}^{1:q}$ appearing in (2.31) are seen as multipliers associated with the q interior point constraints in (2.27) at a regular entry time τ_{en}.

The assumption equivalent to (A2) for the alternative formulation is the following; see Remark 2.11(ii):

(A2$_q$) $\exists\, \alpha > 0,\ \tilde{H}_{uu}(\hat{u}, y(t), p_q(t^\pm), \eta_q(t^\pm)) \geq \alpha$ for all $t \in [0,T]$ and all $\hat{u} \in \mathbb{R}$.

In what follows, we will write (A2)–(A4) (resp., (A2$_q$)–(A4)) to denote the assumptions (A2) (resp., (A2$_q$)), (A3), and (A4).

2.2. Additional conditions. Relations (2.23)–(2.33) due to [8] are necessary, but not sufficient, conditions for regular Pontryagin extremals. This was underlined in [18], where some additional necessary conditions were provided that allowed the authors to show that a trajectory (with a fourth-order state constraint) was not a Pontryagin extremal. We state in Proposition 2.10 the characterization of regular Pontryagin extremals based on the alternative formulation. We need some preliminary lemmas.

LEMMA 2.8. *Let (u, y) be a trajectory, and let $(p_q, \eta_q) \in PC^1_{\mathcal{T}}([0,T]; \mathbb{R}^{n*}) \times PC^0_{\mathcal{T}}[0,T]$ satisfying* (A2$_q$)–(A4) *and* (2.23)–(2.24), (2.26), (2.28). *Then (u, y, p_q, η_q) belongs to the set $PC^q_{\mathcal{T}}[0,T] \times PC^{q+1}_{\mathcal{T}}([0,T]; \mathbb{R}^n) \times PC^{q+1}_{\mathcal{T}}([0,T]; \mathbb{R}^{n*}) \times PC^q_{\mathcal{T}}[0,T]$.*

Proof. By the implicit function theorem, applied to (2.26) on interior arcs, and to (2.26) and (2.28) on boundary arcs, the algebraic variables (u, η_q) can be expressed, on the interior of each arc, as C^q functions of (y, p_q). The result follows. □

LEMMA 2.9. *If constraint regularity* (A3) *holds along a trajectory (u, y), and if $u \in PC^q_{\mathcal{T}}[0,T]$, then, for all $t \in [0,T]$, vectors $(g_y(y(t)), \ldots, g_y^{(q-1)}(y(t)))$ are linearly independent (and hence, $q \leq n$).*

Proof. Since $u \in PC^q_{\mathcal{T}}[0,T]$, the mappings $(A_l)_{0 \leq l \leq q} : [0,T] \setminus \mathcal{T} \to \mathbb{R}^n$ defined inductively by

$$\begin{cases} A_0(t) := f_u(u(t), y(t)), \\ A_l(t) := f_y(u(t), y(t)) A_{l-1}(t) - \dot{A}_{l-1}(t), \qquad l = 1, \ldots, q, \end{cases} \tag{2.34}$$

are well defined, and $A_l \in PC^{q-l}_{\mathcal{T}}([0,T]; \mathbb{R}^n)$ for $l = 0, \ldots, q$. It has been shown in [25] that the following relations hold for all $t \in [0,T]$:

$$\begin{cases} g_y^{(j)}(y(t)) A_l(t^\pm) = 0 \qquad \text{for } j = 0, \ldots, q-2; \qquad l = 0, \ldots, q-2-j, \\ g_u^{(q)}(u(t^\pm), y(t)) = g_y^{(q-l-1)}(y(t)) A_l(t^\pm) \qquad \text{for } l = 0, \ldots, q-1, \end{cases} \tag{2.35}$$

where $t^\pm$ denotes, on both sides of the equality, either t^- or t^+. Denote by C the $n \times q$ matrix $(g_y(y(t))^*, \dots, g_y^{(q-1)}(y(t))^*)$. The above relations imply that the $q \times q$ matrix $D := C^\top(A_{q-1}(t^\pm), \dots, A_0(t^\pm))$ is lower triangular with nonzero diagonal elements equal to $g_u^{(q)}(u(t^\pm), y(t))$ and hence has rank q. Therefore C has rank at least q. The conclusion follows. □

PROPOSITION 2.10. *Let (u, y) be a trajectory satisfying* (A2$_q$)–(A4) *and the alternative formulation* (2.23)–(2.33). *Define the functions η_j, $0 \le j \le q-1$, the costate p, and the jump parameters $\nu_{\mathcal{T}_{en}}$ and $\nu_{\mathcal{T}_{ex}}$ by*

$$\eta_j(t) = (-1)^{q-j} \frac{\mathrm{d}^{q-j}}{\mathrm{d}t^{q-j}} \eta_q(t) \qquad \text{for } j = 0, \dots, q-1; \qquad t \in [0,T] \setminus \mathcal{T}, \tag{2.36}$$

$$p(t) = p_q(t) + \sum_{j=1}^{q} \eta_j(t) g_y^{(j-1)}(y(t)), \qquad t \in [0,T] \setminus \mathcal{T}, \tag{2.37}$$

$$\nu_{\tau_{en}} = \nu^1_{\tau_{en}} - \eta_1(\tau_{en}^+) \text{ for all } \tau_{en} \in \mathcal{T}_{en}; \qquad \nu_{\tau_{ex}} = \eta_1(\tau_{ex}^-) \text{ for all } \tau_{ex} \in \mathcal{T}_{ex}. \tag{2.38}$$

Then (u, y) is a regular Pontryagin extremal that satisfies (2.13)–(2.20) *iff all the following additional conditions are satisfied:*

$$g(y(t)) < 0 \qquad \text{on } [0,T] \setminus (\mathcal{I}_b \cup \mathcal{T}_{to}), \tag{2.39}$$

$$\eta_0(t) = (-1)^q \frac{\mathrm{d}^q}{\mathrm{d}t^q} \eta_q(t) \ge 0 \qquad \text{on } \operatorname{int} \mathcal{I}_b. \tag{2.40}$$

At all entry times τ_{en},

$$\begin{cases} \nu^1_{\tau_{en}} = \eta_1(\tau_{en}^+) & \text{if } q \text{ is odd;} \\ \nu^1_{\tau_{en}} \ge \eta_1(\tau_{en}^+) & \text{if } q \text{ is even;} \end{cases} \qquad \nu^j_{\tau_{en}} = \eta_j(\tau_{en}^+); \qquad j = 2, \dots, q. \tag{2.41}$$

At all exit times τ_{ex},

$$\begin{cases} \eta_1(\tau_{ex}^-) = 0 & \text{if } q \text{ is odd;} \\ \eta_1(\tau_{ex}^-) \ge 0 & \text{if } q \text{ is even;} \end{cases} \qquad \eta_j(\tau_{ex}^-) = 0; \qquad j = 2, \dots, q. \tag{2.42}$$

At all touch times τ_{to},

$$\nu_{\tau_{to}} \ge 0. \tag{2.43}$$

REMARK 2.11.

(i) If (u, y) is a regular Pontryagin extremal solution of (2.13)–(2.20), the functions η_j, $1 \le j \le q$, costate p_q, and alternative jump parameters $\nu^{1:q}_{\mathcal{T}_{en}}$, such that (u, y) satisfies the alternative formulation (2.23)–(2.33) and additional conditions (2.39)–(2.43), can be recovered from p, η_0, and $\nu_{\mathcal{T}}$ as follows. The functions η_j are given by (2.36) by successive integrations of η_0 over boundary arcs, with integration constants determined by the exit time conditions (2.38) for $j = 1$ and (2.42) for $j = 2, \dots, q$. Costate p_q follows then from (2.37), and jump parameters at entry times $\nu^j_{\tau_{en}}$ are given by (2.38) for $j = 1$ and (2.41) for $j = 2, \dots, q$. Jump parameters $\nu_{\mathcal{T}_{to}}$ associated with touch points are the same in both formulations.

(ii) Assumptions (A2) and (A2$_q$) are equivalent, when (2.36)–(2.37) hold, since the constraints are of order q, and hence we have

$$\tilde{H}_{uu}(u,y,p_q,\eta_q) = H_{uu}(u,y,p) - \sum_{j=1}^{q} \eta_j(t) g_y^{(j-1)}(y) f_{uu}(u,y) + \eta_q g_{uu}^{(q)}(u,y)$$
$$= H_{uu}(u,y,p) - \sum_{j=1}^{q-1} \eta_j(t) g_{uu}^{(j)}(y)(u,y) \;=\; H_{uu}(u,y,p).$$

Proof of Proposition 2.10. Since η_q is piecewise C^q by Lemma 2.8, the functions η_j, $0 \le j \le q-1$, are well defined. We show the equivalence between (2.13)–(2.20) and (2.23)–(2.33) augmented with (2.39)–(2.43).

Equivalence between state equations (2.13) and (2.23); final costate conditions (2.15) and (2.25); state constraint equations (2.17) and (2.27), (2.28), (2.30) on boundary arcs; and (2.19) and (2.29) at touch points is obvious. Equivalence between costate equations (2.14) and (2.24), and between control equations (2.16) and (2.26), follows from calculation, using the relations between the functions η_j, p, and p_q and the fact that the state constraint is of order q (see, e.g., [25]).

Additional conditions are necessary to ensure equivalence between complementarity and junction conditions. Obviously, (2.39)–(2.40) are equivalent to (2.18); as well, (2.33) and (2.43) are equivalent to (2.20) for touch points. It remains to check that (2.20) is also equivalent to (2.31)–(2.32) and (2.41)–(2.42) at entry/exit points. Let $\tau_{en} \in \mathcal{T}_{en}$. Expressing $[p_q(\tau_{en})]$, using on the one hand the relationship (2.37) between p and p_q, as well as (2.20), and using on the other hand jump condition (2.31), we obtain

$$[p_q(\tau_{en})] = -\nu_{\tau_{en}} g_y(y(\tau_{en})) - \sum_{j=1}^{q} \eta_j(\tau_{en}^+) g_y^{(j-1)}(y(\tau_{en})), \tag{2.44}$$

$$[p_q(\tau_{en})] = -\sum_{j=1}^{q} \nu_{\tau_{en}}^j g_y^{(j-1)}(y(\tau_{en})). \tag{2.45}$$

By Lemma 2.9 at $t = \tau_{en}$, the right-hand sides of (2.44) and (2.45) are equal iff the coefficients of $g_y^{(j-1)}(y(\tau_{en}))$ for $j = 1, \dots, q$ are equal. Eliminating $\nu_{\tau_{en}}$, which must be nonnegative (and equals zero for odd-order state constraints by Proposition 2.5(i)), we deduce (2.41). Proceeding similarly at exit points, (2.42) follows. ∎

Remark 2.12. Proposition 2.10 slightly improves section 5 of [25], in the sense that we give the complete set of additional conditions for which equivalence between regular Pontryagin extremals and the alternative formulation holds.

Remark 2.13. The sign condition of $\eta_q^{(q)}$ on boundary arcs (2.40) and exit point conditions (2.42) implies that the necessary condition

$$(-1)^{q-j} \frac{\mathrm{d}^{q-j}}{\mathrm{d}t^{q-j}} \eta_q(t) = \eta_j(t) \ge 0 \;\text{ on } \mathcal{I}_b \qquad \text{for } j = 1, \dots, q \tag{2.46}$$

holds as a consequence of (2.40) and (2.42). It is easily seen by induction, since $\dot{\eta}_j = -\eta_{j-1} \le 0$ on $\mathcal{I}_b$ and $\eta_j(\tau_{ex}^-) \ge 0$ for all $\tau_{ex} \in \mathcal{T}_{ex}$. By (2.41), we deduce also that $\nu_{\tau_{en}}^j \ge 0$ for all $\tau \in \mathcal{T}_{en}$ and $j = 1, \dots, q$.

2.3. The shooting algorithm. The shooting algorithm extracts from the necessary optimality conditions a finite-dimensional set of equations (the shooting equations). If its Jacobian is invertible, we obtain a locally convergent algorithm by solving the shooting equations using, say, Newton's method.

In the unconstrained case, the initial value of the costate p_0 is mapped into the final condition (2.25). To handle alternative formulation of Definition 2.7, jump parameters and junction times are introduced as *shooting parameters*. A given set of shooting parameters determines a unique trajectory and multipliers (u, y, p_q, η_q) solution of the coupled state–costate system (2.23)–(2.24) with initial condition $p_q(0) = p_0$; algebraic equations (2.26), (2.28), and (2.30) that give u and η_q as implicit functions of (y, p_q) by (A2)–(A3); and jump conditions (2.31)–(2.33).

We use the shooting formulation of Malanowski and Maurer [23, 24]. Jump parameters $\nu_{\tau_{en}}^{1:q}$ at an entry time τ_{en} are associated with the q interior points conditions (2.27). Necessary optimality conditions for entry and exit points τ_{en} and τ_{ex} and touch points τ_{to} (when $q \geq 2$) are as follows:

$$g^{(q)}(u(\tau_{en}^-), y(\tau_{en})) = 0; \qquad g^{(q)}(u(\tau_{ex}^+), y(\tau_{ex})) = 0, \tag{2.47}$$

$$g^{(1)}(y(\tau_{to})) = 0. \tag{2.48}$$

By Proposition 2.5, the control is continuous along a regular Pontryagin extremal so that (2.47) is a necessary optimality condition for entry/exit times. For a first-order state constraint, we assume in what follows that $\mathcal{T}_{to} = \emptyset$ (see Remark 2.19 below). Since a touch point τ_{to} is a local maximum of $g(y)$ when $q \geq 2$, (2.48) is a necessary optimality condition. Therefore, (2.48) together with the interior point constraint (2.29) provide two conditions associated with τ_{to} and its jump parameter $\nu_{\tau_{to}}$ for each $\tau_{to} \in \mathcal{T}_{to}$.

DEFINITION 2.14. *A trajectory* (u, y) *is a* shooting extremal *if it satisfies both the alternative formulation (Definition* 2.7*) and conditions* (2.47)–(2.48).

Let us show how (2.47) relates to the additional conditions of Proposition 2.10.

PROPOSITION 2.15. *Let* (u, y) *be a trajectory solution of the alternative formulation* (2.23)–(2.33) *and satisfying* (A2$_q$)–(A4). *Then the following two conditions are equivalent:*

(i) *The control* u *is continuous at entry/exit times* τ_{en}, τ_{ex} *(i.e.,* (2.47) *holds).*

(ii) *Those additional conditions in* (2.41)–(2.42) *involving* η_q *are satisfied, i.e.,*

$$\eta_q(\tau_{en}^+) - \nu_{\tau_{en}}^q = 0; \qquad \eta_q(\tau_{ex}^-) = 0. \tag{2.49}$$

Proof. Let $\tau_{en} \in \mathcal{T}_{en}$. By (A3), the function $\hat{u} \mapsto g^{(q)}(\hat{u}, y(\tau_{en}))$ is one-to-one. Since $g^{(q)}(u(\tau_{en}^+), y(\tau_{en})) = 0$, we have that $g^{(q)}(u(\tau_{en}^-), y(\tau_{en})) = 0$ iff the control is continuous at time τ_{en}; the same type of argument holds for exit points. It follows that (2.47) is equivalent to the continuity of the control at entry/exit points.

By (2.26), we have

$$\tilde{H}_u(u(\tau_{en}^-), y(\tau_{en}), p_q(\tau_{en}^-), 0) \;=\; 0 \;=\; \tilde{H}_u(u(\tau_{en}^+), y(\tau_{en}), p_q(\tau_{en}^+), \eta_q(\tau_{en}^+)).$$

We abbreviate $u(\tau_{en}^-)$ to u^- and so on. Using the jump condition of the costate (2.31), it follows that

$$\tilde{H}_u(u^+, y, p_q^+, \eta_q^+) = H_u(u^+, y, p_q^-) - \sum_{j=1}^{q} \nu_{\tau_{en}}^j g_y^{(j-1)}(y) f_u(u^+, y) + \eta_q^+ g_u^{(q)}(u^+, y).$$

The state constraint being of order q, we have $g_y^{(j-1)}(y) f_u(u, y) = g_u^{(j)}(y) = 0$ for $j = 1, \ldots, q-1$, and hence, we obtain

$$0 \;=\; H_u(u^+, y, p_q^-) + (\eta_q^+ - \nu_{\tau_{en}}^q) g_u^{(q)}(u^+, y).$$

Since $g_u^{(q)}(u^+, y) \neq 0$ by (A3), it follows that $H_u(u^+, y, p_q^-) = 0$ iff $\eta_q^+ = \nu_{\tau_{en}}^q$. Since by (A2$_q$) $H_u(u^+, y, p_q^-) = 0$ iff $u^+ = u^-$, we deduce that u is continuous at time τ_{en} iff $\eta_q^+ = \nu_{\tau_{en}}^q$. Similar arguments hold for exit points. The conclusion follows. ∎

REMARK 2.16. We can also check that if (u, y) is a shooting extremal satisfying (A2$_q$)–(A4), then u is continuous at touch points $\tau \in \mathcal{T}_{to}$ if $q \geq 2$. Indeed, (2.26), (2.30), and (2.33) lead to

$$H_u(u^-, y, p_q^-) = 0 = H_u(u^+, y, p_q^+) = H_u(u^+, y, p_q^-) - \nu_\tau g_y(y) f_u(y, u^+).$$

Since $g_y f_u = g_u^{(1)} \equiv 0$, and $H_u(\cdot, y, p_q^-)$ is one-to-one by (A2$_q$), we obtain $u^+ = u^-$.

It follows that if (u, y) is a shooting extremal satisfying (A2$_q$)–(A4), then u is continuous on $[0, T]$, provided that we still assume that $\mathcal{T}_{to} = \emptyset$ if $q = 1$ (see Remark 2.19).

The *structure* of a feasible trajectory is defined as the (finite) *number* of boundary arcs and touch points of the trajectory, and the *order* in which they occur w.r.t. time. Assuming the structure of the optimal trajectory is known, we define the shooting mapping as follows. Denote by N_b and N_{to} the number of boundary arcs and touch points of the trajectory, respectively. The space of shooting parameters is

$$\Theta := \mathbb{R}^n \times \mathbb{R}^{qN_b} \times \mathbb{R}^{N_{to}} \times \mathbb{R}^{N_b} \times \mathbb{R}^{N_b} \times \mathbb{R}^{N_{to}}.$$

With the above notation, and for a given order of boundary arcs and touch points, the shooting mapping $\mathcal{F}$ is defined, over a neighborhood in Θ of shooting parameters associated with a regular Pontryagin extremal, into Θ, by

$$\theta = \begin{pmatrix} p_0^* \\ \nu_{\mathcal{T}_{en}}^{1:q} \\ \nu_{\mathcal{T}_{to}} \\ \mathcal{T}_{en} \\ \mathcal{T}_{ex} \\ \mathcal{T}_{to} \end{pmatrix} \mapsto \begin{pmatrix} p_q(T)^* - \phi_y(y(T))^* \\ g^{(0:q-1)}(y(\mathcal{T}_{en})) \\ g(y(\mathcal{T}_{to})) \\ g^{(q)}(u(\mathcal{T}_{en}^-), y(\mathcal{T}_{en})) \\ g^{(q)}(u(\mathcal{T}_{ex}^+), y(\mathcal{T}_{ex})) \\ g^{(1)}(y(\mathcal{T}_{to})) \end{pmatrix}. \tag{2.50}$$

By construction, a zero of the shooting mapping $\mathcal{F}$ provides a trajectory (u, y) that is a shooting extremal. In view of Propositions 2.10 and 2.15, the following holds.

COROLLARY 2.17. *A shooting extremal satisfying* (A2$_q$)–(A4) *is a regular Pontryagin extremal iff it satisfies the following* minimal additional conditions: (2.39) *on interior arcs,* (2.40) *on boundary arcs,* (2.43) *at touch points, and for all entry points* $\tau_{en} \in \mathcal{T}_{en}$ *and exit points* $\tau_{ex} \in \mathcal{T}_{ex}$,

$$\text{if } q \geq 2 \text{ is even; } \nu_{\tau_{en}}^1 - (-1)^{q-1}\eta_q^{(q-1)}(\tau_{en}^+) \geq 0; \ (-1)^{q-1}\eta_q^{(q-1)}(\tau_{ex}^-) \geq 0; \tag{2.51}$$

$$\begin{cases} \text{if } q \geq 3 \text{ is odd, } j = 1, \ldots, q-1, \text{ and if } q \geq 4 \text{ is even, } j = 2, \ldots, q-1; \\ \nu_{\tau_{en}}^j - (-1)^{q-j}\eta_q^{(q-j)}(\tau_{en}^+) = 0; \quad (-1)^{q-j}\eta_q^{(q-j)}(\tau_{ex}^-) = 0. \end{cases} \tag{2.52}$$

Note that (2.51)–(2.52) is only a reformulation of (2.41)–(2.42), from which we removed the condition corresponding to $j = q$, namely (2.49), since the latter is automatically satisfied by Proposition 2.15. Consequently, when $q = 1$, there remain no additional conditions at entry/exit points for shooting extremals.

REMARK 2.18. It follows that for first- and second-order state constraints, and for constraints of order $q > 2$ having no boundary arcs (see Remark 4.11 concerning existence of boundary arcs for state constraints of order $q \geq 3$), the additional

conditions reduce to the *inequalities* (2.39), (2.40), (2.43), and (2.51) when $q = 2$ at entry/exit points.

REMARK 2.19. For a first-order state constraint, jump parameters $\nu_{\mathcal{T}_{to}}$ associated with touch points are equal to zero along a regular Pontryagin extremal by Proposition 2.5. For this reason, we assume in this paper that $\mathcal{T}_{to} = \emptyset$ if $q = 1$.

REMARK 2.20. The nonlocal hypotheses (A2) (or (A2$_q$)) as well as (2.10) (or (2.12)) are essential in order to prove that the control is continuous. Some of our results remain valid, substituting everywhere *stationary point* for *(regular) Pontryagin extremal*, when the assumptions (A2) and (2.10) in (A3) are replaced by the weaker assumptions that u is continuous over $[0, T]$ and that there exists $\alpha, \gamma > 0$ such that

$$H_{uu}(u(t), y(t), p(t)) \geq \alpha \quad \text{and} \quad |g_u^{(q)}(u(t), y(t))| \geq \gamma \quad \text{for all } t \in [0, T]. \tag{2.53}$$

This holds in particular for Propositions 2.4, 2.5, 2.10, 2.15, Remark 2.16, and Corollary 2.17. The same remark applies for the other results of this paper, i.e., Theorems 3.2, 3.3, 4.3; Corollary 4.10; and Lemmas A.1 and A.2 in the appendix.

3. Well-posedness of the shooting algorithm. We say that the shooting algorithm is locally *well-posed* if the Jacobian of the shooting mapping (2.50) is invertible at some local solution of $(\mathcal{P})$. This allows us to apply locally a Newton method in order to find a shooting extremal; the additional conditions for a Pontryagin extremal have to be checked afterwards.

Let us first give some definitions. Given $u \in \mathcal{U}$, recall that we denote by y_u the (unique) solution in $\mathcal{Y}$ of the state equation (2.2). This well-defined mapping is of class C^k under assumption (A0). Let the cost function be

$$J(u) = \int_0^T \ell(u(t), y_u(t))\mathrm{d}t + \phi(y_u(T)). \tag{3.1}$$

We say that a feasible trajectory $(u, y = y_u)$ is a local solution of $(\mathcal{P})$ satisfying the *quadratic growth condition* if there exists $c, r > 0$ such that

$$J(\tilde{u}) \geq J(u) + c\,\|\tilde{u} - u\|_2^2 \qquad \text{for all } \tilde{u} \in B_\infty(u, r); \quad g(y_{\tilde{u}}(t)) \leq 0 \text{ on } [0, T], \tag{3.2}$$

where B_∞ denotes the open ball in $L^\infty(0, T)$ with center u and radius r. This condition involves two norms, $L^\infty(0, T)$ for the neighborhood and $L^2(0, T)$ for the growth condition.

Let (u, y) be a regular Pontryagin extremal. We make the following strict complementarity assumption (compare to (2.40), (2.51), and (2.43), where large inequalities are replaced by strict inequalities):

(A5) (i) For all boundary arcs $[\tau_{en}, \tau_{ex}]$,

$$(-1)^q \frac{\mathrm{d}^q}{\mathrm{d}t^q}\eta_q(t) > 0 \qquad \text{a.e. on } (\tau_{en}, \tau_{ex}). \tag{3.3}$$

$$\text{If } q \text{ is } \textit{odd}, \qquad \frac{\mathrm{d}^q}{\mathrm{d}t^q}\eta_q(\tau_{en}^+) < 0; \quad \frac{\mathrm{d}^q}{\mathrm{d}t^q}\eta_q(\tau_{ex}^-) < 0. \tag{3.4}$$

$$\text{If } q \text{ is } \textit{even}, \quad \nu^1_{\tau_{en}} + \frac{\mathrm{d}^{q-1}}{\mathrm{d}t^{q-1}}\eta_q(\tau_{en}^+) > 0; \quad \frac{\mathrm{d}^{q-1}}{\mathrm{d}t^{q-1}}\eta_q(\tau_{ex}^-) < 0. \tag{3.5}$$

(ii) For all touch points $\tau_{to} \in \mathcal{T}_{to}$,

$$\nu_{\tau_{to}} > 0. \tag{3.6}$$

Recall that $(-1)^q \frac{\mathrm{d}^q}{\mathrm{d}t^q}\eta_q(t)$ equals η_0, the density of η (see Proposition 2.10). Let $\hat{q} := 2q-1$ if q is *odd* and $\hat{q} := 2q-2$ if q is *even*. By Proposition 2.5, $\hat{q}+1$ is the smallest possible order for which the corresponding time derivative of $g(y(t))$ may be discontinuous at an entry/exit point. Note that $\hat{q} = q$ for $q = 1, 2$.

LEMMA 3.1. *Let (u, y) be a regular Pontryagin extremal satisfying* (A2)–(A4). *For odd (resp., even) q, assumption* (3.4) *(resp.,* (3.5)*) holds iff the following nontangentiality condition at order $\hat{q}+1$ holds: For all entry times $\tau_{en} \in \mathcal{T}_{en}$ and all exit times $\tau_{ex} \in \mathcal{T}_{ex}$,*

$$(-1)^{\hat{q}+1}\frac{\mathrm{d}^{\hat{q}+1}}{\mathrm{d}t^{\hat{q}+1}}g(y(t))|_{t=\tau_{en}^-} < 0; \quad \frac{\mathrm{d}^{\hat{q}+1}}{\mathrm{d}t^{\hat{q}+1}}g(y(t))|_{t=\tau_{ex}^+} < 0. \tag{3.7}$$

Proof. By Proposition 2.10 (see (2.38)), (3.5) is equivalent (when q is even) to the strict positivity of $\nu_\tau > 0$ at entry/exit points $\tau \in \mathcal{T}_{en} \cup \mathcal{T}_{ex}$. The conclusion is then a consequence of Proposition 2.10 and of Lemma A.2, whose (technical) proof is given in the appendix. ∎

Assumption (A5)(ii) implies that if $q = 1$, then $\mathcal{T}_{to} = \emptyset$ by Proposition 2.5(ii). When $q \geq 2$, we assume that all touch points of (u, y) are *reducible* in the following sense:

(A6) For all touch points $\tau_{to} \in \mathcal{T}_{to}$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}g(y(t))|_{t=\tau_{to}} < 0. \tag{3.8}$$

This makes sense, since when $q \geq 2$, we have $\frac{\mathrm{d}^2}{\mathrm{d}t^2}g(y(t)) = g^{(2)}(u, y)$ and u is continuous by Proposition 2.5.

3.1. Statement of main results. Define the quadratic cost function

$$\begin{aligned}\mathcal{J}_q(v, z) \;:=\; & \int_0^T \tilde{H}_{(u,y),(u,y)}(u, y, p_q, \eta_q)((v, z), (v, z))\mathrm{d}t \\ & + \; z(T)^* \phi_{yy}(y(T))z(T) \; + \sum_{\tau\in\mathcal{T}_{en}} \sum_{j=1}^{q} \nu_\tau^j z(\tau)^* g_{yy}^{(j-1)}(y(\tau))z(\tau) \\ & + \sum_{\tau\in\mathcal{T}_{to}} \nu_\tau \left(z(\tau)^* g_{yy}(y(\tau))z(\tau) \; - \; \frac{(g_y^{(1)}(y(\tau))z(\tau))^2}{\frac{\mathrm{d}}{\mathrm{d}t}g^{(1)}(y(t))|_{t=\tau}} \right),\end{aligned} \tag{3.9}$$

where $\tilde{H}$ is the augmented Hamiltonian (2.22) and the set of constraints

$$\dot{z} = f_y(u, y)z + f_u(u, y)v \qquad \text{on } [0, T]; \qquad z(0) = 0, \tag{3.10}$$

$$g_y^{(j)}(y(\tau))z(\tau) = 0 \qquad \text{for } j = 0, \ldots, q-1; \quad \tau \in \mathcal{T}_{en}, \tag{3.11}$$

$$g_{(u,y)}^{(q)}(u(t), y(t))(v(t), z(t)) = 0, \qquad t \in \mathcal{I}_b, \tag{3.12}$$

$$g_y(y(\tau))z(\tau) = 0, \qquad \tau \in \mathcal{T}_{to}. \tag{3.13}$$

Since the state equation and constraints are linear, the cost function is quadratic, and all have bounded coefficients, we may take as linearized control and state spaces $\mathcal{V} := L^2(0, T)$ and $\mathcal{Z} := H^1(0, T; \mathbb{R}^n)$, where $H^1(0, T)$ is the Sobolev space of functions in $L^2(0, T)$ with a weak derivative in $L^2(0, T)$. Let the linear quadratic problem (PQ_q) be defined by

$$(\mathrm{PQ}_q) \qquad \min_{(v,z)\in\mathcal{V}\times\mathcal{Z}} \; \frac{1}{2}\mathcal{J}_q(v, z) \qquad \text{subject to (3.10)–(3.13).} \tag{3.14}$$

Consider the following second-order conditions:

$$(v,z)=0 \text{ is a solution of } (\mathrm{PQ}_q). \tag{3.15}$$

$$(v,z)=0 \text{ is the } \textit{unique} \text{ solution of } (\mathrm{PQ}_q). \tag{3.16}$$

THEOREM 3.2 (no-gap second-order optimality conditions). (i) *Let* (u,y) *be a local solution of* $(\mathcal{P})$ *satisfying* (A2)–(A6). *Then its associated multipliers in the alternative formulation are such that the second-order necessary condition* (3.15) *holds.*

(ii) *Let* (u,y) *be a Pontryagin extremal satisfying* (A2)–(A6). *Then the second-order sufficient condition* (3.16) *holds iff* (u,y) *is a local solution of* $(\mathcal{P})$ *satisfying the quadratic growth condition* (3.2).

THEOREM 3.3 (well-posedness of the shooting algorithm). *Let* (u,y) *be a local solution of* $(\mathcal{P})$ *satisfying* (A2)–(A6). *Then the shooting algorithm is locally well-posed (invertible Jacobian) iff the following two conditions hold:* (i) *If* $q\geq 3$, *the trajectory* (u,y) *does not have boundary arcs.* (ii) *The second-order sufficient condition* (3.16) *holds.*

In general, even for unconstrained problems, the invertibility of the Jacobian of the shooting mapping at a Pontryagin extremal does not imply that the second-order sufficient condition (3.16) holds. We comment on the ill-posedness of the shooting algorithm along boundary arc of order $q\geq 3$ in Remark 4.11.

Combining Theorems 3.2(ii) and 3.3, we obtain that if (u,y) is a local solution of $(\mathcal{P})$ satisfying (A2)–(A6) and condition (i) of Theorem 3.3, then the shooting algorithm is well-posed iff (u,y) satisfies the quadratic growth condition.

3.2. Proof of the no-gap second-order optimality conditions (Theorem 3.2). We use the no-gap second-order optimality conditions established in [6, 5]. Let (u,y) be a regular Pontryagin extremal, with the multiplier $\eta\in\mathcal{M}[0,T]$ given by (2.21). Consider the quadratic cost function

$$\begin{aligned}\mathcal{J}(v,z) \;:=\; & \int_0^T H_{(u,y),(u,y)}(u,y,p)((v,z),(v,z))\mathrm{d}t + z(T)^*\phi_{yy}(y(T))z(T)\\ & + \int_0^T (z^* g_{yy}(y)z)\,\mathrm{d}\eta \;-\; \sum_{\tau\in\mathcal{T}_{to}} \nu_\tau \frac{(g_y^{(1)}(y(\tau))z(\tau))^2}{\frac{\mathrm{d}}{\mathrm{d}t}g^{(1)}(y(t))|_{t=\tau}},\end{aligned} \tag{3.17}$$

where H is the classical Hamiltonian (2.4), and consider the constraint

$$g_y(y(t))z(t)=0 \quad \text{on } \mathcal{I}_b\cup\mathcal{T}_{to}. \tag{3.18}$$

The quadratic problem used in the formulation of the second-order optimality condition in [5] is the following:

$$(\mathrm{PQ}) \qquad \min_{(v,z)\in\mathcal{V}\times\mathcal{Z}} \ \frac{1}{2}\mathcal{J}(v,z) \qquad \text{subject to (3.10) and (3.18).} \tag{3.19}$$

THEOREM 3.4. (i) *If* (u,y) *is a local solution of* $(\mathcal{P})$ *such that* (A2)–(A6) *hold, then* $(v,z)=0$ *is solution of problem* (3.19).

(ii) *If* (u,y) *is a Pontryagin extremal such that* (A2)–(A6) *hold, it is a local solution of* $(\mathcal{P})$ *satisfying the quadratic growth condition* (3.2) *iff problem* (3.19) *has zero for a unique solution.*

Proof. See Corollary 15 and Theorems 18 and 27 in [5], or Theorem 0.1 in [6]. For the sake of completeness, let us recall the main ideas. The proof of the second-order

necessary condition is based on the computation of the curvature term obtained by Kawasaki [19, 20] in an abstract optimization framework. With the junction conditions results of Proposition 2.5 and (A5)(i), we can show that boundary arcs have no zero contribution to the curvature term. For the second-order sufficient condition, a reduction method is used around the finitely many reducible touch points. In fact, the proof of the sufficient condition is very similar to the proof of Lemma 4.9 in the stability analysis below. □

We establish the link between Theorem 3.4 and the second-order conditions (3.15)–(3.16) derived from the alternative formulation. In the end of this section we often omit the time argument when there is no ambiguity. The proof of the next lemma is easy and therefore omitted.

LEMMA 3.5. *Assume that the state constraint is of order q. Then for every trajectory (u, y) and every linearized trajectory $(v, z) \in \mathcal{V} \times \mathcal{Z}$ satisfying* (3.10)*, the following holds:*

$$\frac{\mathrm{d}^j}{\mathrm{d}t^j} g_y(y(t))z(t) = g_y^{(j)}(u, y)z, \qquad j = 1, \dots, q-1, \tag{3.20}$$

$$\frac{\mathrm{d}^q}{\mathrm{d}t^q} g_y(y(t))z(t) = g_y^{(q)}(u, y)z + g_u^{(q)}(u, y)v. \tag{3.21}$$

LEMMA 3.6. *Let (u, y) be a regular Pontryagin extremal satisfying* (A2)–(A4)*, with classical and alternative multipliers (p, η) and $(p_q, \eta_q, \nu^{1:q}_{\mathcal{T}_{en}}, \nu_{\mathcal{T}_{to}})$, respectively, related to each other by* (2.36)–(2.38), (2.41)*, and* (2.21)*. Then the quadratic cost functions $\mathcal{J}$ and $\mathcal{J}_q$, defined, respectively, in* (3.17) *and* (3.9)*, are equal to each other over the space of linearized trajectories $(v, z) \in \mathcal{V} \times \mathcal{Z}$ satisfying* (3.10)*.*

Proof. Let $(v, z) \in \mathcal{V} \times \mathcal{Z}$ satisfy (3.10) and set $\Delta_{PQ} := \mathcal{J}(v, z) - \mathcal{J}_q(v, z)$. Using (2.21), it is easily seen that the terms corresponding to the touch points and to the final time vanish, and hence we get

$$\begin{aligned}\Delta_{PQ} = &\int_0^T (p - p_q) D^2 f(u, y)((v, z), (v, z))\mathrm{d}t + \int_0^T g_{yy}(y)(z, z)\eta_0(t)\mathrm{d}t \\ &- \int_0^T D^2 g^{(q)}(u, y)((v, z), (v, z))\eta_q(t)\mathrm{d}t + \sum_{\tau \in \mathcal{T}_{ex}} \nu_\tau g_{yy}(y)(z, z)(\tau) \\ &+ \sum_{\tau \in \mathcal{T}_{en}} \left(\nu_\tau g_{yy}(y)(z, z)(\tau) \ - \sum_{j=1}^{q} \nu_\tau^j g_{yy}^{(j-1)}(y)(z, z)(\tau) \right).\end{aligned}$$

In what follows we abbreviate the notation $((v, z), (v, z))$ as $((v, z))^2$. Relations (2.36)–(2.37) between p and p_q lead to

$$\begin{aligned}\Delta_{PQ} = &\sum_{j=1}^{q} \int_0^T g_y^{(j-1)}(y) D^2 f(u, y)((v, z))^2 \eta_j(t)\mathrm{d}t + \int_0^T g_{yy}(y)(z, z)\eta_0(t)\mathrm{d}t \\ &- \int_0^T D^2 g^{(q)}(u, y)((v, z))^2 \eta_q(t)\mathrm{d}t + \sum_{\tau \in \mathcal{T}_{ex}} \nu_\tau g_{yy}(y)(z, z)(\tau) \\ &+ \sum_{\tau \in \mathcal{T}_{en}} \left(\nu_\tau g_{yy}(y)(z, z)(\tau) \ - \sum_{j=1}^{q} \nu_\tau^j g_{yy}^{(j-1)}(y)(z, z)(\tau) \right).\end{aligned} \tag{3.22}$$

The constraint being of order q, we have $g^{(j)}(u,y) = g_y^{(j-1)}(y)f(u,y)$ for $j = 0$ to $q-1$. It follows that

$$D^2g^{(j)}(u,y)((v,z))^2 = g_{yyy}^{(j-1)}(f(u,y),z,z) + 2g_{yy}^{(j-1)}(z,Df(u,y)(v,z)) + g_y^{(j-1)}D^2f(u,y)((v,z))^2. \tag{3.23}$$

In addition, by the linearized state equation (3.10), we have, for all $j = 1,\dots,q$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[g_{yy}^{(j-1)}(y(t))(z(t),z(t))\right] = g_{yyy}^{(j-1)}(y)(f(u,y),z,z) + 2g_{yy}^{(j-1)}(y)(z,Df(u,y)(v,z)),$$

which gives, by (3.23), for $j = 1,\dots,q$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[g_{yy}^{(j-1)}(y(t))(z(t),z(t))\right] = D^2g^{(j)}(u,y)((v,z))^2 - g_y^{(j-1)}(y)D^2f(u,y)((v,z))^2. \tag{3.24}$$

Since $g_u^{(j-1)}(u,y) \equiv 0$ for $j = 1,\dots,q$, we have $g_{yy}^{(j-1)}(y)(z,z) = D^2g^{(j-1)}(u,y)((v,z))^2$ for $j = 1,\dots,q$. Multiplying (3.24) by η_j, integrating over $[0,T]$, and integrating by parts on the left-hand side (recall that $\dot\eta_j = -\eta_{j-1}$), we obtain, for $j = 1,\dots,q$,

$$\begin{aligned}\int_0^T D^2g^{(j-1)}(u,y)((v,z))^2\eta_{j-1}(t)\mathrm{d}t &+ \sum_{\tau\in\tau_{ex}} g_{yy}^{(j-1)}(y)(z,z)\eta_j(\tau^-) \\ &- \sum_{\tau\in\tau_{en}} g_{yy}^{(j-1)}(y)(z,z)\eta_j(\tau^+) \\ = \int_0^T D^2g^{(j)}(u,y)((v,z))^2\eta_j(t)\mathrm{d}t &- \int_0^T g_y^{(j-1)}D^2f(u,y)((v,z))^2\eta_j(t)\mathrm{d}t.\end{aligned}$$

Adding the above equalities for $j = 1,\dots,q$, we get, after simplification by the terms $\int_0^T D^2g^{(j)}(u,y)((v,z))^2\eta_j$ for $j = 1,\dots,q-1$, that

$$\begin{aligned}\int_0^T g_{yy}(y)(z,z)\eta_0(t)\mathrm{d}t &+ \sum_{j=1}^q \sum_{\tau\in\tau_{ex}} g_{yy}^{(j-1)}(y)(z,z)\eta_j(\tau^-) \\ &- \sum_{j=1}^q \sum_{\tau\in\tau_{en}} g_{yy}^{(j-1)}(y)(z,z)\eta_j(\tau^+) \\ = \int_0^T D^2g^{(q)}(u,y)((v,z))^2\eta_q(t)\mathrm{d}t &- \sum_{j=1}^q \int_0^T g_y^{(j-1)}D^2f(u,y)((v,z))^2\eta_j(t)\mathrm{d}t.\end{aligned}$$

Substituting into (3.22) gives

$$\begin{aligned}\Delta_{PQ} = &\sum_{\tau\in\tau_{ex}} \left(\nu_\tau g_{yy}(y)(z,z)(\tau) - \sum_{j=1}^q g_{yy}^{(j-1)}(y)(z,z)\eta_j(\tau^-)\right) \\ &+ \sum_{\tau\in\tau_{en}} \left(\nu_\tau g_{yy}(y)(z,z)(\tau) + \sum_{j=1}^q \left(\eta_j(\tau^+) - \nu_\tau^j\right) g_{yy}^{(j-1)}(y)(z,z)(\tau)\right).\end{aligned}$$

Using (2.38) and additional conditions at entry and exit points (2.41)–(2.42), we obtain that $\Delta_{PQ} = 0$. Thus, the cost functions of the two quadratic problems coincide on the feasible set. ∎

Proof of Theorem 3.2. The state constraint being of order q, it follows from (3.20)–(3.21) that (3.11)–(3.13) and (3.18) are equivalent. By Lemma 3.6, problems (PQ_q) and (3.19) have the same feasible set and the same cost function on that feasible set, and hence they also have the same value and the same set of optimal solutions. The conclusion follows then from Theorem 3.4. ∎

3.3. Proof of the well-posedness (Theorem 3.3). We give a sequence of lemmas; some of them will also be used in section 4.

We denote, e.g., by $g_y^{(j)}(y(\mathcal{T}_{en}))z(\mathcal{T}_{en})$, $g_{(u,y)}^{(q)}(u(\mathcal{T}_{en}),y(\mathcal{T}_{en}))(v(\mathcal{T}_{en}^-),z(\mathcal{T}_{en}))$, the vectors in $\mathbb{R}^{N_b}$ of components $g_y^{(j)}(y(\tau))z(\tau)$, $g_{(u,y)}^{(q)}(u(\tau),y(\tau))(v(\tau^-),z(\tau))$, respectively, for $\tau\in\mathcal{T}_{en}$. By $g_y^{(0:q-1)}(y(\mathcal{T}_{en}))z(\mathcal{T}_{en})$ we denote the vector in $\mathbb{R}^{qN_b}$ of component $g_y^{(j)}(y(\tau))z(\tau)$, $0\le j\le q-1$, $\tau\in\mathcal{T}_{en}$.

LEMMA 3.7. *Let (u,y) be a shooting extremal satisfying* (A2$_q$)–(A4)*, with the set of shooting parameters $\theta_0=(p_0^*,\nu_{\mathcal{T}_{en}}^{1:q},\nu_{\mathcal{T}_{to}},\mathcal{T}_{en},\mathcal{T}_{ex},\mathcal{T}_{to})\in\Theta$, such that $\mathcal{F}(\theta_0)=0$ with the shooting mapping $\mathcal{F}$ defined in* (2.50)*. Then $\mathcal{F}$ is of class C^1 on a neighborhood Θ_0 of θ_0, and at the direction*

$$\omega:=(\pi_0^*,\gamma_{\mathcal{T}_{en}}^{1:q},\gamma_{\mathcal{T}_{to}},\sigma_{\mathcal{T}_{en}},\sigma_{\mathcal{T}_{ex}},\sigma_{\mathcal{T}_{to}})\in\Theta, \tag{3.25}$$

the vector $\mathcal{M}:=D\mathcal{F}(\theta_0)\omega$ can be split into $\mathcal{M}=(\mathcal{M}_{\mathcal{Q}}^*,\mathcal{M}_{\mathcal{T}}^*)^*$ given by

$$\mathcal{M}_{\mathcal{Q}}:=\begin{pmatrix}\pi(T)^*-\phi_{yy}(y(T))z(T)\\ g_y^{(0:q-1)}(y(\mathcal{T}_{en}))z(\mathcal{T}_{en})\\ g_y(y(\mathcal{T}_{to}))z(\mathcal{T}_{to})\end{pmatrix}, \tag{3.26}$$

$$\mathcal{M}_{\mathcal{T}}:=\begin{pmatrix}g_{(u,y)}^{(q)}(u(\mathcal{T}_{en}),y(\mathcal{T}_{en}))(v(\mathcal{T}_{en}^-),z(\mathcal{T}_{en}))+\sigma_{\mathcal{T}_{en}}\dfrac{\mathrm{d}}{\mathrm{d}t}g^{(q)}(u,y)|_{t=\mathcal{T}_{en}^-}\\ g_{(u,y)}^{(q)}(u(\mathcal{T}_{ex}),y(\mathcal{T}_{ex}))(v(\mathcal{T}_{ex}^+),z(\mathcal{T}_{ex}))+\sigma_{\mathcal{T}_{ex}}\dfrac{\mathrm{d}}{\mathrm{d}t}g^{(q)}(u,y)|_{t=\mathcal{T}_{ex}^+}\\ g_y^{(1)}(y(\mathcal{T}_{to}))z(\mathcal{T}_{to})+\sigma_{\mathcal{T}_{to}}\dfrac{\mathrm{d}}{\mathrm{d}t}g^{(1)}(y)|_{t=\mathcal{T}_{to}}\end{pmatrix}, \tag{3.27}$$

where (v,z,π,ζ), the linearized control, state, costate, and state constraint multipliers, are the solutions of (omitting arguments (u,y,p_q,η_q) and t)

$$\dot z=f_yz+f_uv \qquad \text{on } [0,T];\qquad z(0)=0, \tag{3.28}$$

$$-\dot\pi=\tilde H_{yy}z+\tilde H_{yu}v+\pi f_y+\zeta g_y^{(q)} \quad \text{on } [0,T]\setminus\mathcal{T}, \tag{3.29}$$

$$0=\tilde H_{uy}z+\tilde H_{uu}v+\pi f_u+\zeta g_u^{(q)} \quad \text{a.e. on } [0,T], \tag{3.30}$$

$$0=g_y^{(q)}z+g_u^{(q)}v \qquad \text{a.e. on } \mathcal{I}_b, \tag{3.31}$$

$$0=\zeta \qquad \text{on } [0,T]\setminus\mathcal{I}_b, \tag{3.32}$$

with initial conditions of π given by $\pi(0)=\pi_0$ and jump conditions of π given by

$$\begin{aligned}[\pi(\tau)] \;=\; &-\sum_{j=1}^{q}\nu_\tau^j z(\tau)^*g_{yy}^{(j-1)}(y(\tau))-\sum_{j=1}^{q}\gamma_\tau^j g_y^{(j-1)}(y(\tau))\\ &-\sigma_\tau\sum_{j=1}^{q-1}\nu_\tau^j g_y^{(j)}(y(\tau));\qquad \tau\in\mathcal{T}_{en},\end{aligned} \tag{3.33}$$

$$[\pi(\tau)] \;=\; 0;\qquad \tau\in\mathcal{T}_{ex}, \tag{3.34}$$

$$[\pi(\tau)] \;=\; -\nu_\tau z(\tau)^*g_{yy}(y(\tau))-\gamma_\tau g_y(y(\tau))-\sigma_\tau\nu_\tau g_y^{(1)}(y(\tau));\quad \tau\in\mathcal{T}_{to}. \tag{3.35}$$

Proof. We detail only how we obtain the jump conditions of the linearized costate π at entry times; the other equations are obvious. In view of (2.31), it is easy to check that the jump of π at $\tau \in \mathcal{T}_{en}$ is given by

$$[\pi(\tau)] = -\sum_{j=1}^{q} \nu_\tau^j z(\tau)^* g_{yy}^{(j-1)}(y(\tau)) - \sum_{j=1}^{q} \gamma_\tau^j g_y^{(j-1)}(y(\tau)) + \sigma_\tau \Delta_\tau,$$

where the vector of sensitivity coefficients Δ_τ on junction time is given by

$$\Delta_\tau = -\sum_{j=1}^{q} \nu_\tau^j g_{yy}^{(j-1)}(y(\tau)) f(u(\tau^-), y(\tau)) + [\tilde{H}_y(u(\tau), y(\tau), p_q(\tau), \eta_q(\tau))].$$

By continuity of u at junction times (Proposition 2.15) and by (2.31) we have (omitting argument τ and setting $\eta_q^+ = \eta_q(\tau^+)$)

$$\Delta_\tau = -\sum_{j=1}^{q} \nu_\tau^j g_{yy}^{(j-1)}(y) f(u,y) - \sum_{j=1}^{q} \nu_\tau^j g_y^{(j-1)}(y) f_y(u,y) + \eta_q^+ g_y^{(q)}(u,y).$$

Since $g_y^{(j)}(u,y) = g_{yy}^{(j-1)}(y) f(u,y) + g_y^{(j-1)}(y) f_y(u,y)$ for $j = 1, \dots, q$, and since by Proposition 2.15 we have $\eta_q(\tau^+) = \nu_\tau^q$, we obtain (3.33). □

We recall that a continuous quadratic form defined over a Hilbert space is a *Legendre form* (see, e.g., [17, 7]) if it is weakly lower semicontinuous and satisfies the following property: For all weakly convergent subsequences $(v_n) \subset L^2(0,T)$, $v_n \rightharpoonup v$, we have that $v_n \to v$ strongly if $Q(v_n) \to Q(v)$.

LEMMA 3.8. *Let (u,y) be a shooting extremal satisfying* (A2$_q$)–(A4). *For all $v \in \mathcal{V}$, define z_v as the (unique) solution in $\mathcal{Z}$ of the linearized state equation* (3.10), *and define the operator $\mathcal{A} : \mathcal{V} \to W := L^2(\mathcal{I}_b) \times \mathbb{R}^{qN_b} \times \mathbb{R}^{N_{to}}$ by*

$$\mathcal{A}v = \begin{pmatrix} (g_y^{(q)}(u(\cdot), y(\cdot)) z_v(\cdot) + g_u^{(q)}(u(\cdot), y(\cdot)) v(\cdot))|_{\mathcal{I}_b} \\ g_y^{(0:q-1)}(y(\mathcal{T}_{en})) z_v(\mathcal{T}_{en}) \\ g_y(y(\mathcal{T}_{to})) z_v(\mathcal{T}_{to}) \end{pmatrix}. \tag{3.36}$$

Then (i) *the continuous linear operator $\mathcal{A}$ is onto, and* (ii) *if in addition the second-order sufficient condition* (3.16) *holds, then there exists $\alpha > 0$ such that*

$$Q(v) := \mathcal{J}_q(v, z_v) \geq \alpha \|v\|_2^2, \quad \text{for all } v \in \operatorname{Ker} \mathcal{A}. \tag{3.37}$$

By $\varphi|_{\mathcal{I}_b}$, we denote the restriction to $\mathcal{I}_b$ of function φ defined over $[0,T]$.

Proof. The continuity of $\mathcal{A}$ follows from that of $\mathcal{V} \to \mathcal{Z}$, $v \mapsto z_v$. By (2.10) and Lemma 3.5, the range of the mapping $\mathcal{V} \to \mathcal{Z}$, $v \mapsto g_y(y(\cdot)) z_v(\cdot)$ is the subspace denoted by H_0^q of functions $\varphi \in H^q(0,T) = W^{q,2}(0,T)$ satisfying $\varphi^{(j)}(0) = 0$ for all $j = 0, \dots, q-1$. Point (i) follows since, by (A4), for all $(\psi(\cdot), b_{\mathcal{T}_{en}}^{1:q}, b_{\mathcal{T}_{to}}) \in W$, there exists $\varphi \in H_0^q$ such that $\varphi^{(q)}(t) = \psi(t)$ a.e. on $\mathcal{I}_b$, $\varphi^{(j-1)}(\mathcal{T}_{en}) = b_{\mathcal{T}_{en}}^j$, $j = 1, \dots, q$, and $\varphi(\mathcal{T}_{to}) = b_{\mathcal{T}_{to}}$.

By (A2$_q$), we can show that $Q(v)$ is a Legendre form over $L^2(0,T)$ (the proof is similar to that of Lemma 21 in [5]). By (3.16), we have $Q(v) > 0$ for all $v \in \operatorname{Ker} \mathcal{A} \backslash \{0\}$, which implies (3.37) by Lemma B.1. □

PROPOSITION 3.9. *Let (u,y) be a shooting extremal satisfying* (A2$_q$)–(A4) *and denote by $\theta_0 \in \Theta$ its set of shooting parameters. Assume that* (i) *the second-order*

sufficient condition (3.16) *is satisfied, and* (ii) *the following holds at junction times:*

$$\frac{\mathrm{d}}{\mathrm{d}t}g^{(q)}(u,y)|_{t=\tau^-} \neq 0 \text{ for all } \tau \in \mathcal{T}_{en}; \quad \frac{\mathrm{d}}{\mathrm{d}t}g^{(q)}(u,y)|_{t=\tau^+} \neq 0 \text{ for all } \tau \in \mathcal{T}_{ex}, \tag{3.38}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}g^{(1)}(y)|_{t=\tau} \neq 0 \quad \text{for all } \tau \in \mathcal{T}_{to}. \tag{3.39}$$

Then the Jacobian $D\mathcal{F}(\theta_0)$ *of the shooting mapping is invertible, and for all* $\delta = (a_T, b^{1:q}_{\mathcal{T}_{en}}, b_{\mathcal{T}_{to}}, c_{\mathcal{T}_{en}}, c_{\mathcal{T}_{ex}}, c_{\mathcal{T}_{to}}) \in \Theta$, *the (unique) solution* $\omega \in \Theta$ *of* $D\mathcal{F}(\theta_0)\omega = \delta$, *with* ω *given by* (3.25), *is as follows. With the notation of Lemma* 3.8, *denote by* (v_δ, w_δ) *with* $w_\delta = (\zeta_\delta, \lambda^{1:q}_{\delta,\mathcal{T}_{en}}, \lambda_{\delta,\mathcal{T}_{to}})$ *the unique solution in* $L^2(0,T) \times W$ *of the first-order optimality system of the problem*

$$(\mathcal{P}^\delta) \qquad \min_{v\in\mathcal{V}} \quad \frac{1}{2}\mathcal{J}_q(v, z_v) \; + \; a_T^* z_v(T) \; + \; \sum_{\tau\in\mathcal{T}_{to}} c_\tau \nu_\tau \frac{g_y^{(1)}(y(\tau))z_v(\tau)}{\frac{\mathrm{d}}{\mathrm{d}t}g^{(1)}(y)|_{t=\tau}}, \tag{3.40}$$
$$\text{subject to} \qquad \mathcal{A}v = (0_{L_2(\mathcal{I}_b)}, b^{1:q}_{\mathcal{T}_{en}}, b_{\mathcal{T}_{to}})^*.$$

Then $\pi_0 = \pi_\delta(0)$, *where* π_δ *is the solution on* $[0,T]\backslash\mathcal{T}$ *of* (3.29) *with* $(v_\delta, \zeta_\delta, z_\delta := z_{v_\delta})$, *with final and jump conditions of* π_δ *being given by*

$$\pi_\delta(T) = z_\delta(T)^* \phi_{yy}(y(T)) + a_T^*, \tag{3.41}$$

$$-[\pi_\delta(\tau)] = \sum_{j=1}^{q} \nu_\tau^j z_\delta(\tau)^* g_{yy}^{(j-1)}(y(\tau)) + \sum_{j=1}^{q} \lambda^j_{\delta,\tau} g_y^{(j-1)}(y(\tau)), \quad \tau \in \mathcal{T}_{en}, \tag{3.42}$$

$$-[\pi_\delta(\tau)] = 0, \quad \tau \in \mathcal{T}_{ex}, \tag{3.43}$$

$$-[\pi_\delta(\tau)] = \nu_\tau z_\delta(\tau)^* g_{yy}(y(\tau)) \; + \; \lambda_{\delta,\tau} g_y(y(\tau))$$
$$- \nu_\tau z_\delta(\tau)^* \frac{g_y^{(1)}(y(\tau))^* g_y^{(1)}(y(\tau))}{\frac{\mathrm{d}}{\mathrm{d}t}g^{(1)}(y)|_{t=\tau}} + c_\tau \nu_\tau \frac{g_y^{(1)}(y(\tau))}{\frac{\mathrm{d}}{\mathrm{d}t}g^{(1)}(y)|_{t=\tau}}, \quad \tau \in \mathcal{T}_{to}, \tag{3.44}$$

and we have $\gamma_{\mathcal{T}_{to}} = \lambda_{\delta,\mathcal{T}_{to}}$,

$$\sigma_\tau = \frac{c_\tau - g_y^{(1)}(y(\tau))z_\delta(\tau)}{\frac{\mathrm{d}}{\mathrm{d}t}g^{(1)}(y)|_{t=\tau}}, \quad \tau \in \mathcal{T}_{to}, \tag{3.45}$$

$$\sigma_\tau = \frac{c_\tau - g^{(q)}_{(u,y)}(u(\tau), y(\tau))(v_\delta(\tau^+), z_\delta(\tau))}{\frac{\mathrm{d}}{\mathrm{d}t}g^{(q)}(u,y)|_{t=\tau^+}}, \quad \tau \in \mathcal{T}_{ex}, \tag{3.46}$$

$$\sigma_\tau = \frac{c_\tau - g^{(q)}_{(u,y)}(u(\tau), y(\tau))(v_\delta(\tau^-), z_\delta(\tau))}{\frac{\mathrm{d}}{\mathrm{d}t}g^{(q)}(u,y)|_{t=\tau^-}}, \quad \tau \in \mathcal{T}_{en}, \tag{3.47}$$

$$\gamma^1_\tau = \lambda^1_{\delta,\tau}, \quad \gamma^j_\tau \; = \; \lambda^j_{\delta,\tau} - \nu^{j-1}_\tau \sigma_\tau, \quad j = 2, \ldots, q, \quad \tau \in \mathcal{T}_{en}. \tag{3.48}$$

Note that $(v_\delta, \zeta_\delta, z_\delta, \pi_\delta)$ satisfies (3.28)–(3.32). It follows by $(\mathrm{A2}_q)$ and (2.10) that $v_\delta, \zeta_\delta \in PC^q_{\mathcal{T}}[0,T]$, and hence v_δ has limits when $t \to \tau^-$ and $t \to \tau^+$ for τ in, respectively, $\mathcal{T}_{en}$ and $\mathcal{T}_{ex}$, so (3.46)–(3.47) make sense.

REMARK 3.10. Note that (3.38) is equivalent to the discontinuity of $\dot{u}$ at entry/exit points and that, when $q = 1, 2$, (3.7) implies (3.38) since $\hat{q} = q$.

REMARK 3.11. The above proposition is an explicit elimination property, valid for any order $q \geq 1$, that enables us to express the solution ω of $D\mathcal{F}(\theta_0)\omega = \delta$ as a function of the optimal solution and multipliers of the quadratic problem $(\mathcal{P}^\delta)$, *independent*

of the variations of junction times. In the case $q = 1$, the term in the factor of the variation of entry time σ_τ in (3.33) is zero so that Lemma 3.9 is nothing but the block decoupling property of the Jacobian already established in [23]. In the case $q \geq 2$, our result differs from the one in [24] since its authors use a quadratic problem depending on the variation of the entry point, leading to an additional assumption, (A.11).

Proof. Let $\delta \in \Theta$. By (i) and Lemma 3.8, Lemma B.2 (with $r = 0$) implies that the first-order optimality system of $(\mathcal{P}^\delta)$ has a unique solution and multipliers. One can easily check that (3.28)–(3.32) and (3.42)–(3.44), together with (3.41) and

$$g_y^{(0:q-1)}(y(\mathcal{T}_{en}))z_\delta(\mathcal{T}_{en}) = b^{1:q}_{\mathcal{T}_{en}}, \qquad g_y(y(\mathcal{T}_{to}))z_\delta(\mathcal{T}_{to}) = b_{\mathcal{T}_{to}}, \tag{3.49}$$

constitute the first-order optimality system of $(\mathcal{P}_\delta)$, with $\lambda^{1:q}_{\delta,\mathcal{T}_{en}}$ and $\lambda_{\delta,\mathcal{T}_{to}}$ the multipliers associated with (3.49), and thus have a unique solution $(v_\delta, z_\delta, \pi_\delta, \zeta_\delta, \lambda^{1:q}_{\delta,\mathcal{T}_{en}}, \lambda_{\delta,\mathcal{T}_{to}})$.

By (ii), define now $\sigma_\mathcal{T}$ by (3.45)–(3.47), and let $\gamma^{1:q}_{\mathcal{T}_{en}}$ and $\gamma_{\mathcal{T}_{to}}$ be related to $\lambda^{1:q}_{\delta,\mathcal{T}_{en}}$ and $\lambda_{\delta,\mathcal{T}_{to}}$ by the invertible relations (3.48) and $\gamma_{\mathcal{T}_{to}} = \lambda_{\delta,\mathcal{T}_{to}}$. Using (3.45) and (3.48) in, respectively, (3.44) and (3.42), it follows that the system of equations (3.28)–(3.32), (3.33)–(3.35), (3.41), (3.49), and (3.45)–(3.47) has a unique solution $(v_\delta, z_\delta, \pi_\delta, \zeta_\delta, \gamma^{1:q}_{\mathcal{T}_{en}}, \gamma_{\mathcal{T}_{to}}, \sigma_\mathcal{T})$. With Lemma 3.7, this implies that $D\mathcal{F}(\theta_0)\omega = \delta$ iff $\pi_0 = \pi_\delta(0)$, and the remaining variables of ω are determined by (3.45)–(3.48). Lipschitz continuity of ω w.r.t. δ is obtained as an easy consequence of Lemma B.2 and the above relations. ∎

Proof of Theorem 3.3. The proof is organized as follows. We first show the sufficiency of the conditions (i) and (ii) for the well-posedness of the shooting algorithm, which is an easy consequence of the above lemmas. After that we show that (i), and then (ii), is also necessary.

Since (A5)(i) implies, by Lemma 3.1, that (3.7) holds, (3.38) is satisfied when $q = 1, 2$ (see Remark 3.10) or trivially when the trajectory (u, y) has no boundary arc, i.e., $\mathcal{T}_{en} = \mathcal{T}_{ex} = \emptyset$. With (A6) and the second-order sufficient condition (3.16), the invertibility of the Jacobian of the shooting mapping follows from Proposition 3.9.

Let us now show the converse. Assume first that (i) does not hold; i.e., $q \geq 3$ and (u, y) has a boundary arc. By Proposition 2.5(i), $\dot{u}$ is continuous at junction times τ_{en} and τ_{ex}. Therefore, the function $\frac{\mathrm{d}}{\mathrm{d}t} g^{(q)}(u(t), y(t))$ depending on $(y, u, \dot{u})$ is also continuous at entry and exit times and vanishes on the boundary arc, so that (3.38) *does not hold* at any of the regular entry/exit times. Then it is easily seen by Lemma 3.7 that we can find some nonzero $\tilde{\omega} \in \Theta$ such that $D\mathcal{F}(\theta_0)\tilde{\omega} = 0$. Indeed, take, e.g., $\tilde{\sigma}_\tau \neq 0$ for $\tau \in \mathcal{T}_{ex}$, and all other components of $\tilde{\omega}$ equal to zero. It follows that the Jacobian of the shooting mapping is singular.

Assume now that (i) is satisfied but (ii) is not. Since (u, y) is a local solution of $(\mathcal{P})$, by Theorem 3.2 the *second-order necessary condition* (3.15) is satisfied. This says that $(v, z) = 0$ is a solution of problem (PQ_q); therefore the value of (PQ_q) is zero, the infimum is attained, and solutions of this problem do exist. If $(v, z) = 0$ is not the unique solution, that is, if the second-order sufficient condition (3.16) does not hold, this means that there exists another optimal solution $(\tilde{v}_0, \tilde{z}_0) \neq 0$ of (PQ_q), and hence a nonzero solution of its first-order optimality conditions (3.10)–(3.13), (3.28)–(3.32), with final and jump conditions of the associated costate $\tilde{\pi}_0$ given by (3.41)–(3.44) with $a_T = 0$ and $c_{\mathcal{T}_{to}} = 0$, and multipliers $(\tilde{\lambda}^{1:q}_{\mathcal{T}_{en}}, \tilde{\lambda}_{\mathcal{T}_{to}})$ associated, respectively, with (3.11) and (3.13).

Setting $\tilde{\pi}_0 := \tilde{\pi}_0(0)$, we claim that $(\tilde{\pi}_0, \tilde{\lambda}^{1:q}_{\mathcal{T}_{en}}, \tilde{\lambda}_{\mathcal{T}_{to}}) \neq 0$. Indeed, suppose that all of them were zero. Eliminating v by (3.30) as a linear function of (z, π), and integrating

from $(z(0), \pi(0)) = 0$ over the first arc the linear differential equations (3.28)–(3.29), we would have $(z, \pi, v, \zeta) = 0$, until the first junction time. If all the jump parameters $\tilde{\lambda}^j_{\mathcal{T}_{en}}$ and $\tilde{\lambda}_{\mathcal{T}_{to}}$ are equal to zero, and (v, ζ) is given by (3.30)–(3.31) on boundary arcs, we obtain $(\tilde{z}_0, \tilde{\pi}_0, \tilde{v}_0, \tilde{\zeta}_0) = 0$ over $[0, T]$, which leads to a contradiction.

Now let $\tilde{\gamma}_{\mathcal{T}_{to}} = \tilde{\lambda}_{\mathcal{T}_{to}}$ and $(\tilde{\sigma}_{\mathcal{T}}, \tilde{\gamma}^{1:q}_{\mathcal{T}_{en}})$ be solution of (3.45)–(3.48) with $c_{\mathcal{T}} = 0$. We have $\tilde{\omega} := (\tilde{\pi}_0, \tilde{\gamma}^{1:q}_{\mathcal{T}_{en}}, \tilde{\gamma}_{\mathcal{T}_{to}}, \tilde{\sigma}_{\mathcal{T}_{en}}, \tilde{\sigma}_{\mathcal{T}_{ex}}, \tilde{\sigma}_{\mathcal{T}_{to}}) \neq 0$, and by Lemma 3.7, $D\mathcal{F}(\theta_0)\tilde{\omega} = 0$. Therefore, the Jacobian of the shooting mapping is singular, which achieves the proof. ∎

4. Sensitivity analysis without strict complementarity at touch points. In this section, we show how to conduct a sensitivity analysis, removing the strict complementarity hypothesis for touch points.

Let us first note that our framework allows us to deal with nonautonomous problems (i.e., when the data f, ℓ, g depend on t) as well, by introducing an additional state variable equal to the time, provided that the data are sufficiently smooth w.r.t. t. When the original problem (2.1)–(2.3) is autonomous, we still can add the time as a state variable. This transformation affects neither the assumptions nor the first- and second-order optimality conditions in sections 2 and 3 and the condition (ii) of Theorem 4.3. Therefore, we will assume w.l.o.g. throughout this section that the problem $(\mathcal{P})$ is written such that the last component of the state variable y_n satisfies

$$\dot{y}_n(t) = 1 \qquad \text{for all } t \in [0, T]; \qquad y_n(0) = 0$$

(i.e., $y_n(t) = t$, for all t). The reason for doing so is to consider in our stability analysis a wide class of perturbations, including nonautonomous perturbations (and possibly a nonautonomous original problem). Allowing nonautonomous perturbations is indeed needed to obtain the equivalence in Theorem 4.3, even when the original problem is autonomous. We shall not repeat this assumption, which intervenes only in the proof of (i) ⇒ (ii) in Theorem 4.3.

Let M_0 be an open subset of a Banach space M (the perturbation space). Consider, for $\mu \in M_0$, the family of perturbed optimal control problems,

$$(\mathcal{P}^\mu) \qquad \min_{(u,y)\in\mathcal{U}\times\mathcal{Y}} \quad \int_0^T \tilde{\ell}(u(t), y(t), \mu)\mathrm{d}t + \tilde{\phi}(y(T), \mu) \quad \text{subject to}$$
$$\dot{y} = \tilde{f}(u(t), y(t), \mu), \quad \text{a.e. } t \in [0, T]; \quad y(0) = \tilde{y}_0(\mu),$$
$$\tilde{g}(y(t), \mu) \leq 0 \quad \text{for all } t \in [0, T],$$

where $\tilde{\ell} : \mathbb{R} \times \mathbb{R}^n \times M_0 \to \mathbb{R}$, $\tilde{\phi} : \mathbb{R}^n \times M_0 \to \mathbb{R}$, $\tilde{f} : \mathbb{R} \times \mathbb{R}^n \times M_0 \to \mathbb{R}^n$, $\tilde{g} : \mathbb{R}^n \times M_0 \to \mathbb{R}$, and $\tilde{y}_0 : M_0 \to \mathbb{R}^n$ are at least C^2 mappings. We denote $y^\mu_0 := \tilde{y}_0(\mu)$, $\ell^\mu(u, y) := \tilde{\ell}(u, y, \mu)$, etc., and identify $(\ell^\mu, \phi^\mu, f^\mu, g^\mu, y^\mu_0)$ with problem $(\mathcal{P}^\mu)$.

We say that $(\mathcal{P}^\mu)$ is a *q-stable extension* of $(\mathcal{P})$ if (i) there exists $\mu_0 \in M_0$ such that $(\mathcal{P}^{\mu_0}) = (\mathcal{P})$ (i.e., $\ell^{\mu_0} \equiv \ell$, etc.); (ii) the mappings $\tilde{\ell}, \tilde{\phi}, \tilde{f}, \tilde{g}$ are C^{2q}, where q is the order of the state constraint of problem $(\mathcal{P})$; (iii) the state constraints are of order q for all $\mu \in M_0$; and (iv) the mappings f^μ are Lipschitz continuous over $\mathbb{R} \times \mathbb{R}^n$, uniformly over $\mu \in M_0$.

For each $\mu \in M_0$, problem $(\mathcal{P}^\mu)$ satisfies (A0); taking if necessary a smaller neighborhood of μ_0, we may assume that (A1) holds as well. Given $(\mu, u, v) \in M_0 \times \mathcal{U} \times \mathcal{V}$, denote by $(y^\mu_u, z^\mu_{u,v}) \in \mathcal{Y} \times \mathcal{Z}$ the state and linearized state solutions of

$$\dot{y}^\mu_u = f^\mu(u, y^\mu_u); \quad y^\mu_u(0) = y^\mu_0, \tag{4.1}$$
$$\dot{z}^\mu_{u,v} = f^\mu_y(u, y^\mu_u) z^\mu_{u,v} + f^\mu_u(u, y^\mu_u) v; \quad z^\mu_{u,v}(0) = 0, \tag{4.2}$$

and let $J^\mu(u) := \int_0^T \ell^\mu(u(t), y_u^\mu(t))\mathrm{d}t + \phi^\mu(y_u^\mu(T))$.

In what follows, $(\bar u, \bar y)$ denotes a Pontryagin extremal of $(\mathcal{P}) \equiv (\mathcal{P}^{\mu_0})$, with associated multipliers $(\bar p, \bar\eta)$. We denote by $\theta_0 \in \Theta$ the vector of shooting parameters associated with $(\bar u, \bar y)$.

We say that a feasible trajectory (u, y) for $(\mathcal{P}^\mu)$ has a *neighboring structure* to that of $(\bar u, \bar y)$ if the structure of (u, y) (number and order of boundary arcs and touch points) differs from that of $(\bar u, \bar y)$ only by possibly removing some nonessential touch points. With a trajectory (u, y) having a neighboring structure to that of $(\bar u, \bar y)$ is naturally associated a set of shooting parameters $\hat\theta$, but the latter may have a lower dimension than θ_0 if (u, y) has (strictly) fewer touch points than $(\bar u, \bar y)$. We can show (and this is precisely the idea of reduction methods; see further) that when $\|u - \bar u\|_\infty$ and $\|\mu - \mu_0\|$ are small enough and $q \geq 2$, for every touch point τ_{to} of $(\bar u, \bar y)$ satisfying (3.8), the function $g^\mu(y(\cdot))$ reaches its maximum over a small neighborhood of τ_{to} at a unique time denoted τ'_{to}. Then adding to $\hat\theta$ this time τ'_{to} and a zero jump parameter, and doing so for each touch point of $(\bar u, \bar y)$ that is inactive for (u, y), we obtain an *augmented* vector of shooting parameters θ having the same dimension as θ_0. Therefore the following definition makes sense.

Definition 4.1. *We say that the* uniform second-order quadratic growth *condition holds if, for every q-stable extension $(\mathcal{P}^\mu)$ there exist $c > 0$ and open neighborhoods $V_\mu \times V_u \times V_\theta$ of $(\mu_0, \bar u, \theta_0)$ in $M_0 \times \mathcal{U} \times \Theta$, such that for all $\mu \in V_\mu$, there exists a unique stationary point $(u^\mu, y^\mu := y^\mu_{u^\mu}) \in V_u \times \mathcal{Y}$ of $(\mathcal{P}^\mu)$ having a neighboring structure to that of $(\bar u, \bar y)$ with its augmented shooting parameters in V_θ, and that point satisfies*

$$J^\mu(u) \geq J^\mu(u^\mu) + c\|u - u^\mu\|_2^2, \qquad \text{for all } u \in V_u, \quad g^\mu(y_u^\mu) \leq 0 \text{ on } [0, T]. \tag{4.3}$$

As a consequence of the definition of the uniform growth condition, we have $\bar u = u^{\mu_0}$ and $\bar y = y^{\mu_0}$.

Note that in the uniform growth condition (4.3), the neighborhood (in L^∞) on which u^μ satisfies the quadratic growth condition is independent of μ. Our definition of uniform quadratic growth is different from the one in [7, section 5.1], since the latter implies the local uniqueness of solutions of the first-order optimality system (stationary points). Here, since our stability analysis is based on the shooting formulation, we can argue only the uniqueness of the stationary point among the feasible trajectories that have their structure and shooting parameters "in the neighborhood" of those of $(\bar u, \bar y)$. The uniqueness of the stationary point, in a certain sense, is needed to prove the implication (i) $\Rightarrow$ (ii) in Theorem 4.3 below.

We will use the assumption below, which is a modification of (A5).

(A5′) (i) If $q \leq 2$, the following strengthening of (3.3)–(3.4) holds:

$$\exists \beta > 0 \qquad (-1)^q \frac{\mathrm{d}^q}{\mathrm{d}t^q} \bar\eta_q(t) \geq \beta \qquad \text{for all } t \in \operatorname{int} \mathcal{I}_b; \tag{4.4}$$

if $q = 2$, (3.5) holds; if $q > 2$, the trajectory $(\bar u, \bar y)$ has no boundary arc.

(ii) If $q = 1$, $(\bar u, \bar y)$ has no (nonessential) touch points.

Assumption (A5′)(i) is a strengthening of (A5)(i). It requires, in addition to (A5)(i), *uniform* strict complementarity on boundary arcs, which is stronger than (3.3) (and implies (3.4)), and that $(\bar u, \bar y)$ have no boundary arc if $q \geq 3$. Assumption (A5′)(ii) is weaker than (A5)(ii) since it allows nonessential touch points for constraints of order $q \geq 2$ only.

Define the set of increasing times in $(0,T)$ of cardinal N as

$$IT_N := \{\tau \in \mathbb{R}^N;\ 0 < \tau_1 < \cdots < \tau_N < T\}. \tag{4.5}$$

Set $\tau_0 := 0$ and $\tau_{N+1} := T$. Given $\mathcal{S} \subset IT_N$, we have a natural isomorphism between $PC^k_{\mathcal{S}}[0,T]$ and $C^k([0,1];\mathbb{R}^{N+1})$, defined by

$$\left\{\begin{array}{ll} \hat\varphi_i(s) = & \varphi(\tau_i + (\tau_{i+1}-\tau_i)s) \quad \text{for all } \ s \in (0,1), \\ \hat\varphi_i(0) = & \varphi(\tau_i^+),\ \hat\varphi_i(1) = \varphi(\tau_{i+1}^-), \end{array}\right. \qquad i = 0,\dots,N. \tag{4.6}$$

We may therefore identify the set $PC^k_N[0,T] := \cup\{PC^k_{\mathcal{S}}[0,T]; \mathcal{S} \in IT_N\}$ of all possible N-piecewise k-times continuously differentiable functions, with $C^k([0,1];\mathbb{R}^{N+1}) \times IT_N$. The corresponding notion of convergence follows: A sequence $\varphi^n \in PC^k_{\mathcal{S}^n}[0,T]$ converges to $\varphi \in PC^k_{\mathcal{S}}[0,T]$ if $\mathcal{S}^n \to \mathcal{S}$ in $\mathbb{R}^N$ and $\hat\varphi^n \to \hat\varphi$ in $C^k([0,1];\mathbb{R}^{N+1})$. Similarly, a mapping defined over an open subset W of a Banach space, $W \to PC^k_N$, $w \mapsto \varphi^w \in PC^k_{\mathcal{S}^w}$ is of class C^k if the mapping $W \to C^k([0,1];\mathbb{R}^{N+1}) \times \mathbb{R}^N$, $w \mapsto (\hat\varphi^w, \mathcal{S}_w)$ is C^k. We denote by $PC^{k,r}_N[0,T] = PC^k_N[0,T] \cap C^r[0,T]$ the subset of $PC^k_N[0,T]$ of functions having continuous derivatives on $[0,T]$ until order $r \geq 0$. The next lemma is elementary and will be used at the end of this section.

LEMMA 4.2. *Let W be an open subset of a Banach space, and $W \to PC^{1,0}_N$, $w \mapsto \varphi^w \in PC^{1,0}_{\mathcal{S}^w}$ a C^1 mapping. Then the mapping $w \mapsto \varphi^w$ is C^1 in $L^r(0,T)$ for all $1 \leq r < \infty$. More precisely, for $w \in W$, let $\mathcal{S}^w := \{\tau_1^w < \cdots < \tau_N^w\}$ and denote by $(\hat\xi^w, \sigma^w)$ the directional derivative in $C^1([0,1];\mathbb{R}^{N+1}) \times IT_N$ of the mapping $w \mapsto (\hat\varphi^w, \tau^w)$ at point w in direction $\delta w \in W$. Then the directional derivative $\tilde\xi^w$ in $L^r(0,T)$ is given by*

$$\tilde\xi^w(t) = \hat\xi_i^w\left(\frac{t-\tau_i^w}{\tau_{i+1}^w - \tau_i^w}\right) - \dot\varphi^w(t)\left(\sigma_i^w + \frac{t-\tau_i^w}{\tau_{i+1}^w-\tau_i^w}(\sigma_{i+1}^w - \sigma_i^w)\right) \text{ on } (\tau_i^w, \tau_{i+1}^w).$$

By Proposition 2.5, a regular Pontryagin extremal and its multipliers $(u^\mu, y^\mu, p^\mu, \eta^\mu)$ satisfying (A2)–(A4) belong to the product space

$$\mathcal{X}_{\mathcal{S}} := PC^{q,0}_{\mathcal{S}}[0,T] \times PC^{q+1,1}_{\mathcal{S}}([0,T];\mathbb{R}^n) \times PC^1_{\mathcal{S}}([0,T];\mathbb{R}^{n*}) \times PC^1_{\mathcal{S}}[0,T], \tag{4.7}$$

with here $\mathcal{S} = \mathcal{T}$, which is the finite set of its junction times assumed to be of cardinal N. So let us define the union $\mathcal{X}_N$ of all such spaces, and define as well some other sets needed later:

$$\begin{aligned} &\mathcal{X}_N := \cup\{\mathcal{X}_{\mathcal{S}};\ \mathcal{S} \in IT_N\}, \\ &\mathcal{X}^q_{\mathcal{S}} := PC^q_{\mathcal{S}}[0,T] \times PC^{q+1,0}_{\mathcal{S}}([0,T];\mathbb{R}^n) \times PC^{q+1}_{\mathcal{S}}([0,T];\mathbb{R}^{n*}) \times PC^q_{\mathcal{S}}[0,T], \\ &\mathcal{X}^1_{\mathcal{S}} := PC^q_{\mathcal{S}}[0,T] \times PC^{q+1,0}_{\mathcal{S}}([0,T];\mathbb{R}^n) \times PC^1_{\mathcal{S}}([0,T];\mathbb{R}^{n*}) \times PC^1_{\mathcal{S}}[0,T], \\ &\mathcal{X}^q_N := \cup\{\mathcal{X}^q_{\mathcal{S}};\ \mathcal{S} \in IT_N\}, \qquad \mathcal{X}^1_N := \cup\{\mathcal{X}^1_{\mathcal{S}};\ \mathcal{S} \in IT_N\}. \end{aligned}$$

The main result of this section is the next theorem, which gives stability results for the optimal control problem $(\mathcal{P})$ without assuming strict complementarity at the touch points. Therefore we cannot directly apply the implicit function theorem as it was done in [23, 24] and our section 3.

THEOREM 4.3. *Let $(\bar u, \bar y)$ be a Pontryagin extremal of $(\mathcal{P})$ satisfying* (A2)–(A4), (A5′) and (A6). *Then the following statements are equivalent:*

(i) *The uniform second-order quadratic growth condition (Definition* 4.1*) holds. Denote by $u^\mu \in V_u$ the solution of* (4.3) *for $\mu \in V_\mu$, and set $y^\mu := y^\mu_{u^\mu}$. With (u^μ, y^μ)*

are associated a unique costate p^μ *and state constraint multiplier* η^μ*, and the mapping* $\mu \mapsto (u^\mu, y^\mu, p^\mu, \eta^\mu) \in \mathcal{X}_N$ *is Lipschitz continuous over* V_μ.

(ii) *The following strong second-order sufficient condition holds:*

$$\begin{aligned} &\mathcal{J}(v,z) > 0 \text{ for all } (v,z) \in \mathcal{V} \times \mathcal{Z} \setminus \{0\} \text{ satisfying (3.10) and} \\ &g_y(\bar{y}(t))z(t) = 0 \quad \text{for all } t \in \mathcal{I}_b \cup \mathcal{T}_{to}^{ess}. \end{aligned} \tag{4.8}$$

REMARK 4.4. Note that condition (ii) is stronger than the following second-order characterization of quadratic growth (3.2) (see [5]):

$$\begin{aligned} &\mathcal{J}(v,z) > 0 \text{ for all } (v,z) \in \mathcal{V} \times \mathcal{Z} \setminus \{0\} \text{ satisfying (3.10), (4.8), and} \\ &g_y(\bar{y}(\tau))z(\tau) \le 0 \quad \text{for all } \tau \in \mathcal{T}_{to} \setminus \mathcal{T}_{to}^{ess}. \end{aligned}$$

We need the following notation. Denote by $\mathcal{T}_{to}^{nes} := \mathcal{T}_{to} \setminus \mathcal{T}_{to}^{ess}$ the subset of *nonessential touch points* of the trajectory $(\bar{u}, \bar{y})$. For μ close to μ_0, let $\mathcal{F}(\cdot, \mu)$ be the shooting mapping (2.50) for problem $(\mathcal{P}^\mu)$, with the *same structure* as the trajectory $(\bar{u}, \bar{y})$, i.e., the same number of boundary arcs and touch points and the same order of their occurrence w.r.t. time. Thus nonessential touch points are present in the shooting mapping and may be active or inactive for the perturbed problem. Let $\bar{N} := n + (q+2)N_b + 2N_{to}$ denote the dimension of the shooting mapping, with $N_b = \operatorname{Card} \mathcal{T}_{en} = \operatorname{Card} \mathcal{T}_{ex}$ and $N_{to} = \operatorname{Card} \mathcal{T}_{to}$, and denote by N_0 the cardinal of $\mathcal{T}_{to}^{nes}$, the set of nonessential touch points. Split $\mathcal{F}$ into two components such that $\mathcal{F}(\cdot, \mu) = (\Phi(\cdot, \mu)^*, \Psi(\cdot, \mu)^*)^*$ and Ψ corresponds to the component $g^\mu(y(\mathcal{T}_{to}^{nes})) \in \mathbb{R}^{N_0}$. We consider the following problem for μ close to μ_0: Find

$$\theta = (p_0^{\mu *}, \nu_{\mathcal{T}_{en}}^{\mu,1:q}, \nu_{\mathcal{T}_{to}^{ess}}^{\mu}, \nu_{\mathcal{T}_{to}^{nes}}^{\mu}, \mathcal{T}_{en}^{\mu}, \mathcal{T}_{ex}^{\mu}, \mathcal{T}_{to}^{\mu,ess}, \mathcal{T}_{to}^{\mu,nes}) \in \Theta \tag{4.9}$$

such that

$$\Phi(\theta, \mu) = 0; \quad \Psi(\theta, \mu) \in \mathbb{R}_-^{N_0} \cap (\nu_{\mathcal{T}_{to}^{nes}}^{\mu})^\perp; \quad \nu_{\mathcal{T}_{to}^{nes}}^{\mu} \in \mathbb{R}_+^{N_0}. \tag{4.10}$$

In (4.10), we express the complementarity condition for nonessential touch points only. The complementarity condition at essential touch points and boundary arcs, where strict complementarity is satisfied, will hold by continuity, since we perform a local analysis (see further Lemmas 4.6–4.7).

The point θ_0, solution of (4.10) for $\mu = \mu_0$, is said to be *strongly regular* (see Robinson [28]) if there exists a neighborhood $V_\theta' \times V_\delta$ in $\mathbb{R}^{\bar{N}} \times \mathbb{R}^{\bar{N}}$ of $(\theta_0, 0)$ such that for all $\delta \in V_\delta$, $\delta = (\delta_1, \delta_2) \in \mathbb{R}^{\bar{N}-N_0} \times \mathbb{R}^{N_0}$, there exists a unique solution θ in V_θ' of

$$\begin{aligned} &D_\theta \Phi(\theta_0, \mu_0)(\theta - \theta_0) - \delta_1 = 0, \\ &D_\theta \Psi(\theta_0, \mu_0)(\theta - \theta_0) - \delta_2 \in \mathbb{R}_-^{N_0} \cap \nu_{\mathcal{T}_{to}^{nes}}^{\perp}; \quad \nu_{\mathcal{T}_{to}^{nes}} \in \mathbb{R}_+^{N_0}, \end{aligned} \tag{4.11}$$

and the mapping $\Xi : \delta \mapsto \theta(\delta)$ is Lipschitz continuous over V_δ. If θ_0 is strongly regular, then by [28], there exists a neighborhood $V_\theta \times V_\mu$ of (θ_0, μ_0) such that for each $\mu \in V_\mu$, (4.10) has in V_θ a unique solution θ^μ and there exists $\kappa > 0$ such that for all $\mu, \mu' \in V_\mu$,

$$|\theta^\mu - \theta^{\mu'}| \le \kappa \|\mu - \mu'\|. \tag{4.12}$$

In addition, the following expansion of θ^μ holds (see [7, eq. (5.41), p. 413]):

$$\theta^\mu = \Xi(-D_\mu \mathcal{F}(\theta_0, \mu_0)(\mu - \mu_0)) + o(\|\mu - \mu_0\|). \tag{4.13}$$

4.1. Stability analysis (proof of Theorem 4.3). The first step in the proof of (ii) ⇒ (i) in Theorem 4.3 is to show that (ii) implies the strong regularity property (Lemma 4.5). The existence of a (locally unique) shooting extremal (u^μ, y^μ) for problem $(\mathcal{P}^\mu)$ having its shooting parameters in the neighborhood of those of $(\bar u, \bar y)$ follows (Lemma 4.6). The next step is to check the additional conditions of Corollary 2.17, implying that (u^μ, y^μ) is a stationary point (Lemma 4.7). We end the proof by checking that u^μ satisfies the uniform quadratic growth condition (4.3) (Lemmas 4.8–4.9).

LEMMA 4.5. *Under the assumptions of Theorem* 4.3*, condition* (ii) *of Theorem* 4.3 *implies that* θ_0 *is a strongly regular solution of* (4.10) *for* $\mu = \mu_0$.

Proof. The proof is somewhat similar to that of Proposition 3.9. Let $\delta = (\delta_1, \delta_2) \in \mathbb{R}^{\bar N - N_0} \times \mathbb{R}^{N_0}$ with

$$\delta_1 = (a_T, b^{1:q}_{\mathcal{T}_{en}}, b_{\mathcal{T}^{ess}_{to}}, c_{\mathcal{T}_{en}}, c_{\mathcal{T}_{ex}}, c_{\mathcal{T}^{ess}_{to}}, c_{\mathcal{T}^{nes}_{to}}); \quad \delta_2 = b_{\mathcal{T}^{nes}_{to}}.$$

Let us show that there exists a unique $\omega \in \Theta$,

$$\omega = (\pi^*_0, \gamma^{1:q}_{\mathcal{T}_{en}}, \gamma_{\mathcal{T}^{ess}_{to}}, \gamma_{\mathcal{T}^{nes}_{to}}, \sigma_{\mathcal{T}_{en}}, \sigma_{\mathcal{T}_{ex}}, \sigma_{\mathcal{T}^{ess}_{to}}, \sigma_{\mathcal{T}^{nes}_{to}}),$$

solution of the following relation, equivalent to (4.11) with $\omega = \theta - \theta_0$:

$$\begin{aligned} &D_\theta \Phi(\theta_0, \mu_0)\omega - \delta_1 = 0, \\ &D_\theta \Psi(\theta_0, \mu_0)\omega - \delta_2 \in \mathbb{R}^{N_0}_- \cap \gamma^\perp_{\mathcal{T}^{nes}_{to}}; \quad \gamma_{\mathcal{T}^{nes}_{to}} \in \mathbb{R}^{N_0}_+. \end{aligned} \tag{4.14}$$

Consider the following linear quadratic optimal control problem:

$$\begin{aligned} (\mathcal{P}^\delta) \qquad \min_{v \in \mathcal{V}} \quad & \frac{1}{2}\mathcal{J}_q(v, z_v) \;+\; a_T^* z_v(T) \;+\; \sum_{\tau \in \mathcal{T}_{to}} c_\tau \nu_\tau \frac{g_y^{(1)}(y(\tau)) z_v(\tau)}{\frac{\mathrm{d}}{\mathrm{d}t} g^{(1)}(y)|_{t=\tau}} \\ \text{subject to} \quad & Av \;=\; (0_{L_2(\mathcal{I}_b)}, b^{1:q}_{\mathcal{T}_{en}}, b_{\mathcal{T}^{ess}_{to}})^*; \; Bv \;\le\; b_{\mathcal{T}^{nes}_{to}}, \end{aligned} \tag{4.15}$$

where $\mathcal{J}_q(v, z_v)$ is defined by (3.9) and the linear operators A, B are defined by

$$\begin{aligned} Av \;&:=\; \begin{pmatrix} (g_y^{(q)}(u(\cdot), y(\cdot)) z_v(\cdot) + g_u^{(q)}(u(\cdot), y(\cdot)) v(\cdot))|_{\mathcal{I}_b} \\ g_y^{(0:q-1)}(y(\mathcal{T}_{en})) z_v(\mathcal{T}_{en}) \\ g_y(y(\mathcal{T}^{ess}_{to})) z_v(\mathcal{T}^{ess}_{to}) \end{pmatrix}, \\ Bv \;&:=\; g_y(y(\mathcal{T}^{nes}_{to})) z_v(\mathcal{T}^{nes}_{to}). \end{aligned} \tag{4.16}$$

Being equal to $\mathcal{A}$ defined in (3.36), the operator (A, B) is onto by Lemma 3.8. By Lemma B.1, the Legendre form $\bar Q(v) := \mathcal{J}_q(v, z_v)$ is coercive over $\operatorname{Ker} A$. It follows from Lemma B.2 that the first-order optimality system of problem $(\mathcal{P}^\delta)$ has a unique solution $v_\delta \in \mathcal{V}$, with a unique associated Lagrange multiplier $(\zeta_\delta, \lambda^{1:q}_{\delta, \mathcal{T}_{en}}, \lambda_{\delta, \mathcal{T}^{ess}_{to}}, \lambda_{\delta, \mathcal{T}^{nes}_{to}})$ in $L^2(\mathcal{I}_b) \times \mathbb{R}^{qN_b} \times \mathbb{R}^{N_{to} - N_0} \times \mathbb{R}^{N_0}$, and the mapping $\delta \mapsto (v_\delta, \zeta_\delta, \lambda^{1:q}_{\delta, \mathcal{T}_{en}}, \lambda_{\delta, \mathcal{T}^{ess}_{to}}, \lambda_{\delta, \mathcal{T}^{nes}_{to}})$ is Lipschitz continuous. Now, defining as in Proposition 3.9 $\sigma_{\mathcal{T}}$ by (3.45)–(3.47) and defining $\gamma^{1:q}_{\mathcal{T}_{en}}$, $\gamma_{\mathcal{T}^{ess}_{to}}$, $\gamma_{\mathcal{T}^{nes}_{to}}$ by the invertible relations (3.48), $\gamma_{\mathcal{T}^{ess}_{to}} = \lambda_{\delta, \mathcal{T}^{ess}_{to}}$ and $\gamma_{\mathcal{T}^{nes}_{to}} = \lambda_{\delta, \mathcal{T}^{nes}_{to}}$, this implies that the system of equations (3.28)–(3.32), (3.33)–(3.35), (3.41), (3.45)–(3.47), together with the constraints and complementarity conditions of $(\mathcal{P}^\delta)$

$$\begin{gathered} g_y^{(0:q-1)}(y(\mathcal{T}_{en})) z_\delta(\mathcal{T}_{en}) = b^{1:q}_{\mathcal{T}_{en}}, \qquad g_y(y(\mathcal{T}^{ess}_{to})) z_\delta(\mathcal{T}^{ess}_{to}) = b_{\mathcal{T}^{ess}_{to}}, \\ g_y(y(\mathcal{T}^{nes}_{to})) z_\delta(\mathcal{T}^{nes}_{to}) \le b_{\mathcal{T}^{nes}_{to}}, \quad \gamma_{\mathcal{T}^{nes}_{to}} \ge 0, \quad (g_y(y(\mathcal{T}^{nes}_{to})) z_\delta(\mathcal{T}^{nes}_{to}) - b_{\mathcal{T}^{nes}_{to}}) \perp \gamma_{\mathcal{T}^{nes}_{to}}, \end{gathered}$$

has a unique solution $(v_\delta, z_\delta, \pi_\delta, \zeta_\delta, \gamma^{1:q}_{\mathcal{T}_{en}}, \gamma_{\mathcal{T}^{ess}_{to}}, \gamma_{\mathcal{T}^{nes}_{to}}, \sigma_{\mathcal{T}})$. Thus by Lemma 3.7, we obtain that ω is a solution of (4.14) iff $\pi_0 = \pi_\delta(0)$ and the other variables of ω are given as above. The existence and uniqueness of ω follows, and it is not difficult to check the Lipschitz continuity of ω w.r.t. δ. □

By strong regularity, there exist neighborhoods V_μ and V_θ of μ_0 and θ_0 such that, for all $\mu \in V_\mu$, there exists in V_θ a unique solution θ^μ of (4.10):

$$\theta^\mu = (p_0^{\mu *}, \nu^{\mu,1:q}_{\mathcal{T}_{en}}, \nu^{\mu}_{\mathcal{T}^{ess}_{to}}, \nu^{\mu}_{\mathcal{T}^{nes}_{to}}, \mathcal{T}^{\mu}_{en}, \mathcal{T}^{\mu}_{ex}, \mathcal{T}^{\mu,ess}_{to}, \mathcal{T}^{\mu,nes}_{to}) \in V_\theta \subset \mathbb{R}^{\bar{N}}.$$

Denote the associated trajectory and multipliers by $(u^\mu, y^\mu, p^\mu_q, \eta^\mu_q) \in \mathcal{X}^q_N$. Recall that $\Psi(\theta^\mu, \mu) = g^\mu(y^\mu(\mathcal{T}^{\mu,nes}_{to}))$ and set

$$\mathcal{T}^\mu_{to} := \mathcal{T}^{\mu,ess}_{to} \cup \{\tau \in \mathcal{T}^{\mu,nes}_{to} \ ; \ g^\mu(y^\mu(\tau)) = 0\}.$$

By the definition of (4.10), we have $g^\mu(y^\mu(\tau)) < 0$ and $\nu^\mu_\tau = 0$ if $\tau \notin \mathcal{T}^\mu_{to}$. Hence $(u^\mu, y^\mu, p^\mu_q, \eta^\mu_q)$ is a shooting extremal for $(\mathcal{P}^\mu)$, with jump parameters $(\nu^{\mu,1:q}_{\mathcal{T}_{en}}, \nu^{\mu}_{\mathcal{T}^\mu_{to}})$ and junction times $(\mathcal{T}^\mu_{en}, \mathcal{T}^\mu_{ex}, \mathcal{T}^\mu_{to})$.

In order to show now that the mapping $\mu \mapsto (u^\mu, y^\mu, p^\mu, \eta^\mu)$ is Lipschitz continous, where (p^μ, η^μ) is given by (2.36)–(2.38) and (2.21), consider the mapping

$$V_\mu \times V_\theta \to \mathcal{X}^q_N, \qquad (\mu, \theta) \mapsto (u^{\mu,\theta}, y^{\mu,\theta}, p^{\mu,\theta}_q, \eta^{\mu,\theta}_q), \tag{4.17}$$

where $(u^{\mu,\theta}, y^{\mu,\theta}, p^{\mu,\theta}_q, \eta^{\mu,\theta}_q)$ is the solution of (2.23)–(2.24), (2.26), (2.28), (2.30), and (2.31)–(2.33) for $(\mathcal{P}^\mu)$, with initial value of the costate, jump parameters and junction times given by argument θ. By the Cauchy–Lipschitz theorem, this mapping is well defined and of class C^q on neighborhoods $V_\mu \times V_\theta$ of (μ_0, θ_0). Therefore the mapping

$$V_\mu \times V_\theta \to \mathcal{X}^1_N, \qquad (\mu, \theta) \mapsto (u^{\mu,\theta}, y^{\mu,\theta}, p^{\mu,\theta}, \eta^{\mu,\theta}), \tag{4.18}$$

where $\eta^{\mu,\theta}_j$, $0 \le j \le q-1$, $p^{\mu,\theta}$, and $\eta^{\mu,\theta}$ are defined by (2.36)–(2.38) and (2.21), is of class C^1.

LEMMA 4.6. *Under assumptions and condition* (ii) *of Theorem* 4.3*, there exists a neighborhood V_μ of μ_0 such that the mapping $V_\mu \to \mathcal{X}_N$, $\mu \mapsto (u^\mu, y^\mu, p^\mu, \eta^\mu)$, is well defined and Lipschitz continuous on V_μ.*

Proof. Since strong regularity holds by Lemma 4.5, the mapping $\mu \mapsto \theta^\mu$ solution of (4.10) is well defined on a neighborhood of μ and Lipschitz continuous by (4.12). By continuity of the mappings (4.18) and $\mu \mapsto \theta^\mu$, the mapping $\mu \mapsto (u^\mu, y^\mu, p^\mu, \eta^\mu)$ is continuous $V_\mu \to \mathcal{X}^1_N$. Let us show now that u^μ is continuous. By (A2)–(A3), reducing V_μ if necessary, we have $H^\mu_{uu}(\hat{u}, y^\mu(t), p^\mu(t^\pm)) \ge \alpha/2$ and $|(g^\mu)^{(q)}_u(\hat{u}, y^\mu(t))| \ge \gamma/2$ for all t and all $\hat{u}$ in the segment $[u^\mu(t^-), u^\mu(t^+)] := \{\sigma u^\mu(t^+) + (1-\sigma)(u(t^-)),\ \sigma \in [0,1]\}$. By arguments similar to those used in the proof of Proposition 2.15(i) and in Remark 2.16, this is enough to show that u^μ is continuous, and hence, $(u^\mu, y^\mu) \in PC^{q,0}_{\mathcal{T}^\mu}[0,T] \times PC^{q+1,1}_{\mathcal{T}^\mu}([0,T]; \mathbb{R}^n)$. Reducing V^μ if necessary, by composition of $\mu \mapsto \theta^\mu$ with the C^1 mapping (4.18), we deduce that the mapping $\mu \mapsto (u^\mu, y^\mu, p^\mu, \eta^\mu) \in \mathcal{X}_N$ is Lipschitz continuous on a neighborhood of μ. □

LEMMA 4.7. *Under assumptions and condition* (ii) *of Theorem* 4.3*, the shooting extremal (u^μ, y^μ) is a stationary point for problem $(\mathcal{P}^\mu)$.*

Proof. By Corollary 2.17 and Remark 2.20, we need to check (2.39), (2.40), (2.43), and also, when $q = 2$, (2.51). By (A5′) and Lemma 4.6, (2.40) follows from (4.4). If $q = 2$, (2.51) follows from (3.5). By continuity of jumps at essential touch points

and the definition of (4.10), we obtain (2.43). It remains to prove (2.39). Near an entry/exit point τ^μ (when $q = 1$ or 2) this is a consequence of hypothesis (3.7) and continuity w.r.t. μ of $u(\tau^{\mu\pm})$. Similarly, near touch points, this follows from the reducibility hypothesis (3.8). Finally, outside a small neighborhood of contact points, we obtain that $g^\mu(y^\mu) < 0$ by a standard compactness argument. ∎

The next two lemmas extend those in [5, section 4] to the setting of perturbed optimal control problems. In what follows we denote by $\mathrm{supp}(\mathrm{d}\eta)$ the support of the measure η in $\mathcal{M}[0,T]$.

LEMMA 4.8. *Assume that the assumptions and condition* (ii) *of Theorem* 4.3 *hold. Let* $(\mathcal{P}^\mu)$ *be a q-stable extension, and* $\mu_n \to \mu_0$ *with its associated shooting extremal* (u_n, y_n) *and multipliers* (p_n, η_n). *For* $v \in \mathcal{V}$, *define* $Q^n(v) := \mathcal{J}^{\mu_n}(v, z^{\mu_n}_{u_n,v})$, *where* $\mathcal{J}^{\mu_n}(\cdot,\cdot)$ *is given by* (3.17) *for* $(\mathcal{P}^{\mu_n})$ *and* $z^{\mu_n}_{u_n,v}$ *is defined by* (4.2). *Define similarly* $\bar{Q}(v) := \mathcal{J}^{\mu_0}(v, z^{\mu_0}_{\bar{u},v})$. *Let* $v_n \rightharpoonup \bar{v} \in L^2$. *Then it holds that*

$$\bar{Q}(\bar{v}) \;\le\; \liminf Q^n(v_n) \quad \textit{and} \quad v_n \to \bar{v} \textit{ strongly if } Q^n(v_n) \to \bar{Q}(\bar{v}). \tag{4.19}$$

Set $z_n := z^{\mu_n}_{u_n,v_n}$, *and assume in addition that* $g^{\mu_n}_y(y_n(t))z_n(t) \le r_n$, *where* $\|r_n\|_\infty \to 0$ *for all* $t \in \mathrm{supp}(\mathrm{d}\eta_n)$ *and all* n. *Let* $\bar{z} := z^{\mu_0}_{\bar{u},\bar{v}}$. *Then*

$$g_y(\bar{y}(t))\bar{z}(t) \;\le\; 0 \quad \textit{on } \mathrm{supp}(\mathrm{d}\bar{\eta}). \tag{4.20}$$

Proof. Since by Lemma 4.6, (u_n, y_n) converges uniformly to $(\bar{u}, \bar{y})$, and $v_n \rightharpoonup v$, we have that (z_n) converges weakly in H^1 to $\bar{z}$ and hence uniformly. Relation (4.20) follows from the convergence of η_n in PC^1_N, strict complementarity (4.4), and uniform convergence of $g^{\mu_n}_y(y_n)z_n$. Let us now show (4.19).

Set $Q^0_n(v_n) := \int_0^T v_n^* H^{\mu_n}_{uu}(u_n, y_n, p_n)v_n \mathrm{d}t$. By Lemma 4.6, uniform convergence of z_n, and convergence in $\mathcal{X}_N$ of $H^{\mu_n}_{uy}(u_n, y_n, p_n)$ and $H^{\mu_n}_{yy}(u_n, y_n, p_n)$, it follows easily that $Q_n(v_n) - Q^0_n(v_n) \to \bar{Q}(\bar{v}) - \bar{Q}^0(\bar{v})$. Writing $Q^0_n(v_n) = \bar{Q}^0(v_n) + \epsilon_n$ with $\epsilon_n = \int_0^T v_n^*(H^{\mu_n}_{uu}(u_n, y_n, p_n) - H_{uu}(\bar{u}, \bar{y}, \bar{p}))v_n \mathrm{d}t$, by continuity of $H^{\mu_n}_{uu}$ at junction times (Lemma A.1 and Remark 2.20), Lemma 4.6 implies that $H^{\mu_n}_{uu}(u_n, y_n, p_n) \to H_{uu}(\bar{u}, \bar{y}, \bar{p})$ uniformly, and hence, $\epsilon_n \to 0$. Since by (A2), $\bar{Q}^0 : v \mapsto \int_0^T v^* H_{uu}(\bar{u}, \bar{y}, \bar{p})v$ is a Legendre form, (4.19) follows. ∎

We recall the reduction approach of [5, section 5.2]. When $q \ge 2$, with all touch points of the trajectory $(\bar{u}, \bar{y})$ being reducible by (A6), let $\varepsilon, \delta > 0$ and V_μ be small enough so that, for all $\|u - \bar{u}\|_\infty \le \delta$, all $\mu \in V_\mu$, and all $\tau_{to} \in \mathcal{T}_{to}$, the function $g^\mu(y^\mu_u)$ attains its maximum over $[\tau_{to} - \varepsilon, \tau_{to} + \varepsilon]$ at a unique point $\tau^\mu_u \in (\tau_{to} - \varepsilon, \tau_{to} + \varepsilon)$. Set $\bar{I}_{to} := \cup_{\tau_{to} \in \mathcal{T}_{to}}(\tau_{to} - \varepsilon, \tau_{to} + \varepsilon)$ and $\bar{I}_b := [0,T] \setminus \bar{I}_{to}$. When $q = 1$, set $\bar{I}_b := [0,T]$ and $\bar{I}_{to} := \emptyset$. Then the following *reduced problem* is well defined and locally equivalent to $(\mathcal{P}^\mu)$:

$$(\mathcal{P}^\mu_{red}) \qquad \min_{u \in B_\infty(\bar{u},\delta)} J^\mu(u) \qquad \text{subject to}$$

$$\mathcal{G}^\mu(u) := \begin{pmatrix} g(y_u)|_{\bar{I}_b} \\ g^\mu(y^\mu_u(\tau^{\mu,1}_u)) \\ \vdots \\ g^\mu(y^\mu_u(\tau^{\mu,N_{to}}_u)) \end{pmatrix} \in \mathcal{K} := C_-[\bar{I}_b] \times \mathbb{R}^{N_{to}}_-. \tag{4.21}$$

The Lagrangian $\mathcal{L}^\mu$ of the reduced problem (4.21) is given, for $u \in B_\infty(\bar{u}, \delta)$ and a multiplier $\lambda = (\eta_b, \nu) \in \mathcal{M}_+[\bar{I}_b] \times \mathbb{R}^{N_{to}}_+$, by

$$\mathcal{L}^\mu(u, \lambda) = J^\mu(u) + \int_{\bar{I}_b} g^\mu(y^\mu_u(t))\mathrm{d}\eta_b(t) + \sum_{i=1}^{N_{to}} \nu_i g^\mu(y^\mu_u(\tau^{\mu,i}_u)). \tag{4.22}$$

Multipliers η^μ and $\lambda^\mu = (\eta_b^\mu, \nu^\mu)$ associated with u^μ in, respectively, problem $(\mathcal{P}^\mu)$ and its reduced form $(\mathcal{P}^\mu_{red})$, are related by

$$\mathrm{d}\eta^\mu(t) = \mathrm{d}\eta_b^\mu(t) \text{ on } \bar{I}_b; \qquad \mathrm{d}\eta^\mu(t) = \sum_{i=1}^{N_{to}} \nu^\mu_{\tau_i} \delta_{\tau_u^{\mu,i}}(t) \text{ on } \bar{I}_{to}. \tag{4.23}$$

In addition, we can show that the reduced Lagrangian (4.22) is twice Fréchet differentiable at u^μ, and its second-order derivative satisfies, for $v \in \mathcal{V}$,

$$D^2_{uu}\mathcal{L}^\mu(u^\mu, \lambda^\mu)(v, v) = \mathcal{J}^\mu(v, z^\mu_{u,v}), \tag{4.24}$$

with $\mathcal{J}^\mu$ given by (3.17), and that the remainder $r(v)$ in the second-order expansion

$$\mathcal{L}^\mu(u^\mu + v, \lambda^\mu) = \mathcal{L}^\mu(u^\mu, \lambda^\mu) + D_u\mathcal{L}^\mu(u^\mu, \lambda^\mu)v + \tfrac{1}{2}D^2_{uu}\mathcal{L}^\mu(u^\mu, \lambda^\mu)(v, v) + r(v)$$

satisfies

$$r(v)/\|v\|_2^2 \;\to\; 0 \;\text{ when }\; \|v\|_\infty \;\to\; 0. \tag{4.25}$$

In what follows, $T_{\mathcal{K}}(x)$ and $N_{\mathcal{K}}(x)$ denote, respectively, the tangent and normal cones to $\mathcal{K}$ at point $x \in \mathcal{K}$ (in the sense of convex analysis).

LEMMA 4.9. *Under assumptions and condition* (ii) *of Theorem* 4.3*, there exists an open neighborhood V_μ of μ_0 such that the shooting extremal (u^μ, y^μ) associated with $(\mathcal{P}^\mu)$ for $\mu \in V_\mu$ satisfies the uniform quadratic growth condition, and hence, is a local solution of $(\mathcal{P}^\mu)$.*

Proof. If the conclusion does not hold, then there exists a q-stable extension $(\mathcal{P}^\mu)$, a sequence $\mu_n \to \mu_0$, with associated shooting extremal and multipliers (u_n, y_n, p_n, η_n) converging to $(\bar{u}, \bar{y}, \bar{p}, \bar{\eta})$ in $\mathcal{X}_N$ by Lemma 4.6 (which implies in particular $u_n \to \bar{u}$ in L^∞), and a point $\tilde{u}_n \in \mathcal{U}$ feasible for (P^{μ_n}), $\tilde{u}_n \neq u_n$, $\tilde{u}_n \to \bar{u}$ in L^∞, satisfying for all n,

$$J^{\mu_n}(\tilde{u}_n) \le J^{\mu_n}(u_n) + o(\|\tilde{u}_n - u_n\|_2^2). \tag{4.26}$$

Since $\lambda_n \in N_{\mathcal{K}}(\mathcal{G}^{\mu_n}(u_n))$, we have (for the appropriate duality products)

$$\langle \lambda_n, \mathcal{G}^{\mu_n}(\tilde{u}_n) - \mathcal{G}^{\mu_n}(u_n) \rangle \;\le\; 0,$$

and thus

$$\mathcal{L}^{\mu_n}(\tilde{u}_n, \lambda_n) - \mathcal{L}^{\mu_n}(u_n, \lambda_n) \;\le\; o(\|\tilde{u}_n - u_n\|_2^2). \tag{4.27}$$

Let $0 < \varepsilon_n := \|\tilde{u}_n - u_n\|_2 \to 0$ and $v_n := \varepsilon_n^{-1}(\tilde{u}_n - u_n)$. Since $\|v_n\|_2 = 1$ for all n, taking a subsequence if necessary, we may assume that $v_n \rightharpoonup \bar{v} \in \mathcal{V}$. With the notation of Lemma 4.8, we deduce from this lemma that (4.19) holds. Combining $D_u\mathcal{L}^{\mu_n}(u_n, \lambda_n) = 0$ and (4.24) with (4.27) and (4.25), we get

$$Q^n(v_n) = D_{uu}\mathcal{L}^{\mu_n}(u_n, \lambda_n)(v_n, v_n) \;\le\; o(1), \tag{4.28}$$

and thus $\bar{Q}(\bar{v}) \le 0$ by (4.19). Now

$$\mathcal{K} \ni \mathcal{G}^{\mu_n}(\tilde{u}_n) = \mathcal{G}^{\mu_n}(u_n) + \varepsilon_n D\mathcal{G}^{\mu_n}(u_n)v_n + \varepsilon_n r_n,$$

where $\|r_n\|_\infty = o(1)$, and therefore $D\mathcal{G}^{\mu_n}(u_n)v_n + r_n \in T_{\mathcal{K}}(\mathcal{G}^{\mu_n}(u_n))$, implying $g_y^{\mu_n}(y_n)z_n + r_n \le 0$ on $\mathrm{supp}(\mathrm{d}\eta_n)$. Thus (4.20) is satisfied by Lemma 4.8. Also, by (4.26), $DJ^{\mu_n}(u_n)v_n \le o(1)$, and hence,

$$\langle \eta_n, g_y^{\mu_n}(y_n)z_n\rangle = \langle \lambda_n, D\mathcal{G}^{\mu_n}(u_n)v_n\rangle \ge o(1).$$

Passing to the limit, we obtain $\langle \bar\eta, g_y(\bar y)\bar z\rangle \ge 0$. By (4.20) and $\mathrm{d}\bar\eta \ge 0$, we deduce that $g_y(\bar y)\bar z \in \mathrm{supp}(\mathrm{d}\bar\eta)^\perp$; thus $\bar v$ and its associated linearized state $\bar z$ satisfy (3.10) and (4.8). Therefore condition (ii) and $\bar Q(\bar v) \le 0$ imply $\bar v = 0$. Since by (4.28), $\limsup Q^n(v_n) \le 0$, it follows from (4.19) that $Q^n(v_n) \to 0 = \bar Q(\bar v)$, and hence, $v_n \to \bar v = 0$, contradicting $\|v_n\|_2 = 1$ for all n. ∎

Proof of Theorem 4.3. (ii) ⇒ (i) is a consequence of Lemmas 4.5–4.9. Let us show (i) ⇒ (ii). Let ρ be a C^∞ function over $\mathbb{R}$ such that $\mathrm{supp}(\rho) \subset [-1,1]$ and ρ is positive over $(-1,1)$. The function ψ^μ defined by $\psi^\mu(s) := \sum_{\tau\in\mathcal{T}_{to}^{nes}} \mu^{4q+2}\rho(\frac{s-\tau}{\mu})$ for $\mu \ne 0$, and $\psi^0(s) = 0$, for all $s \in [0,T]$, is of class C^{2q} w.r.t. its arguments s and μ and has support in $\cup_{\tau\in\mathcal{T}_{to}^{nes}}[\tau - |\mu|, \tau + |\mu|]$ for $\mu \ne 0$. Consider the perturbed constraint mapping $g^\mu(y) := g(y) - \psi^\mu(y_n)$ (recall that we assume that $(\mathcal{P})$ is written such that $y_n(t) = t$). Observe that $g^0 = g$ and g^μ is of order q for all μ; therefore $(\mathcal{P}^\mu) \equiv (\ell, \phi, f, g^\mu, y_0)$ is a q-stable extension of $(\mathcal{P}^0) = (\mathcal{P})$ with $\mu_0 = 0$. In addition, $g^\mu(y) = g(y)$ for all y such that $y_n \notin \cup_{\tau\in\mathcal{T}_{to}^{nes}}(\tau - |\mu|, \tau + |\mu|)$, and $g^\mu(\bar y(t)) < 0$ on $(\tau - |\mu|, \tau + |\mu|)$ for all $\tau \in \mathcal{T}_{to}^{nes}$. Since the touch points are isolated, we have for $|\mu| > 0$ small enough $g^\mu(\bar y) = g(\bar y)$ on $\mathcal{I}_b \cup \mathcal{T}_{to}^{ess} = \mathrm{supp}(\mathrm{d}\bar\eta)$, and it is easy to see that $(\bar u, \bar y)$ is a stationary point for $(\mathcal{P}^\mu)$, with the same Lagrange multiplier $\bar\eta$ and the same costate $\bar p$. In addition, the stationary point $(\bar u, \bar y)$ for $(\mathcal{P}^\mu)$ has a neighboring structure to that of $(\bar u, \bar y)$ for $(\mathcal{P}^0)$ (all nonessential touch points are removed). Therefore, by (i) and Definition 4.1, for $|\mu|$ small enough, $(\bar u, \bar y)$ satisfies the uniform quadratic growth condition (4.3) for $(\mathcal{P}^\mu)$. Since assumptions (A2)–(A6) are satisfied for $(\mathcal{P}^\mu)$, it follows from Theorem 3.4(ii) that the sufficient condition (ii) holds, which achieves the proof. ∎

4.2. Sensitivity analysis. If strong regularity holds, the mapping $\Xi : V_\delta \to V_\theta$, $\delta \mapsto \theta(\delta)$ is given by $\Xi(\delta) = \theta_0 + \omega(\delta)$, where $\omega(\delta)$ is the solution of (4.14). It follows then from (4.13) that

$$\theta^\mu = \theta_0 + \omega(-D_\mu\mathcal{F}(\theta_0,\mu_0)(\mu - \mu_0)) + o(\|\mu - \mu_0\|).$$

Since the mapping $\mathbb{R}^{\bar N} \to \Theta$, $\delta \mapsto \omega(\delta)$ is positively homogeneous of degree one, the mapping $\mu \mapsto \theta^\mu$ is Fréchet directionally differentiable. The directional derivatives in direction d are obtained by substituting $-D_\mu\mathcal{F}(\theta_0,\mu_0)d$ for δ in (4.14). Therefore,

$$\theta^{\mu_0+d} = \theta_0 + \omega_d + o(\|d\|), \tag{4.29}$$

where $\omega_d = (\pi^*_{d,0}, \gamma^{1:q}_{d,\mathcal{T}_{en}}, \gamma_{d,\mathcal{T}_{to}}, \sigma_{d,\mathcal{T}_{en}}, \sigma_{d,\mathcal{T}_{ex}}, \sigma_{d,\mathcal{T}_{to}})$ is as follows. Denote by (v_d, z_d) and $(\zeta_d, \pi_d, \lambda^{1:q}_{d,\mathcal{T}_{en}}, \lambda_{d,\mathcal{T}_{to}})$ the (unique) optimal solution and multipliers of the quadra-

tic problem below:

$$
\begin{aligned}
(\mathcal{P}_d)\quad \min_{(v,z)\in\mathcal{V}\times\mathcal{Z}}\ & \frac{1}{2}\int_0^T D^2_{(u,y,\mu),(u,y,\mu)}\tilde{H}(\bar{u},\bar{y},\bar{p}_q,\bar{\eta}_q,\mu_0)((v,z,d),(v,z,d))\mathrm{d}t\\
&+\frac{1}{2}D^2\tilde{\phi}(\bar{y}(T),\mu_0)((z(T),d),(z(T),d))\\
&+\frac{1}{2}\sum_{\tau\in\mathcal{T}_{en}}\sum_{j=1}^{q}\nu_\tau^j D^2\tilde{g}^{(j-1)}(\bar{y}(\tau),\mu_0)((z(\tau),d),(z(\tau),d))\\
&+\frac{1}{2}\sum_{\tau\in\mathcal{T}_{to}}\nu_\tau\left(D^2\tilde{g}(\bar{y}(\tau),\mu_0)((z(\tau),d),(z(\tau),d))-\frac{(D\tilde{g}^{(1)}(\bar{y}(\tau),\mu_0)(z(\tau),d))^2}{\frac{\mathrm{d}}{\mathrm{d}t}\tilde{g}^{(1)}(\bar{y}(t),\mu_0)|_{t=\tau}}\right)\\
\text{subject to}\quad & \begin{cases}
\dot{z}(t)=D\tilde{f}(\bar{u},\bar{y},\mu_0)(v,z,d) & \text{on } [0,T],\quad z(0)=D\tilde{y}_0(\mu_0)d,\\
D\tilde{g}^{(0:q-1)}(\bar{y}(\tau),\mu_0)(z(\tau),d)=0, & \tau\in\mathcal{T}_{en},\\
D\tilde{g}(\bar{y}(\tau),\mu_0)(z(\tau),d)=0, & \tau\in\mathcal{T}_{to}^{ess},\\
D\tilde{g}(\bar{y}(\tau),\mu_0)(z(\tau),d)\le 0, & \tau\in\mathcal{T}_{to}^{nes},\\
D\tilde{g}^{(q)}(\bar{u},\bar{y},\mu_0)(v,z,d)=0 & \text{on } \mathcal{I}_b.
\end{cases}
\end{aligned}
$$

Then ω_d is given by $\pi_{d,0}=\pi_q(0)$, $\gamma_{d,\mathcal{T}_{to}}=\lambda_{d,\mathcal{T}_{to}}$,

$$\sigma_{d,\tau}=-\frac{D\tilde{g}^{(1)}(\bar{y}(\tau),\mu_0)(z_d(\tau),d)}{\frac{\mathrm{d}}{\mathrm{d}t}\tilde{g}^{(1)}(\bar{y},\mu_0)|_{t=\tau}},\ \tau\in\mathcal{T}_{to}, \tag{4.30}$$

$$\sigma_{d,\tau}=-\frac{D\tilde{g}^{(q)}(\bar{u}(\tau),\bar{y}(\tau),\mu_0)(v_d(\tau^+),z_d(\tau),d)}{\frac{\mathrm{d}}{\mathrm{d}t}\tilde{g}^{(q)}(\bar{u},\bar{y},\mu_0)|_{t=\tau^+}},\ \tau\in\mathcal{T}_{ex}, \tag{4.31}$$

$$\sigma_{d,\tau}=-\frac{D\tilde{g}^{(q)}(\bar{u}(\tau),\bar{y}(\tau),\mu_0)(v_d(\tau^-),z_d(\tau),d)}{\frac{\mathrm{d}}{\mathrm{d}t}\tilde{g}^{(q)}(\bar{u},\bar{y},\mu_0)|_{t=\tau^-}},\ \tau\in\mathcal{T}_{en}, \tag{4.32}$$

$$\gamma_{d,\tau}^1=\lambda_{d,\tau}^1,\quad \gamma_{d,\tau}^j\ =\ \lambda_{d,\tau}^j-\nu_\tau^{j-1}\sigma_{d,\tau},\quad j=2,\dots,q,\ \tau\in\mathcal{T}_{en}. \tag{4.33}$$

Once we have the expressions for the directional derivatives of the shooting parameters, by composition with the Fréchet derivatives of the C^1 mapping (4.18) in direction (d,ω_d), we obtain the expressions of the directional derivatives, in $\mathcal{X}_N$, of the mapping $\mu\mapsto(u^\mu,y^\mu,p^\mu,\eta^\mu)$. By Lemma 4.2, we then easily obtain the expression of the directional derivatives of the control and state in $L^r(0,T)\times W^{1,r}(0,T;\mathbb{R}^n)$ for all $1\le r<\infty$.

COROLLARY 4.10. *If either point* (i) *or* (ii) *of Theorem* 4.3 *is satisfied, then there exists a neighborhood V_μ of μ such that the mapping $V_\mu\to\mathcal{X}_N$, $\mu\mapsto(u^\mu,y^\mu,p^\mu,\eta^\mu)$, is Fréchet directionally differentiable on V_μ. In addition, the directional derivative in $L^r(0,T)\times W^{1,r}(0,T;\mathbb{R}^n)$, $1\le r<\infty$, of the mapping $\mu\mapsto(u^\mu,y^\mu)$ at point μ_0 in direction d, is the optimal solution (v_d,z_d) of problem $(\mathcal{P}_d)$.*

We end the paper with a remark related to the ill-posedness of the shooting algorithm for a state constraint of order $q\ge 3$ when boundary arcs are present (see Theorem 3.3).

REMARK 4.11 (existence of regular boundary arcs for constraints of order $q\ge 3$). Contrary to some conjectures in the literature, regular boundary arcs *can occur* for state constraints of *all* orders. Take, for example, the problem

$$
\begin{aligned}
(P_q)\qquad & \min_{(u,y)\in L^\infty(0,T)\times W^{q,\infty}(0,T)}\int_0^T\left(y(t)+\frac{u^2(t)}{2}\right)\mathrm{d}t\\
\text{subject to}\quad & y^{(q)}(t)=u(t);\ y(0)=y_1^0;\ \dot{y}(0)=y_2^0;\dots;\ y^{(q-1)}(0)=y_q^0;\\
& y(t)\ge 0,\quad t\in[0,T].
\end{aligned}
$$

It is easy to check that, for $\tau \in (0,T)$, y defined by $y(t) = 0$ on $[\tau, T]$ and

$$y(t) = \begin{cases} \dfrac{(t-\tau)^{2q}}{(2q)!} & \text{if } q \text{ is odd} \\ -\dfrac{(t-\tau)^{2q}}{(2q)!} - \nu\dfrac{(t-\tau)^{2q-1}}{(2q-1)!} & \text{if } q \text{ is even} \end{cases} \quad \text{on } [0,\tau],$$

is, for $\nu > \tau/2q$ if q is even and for *appropriate initial conditions* when $q \geq 3$, a solution that satisfies all necessary optimality conditions, and hence, by convexity of the problem, an optimal solution with a regular entry point τ. Strict complementarity holds since $\eta_0(t) = 1$ on $(\tau, T]$.

Robbins in [27] studies this example when $q = 3$ for generic initial conditions and shows that the optimal trajectory has a boundary arc, whose entry point is not regular, being the limit of an infinite number of touch points, with a geometric decreasing of the length of the interior arcs. Regular boundary arcs correspond to the case when the multiplier of the geometric sequence is equal to zero for a specific subset of initial conditions. Therefore, we see in that example, though satisfying all regularity assumptions (A0)–(A3), that the *structure of boundary arcs* is *not stable* under perturbations of the initial condition when $q \geq 3$, which illustrates why the shooting algorithm should be ill-posed in that case.

Appendix A. The next two lemmas follow immediately from the junction conditions established in [18, 25].

LEMMA A.1. *Let (u,y) be a regular Pontryagin extremal satisfying* (A2)–(A4). *Then the function $t \mapsto H_{uu}(u(t), y(t), p(t))$ is continuous on $[0,T]$.*

Proof. Let $\tau \in \mathcal{T}$. Since u is continuous by Proposition 2.5, we have

$$[H_{uu}(u(\tau), y(\tau), p(\tau))] = [p(\tau)] f_{uu}(u(\tau), y(\tau)) = -\nu_\tau g^{(1)}_{uu}(u(\tau), y(\tau)) = 0,$$

since either $\nu_\tau = 0$ when $q = 1$ by Proposition 2.5, or $g^{(1)}_u \equiv 0$ when $q > 1$. ∎

LEMMA A.2. *Let (u,y) be a regular Pontryagin extremal, satisfying* (A2)–(A4), *and let $\tau \in \mathcal{T}_{en} \cup \mathcal{T}_{ex}$ be an entry/exit time. The following conditions are equivalent:*

(i) (3.7) *holds at τ;*

(ii) *if q is* odd, $\lim_{t \to \tau;\, t \in \mathcal{I}_b} \eta_0(t) > 0$; *if q is* even, $\nu_\tau > 0$.

Proof. Define the mappings $(A_l)_{0 \leq l \leq q} : [0,T] \setminus \mathcal{T} \to \mathbb{R}^n$ by (2.34) and $(a_l)_{0 \leq l \leq q} : [0,T] \setminus \mathcal{T} \to \mathbb{R}$ by

$$a_0(t) = \ell_u(u(t), y(t)); \quad a_l(t) = \ell_y(u(t), y(t)) A_{l-1}(t) - \dot{a}_{l-1}(t), \quad l = 1, \ldots, q.$$

Then it can be seen by (2.35) (see [25]) that for all $t \in [0,T] \setminus \mathcal{T}$, we have

$$0 = \frac{\mathrm{d}^j}{\mathrm{d}t^j} H_u(u(t), y(t), p(t)) = (-1)^j (a_j(t) + p(t) A_j(t)); \qquad j = 0, \ldots, q-1, \tag{A.1}$$

$$0 = \frac{\mathrm{d}^q}{\mathrm{d}t^q} H_u(u(t), y(t), p(t)) = (-1)^q \left(a_q(t) + p(t) A_q(t) + \frac{\mathrm{d}\eta}{\mathrm{d}t} g^{(q)}_u(u(t), y(t)) \right). \tag{A.2}$$

Since the derivatives of the control are continuous until order $q-2$, the functions a_j and A_j are continuous for $j = 0, \ldots, q-2$, and it is then easily seen, since u is continuous, that the jumps of A_{q-1} and a_{q-1} at $\tau \in \mathcal{T}$, when q is even, are given, respectively, by

$$[A_{q-1}(\tau)] = (-1)^{q-1} f_{uu}(u(\tau), y(\tau)) [u^{(q-1)}(\tau)],$$
$$[a_{q-1}(\tau)] = (-1)^{q-1} \ell_{uu}(u(\tau), y(\tau)) [u^{(q-1)}(\tau)].$$

Taking the jump in (A.1) at τ for $j = q-1$ then yields

$$0 \;=\; (-1)^{q-1} H_{uu}(u(\tau), y(\tau), p(\tau^+))[u^{(q-1)}(\tau)] - \nu_\tau g_y(y(\tau)) A_{q-1}(\tau^-).$$

By (2.35), we have $g_y(y(\tau))A_{q-1}(\tau^\pm) = g_u^{(q)}(u(\tau), y(\tau))$, so we obtain, when q is even,

$$\nu_\tau = (-1)^{q-1} \frac{H_{uu}(u(\tau), y(\tau), p(\tau^+))[u^{(q-1)}(\tau)]}{g_u^{(q)}(u(\tau), y(\tau))}. \tag{A.3}$$

It follows that $\nu_\tau > 0$ iff $u^{(q-1)}$ is discontinuous at τ, which is equivalent to saying that (3.7) holds (when q is even). When q is odd, $u^{(q-1)}$, a_{q-1}, and A_{q-1} are continuous (and $\nu_\tau = 0$). Taking the jump in (A.2), we obtain

$$0 = (-1)^q H_{uu}(u(\tau), y(\tau), p(\tau))[u^{(q)}(\tau)] + [\eta_0(\tau)]\, g_u^{(q)}(u(\tau), y(\tau)).$$

Consequently, we have $\eta_0(\tau^\pm) > 0$ at an entry/exit point, where $\tau^\pm$ stands for τ^+ if $\tau \in \mathcal{T}_{en}$ and τ^- if $\tau \in \mathcal{T}_{ex}$ iff $u^{(q)}$ is discontinuous at τ, and hence iff (3.7) holds. ∎

Appendix B. The next two lemmas recall classical results. For the second one see related results by Aubin [1].

LEMMA B.1. *Let X be a Hilbert space and Q a Legendre form over X. Let A be a continuous linear operator over X. The following assertions are equivalent:*

(i) *$Q(v) > 0$, for all $v \in \operatorname{Ker} A \setminus \{0\}$.*

(ii) *There exists $\alpha > 0$ such that $Q(v) \geq \alpha \|v\|_2^2$ for all $v \in \operatorname{Ker} A$.*

LEMMA B.2. *Let X be a Hilbert space and Y a Banach space, $H : X \to X^* \equiv X$ a self-adjoint continuous linear operator, and $A : X \to Y$ and $B : X \to \mathbb{R}^r$, $r \in \mathbb{N}$, continuous linear operators. Assume that*

$$\text{(i)} \quad \exists \alpha > 0, \quad \langle Hx, x\rangle \geq \alpha \|x\|^2 \quad \textit{for all } x \in \operatorname{Ker} A.$$
$$\text{(ii)} \quad \textit{The operator } (A, B) : X \to Y \times \mathbb{R}^r \textit{ is onto.}$$

Then, for all $(x^*, y, \delta) \in X^* \times Y \times \mathbb{R}^r$, there exists a unique $(x, y^*, \nu) \in X \times Y^* \times \mathbb{R}^{r*}$, solution of

$$\left\{ \begin{array}{l} Hx + A^* y^* + B^* \nu \;=\; x^*, \\ Ax \;=\; y, \\ Bx \;\leq\; \delta, \quad \nu \geq 0, \quad \nu(Bx - \delta) = 0, \end{array} \right. \tag{B.1}$$

and the mapping $(x^*, y, \delta) \mapsto (x, y^*, \nu)$, where (x, y^*, ν) is solution of (B.1)*, is Lipschitz continuous.*

Acknowledgments. The authors thank the anonymous referees for their useful remarks.

REFERENCES

[1] J. P. AUBIN, *Comportement lipschitzien des solutions de problèmes de minimisation convexes*, C. R. Acad. Sci. Paris Sér. I Math., 295 (1982), pp. 235–238.

[2] D. AUGUSTIN AND H. MAURER, *Computational sensitivity analysis for state constrained optimal control problems*, Ann. Oper. Res., 101 (2001), pp. 75–99.

[3] P. BERKMANN AND H. J. PESCH, *Abort landing in windshear: Optimal control problem with third-order state constraint and varied switching structure*, J. Optim. Theory Appl., 85 (1995), pp. 21–57.

[4] J. F. Bonnans and A. Hermant, *Second-Order Analysis for Optimal Control Problems with Pure and Mixed State Constraints*, Research Report 6199, INRIA, Le Chesnay, France, 2007.
[5] J. F. Bonnans and A. Hermant, *No-gap second-order optimality conditions for optimal control problems with a single state constraint and control*, Math. Program. Ser. B, to appear.
[6] J. F. Bonnans and A. Hermant, *Conditions d'optimalité du second-ordre nécessaires ou suffisantes pour les problèmes de commande optimale avec une contrainte sur l'état et une commande scalaires*, C. R. Math. Acad. Sci. Paris Sér. I, 343 (2006), pp. 473–478.
[7] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[8] A. E. Bryson, W. F. Denham, and S. E. Dreyfus, *Optimal programming problems with inequality constraints* I: *Necessary conditions for extremal solutions*, AIAA J., 1 (1963), pp. 2544–2550.
[9] A. E. Bryson and Y.-C. Ho, *Applied Optimal Control*, Hemisphere Publishing, New York, 1975.
[10] R. Bulirsch, F. Montrone, and H. J. Pesch, *Abort landing in the presence of windshear as a minimax optimal control problem.* I. *Necessary conditions*, J. Optim. Theory Appl., 70 (1991), pp. 1–23.
[11] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
[12] A. L. Dontchev and W. W. Hager, *Lipschitzian stability for state constrained nonlinear optimal control*, SIAM J. Control Optim., 36 (1998), pp. 698–718.
[13] A. L. Dontchev and W. W. Hager, *The Euler approximation in state constrained optimal control*, Math. Comp., 70 (2001), pp. 173–203.
[14] H. Frankowska and B. Kaŝkosz, *A maximum principle for differential inclusion problems with state constraints*, Systems Control Lett., 11 (1988), pp. 189–194.
[15] A. Haraux, *How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities*, J. Math. Soc. Japan, 29 (1977), pp. 615–631.
[16] R. F. Hartl, S. P. Sethi, and R. G. Vickson, *A survey of the maximum principles for optimal control problems with state constraints*, SIAM Rev., 37 (1995), pp. 181–218.
[17] A. D. Ioffe and V. M. Tihomirov, *Theory of Extremal Problems*, North–Holland Publishing Company, Amsterdam, 1979; Russian edition, Nauka, Moscow, 1974.
[18] D. H. Jacobson, M. M. Lele, and J. L. Speyer, *New necessary conditions of optimality for control problems with state-variable inequality constraints*, J. Math. Anal. Appl., 35 (1971), pp. 255–284.
[19] H. Kawasaki, *An envelope-like effect of infinitely many inequality constraints on second order necessary conditions for minimization problems*, Math. Programming, 41 (1988), pp. 73–96.
[20] H. Kawasaki, *Second order necessary optimality conditions for minimizing a sup-type function*, Math. Programming, 49 (1990/91), pp. 213–229.
[21] K. Malanowski, *Stability and sensitivity of solutions to nonlinear optimal control problems*, J. Appl. Math. Optim., 32 (1995), pp. 111–141.
[22] K. Malanowski, *Sufficient optimality conditions for optimal control subject to state constraints*, SIAM J. Control Optim., 35 (1997), pp. 205–227.
[23] K. Malanowski and H. Maurer, *Sensitivity analysis for state constrained optimal control problems*, Discrete Contin. Dynam. Systems, 4 (1998), pp. 241–272.
[24] K. Malanowski and H. Maurer, *Sensitivity analysis for optimal control problems subject to higher order state constraints. Optimization with data perturbations*, II, Ann. Oper. Res., 101 (2001), pp. 43–73.
[25] H. Maurer, *On the Minimum Principle for Optimal Control Problems with State Constraints*, Tech. report, Schriftenreihe des Rechenzentrum 41, Universität Münster, Münster, Germany, 1979.
[26] H. J. Pesch, *A practical guide to the solution of real-life optimal control problems*, Control Cybernet., 23 (1994), pp. 7–60.
[27] H. M. Robbins, *Junction phenomena for optimal control with state-variable inequality constraints of third order*, J. Optim. Theory Appl., 31 (1980), pp. 85–99.
[28] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[29] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1993.

CORRECTION TO "WELL-POSEDNESS OF THE SHOOTING ALGORITHM FOR STATE CONSTRAINED OPTIMAL CONTROL PROBLEMS WITH A SINGLE CONSTRAINT AND CONTROL"

- page 1410, proof of Lemma 3.1: Delete > 0 after ν_τ. The sentence should read "By Proposition 2.10 (see (2.38), (3.5) is equivalent (when q is even) to the strict positivity of ν_τ at entry/exit points $\tau \in \mathcal{T}_{en} \cup \mathcal{T}_{ex}$)."
- page 1411, Theorem 3.4, line 5: Delete "a" before "unique solution." The sentence should read "If (u, y) is a Pontryagin extremal such that (A2)–(A6) hold, it is a local solution of $(\mathcal{P})$ satisfying the quadratic growth condition (3.2) iff problem (3.19) has zero for unique solution."
- page 1412, line 4: delete the word "zero." The sentence should read "With the junction conditions results of Proposition 2.5 and (A5)(i), we can show that boundary arcs have no contribution to the curvature term."

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1431–1460

ON FEEDBACK CLASSIFICATION OF CONTROL-AFFINE SYSTEMS WITH ONE- AND TWO-DIMENSIONAL INPUTS*

ANDREI AGRACHEV† AND IGOR ZELENKO†

Abstract. The paper is devoted to the local classification of generic germs of control-affine systems on an n-dimensional manifold with scalar input for any $n \geq 4$ or with two inputs for $n = 4$ and $n = 5$, up to state-feedback transformations, preserving the affine structure (in the C^∞ category for $n = 4$ and the C^ω category for $n \geq 5$). First, using the Poincaré series of moduli numbers, we introduce the intrinsic numbers of functional moduli of each prescribed number of variables on which a classification problem depends. In order to classify generic germs of affine systems with scalar input we associate with such a system the canonical frame by normalizing some structural functions in a commutative relation of the vector fields, which define our control system. Then, using this canonical frame, we introduce the canonical coordinates and find a complete system of state-feedback invariants of the system. It also automatically gives the local in state-input space classification of generic germs of nonaffine n-dimensional control systems with scalar input for $n \geq 3$ (in the C^∞ category for $n = 3$ and in the C^ω category for $n \geq 4$). Further, we show how the problem of feedback equivalence of generic germs of affine systems with two-dimensional input in state space of dimensions 4 and 5 can be reduced to the same problem for affine systems with scalar input. In order to make this reduction we distinguish the subsystem of our control system, consisting of the directions of all extremals in dimension 4 and all abnormal extremals in dimension 5 of the time optimal problem, defined by the original control system. In each classification problem under consideration we find the intrinsic numbers of functional moduli of each prescribed number of variables according to its Poincaré series.

Key words. state-feedback equivalence, control-affine systems, Poincaré series, extremals

AMS subject classifications. 93B29, 53A55

DOI. 10.1137/050623711

1. Introduction.

For the convenience of presentation, all objects are C^∞ unless otherwise noted, although all constructions and some statements remain valid in an obvious way also in the C^k category for an appropriate finite k. On the other hand, some of the statements are known to be true only in the real analytic category, which will be indicated explicitly.

Let M be an n-dimensional manifold, and let $f_0, f_1, \dots, f_r$ be vector fields on M, $r < n$. Consider the following control-affine system with r inputs on M:

$$\dot{q} = f_0(q) + \sum_{k=1}^{r} u_k f_k(q), \quad q \in M, \quad u_1, \dots, u_r \in \mathbb{R}. \tag{1.1}$$

We say that a system of type (1.1) is an (r, n) control-affine system. Fix a point $q_0 \in M$. Throughout the paper we will assume that at the point q_0

$$\dim \operatorname{span}\big(f_0(q_0), f_1(q_0), \dots, f_r(q_0)\big) = r + 1. \tag{1.2}$$

Consider the group FB_{q_0} of state-feedback transformations that preserve an affine structure and the point q_0, i.e., transformations of the type

$$\begin{cases} q = \Phi(\tilde{q}), \\ u = \mathcal{B}(\tilde{q})\tilde{u} + \mathcal{A}(\tilde{q}), \\ q_0 = \Phi(q_0), \end{cases} \qquad u = \begin{pmatrix} u_1 \\ \vdots \\ u_r \end{pmatrix}, \tag{1.3}$$

*Received by the editors February 1, 2005; accepted for publication (in revised form) March 17, 2007; published electronically September 26, 2007.
http://www.siam.org/journals/sicon/46-4/62371.html
†S.I.S.S.A., Via Beirut 2-4, 34014 Trieste, Italy (agrachev@sissa.it, zelenko@sissa.it).

where Φ is a diffeomorphism in a neighborhood of q_0, $\mathcal{A}(q) \in \mathbb{R}^r$, $\mathcal{B}(q)$ is an $(r \times r)$-matrix, and $\det \mathcal{B}(q_0) \neq 0$. This group of transformations acts naturally on the set of germs at q_0 of systems of type (1.1) and defines an equivalence relation, called the *state-feedback equivalence*.[1] The state-feedback equivalence can be described in more geometric terms: any control-affine system on M defines the affine subbundle of the tangent bundle TM of M. Two germs of control-affine systems are state-feedback equivalent if and only if the corresponding germs of affine subbundles belong to the same orbit w.r.t. the natural action of the group of germs of diffeomorphisms of M on the set of germs of affine subbundles of TM. Thus, the problem of state-feedback equivalence of control-affine systems is a particular case of the classical equivalence problem of geometric structures on manifolds.

In the case of corank 1 control-affine systems, i.e., when $r = n - 1$, all generic germs of control-affine systems are state-feedback equivalent to each other. Indeed, under assumption (1.2) there is the natural one-to-one correspondence between the set of control-affine systems, up to feedback transformations, and the set of differential 1-forms on the ambient manifold: To any affine system (1.1) one can assign a unique differential 1-form ω such that $\omega(f_i) = 0$ for $i = 1, \dots, n-1$ and $\omega(f_0) = 1$. Thus, the state-feedback classification of corank 1 control-affine systems satisfying (1.2) is equivalent to the well-known classification of differential 1-forms w.r.t. the action of the group of diffeomorphisms (see, for example, [19, section 3 and Appendix C]). In particular, all germs of (1.1) such that the underlying vector distribution $\operatorname{span}(f_1, \dots, f_{n-1})$ is contact for odd n or quasi-contact for even n (which is a generic assumption) are state-feedback equivalent to the control-affine system, corresponding to the classical Darboux model (note also that in [19] normal forms for codimension 1 singularities are given, too).

Now suppose that $r < n-1$. Let us roughly estimate the "number of parameters" in the considered equivalence problem. The set of r-dimensional affine subspaces in $\mathbb{R}^n$ forms an $(r+1)(n-r)$-dimensional manifold. Therefore, if the coordinates on M are fixed, then the control system of type (1.1) can be defined by $(r+1)(n-r)$ functions of n variables. The group of the coordinate changes on M is parameterized by n functions of n variables. Thus, by a coordinate change one can "normalize," in general, only n functions among those $(r+1)(n-r)$ functions defining our control system. Thus we may expect that the set of orbits of generic germs of systems (1.1) at $q_0 \in M$ w.r.t. the action of the group of transformations of type (1.3) can be parameterized by $(r+1)(n-r) - n = r(n-r-1)$ arbitrary germs of functions of n variables and a number of germs of functions, depending on less than n variables (see the next section for the discussion about the number of these additional functional invariants).

According to the last estimate, for $r < n-1$ generic germs of (r, n) control-affine systems have functional invariants (see also [8, Proposition 3.12]). How does one construct such invariants and, more generally, a complete system of such invariants? These and related questions were the subject of intensive study in the last two decades, and numerous important contributions were made by different authors. Classical methods of differential geometry such as the Cartan method of equivalence (see, for example, [7], [18], [14]) were used, and new methods coming from Pontryagin's maximum principle, such as the theory of Jacobi curves of extremals [1], [2], [3], [4], [16] and the method of critical Hamiltonians [9], were developed. Different criteria

[1] Some authors refer to the transformations (1.3) and the corresponding equivalence by the name "feedback," while in the case $\Phi = \mathrm{Id}$ they use the term "pure feedback."

for the state-feedback equivalence in terms of the associated system of extremals were given as well [6], [10]. Also the important case of the so-called equilibrium point, i.e., when $f_0(q_0) \in \operatorname{span}\big(f_1(q_0), \dots, f_r(q_0)\big)$, was intensively studied for corank 1 control-affine systems, using singularity theory (see, for example, [11], [15], and [20]), and for control-affine systems with scalar input, using the formal approach similar to the classical Poincaré–Dulac procedure [12], [13], [17].

However, as far as we know, except for the case $r = n - 1$, the set of orbits of generic germs of (r, n) control-affine systems w.r.t. the action of the group of state-feedback transformations was parameterized by certain tuples of invariants only in the case $r = 1$, $n = 3$, i.e., in the case of the smallest dimensions, when functional parameters appear [2, section 3, Proposition 3.2]. In particular, it was shown there that in this case this set of orbits can be parameterized by one arbitrary function of three variables, two arbitrary functions of two variables, and the discrete invariant from the set $\{-1, 1\}$.

Remark 1. Actually, in Proposition 3.2 of [2] the two functions of two variables satisfy certain conditions on coordinate subspaces of some special coordinates, which are canonical up to some reflections, but instead of these functions one can take their appropriate partial derivatives, which are already arbitrary. The functions of the parameterization are state-feedback invariants up to some reflections in the coordinates. □

In the present paper we make a classification of generic germs of systems of type (1.1), up to state-feedback equivalence, in the following cases:

1. $r = 1$, $n = 4$;
2. $r = 1$, $n \geq 5$ in the real analytic category;
3. $r = 2$, $n = 4$;
4. $r = 2$, $n = 5$ in the real analytic category.

In general, statements of the kind "the classification problem depends on the tuple of functional invariants, consisting of certain number of functions of each number of variables" need to be clarified: These numbers could be changed rather arbitrarily by mixing, combining, or separating the formal Taylor series of these functional invariants without losing any information (at least if we work in the category of formal Taylor series or in the real analytic category). In [5, section 1] the author proposed using the so-called *Poincaré series of the moduli numbers of the classification problem* in order to determine intrinsically the number of functions of each prescribed number of variables on which some classification problem depends. In section 2 below, using the Poincaré series, we give a canonical selection of these numbers. Throughout this section we demonstrate all our notions and constructions on the problem of classification of germs of Riemannian metrics on a two-dimensional manifold. The way in which the canonical parameterization is obtained indicates the presence of an interesting algebraic structure on the set of all tuples of fundamental invariants parameterizing a given classification problem. For the moment, this algebraic structure remains hidden and needs further research.

In the case of scalar input, our method of the classification is similar to the procedure used in [2] for the case $n = 3$ and is described in section 3. It basically consists of the following two steps: first, for any control system, satisfying some genericity assumptions, we construct the canonical frame by normalizing some structural functions in certain commutative relations of the vector fields, which define our control-affine systems; then, using this canonical frame, we introduce the canonical coordinates and find the complete system of state-feedback invariants of the system.

In addition, to any control system

$$\dot{y} = \mathcal{F}(y, v), \quad y \in S, \quad v \in V, \tag{1.4}$$

on an m-dimensional manifold S (the state space) with one-dimensional control space V one can assign the following control-affine system on the $(m+1)$-dimensional state-space $S \times V$:

$$\begin{cases} \dot{y} = \mathcal{F}(y, v), \\ \dot{v} = u, \end{cases} \tag{1.5}$$

where $v \in \mathbb{R}$ (here we look on v as on a new state variable, u is the new control, $f_0 = (\mathcal{F}(y, v), 0)^T$, and $f_1 = (0, 1)^T$ in the notation of (1.1)). It turns out that having the local classification of generic germs of $(m + 1)$-dimensional control-affine systems with scalar input, one also gets the local in state-input space classification of generic germs of nonaffine m-dimensional control system with scalar input (see Remark 6).

Further, in section 4, we show that the problem of state-feedback classification of the control-affine systems with two-dimensional input in dimensions 4 and 5 can be reduced to the previous problem for the control-affine systems with scalar input in the same dimensions. In order to make this reduction, we distinguish the subsystem, consisting of the directions of all extremals in dimension 4 and all abnormal extremals in dimension 5 of the time optimal problem, defined by the original control system.

In each classification problem under consideration we find the intrinsic numbers of functional moduli of each prescribed number of variables according to its Poincaré series.

Finally note that the previously mentioned papers [12], [17] were devoted to the classification of control-affine systems with scalar control in a neighborhood of an equilibrium point. Here we classify the control-affine systems with scalar control near a nonequilibrium point. In general the feedback invariants, constructed here for generic germs, could also be used for the problem with equilibrium points by passing to the limit. The relation of the invariants obtained in this way with the invariants obtained in [12], [17] is the subject of further study.

2. Poincaré series and the intrinsic number of functional invariants. We start with some terminology. Let M be a smooth manifold. Fix a point $q_0 \in M$. Consider a set $\mathcal{O}$ of germs at q_0 of smooth objects on M (for example, Riemannian metrics, vector distributions, control-affine systems) such that the group of local diffeomorphisms Diff_{q_0}, preserving the point q_0, acts naturally on it. This action defines the equivalence relation on $\mathcal{O}$.

Denote by $J^k(\mathcal{O})$ the space of all k-jets at q_0 of objects from the set $\mathcal{O}$. We say that the set $\widetilde{\mathcal{O}} \subset \mathcal{O}$ is *a generic subset* of $\mathcal{O}$ if there exists an integer $k \geq 0$ and a Zariski open set U in $J^k(\mathcal{O})$ such that

$$\widetilde{\mathcal{O}} = \{\mathfrak{b} \in \mathcal{O} : k\text{-jet of } \mathfrak{b} \text{ belongs to } U\}.$$

By "classification problem on $\mathcal{O}$" we mean the problem of finding a system of fundamental invariants for objects from some generic subset of $\mathcal{O}$ such that two generic objects are equivalent if and only if they have the same systems of fundamental invariants.

Let $\widetilde{\mathcal{O}}$ be a generic subset of $\mathcal{O}$, which is invariant w.r.t. the action of the group Diff_{q_0}.

DEFINITION 1. *A mapping F from the set $\widetilde{\mathcal{O}}$ to the set $C_0^\infty(\mathbb{R}^l,\mathbb{R})$ of germs at 0 of smooth functions in $\mathbb{R}^l$, which is invariant w.r.t. the action of the group Diff_{q_0} on $\widetilde{\mathcal{O}}$, is called a functional invariant of l variables of a generic subset of objects from $\mathcal{O}$ (or, for short, the functional invariant of $\mathcal{O}$).*

When the object $\mathfrak{b}\in\widetilde{\mathcal{O}}$ is fixed, we will mean by the functional invariant also the value of the mapping F at $\mathfrak{b}$. We will denote this germ of function by the same letter F.

Let $\mathrm{Orb}(\widetilde{\mathcal{O}})$ be the set of orbits of $\widetilde{\mathcal{O}}$ w.r.t. the action of Diff_{q_0}. Then any functional invariant $F:\widetilde{\mathcal{O}}\mapsto C_0^\infty(\mathbb{R}^l,\mathbb{R})$ induces the mapping $\widehat{F}:\mathrm{Orb}(\widetilde{\mathcal{O}})\mapsto C_0^\infty(\mathbb{R}^l,\mathbb{R})$ in the obvious way.

DEFINITION 2. *Let $\{F_i\}_{i=1}^s$ be the tuple of functional invariants of $\mathcal{O}$ defined on $\widetilde{\mathcal{O}}$, where each F_i is a functional invariant of l_i variables. We say that the tuple $\{F_i\}_{i=1}^s$ defines a parameterization of the classification problem on $\mathcal{O}$ if the mapping*

$$(\widehat{F}_1,\dots,\widehat{F}_s):\mathrm{Orb}(\widetilde{\mathcal{O}})\mapsto C_0^\infty(\mathbb{R}^{l_1},\mathbb{R})\times\dots\times C_0^\infty(\mathbb{R}^{l_s},\mathbb{R})$$

is injective and has an open image in $C_0^\infty(\mathbb{R}^{l_1},\mathbb{R})\times\dots\times C_0^\infty(\mathbb{R}^{l_s},\mathbb{R})$.

Example 1. Let $\mathcal{O}_1$ be a set of germs of Riemannian metrics at a point q_0 on an oriented two-dimensional manifold M. Given a germ of metric G, let K_G be its Gaussian curvature. Let $\widetilde{\mathcal{O}}_1$ be the set of all germs G of Riemannian metrics at q_0 such that $dK_G(q_0)\neq 0$. Obviously, $\widetilde{\mathcal{O}}_1$ is a generic subset of $\mathcal{O}_1$. For any $G\in\mathcal{O}_1$ consider the geodesic polar coordinates (r,φ) centered at q_0 and in accordance with the orientation such that the vector $\mathrm{grad}\,K_G(q_0)$ is in the direction of the ray $\{\varphi=0\}$. We will call these coordinates *the canonical polar coordinates of* G. The corresponding Cartesian coordinates (x_1,x_2) will be called *the canonical coordinates of the metric* G *at* q_0. Then the mapping $\mathfrak{K}:\widetilde{\mathcal{O}}_1\mapsto C_0^\infty(\mathbb{R}^2,\mathbb{R})$ such that

$$\mathfrak{K}(G)=K_G(x_1,x_2) \tag{2.1}$$

is a functional invariant of two variables of $\mathcal{O}_1$.

Now let us construct the parameterization of the classification problem on $\mathcal{O}_1$. By construction the image of $\mathfrak{K}$ lies in the set

$$\mathcal{N}=\{f\in C_0^\infty(\mathbb{R}^2,\mathbb{R}):f_{x_2}(0,0)=0,\ f_{x_1}(0,0)>0\}.$$

Actually, $\mathfrak{K}$ defines the one-to-one correspondence between the set $\mathrm{Orb}\big(\widetilde{\mathcal{O}}_1\big)$ of orbits of $\widetilde{\mathcal{O}}_1$ w.r.t. the action of Diff_{q_0} and the set $\mathcal{N}$. Indeed, for any $f\in\mathcal{N}$ take the metric G, which can be written in some polar coordinates centered at q_0 in the following way:

$$G=dr^2+\big(B(r,\varphi)\big)^2d\varphi^2, \tag{2.2}$$

where

$$\begin{cases}\dfrac{\partial^2}{\partial r^2}B+f(r\cos\varphi,r\sin\varphi)B=0,\\[2mm] B(0,\phi)=0,\quad \dfrac{\partial}{\partial r}B(0,\phi)=1.\end{cases} \tag{2.3}$$

Then $\mathfrak{K}(G)=f$, and any germ of metric $\overline{G}$ at q_0 such that $\mathfrak{K}(\overline{G})=f$ is isometric to G.

However, $\mathfrak{K}$ does not define the parameterization of the considered classification problem in the sense of Definition 2, because the set $\mathcal{N}$ is not an open subset of $C_0^\infty(\mathbb{R}^2,\mathbb{R})$. But instead of $\mathfrak{K}$ we can consider the following two functional invariants of one variable and one functional invariant of two variables:

$$\mathfrak{K}_1(G)=K_G(x_1,0),\quad \mathfrak{K}_2(G)=\frac{\partial^2}{\partial x_2^2}K_G(0,x_2),\quad \mathfrak{K}_3(G)=\frac{\partial^2}{\partial x_1\partial x_2}K_G(x_1,x_2), \tag{2.4}$$

where (x_1,x_2) are the canonical coordinates of the metric G at q_0. The image of the mapping $(\mathfrak{K}_1,\mathfrak{K}_2,\mathfrak{K}_3)$ is open in $C_0^\infty(\mathbb{R},\mathbb{R})\times C_0^\infty(\mathbb{R},\mathbb{R})\times C_0^\infty(\mathbb{R}^2,\mathbb{R})$; namely, it is equal to

$$\mathcal{N}_1=\{f_1(x_1),f_2(x_2),f_3(x_1,x_2)\in C_0^\infty(\mathbb{R},\mathbb{R})\times C_0^\infty(\mathbb{R},\mathbb{R})\times C_0^\infty(\mathbb{R}^2,\mathbb{R}) : f_1'(0)>0\}.$$

In addition, the original functional invariant $\mathfrak{K}$ can be uniquely recovered from the tuple $(\mathfrak{K}_1,\mathfrak{K}_2,\mathfrak{K}_3)$. Thus, the tuple $(\mathfrak{K}_1,\mathfrak{K}_2,\mathfrak{K}_3)$ defines the one-to-one correspondence between the set $\mathrm{Orb}(\widetilde{\mathcal{O}}_1)$ and the set $\mathcal{N}_1$. Hence this tuple defines the parameterization of the considered classification problem in the sense of Definition 2. □

Now let us describe the Poincaré series of the moduli numbers of the classification problem. The action of the group Diff_{q_0} induces the action $\mathcal{A}_k$ of some finite dimensional Lie group G_k on the space $J^k(\mathcal{O})$ for any integer $k\geq 0$. Thus, $\mathcal{A}_k$ is a mapping from $G_k\times J^k(\mathcal{O})$ to $J^k(\mathcal{O})$. Given any $\mathfrak{b}\in J^k(\mathcal{O})$, let $\mathcal{A}_k^{\mathfrak{b}}$ be the mapping from G_k to $J^k(\mathcal{O})$ such that $\mathcal{A}_k^{\mathfrak{b}}(\cdot)=\mathcal{A}_k(\cdot,\mathfrak{b})$. Let e_k be the identity of the group G_k. Set

$$m(k)=\dim J^k(\mathcal{O})-\max_{b\in J^k(\mathcal{O})}\operatorname{rank} d\mathcal{A}_k^{\mathfrak{b}}(e_k)=\min_{b\in J^k(\mathcal{O})}\operatorname{corank} d\mathcal{A}_k^{\mathfrak{b}}(e_k) \tag{2.5}$$

(here $\operatorname{rank} d\mathcal{A}_k^{\mathfrak{b}}(e_k)$ and $\operatorname{corank} d\mathcal{A}_k^{\mathfrak{b}}(e_k)$ are the rank and the corank of the differential of $\mathcal{A}_k^{\mathfrak{b}}$ at e_k, respectively). Roughly speaking, $m(k)$ is the dimension of the space of orbits w.r.t. the last action $\mathcal{A}_k$. The number $m(k)$ is called the *moduli number of the k-jets. The Poincaré series of the moduli numbers of the classification problem* (or, for short, *the Poincaré series of the classification problem*) is by definition the following function:

$$M(t)=\sum_{k=0}^{\infty}m(k)t^k. \tag{2.6}$$

Remark 2. Since the integer-valued function $\mathfrak{b}\mapsto\operatorname{rank} d\mathcal{A}_k^{\mathfrak{b}}(e_k)$ takes its maximal value at a Zariski open set, in (2.5) we can replace $\mathcal{O}$ by any of its generic subsets $\widetilde{\mathcal{O}}$. □

The Poincaré series could be useful in evaluating the number of the functional invariants of the given number of variables, on which the given classification problem depends, because of the following well-known fact: If one denotes by $j_l(k)$ the dimension of the space $J^k(\mathbb{R}^l,\mathbb{R})$ of k-jets of functions of l-variables, then the corresponding Poincaré series of numbers $j_l(k)$ satisfies

$$\sum_{k=0}^{\infty}j_l(k)t^k=\frac{1}{(1-t)^{l+1}} \tag{2.7}$$

(here one uses that $j_l(k)=\frac{(l+k)!}{l!k!}$). So if, for example, the Poincaré series of some classification problem is equal to

$$M(t) = t^w \sum_{i=1}^{n} \frac{p_i}{(1-t)^{i+1}}, \tag{2.8}$$

where all p_i are nonnegative integers, then it is natural to conclude that this problem depends on the tuple, consisting of p_i functional invariants of i variables for each $1 \le i \le n$, while the parameter w (i.e., the order of zero of the Poincaré series $M(t)$ at $t = 0$) is equal to the minimal $k \ge 0$ such that the action of the group G_k on the space of $J^k(\mathcal{O})$ has a nondiscrete set of orbits.

Until now nothing was known about the form of the functions $M(t)$ for general classification problems. For example, the following open question is stated in [5]: Is it true that the Poincaré series of moduli numbers are rational functions in most classification problems? In the next sections we will show by direct computations that in all classification cases 1–4 listed in the introduction it is true and, moreover, the function $M(t)$ has a unique pole at $t = 1$. On the other hand, in all considered cases (except the cases $r = 1$, $n = 3$ or 5) the Poincaré series has no representation of type (2.8) with nonnegative p_i. Below we give an algorithm to extract the number of functional invariants from Poincaré series also in these cases.

From now on we will suppose that the Poincaré series $M(t)$ of the classification problem is a rational function with a unique pole at $t = 1$. Let w_0 be the order of zero of the function $M(t)$ at $t = 0$.

LEMMA 1. *For any integers $w \ge w_0$ and l there exist a unique polynomial $R(t)$ with*

$$\deg R(t) < w - w_0 \tag{2.9}$$

and a unique rational function $Q(t)$ with the unique pole at $t = 1$ such that

$$M(t) = \frac{t^{w_0} R(t)}{(1-t)^{l+1}} + t^w Q(t). \tag{2.10}$$

Proof. Let us fix $l \in \mathbb{Z}$ and prove the existence of a representation of type (2.10) for any $w \ge w_0$ by induction in w.

If $w = w_0$, then from the condition (2.9) it follows that $R(t) \equiv 0$. Then by definition of order of zero the function $Q(t) = \frac{1}{t^{w_0}} M(t)$ is rational with the unique pole at $t = 1$, which implies (2.10).

Now suppose that a representation of type (2.10) exists for some $w = \bar{w}$, $\bar{w} \ge w_0$, and prove its existence for $w = \bar{w} + 1$. For this let $Q(t)$ and $R(t)$ be as in the representation (2.10) for $w = \bar{w}$. Let

$$Q_1(t) = \frac{1}{t}\left(Q(t) - \frac{Q(0)}{(1-t)^{l+1}}\right). \tag{2.11}$$

Then by construction Q_1 is also the rational function with the unique pole at $t = 1$. Expressing $Q(t)$ from (2.11) and substituting it into (2.10), one has

$$M(t) = \frac{t^{w_0}\big(R(t) + Q(0)t^{\bar{w}-w_0}\big)}{(1-t)^{l+1}} + t^{\bar{w}+1} Q_1(t). \tag{2.12}$$

Since $\deg\big(R(t)+Q(0)t^{\bar{w}-w_0}\big) < \bar{w}-w_0+1$, it implies the existence of a representation (2.10) also for $w = \bar{w}+1$. This completes the proof by induction of the existence part of the lemma.

Now let us prove the uniqueness part. If there exists another representation of $M(t)$ of type (2.10) with a polynomial $\bar{R}(t)$, $\deg \bar{R}(t) < w - w_0$, and a rational function $\bar{Q}(t)$ instead of $R(t)$ and $Q(t)$, then we have the following identity:

$$\bar{R}(t) - R(t) = t^{w-w_0}(1-t)^{l+1}\big(Q(t) - \bar{Q}(t)\big).$$

It implies that the polynomial $\bar{R}(t) - R(t)$ has zero of order not less than $w - w_0$. On the other hand, by assumptions $\deg\big(R(t) - \bar{R}(t)\big) < w - w_0$, which implies that $R(t) \equiv \bar{R}(t)$ and then also $Q(t) \equiv \bar{Q}(t)$. □

We will call the representation (2.10) (with $R(t)$ satisfying (2.9)) the (w,l)-*representation of the function* $M(t)$. Let N be the order of pole of $(1-t)M(t)$ at $t = 1$.

Definition 3. *The* (w,l)*-representation* (2.10) *of* $M(t)$ *with* $R(t)$ *and* $Q(t)$ *satisfying*

$$R(t) = \sum_{i=0}^{w-w_0-1} r_i t^i, \quad Q(t) = \sum_{j=l_1}^{N} \frac{q_j}{(1-t)^{j+1}}, \quad q_{l_1} \neq 0 \tag{2.13}$$

is called nice if $1 \le l \le N$, $l_1 \ge l$, *and all coefficients* r_i, q_j *in* (2.13) *are nonnegative integers.*

Of course, in general a rational function $M(t) = \frac{t^{w_0} Z(t)}{(1-t)^{N+1}}$, where $Z(t)$ is a polynomial (even with integer coefficients), may not have any nice (w,l)-representation. But if the function $M(t)$ is the Poincaré series of a classification problem, which can be parameterized by functional invariants in some reasonable way, then $M(t)$ has at least one nice representation.

To be more precise and to explain why the nice representation of the Poincaré series is interesting, let us introduce additional terminology. Let F be a functional invariant of l variables of a generic subset $\widetilde{\mathcal{O}}$ of objects from $\mathcal{O}$. Denote by Π_k the natural projection from the set $\widetilde{\mathcal{O}}$ to the space $J^k(\widetilde{\mathcal{O}})$. Let $(x_1, \dots, x_l)$ be the standard coordinates in $\mathbb{R}^l$. As before, let $C_0^\infty(\mathbb{R}^l, \mathbb{R})$ be the set of germs at 0 of smooth functions in $\mathbb{R}^l$ and let $J^k(\mathbb{R}^l, \mathbb{R})$ be the space of k-jets of germs of these functions at 0. Denote also by π_k^l the natural projection from $C_0^\infty(\mathbb{R}^l, \mathbb{R})$ to $J^k(\mathbb{R}^l, \mathbb{R})$.

Definition 4. *The weight of the functional invariant* F *of* l *variables, defined on a generic subset* $\widetilde{\mathcal{O}} \subset \mathcal{O}$, *is the minimal nonnegative integer* w *with the following property: For any integer* $k \ge w$ *there exists a mapping* $\mathfrak{F}_k : J^k(\widetilde{\mathcal{O}}) \mapsto J^{k-w}(\mathbb{R}^l, \mathbb{R})$ *such that the diagram*

$$\begin{array}{ccc} \widetilde{\mathcal{O}} & \xrightarrow{\quad F \quad} & C_0^\infty(\mathbb{R}^l, \mathbb{R}) \\ {\scriptstyle \Pi_k}\downarrow & & \downarrow{\scriptstyle \pi^l_{k-w}} \\ J^k(\widetilde{\mathcal{O}}) & \xrightarrow[\quad \mathfrak{F}_k \quad]{} & J^{k-w}(\mathbb{R}^l, \mathbb{R}) \end{array} \tag{2.14}$$

is commutative. If such w *does not exist, we will say that* F *has an infinite weight.*

Suppose that the tuple $\{F_i\}_{i=1}^s$ defines a parameterization of the classification problem on $\mathcal{O}$ such that each F_i is a functional invariant of $l_i \ge 0$ variables and the finite weight ν_i, $\nu_i \le \nu_{i+1}$. Let $\widetilde{\mathcal{O}}$ be the common domain of definition of all invariants F_i, $1 \le i \le s$. For any $k \ge 0$ set $\mu_k = \max\{i \in \{1, \dots, s\} : \nu_i \le k\}$. For a given functional invariant F_i and any $k \ge \nu_i$ denote by $\mathfrak{F}_{i,k} : J^k(\widetilde{\mathcal{O}}) \mapsto J^{k-\nu_i}(\mathbb{R}^{l_i}, \mathbb{R})$ the corresponding mapping in the commutative diagram (2.14) for F_i. Let $\mathrm{Orb}\big(J^k(\widetilde{\mathcal{O}})\big)$

be the set of orbits of $J^k(\widetilde{\mathcal{O}})$ w.r.t. the action $\mathcal{A}_k$ of the group G_k (recall that $\mathcal{A}_k$ is the action on $J^k(\mathcal{O})$ induced by the action of the group Diff_{q_0} on $\mathcal{O}$). Then the mapping $\mathfrak{F}_{i,k} : J^k(\widetilde{\mathcal{O}}) \mapsto J^{k-\nu_i}(\mathbb{R}^{l_i}, \mathbb{R})$ induces the mapping $\widehat{\mathfrak{F}}_{i,k} : \mathrm{Orb}\big(J^k(\widetilde{\mathcal{O}})\big) \mapsto J^{k-\nu_i}(\mathbb{R}^{l_i}, \mathbb{R})$ in the obvious way. Also let $\mathcal{U}_i$ be the image of $\widetilde{\mathcal{O}}$ under F_i. From Definition 2 it follows that the set $\mathcal{U}_i$ is an open subset of $C_0^\infty(\mathbb{R}^{l_i}, \mathbb{R})$. Then, according to the definition of the weight, for any $k \geq 0$ the mapping

$$(\widehat{\mathfrak{F}}_{1,k}, \dots, \widehat{\mathfrak{F}}_{\mu_k,k}) : \mathrm{Orb}\big(J^k(\widetilde{\mathcal{O}})\big) \mapsto J^{k-\nu_1}(\mathbb{R}^{l_1}, \mathbb{R}) \times \cdots \times J^{k-\nu_{\mu_k}}(\mathbb{R}^{l_{\mu_k}}, \mathbb{R}) \tag{2.15}$$

is well defined and has an open image equal to $\pi^{l_1}_{k-\nu_1}(\mathcal{U}_1) \times \cdots \times \pi^{l_{m_k}}_{k-\nu_{\mu_k}}(\mathcal{U}_{\mu_k})$.

Definition 5. *The parameterization, defined by the tuple $\{F_i\}_{i=1}^s$ as above, is called regular if for any $k \geq 0$ the mappings* (2.15) *are injective. We also say that a classification problem is regular if it can be parameterized by a regular parameterization.*

Suppose that w_1 and w_2 are, respectively, the minimal and maximal weights of functional invariants, defining some regular reparameterization, while n_1 and n_2 are, respectively, the minimal and maximal number of variables of these invariants. Then to such regular parameterization one can assign an $(n_2-n_1+1)\times(w_2-w_1+1)$ matrix P such that its (i,j)-entry p_{ij} is equal to the number of the functional invariants of $i+n_1-1$ variables and weight $j+w_1-1$ in this parameterization. Directly from the definitions of regular parameterization, formula (2.7), and Remark 2 it follows that the Poincaré series of the classification problem, admitting such regular parameterization, satisfies

$$M(t) = \frac{t^{w_1-1}}{(1-t)^{n_1}} \sum_{j=1}^{w_2-w_1+1} t^j \left(\sum_{i=1}^{n_2-n_1+1} \frac{p_{ij}}{(1-t)^{i+1}} \right). \tag{2.16}$$

Note that $w_1 = w_0$ and $n_2 = N$, where as before w_0 is the order of zero of $M(t)$ at $t = 0$ and N is the order of pole of $(1-t)M(t)$ at $t = 1$, but all other parameters appearing in (2.16) cannot be uniquely recovered from the Poincaré series $M(t)$. In the considered situation we will say that the classification problem admits *a regular (w_2, n_1)-parameterization with parameterization matrix P.*

The continuation of Example 1. Consider again the situation described in Example 1. From the normal form (2.2)–(2.3) it follows easily that the functional invariant $\mathfrak{K}$, defined by (2.1), has weight 2, while invariants $\mathfrak{K}_1$, $\mathfrak{K}_2$, and $\mathfrak{K}_3$, defined by (2.4), have weights 2, 4, and 4, respectively. Moreover, the tuple of invariants $(\mathfrak{K}_1, \mathfrak{K}_2, \mathfrak{K}_3)$ defines the regular $(4,1)$-parameterization with parameterization matrix

$$P = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.17}$$

The Poincaré series of the considered classification problem satisfies

$$M(t) = \frac{t^2}{(1-t)^2} + t^4 \left(\frac{1}{(1-t)^3} + \frac{1}{(1-t)^2} \right). \qquad \square \tag{2.18}$$

Given some regular (w_2, n_1)-parameterization of the classification problem, one can easily build a new regular (w_2, n_1)-parameterization with another parameterization matrix. Indeed, take some functional invariant F of the weight j_0, depending on i_0 variables, say $x_1, x_2, \dots, x_{i_0}$, where $2 \leq i_0 \leq n_2 - n_1 + 1$ and $1 \leq j_0 \leq w_2 - w_1$. Let $G(x_1, \dots, x_{i_0})$ be the function such that

$$F(x_1,\dots,x_{i_0}) = F(x_1,\dots,x_{i_0-1},0) + x_{i_0}G(x_1,\dots,x_{i_0}). \tag{2.19}$$

Then we can obtain the new parameterization of the classification problem by replacing the functional invariant $F(x_1,\dots,x_{i_0})$ by two functional invariants $F(x_1,\dots,x_{i_0-1},0)$ and $G(x_1,\dots,x_{i_0})$. Obviously, the first invariant has weight j_0 and depends on i_0-1 variables, while the second one has weight j_0+1 and depends on i_0 invariants. The matrix of the new parameterization is obtained from the original one by decreasing the (i_0,j_0)-entry by 1 and increasing both the (i_0-1,j_0)-entry and the (i_0,j_0+1)-entry by 1. Such transformation on the set of $(N-n_1+1)\times(w_2-w_0+1)$ matrices will be called *an elementary transformation*. Conversely, given two functional invariants G_1 and G_2 such that G_1 depends on i_0-1 variables, say $x_1,\dots,x_{i_0-1}$, and has the weight j_0, while G_2 depends on i_0 variables, say $x_1,\dots,x_{i_0}$, and has the weight j_0+1 (here again $2\le i_0\le n_2-n_1+1$ and $1\le j_0\le w_2-w_1$), one can build the new parameterization by replacing the invariants G_1 and G_2 by one invariant $G_1+x_{i_0}G_2$, which depends on i_0 variables and has the weight j_0. Of course, in this case the matrix of the new parameterization is obtained from the original one by the transformation, which is inverse to the elementary one.

Now for convenience denote

$$K_1 = N-n_1+1, \quad K_2 = w_2-w_0+1. \tag{2.20}$$

Note that among all matrices, which can be obtained from the given $K_1\times K_2$ matrix P with integer entries by a composition of a finite number of elementary transformations and their inverses, there exists a unique matrix, denoted by $\mathrm{Norm}(P)$, such that all its entries, except those lying on the first row and the last column, are equal to zero. To prove the existence of $\mathrm{Norm}(P)$ one can vanish the entries of the matrix P by a composition of elementary transformations and their inverses step by step, starting from the entry in the lower left corner, going along the first column from the bottom to the top until the entry on the second row, then passing to the bottom of the second column, going along it from the bottom to the top until the entry on the second row, and so on until the column before the last one. The uniqueness follows from the fact that if we put the entries of the matrix $\mathrm{Norm}(P)$ instead of the entries of P into the representation (2.16), then we obtain the (w_2,n_1)-representation of the Poincaré function $M(t)$. This representation is unique according to Lemma 1, and the matrix $\mathrm{Norm}(P)$ is obviously uniquely recovered from it. Also it is not difficult to express all nontrivial entries of $\mathrm{Norm}(P)$ by the entries of P:

(2.21a)
$$(\mathrm{Norm}(P))_{1j} = p_{1j} + \sum_{l=0}^{j-1}\sum_{k=1}^{K_1-1}\binom{k+l-1}{l}p_{k+1,j-l}, \qquad 1\le j\le K_2-1,$$

(2.21b)
$$(\mathrm{Norm}(P))_{i,K_2} = p_{i,K_2} + \sum_{l=0}^{K_1-i+1}\sum_{k=1}^{K_2-1}\binom{K_2-k+l-1}{l}p_{i+l,k}, \qquad 2\le i\le K_1,$$

(2.21c)
$$(\mathrm{Norm}(P))_{1,K_2} = p_{1,K_2}.$$

The last relation can be proved, for example, using the procedure of passing from P to $\mathrm{Norm}(P)$, described above, and the following well-known combinatorial identity:

$$\sum_{i=1}^{n}\binom{i+k-1}{k} = \binom{n+k}{k+1}.$$

If the matrix P has only nonnegative integer entries, then the matrix $\mathrm{Norm}(P)$ is obtained from P by a finite composition of elementary transformations (without using their inverses) and also has only nonnegative integer entries (which follows also from relations (2.21)). Moreover, if we put the entries of the matrix $\mathrm{Norm}(P)$ instead of the entries of P into the representation (2.16), then we obtain the nice (w_2, n_1)-representation of the Poincaré function $M(t)$ of our classification problem. We also say that this nice representation *corresponds to the matrix* $\mathrm{Norm}(P)$. We can summarize all of the above in the following proposition.

PROPOSITION 1. *If the classification problem admits a regular (w_2, n_1)-parameterization with parameterization matrix P, then it admits a regular (w_2, n_1)-parameterization with parameterization matrix $\mathrm{Norm}(P)$, and its Poincaré series has the nice (w_2, n_1)-representation, which corresponds to the matrix $\mathrm{Norm}(P)$.*

The last proposition indicates that the nice representations of the Poincaré series (if they exist) may be used in the definition of the intrinsic number of functional invariants of each number of variables and weight, on which the given classification problem depends. Suppose that the Poincaré series $M(t)$ has the nice representation for some (w, l). Thus, the set

$$\mathrm{NS}\big(M(t)\big) \stackrel{def}{=} \{(w,l) : \text{the } (w,l)\text{-representation of } M(t) \text{ is nice}\} \tag{2.22}$$

is not empty. The natural question is what pair to choose from $\mathrm{NS}\big(M(t)\big)$. To answer this question we propose to introduce the order $\prec$ on the set of ordered pairs (w, l) in the following way: $(w, l) \prec (\bar{w}, \bar{l})$ if and only if $w < \bar{w}$ or $w = \bar{w}$, but $l > \bar{l}$. By Definition 3,

$$\mathrm{NS}\big(M(t)\big) \subset \{(w,l) : w \geq w_0,\, l \leq N\}, \tag{2.23}$$

which implies immediately that the set $\mathrm{NS}\big(M(t)\big)$ contains the minimal element w.r.t. the introduced order $\prec$. This minimal element will be called *the characteristic pair of the classification problem*. Denote it by $(\bar{w}, \bar{l})$. Let $\mathcal{C}$ be the $(N - \bar{l} + 1) \times (\bar{w} - w_0 + 1)$ matrix such that the $(\bar{w}, \bar{l})$-representation of $M(t)$ corresponds to the matrix $\mathcal{C}$.

DEFINITION 6. *The (i, j)-entry of the matrix $\mathcal{C}$ is called the intrinsic number of the functional invariants of $i+\bar{l}-1$ variables and the weight $j+w_0-1$ of the considered classification problem. The matrix $\mathcal{C}$ is called the characteristic matrix of the classification problem. Any regular $(\bar{w}, \bar{l})$-parameterization of the problem (if it exists) with the parameterization matrix $\mathcal{C}$ is called the characteristic regular parameterization.*

In general, it is better to have a parameterization consisting of invariants which have minimal possible weight and depend on the maximal possible number of variables. Our definition of characteristic parameterization is in accordance with this goal. Actually the maximal weight of invariants appearing in a characteristic regular parameterization is not greater than the maximal weight of invariants appearing in any other regular parameterization. Besides, the minimal number of variables in invariants of a characteristic regular parameterization is not less than the minimal number of variables in invariants of any other regular parameterization having the same maximal weight of invariants as a characteristic one.

One can improve the formula (2.23) for the localization of the set $\mathrm{NS}\big(M(t)\big)$. Indeed, let d be the degree of the rational function $M(t)$ at infinity. Namely, if $M(t) = \frac{Q_1(t)}{Q_2(t)}$, where $Q_1(t)$ and $Q_2(t)$ are polynomials, then $d = \deg Q_1(t) - \deg Q_2(t)$. Then

$$\mathrm{NS}\big(M(t)\big) \subset \{(w,l) : w \geq w_0,\ 1 \leq l \leq \min(w-d-1, N)\}. \tag{2.24}$$

To prove (2.24) we actually have to prove that if the pair $(w,l) \in \mathrm{NS}\big(M(t)\big)$, then $l \leq w-d-1$. Indeed, from (2.10) and (2.13) it follows that $d = \max(w-l_1-1,\ w_0+\deg R-l_1-1)$. But first, since $l_1 > l$, we have $w-l_1-1 \leq w-l-1$, and second, since $\deg R < w-w_0$, we have $w_0+\deg R-l_1-1 < w-l-1$. Therefore $d \leq w-l-1$. □

From (2.24) it follows also that

$$\mathrm{NS}\big(M(t)\big) \subset \{(w,l) : w \geq \max(w_0, d+2)\}. \tag{2.25}$$

The relations (2.24) and (2.25) may be useful in searching for the characteristic pair of the classification problem.

Conclusion about Example 1. The representation (2.18) of the Poincaré series of the classification problem considered in Example 1 is its nice $(4,1)$-representation. Let us show that $(4,1)$ is the characteristic pair of the considered classification problem. Indeed, from (2.18) in the considered case $w_0 = 2$, $N = 2$, and $d = 2$. Hence from (2.25) it follows that

$$\mathrm{NS}\big(M(t)\big) \subset \{(w,l) : w \geq 4\}. \tag{2.26}$$

Further, using (2.24), one has $\mathrm{NS}\big(M(t)\big) \cap \{(w,l) : w = 4\} = \{(4,1)\}$, which together with (2.26) implies that $(4,1)$ is the minimal element of $\mathrm{NS}\big(M(t)\big)$. In other words, $(4,1)$ is the characteristic pair of our classification problem. Also, it implies that the characteristic matrix $\mathcal{C}$ of the problem is equal to the matrix P from (2.17) (note that in this case $\mathrm{Norm}(P) = P$). Thus, *the characteristic parameterization of a set of germs of Riemannian metrics on an oriented two-dimensional Riemannian manifold consists of one functional invariant of one variable and the weight* 2*, one functional invariant of one variable and the weight* 4*, and one functional invariant of two variables and the weight* 4.

Another useful property of the set $\mathrm{NS}\big(M(t)\big)$ can be formulated as follows.

Lemma 2. *Assume that the function $M(t)$ has the nice (w,l)-representation* (2.10)*, the functions $R(t)$, $Q(t)$, and the number l_1 are as in* (2.13)*, and $l_1 = l$ (or, equivalently, $q_l > 0$); then $(w-1,\ l-1) \notin \mathrm{NS}\big(M(t)\big)$.*

Proof. Let $S(t)$ be the polynomial such that $M(t) = \frac{S(t)}{(1-t)^{N+1}}$. Then, using the assumption $l = l_1$, it is easy to get

$$\deg S(t) = w + N - l. \tag{2.27}$$

Moreover, directly from (2.10) and (2.13) one can obtain that

$$\frac{d^{w+N-l}S}{d\,t^{w+N-l}} = (-1)^{N-l}(w+N-l)!\,q_l. \tag{2.28}$$

On the other hand, if the $(w-1,\ l-1)$-representation of $M(t)$ has the form

$$M(t) = \frac{t^{w_0}\sum_{i=0}^{w-w_0-2}\bar{r}_i t^i}{(1-t)^l} + t^{w-1}\sum_{j=l_2}^{N}\frac{\bar{q}_j}{(1-t)^{j+1}}, \quad \bar{q}_{l_2} \neq 0, \tag{2.29}$$

then $\deg S(t) = \max(w-1+N-l_2,\ w+N-l-1)$. Comparing this with (2.27) one gets easily that $l_2 = l-1$. But then by analogy with (2.28) (applied for the

$(w-1,\, l-1)$-representation instead of the (w,l)-representation), one has

$$\frac{d^{w+N-l}S}{d\,t^{w+N-l}} = (-1)^{N-l+1}(w+N-l)!\,\bar{q}_{l-1}. \tag{2.30}$$

Comparing (2.28) and (2.30), we obtain that $\bar{q}_{l-1} = -q_l$. Hence $\bar{q}_{l-1} < 0$, and the $(w-1,\, l-1)$-representation (2.29) is not nice. □

As a direct consequence of Proposition 1, the previous lemma, and the relation (2.21c) one has the following corollary.

COROLLARY 1. *If the classification problem with the Poincaré series $M(t)$ admits the regular (w,l)-parameterization with the parameterization matrix P such that the entry in the upper right corner of P is positive (i.e., in the previous notation $p_{1,w-w_0+1} > 0$), then $(w-1,\, l-1) \notin \mathrm{NS}\big(M(t)\big)$.*

In what follows we will show that all of the classification problems 1–4 listed in the introduction are regular. For each of these problems we will describe explicitly some of its regular (w,l)-parameterization such that (w,l) is the characteristic pair of the problem and find its characteristic matrix, which also enables us to obtain the characteristic parameterization.

Remark 3. In the classification problem of Example 1 and, as we will see later, in all of the classification problems 1–4 listed in the introduction, one can assign to any generic object the canonical coordinates. This allows us to construct functional invariants in the sense of Definition 1 from the invariants defined on the ambient manifold (in the same manner as in Example 1 where we constructed the functional invariant $\mathfrak{K}$ from the Gaussian curvature). Sometimes (as in the case of $(1,3)$ control-affine systems and in the case of Riemannian metrics on a nonoriented two-dimensional manifold) the canonical coordinates (even for generic objects) are defined up to some discrete group of transformations (in the mentioned cases a group of reflections). Such classification problems can be treated in a similar way after slight modifications of definitions of the functional invariants, the parameterization, the weight, and the regular parameterization.

Indeed, a discrete group Γ on $\mathbb{R}^l$ induces the action $\mathfrak{S}$ on $C_0^\infty(\mathbb{R}^l,\mathbb{R})$ in the obvious way. By the functional quasi invariant we mean the mapping from the generic subset $\widetilde{\mathcal{O}}$ of the set of considered objects to the set of orbits w.r.t. the action $\mathfrak{S}$ on $C_0^\infty(\mathbb{R}^l,\mathbb{R})$ such that this mapping is invariant w.r.t. the action of the group Diff_{q_0} on $\widetilde{\mathcal{O}}$. The notions of the weight of the functional quasi invariants, the parameterization, and the regular parameterization by functional quasi invariants can be defined in the natural way. Besides, in parameterizations we can admit invariants taking their values in some discrete set (or, for short, discrete invariants) and define the regular parameterization by functional (quasi) invariants and discrete invariants in the natural way. Permission of the functional quasi invariant and the discrete invariants does not affect the formula (2.16) for the Poincaré series in terms of the parameterization matrix containing the information about the number of functional (quasi) invariants of the given number of variables and weight in the parameterization. Thus, also the characteristic pair, the characteristic matrix, and the characteristic regular parameterization can be defined in this case, too. In the case of Riemannian metrics on the nonoriented manifolds the canonical coordinates are defined up to the reflection $(x_1,x_2) \mapsto (x_1,-x_2)$, the characteristic pair and the characteristic matrix are as in the oriented case, and the characteristic parameterization is defined by formula (2.4), where we take into account both canonical coordinates (x_1,x_2) and $(x_1,-x_2)$.

Further, according to Proposition 3.2 of [2] (see also Remark 1 and the paragraph before it in the introduction), the state-feedback classification problem for the $(1,3)$

control-affine system admits regular $(2,2)$-parameterization by one discrete invariant taking values in $\{-1,1\}$ and the functional quasi invariants with the 2×2 parameterization matrix $P=\left(\begin{smallmatrix}2&0\\0&1\end{smallmatrix}\right)$. Besides, $w_0=1$ and $N=3$. By the inverse to the elementary transformation one can transform P into the matrix $\widetilde{P}=\left(\begin{smallmatrix}1&0\\1&0\end{smallmatrix}\right)$. The Poincaré series $M(t)$ of the problem satisfies

$$M(t)=t\left(\frac{1}{(1-t)^3}+\frac{1}{(1-t)^4}\right).$$

It is not difficult to see that by erasing the last column of the matrix $\widetilde{P}$ one obtains the characteristic matrix $\mathcal{C}=(1,1)^T$ of the considered classification problem and the characteristic pair is equal to $(1,2)$. Thus, *the characteristic regular parameterization of the state-feedback classification problem for the* $(1,3)$ *control-affine system consists of one functional quasi invariant of three variables and the weight* 1*, one functional quasi invariant of two variables and the weight* 1*, and the discrete invariant from the set* $\{-1,1\}$. This parameterization is obtained from the original one by a rearrangement of the invariants, which corresponds to the inverse to elementary transformation, transforming the matrix P into the matrix $\widetilde{P}$ (such rearrangements were described in the paragraph after the formula (2.19)). □

3. Classification of $(1,n)$ control-affine systems for $n\geq 4$. For $r=1$ the system (1.1) has the form

$$\dot{q}=f_0(q)+u\,f_1(q),\quad q\in M,\quad u\in\mathbb{R}. \tag{3.1}$$

Our genericity assumptions are

$$\dim\operatorname{span}(f_0,f_1,[f_1,f_0],\dots,(\operatorname{ad}f_1)^{n-2}f_0)=n, \tag{3.2}$$

$$\dim\operatorname{span}\big(f_1,\big[f_0,[f_0,f_1]\big],[f_1,f_0],\dots,(\operatorname{ad}f_1)^{n-2}f_0\big)=n, \tag{3.3}$$

and for $n\geq 5$ also

$$\dim\operatorname{span}\big(f_0,f_1,\big[f_0,[f_0,f_1]\big],[f_1,f_0],\dots,(\operatorname{ad}f_1)^{n-3}f_0\big)=n. \tag{3.4}$$

The group of feedback transformations

$$u=\beta(q)\tilde{u}+\alpha(q),\quad \alpha(q),\beta(q)\in\mathbb{R},\quad \beta(q)\neq 0 \tag{3.5}$$

acts naturally on the set of pairs of vector fields (f_0,f_1). The orbit w.r.t. this action is

$$\mathcal{O}_{(f_0,f_1)}=\{(f_0+\alpha f_1,\beta f_1):\alpha,\beta:M\mapsto\mathbb{R}\text{ are functions, }\beta\neq 0\}. \tag{3.6}$$

The first observation is given by the following proposition.

PROPOSITION 2. *If the pair* (f_0,f_1) *satisfies conditions* (3.2) *and* (3.3)*, then there exists a unique pair* $(F_0,F_1)\in\mathcal{O}_{(f_0,f_1)}$ *such that*

$$\big[F_0,[F_0,F_1]\big]=F_0+I_1F_1+I_2[F_1,F_0]+\sum_{k=3}^{n-2}I_k(\operatorname{ad}F_1)^kF_0. \tag{3.7}$$

Proof. By assumption (3.2) the vector fields $f_0, f_1, [f_1, f_0], \dots, (\operatorname{ad} f_1)^{n-2} f_0$ constitute the frame on M. Therefore there are functions E, N, $J_1, \dots, J_{n-2}$ such that

$$\big[f_0, [f_0, f_1]\big] = Ef_0 + J_1 f_1 + J_2[f_1, f_0] + Z\big[f_1, [f_1, f_0]\big] + \sum_{k=3}^{n-2} J_k (\operatorname{ad} f_1)^k f_0. \tag{3.8}$$

Take some pair $(\tilde f_0, \tilde f_1) \in \mathcal{O}_{(f_0, f_1)}$; then

$$\tilde f_0 = f_0 + \alpha f_1, \quad \tilde f_1 = \beta f_1. \tag{3.9}$$

Suppose that

$$\big[\tilde f_0, [\tilde f_0, \tilde f_1]\big] = \widetilde{E} \tilde f_0 + \widetilde{J}_1 \tilde f_1 + \widetilde{J}_2[\tilde f_1, \tilde f_0] + \widetilde{Z}\big[\tilde f_1, [\tilde f_1, \tilde f_0]\big] + \sum_{k=3}^{n-2} J_k (\operatorname{ad} \tilde f_1)^k \tilde f_0. \tag{3.10}$$

First note that

$$\widetilde{E} = \beta E. \tag{3.11}$$

It follows immediately from (3.9) and the following relations:

$$\big[\tilde f_0, [\tilde f_0, \tilde f_1]\big] \equiv \beta \big[f_0, [f_0, f_1]\big] - \alpha\beta \big[f_1, [f_1, f_0]\big] \bmod \operatorname{span}(f_1, [f_0, f_1]), \tag{3.12}$$

$$(\operatorname{ad} f_1)^k f_0 \in \operatorname{span}\big(\tilde f_1, \dots, (\operatorname{ad} \tilde f_1)^k \tilde f_0\big), \quad k \in \mathbb{N}. \tag{3.13}$$

From assumption (3.3) it follows that $E \neq 0$. Therefore, taking $\beta = \frac{1}{E}$, we have

$$\widetilde{E} = 1. \tag{3.14}$$

Let us denote by $\overline{\mathcal{O}}_{(f_0, f_1)}$ the set of all pairs $(\tilde f_0, \tilde f_1)$ satisfying (3.14). We can assume from the beginning that the original pair (f_0, f_1) belongs to $\overline{\mathcal{O}}_{(f_0, f_1)}$; i.e., $E = 1$ (we make this assumption just to avoid extra notation). If $(\tilde f_0, \tilde f_1) \in \overline{\mathcal{O}}_{(f_0, f_1)}$, then also $\widetilde{E} = 1$. Hence $\beta = 1$ or, equivalently, $f_1 = \tilde f_1$. In other words, condition (3.14) normalizes the vector field f_1 or the direction defining the straight line in the set of admissible velocities of the system (3.1) at any point.

Further, from (3.9), taking into account that $\beta = 1$, it follows easily that

$$(\operatorname{ad} f_1)^k f_0 \equiv (\operatorname{ad} \tilde f_1)^k \tilde f_0 \bmod \operatorname{span}(f_1), \quad k \in \mathbb{N}.$$

This and relation (3.12) imply that

$$\widetilde{Z} = Z - \alpha. \tag{3.15}$$

Setting $\alpha = Z$, we make $\widetilde{Z} = 0$, which normalizes the drift $\tilde f_0$. Thus, we have proved that there is a unique $(\tilde f_0, \tilde f_1) \in \mathcal{O}_{(f_0, f_1)}$ such that $\widetilde{E} = 1$ and $\widetilde{Z} = 0$, which completes the proof of the proposition. □

Remark 4. The mappings $I_1, \dots, I_{n-2}$ from M to $\mathbb{R}$, defined by identity (3.7), are state-feedback invariants of the control system (3.1). □

The vector field F_0 and the pair of vector fields (F_0, F_1) from Proposition 2 are called *the canonical drift* and *the canonical pair* of the system (3.1), respectively.

Remark 5. Actually, in the case $n = 4$, the vector $F_0(q)$ is the velocity of the unique abnormal extremal starting at q of the time optimal problem defined by system (3.1). □

Now fix some point $q_0 \in M$. Denote by e^{tf} the flow generated by the vector field f and by $q \circ e^{tf}$ the image of the point q w.r.t. this flow. Let $\Phi_n : \mathbb{R}^n \mapsto M$ be the following mapping:

$$\Phi_4(x_1, x_2, x_3, x_4) = q_0 \circ e^{x_4 \left[F_1,[F_1,F_0]\right]} \circ e^{x_3[F_1,F_0]} \circ e^{x_2 F_1} \circ e^{x_1 F_0}, \tag{3.16}$$

$$\begin{aligned}\Phi_n(x_1, \dots, x_n) = q_0 \circ e^{x_n (\operatorname{ad} F_1)^{n-3} F_0} \circ \dots \circ e^{x_5 \left[F_1,[F_1,F_0]\right]} \\ \circ e^{x_4 \left[F_0,[F_0,F_1]\right]} \circ e^{x_3[F_1,F_0]} \circ e^{x_2 F_0} \circ e^{x_1 F_1}, \quad n \geq 5.\end{aligned} \tag{3.17}$$

From assumption (3.2) in the case $n = 4$ or assumption (3.4) in the case $n \geq 5$ it follows that $\Phi_n'(0)$ is bijective. Hence Φ_n^{-1} defines *the canonical coordinates in a neighborhood of* q_0 (or, for short, *the canonical coordinates at* q_0). Denote

$$\mathcal{I}_k = I_k \circ \Phi_n, \quad k = 1, \dots, n-2. \tag{3.18}$$

Assigning to any generic germ at q_0 of control-affine systems (3.1) the function $\mathcal{I}_k$, we obtain the functional invariant of n variables of this set of objects in the sense of Definition 1 for any $1 \leq k \leq n-2$.

Now let us consider the cases $n = 4$ and $n \geq 5$ separately.

(a) *The case* $n = 4$. By (3.16) and (3.7), in the canonical coordinates the vector fields F_0 and F_1 have the form

$$F_0 = \frac{\partial}{\partial x_1}, \quad F_1 = \sum_{k=1}^{4} a_k \frac{\partial}{\partial x_k}, \tag{3.19}$$

where the components of F_1 satisfy the following second order linear ordinary differential equations w.r.t. the variable x_1:

$$\frac{\partial^2 a_k}{\partial x_1^2} + \mathcal{I}_2 \frac{\partial a_k}{\partial x_1} - \mathcal{I}_1 a_k - \delta_{1,k} = 0, \quad k = 1, 2, 3, 4, \tag{3.20}$$

with the following restrictions on the initial conditions for any $k = 1, 2, 3, 4$:

$$a_k(0, x_2, x_3, x_4) \equiv \delta_{2k}, \tag{3.21a}$$

$$\frac{\partial a_k}{\partial x_1}(0, 0, x_3, x_4) \equiv -\delta_{3k}, \tag{3.21b}$$

$$\frac{\partial^2 a_k}{\partial x_1 \partial x_2}(0, 0, 0, x_4) \equiv -\delta_{4k}, \tag{3.21c}$$

where δ_{ij} is the Kronecker symbol. Let for any $k = 1, 2, 3, 4$

$$\beta_k(x_2, x_3, x_4) \stackrel{def}{=} \frac{\partial^3 a_k}{\partial x_1 \partial x_2^2}(0, x_2, x_3, x_4), \tag{3.22a}$$

$$\psi_k(x_3, x_4) \stackrel{def}{=} \frac{\partial^3 a_k}{\partial x_1 \partial x_2 \partial x_3}(0, 0, x_3, x_4). \tag{3.22b}$$

Thus, with any germ at q_0 of a four-dimensional affine system (3.1) satisfying genericity assumptions (3.2) and (3.3), one can associate the ordered tuple

$$(\mathcal{I}_1, \mathcal{I}_2, \beta_1, \beta_2, \beta_3, \beta_4, \psi_1, \psi_2, \psi_3, \psi_4) \tag{3.23}$$

of state-feedback functional invariants, consisting of two germs $\mathcal{I}_1$ and $\mathcal{I}_2$ of functions of four variables at 0, four germs $\beta_1, \beta_2, \beta_3, \beta_4$ of functions of three variables at 0, and four germs $\psi_1, \psi_2, \psi_3, \psi_4$ of functions of three variables at 0. We call it *the tuple of the primary invariants of the* $(1,4)$ *control-affine system* (3.1) *at the point* q_0. Note that by (3.22) the functional invariants β_k and ψ_k have the weight 3 for any $1 \le k \le 4$, while by (3.20) and (3.21) the functional invariants $\mathcal{I}_1$ and $\mathcal{I}_2$ have the weight 2.

Further, fixing β_k and ψ_k and using (3.21b) and (3.21c), we can find $\frac{\partial a_k}{\partial x_1}(0, x_2, x_3, x_4)$ for any $1 \le k \le 4$ by the appropriate integrations (see (3.31) below). If in turn we fix also $\mathcal{I}_1$ and $\mathcal{I}_2$, then from the knowledge of $\frac{\partial a_k}{\partial x_1}(0, x_2, x_3, x_4)$, condition (3.21a), and differential equation (3.20), we can recover the functions $a_k(x_1, x_2, x_3, x_4)$ and therefore our control-affine system itself, just using the standard existence and uniqueness results from the theory of ordinary differential equations. We summarize all of the above in the following theorem.

THEOREM 1. *Given two arbitrary germs* $\mathcal{I}_1$ *and* $\mathcal{I}_2$ *of functions of four variables at* 0*, four arbitrary germs* $\beta_1, \beta_2, \beta_3, \beta_4$ *of functions of three variables at* 0*, and four arbitrary germs* $\psi_1, \psi_2, \psi_3, \psi_4$ *of functions of two variables at* 0*, there exists a unique, up to state-feedback transformation of type* (1.3)*, four-dimensional control-affine system with scalar input, satisfying genericity assumptions* (3.2) *and* (3.3)*, such that the tuple* $(\mathcal{I}_1, \mathcal{I}_2, \beta_1, \beta_2, \beta_3, \beta_4, \psi_1, \psi_2, \psi_3, \psi_4)$ *is its tuple of the primary invariants at the given point* q_0*. In other words, the tuples of the primary invariants give the regular* $(3,2)$*-parameterization of the considered classification problem with the following* 3×2 *parameterization matrix* P*:*

$$P = \begin{pmatrix} 0 & 4 \\ 0 & 4 \\ 2 & 0 \end{pmatrix}. \tag{3.24}$$

The Poincaré series $M(t)$ *of the considered classification problem satisfies*

$$M(t) = \frac{2t^2}{(1-t)^5} + t^3 \left(\frac{4}{(1-t)^4} + \frac{4}{(1-t)^3} \right). \tag{3.25}$$

It turns out that $(3,2)$ is the characteristic pair of the considered classification problem. Indeed, as before let w_0 be the order of zero of $M(t)$ at $t = 0$, N be the order of pole of $(1-t)M(t)$ at $t = 1$, and d be the degree of $M(t)$ (at infinity). Then from (3.25) it follows that $w_0 = 2$, $N = 4$, and $d = 0$. Hence from (2.23) (or (2.25)) it follows that

$$\mathrm{NS}\big(M(t)\big) \subset \{(w,l) : w \ge 2\}. \tag{3.26}$$

Further, using (2.24), one has

(3.27)

$$\mathrm{NS}\big(M(t)\big) \cap \{(w,l) : w = 2\} \subset \{(2,1)\}, \quad \mathrm{NS}\big(M(t)\big) \cap \{(w,l) : w = 3\} \subset \{(3,1), (3,2)\}.$$

But by the previous theorem our classification problem admits $(3,2)$-parameterization such that its parameterization matrix has a positive entry in the upper right corner.

Therefore from Corollary 1 it follows that $(2,1) \notin \mathrm{NS}\big(M(t)\big)$. Since $(3,2) \prec (3,1)$ we can conclude from (3.26) and (3.27) that $(3,2)$ is the minimal element of $\mathrm{NS}\big(M(t)\big)$. In other words, $(3,2)$ is the characteristic pair of our classification problem. Also, it implies that the characteristic matrix $\mathcal{C}$ of the problem is equal to $\mathrm{Norm}(P)$, which can be found easily by the series of elementary transformations. Namely,

$$\mathcal{C} = \mathrm{Norm}(P) = \begin{pmatrix} 2 & 4 \\ 0 & 6 \\ 0 & 2 \end{pmatrix}. \tag{3.28}$$

Conclusion 1. *The characteristic parameterization of* $(1,4)$ *control-affine systems, up to state-feedback transformations, consists of two functional invariants of four variables and the weight* 3*, six functional invariants of three variables and the weight* 3*, two functional invariants of two variables and the weight* 2*, and four functional invariants of two variables and the weight* 3*.*

In order to obtain a characteristic parameterization from the parameterization by the tuple of the primary invariants one can implement some series of rearrangement of the primary invariants according to the series of elementary transformations from the matrix P to $\mathrm{Norm}(P)$, as described in the previous section (see, for example, formula (2.19) and the paragraph that follows it).

We finish the treatment of the case of $(1,4)$-affine control systems by writing the local normal form of such systems, up to state-feedback transformation, in terms of the tuple of their primary invariants: Let N be the solution of the following nonhomogeneous second order linear ordinary differential equation w.r.t. the variable x_1 with prescribed initial values:

$$\begin{cases} \dfrac{\partial^2 N}{\partial x_1^2} + \mathcal{I}_2 \dfrac{\partial N}{\partial x_1} - \mathcal{I}_1 N - 1 = 0; \\ N(0, x_2, x_3, x_4) \equiv 0, \quad \dfrac{\partial N}{\partial x_1}(x_1, x_2, x_3, x_4)\Big|_{x_1=0} \equiv 0. \end{cases} \tag{3.29}$$

Then let the functions ρ_1, ρ_2 be the solution of the following homogeneous second order linear ordinary differential equations w.r.t. the variable x_1 with prescribed initial values:

$$\begin{cases} \dfrac{\partial^2 \rho_i}{\partial x_1^2} + \mathcal{I}_2 \dfrac{\partial \rho_i}{\partial x_1} - \mathcal{I}_1 \rho_i = 0, \quad i = 1, 2; \\ \begin{pmatrix} \rho_1(0, x_2, x_3, x_4) & \rho_1(0, x_2, x_3, x_4) \\ \frac{\partial}{\partial x_1}\rho_1(0, x_2, x_3, x_4) & \frac{\partial}{\partial x_1}\rho_2(0, x_2, x_3, x_4) \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \end{cases} \tag{3.30}$$

Let also

$$B_k(x_2, x_3, x_4) = -\delta_{3k} + x_2 \left(-\delta_{4k} + \int_0^{x_3} \psi_k(y, x_4)\, dy \right) + \int_0^{x_2} (x_2 - y) \beta_k(y, x_3, x_4)\, dy \tag{3.31}$$

for $1 \le k \le 4$ (actually $B_k(x_2, x_3, x_4) = \frac{\partial a_k}{\partial x_1}(0, x_2, x_3, x_4)$, where the functions a_k are as in (3.19)). Then the generic germs at q_0 of four-dimensional control-affine system (3.1) with the tuple of the primary invariants $(\mathcal{I}_1, \mathcal{I}_2, \beta_1, \beta_2, \beta_3, \beta_4, \psi_1, \psi_2, \psi_3, \psi_4)$ has the following form in the canonical coordinates (x_1, x_2, x_3, x_4), up to a feedback transformation:

$$\begin{cases} \dot{x}_1 = 1 + (N + B_1\rho_2)u, \\ \dot{x}_2 = (\rho_1 + B_2\rho_2)u, \\ \dot{x}_i = B_i\rho_2 u, \quad i = 3, 4, \end{cases} \tag{3.32}$$

where $u \in \mathbb{R}$.

(b) *The case* $n \geq 5$. By (3.17) and (3.7), in the canonical coordinates the vector fields F_0 and F_1 have the form

$$F_0 = \sum_{k=1}^{n} a_k \frac{\partial}{\partial x_k}, \quad F_1 = \frac{\partial}{\partial x_1}, \tag{3.33}$$

where the components a_k of F_0 satisfy the system of partial differential equations

(3.34)

$$\mathcal{I}_{n-2}\frac{\partial^{n-2}a_k}{\partial x_1^{n-2}} + \sum_{j=3}^{n-3}\mathcal{I}_j\frac{\partial^j a_k}{\partial x_1^j} + \sum_{l=1}^{n}\left(a_l\frac{\partial^2 a_k}{\partial x_1^2} - \frac{\partial a_l}{\partial x_1}\frac{\partial a_k}{\partial x_l}\right) + \mathcal{I}_2\frac{\partial a_k}{\partial x_1} + a_k + \mathcal{I}_1\delta_{1,k} = 0,$$

$k = 1, \ldots, n,$

with the following restrictions on the boundary conditions for any $1 \leq k \leq n$:

$$a_k(0, x_2, \ldots, x_n) \equiv \delta_{2k}, \tag{3.35a}$$

$$\frac{\partial a_k}{\partial x_1}(0, 0, x_3 \ldots, x_n) \equiv \delta_{3k}, \tag{3.35b}$$

$$\frac{\partial^2 a_k}{\partial x_1 \partial x_2}(0, 0, 0, x_4, \ldots, x_n) \equiv -\delta_{4k}, \tag{3.35c}$$

$$\frac{\partial^j a_k}{\partial x_1^j}(0, \ldots, 0, x_{j+1}, \ldots, x_n) \equiv \delta_{j+3,k}, \quad 2 \leq j \leq n-3, \tag{3.35d}$$

where δ_{ij} is the Kronecker symbol. Note also that the genericity assumption (3.4) implies that

$$\mathcal{I}_{n-2} \neq 0. \tag{3.36}$$

Let us introduce the following functions for any $1 \leq k \leq n$:

(3.37a)

$$\beta_k(x_2, \ldots, x_n) \stackrel{def}{=} \frac{\partial^3 a_k}{\partial x_1 \partial x_2^2}(0, x_2, \ldots, x_n),$$

(3.37b)

$$\psi_k(x_3, \ldots, x_n) \stackrel{def}{=} \frac{\partial^3 a_k}{\partial x_1 \partial x_2 \partial x_3}(0, 0, x_3, \ldots, x_n),$$

(3.37c)

$$\phi_{kjl}(x_l, \ldots, x_n) \stackrel{def}{=} \frac{\partial^{j+1} a_k}{\partial x_1^j \partial x_l} a_k(0, \ldots, 0, x_l, \ldots x_n), \quad 2 \leq j \leq n-3, \quad 2 \leq l \leq j+2.$$

Thus, with any germ at q_0 of an n-dimensional affine system (3.1), satisfying genericity assumptions (3.2) and (3.3), one can associate the ordered tuple

$$\Big(\{\mathcal{I}_s(x_1, \ldots, x_n)\}_{s=1}^{n-2}, \{\beta_k(x_2, \ldots, x_n)\}_{k=1}^{n}, \{\psi_k(x_3, \ldots, x_n)\}_{k=1}^{n}, \\ \{\phi_{kjl}(x_l, \ldots, x_n) : 1 \leq k \leq n, 2 \leq j \leq n-3, 2 \leq l \leq j+2\}\Big) \tag{3.38}$$

of state-feedback invariants. We call it *the tuple of the primary invariants of the* $(1,n)$*-affine control system* (3.1) *with* $n>4$ *at the point* q_0. Note that by (3.22) for any $1\le k\le n$ the functional invariants β_k and ψ_k have the weight 3, and the functional invariants ϕ_{kjl} have the weight $j+1$, while by (3.20) and (3.21) for any $1\le s\le n-2$ the functional invariants $\mathcal{I}_s$ have the weight $n-2$.

Further, fixing β_k and ψ_k and using (3.35b) and (3.35c), one finds $\frac{\partial a_k}{\partial x_1}(0,x_2,\dots,x_n)$ for any $1\le k\le n$ by the appropriate integrations. Similarly, fixing $\{\phi_{kjl}\}_{l=2}^{j+2}$ for given j, $2\le j\le n-3$, and using (3.35d), one can find $\frac{\partial^j a_k}{\partial x_1^j}(0,x_2,\dots,x_n)$ for any $1\le k\le n$ by the appropriate integrations. Finally, if we suppose that all functions β_k, ψ_k, and ϕ_{kjl} are real analytic and fix also real analytic $\{\mathcal{I}_s\}_{s=1}^{n-2}$, then from the knowledge of $\frac{\partial^j a_k}{\partial x_1^j}(0,x_2,\dots,x_n)$ for all $1\le j\le n-3$, condition (3.35a), and differential equation (3.34), we can recover the functions $a_k(x_1,\dots,x_n)$ and therefore our affine control system itself, just by using the classical Cauchy–Kowalewsky theorem for system (3.34). We summarize all of the above in the following theorem.

THEOREM 2. *If* $n\ge 5$*, then given an arbitrary tuple* (3.38) *of germs of real analytic functions, there exists a unique, up to state-feedback real analytic transformation of type* (1.3)*,* n*-dimensional real analytic control-affine system with scalar input, satisfying genericity assumptions* (3.2)*,* (3.3)*, and* (3.4)*, such that the tuple* (3.38) *is its tuple of the primary invariants at the given point* q_0*. In other words, the tuples of the primary invariants give the regular* $(n-2,2)$*-parameterization of the considered classification problem in the real analytic category with the* $(n-1)\times(n-4)$*-parameterization matrix* P *such that for* $n=5$

$$P=(5,10,10,3)^T, \tag{3.39}$$

and for $n>5$

$$P=\begin{pmatrix} 0 & \dots\dots\dots & 0 & n\\ \vdots & & \iddots\ \iddots & \vdots\\ \vdots & \iddots\ \iddots & & \vdots\\ \vdots & \iddots\ \iddots & & \vdots\\ 0 & n & \dots\dots\dots & n\\ n & n & \dots\dots\dots & n\\ 2n & n & \dots\dots\dots & n\\ 2n & n & \dots\dots\dots & n\\ 0 & \dots\dots\dots & 0 & n-2 \end{pmatrix} \tag{3.40}$$

(in the matrix P *all entries in the triangle with vertices in the* $(1,1)$*-entry, the* $(1,n-5)$*-entry, and the* $(n-5,1)$*-entry are equal to* 0*, while all entries in the triangle with vertices in the* $(1,n-4)$*-entry, the* $(n-4,1)$*-entry, and the* $(n-4,\,n-4)$*-entry are equal to* n*). The Poincaré series* $M(t)$ *of the considered classification problem satisfies*

$$\begin{aligned} M(t)&=nt^3\left(\frac{2}{(1-t)^n}+\frac{2}{(1-t)^{n-1}}+\frac{1}{(1-t)^{n-2}}\right)\\ &\quad+n\sum_{i=4}^{n-2}t^i\sum_{j=n-i}^{n-1}\frac{1}{(1-t)^{j+1}}+\frac{(n-2)t^{n-2}}{(1-t)^{n+1}}. \end{aligned} \tag{3.41}$$

It turns out that $(n-2,2)$ is the characteristic pair of the considered classification problem. Indeed, as before let w_0 be the order of zero of $M(t)$ at $t=0$, let N be the order of pole of $(1-t)M(t)$ at $t=1$, and let d be the degree of $M(t)$ (at infinity). Then from (3.25) it follows that $w_0=3$, $N=n$, and $d=n-5$. Hence from (2.25) it follows that

$$\mathrm{NS}\big(M(t)\big) \subset \{(w,l) : w \geq n-3\}. \tag{3.42}$$

Further, using (2.24), one has

$$\begin{aligned} &\mathrm{NS}\big(M(t)\big) \cap \{(w,l) : w = n-3\} \subset \{(n-3,1)\},\\ &\mathrm{NS}\big(M(t)\big) \cap \{(w,l) : w = n-2\} \subset \{(n-2,1),(n-2,2)\}. \end{aligned} \tag{3.43}$$

But by the previous theorem our classification problem admits $(n-2,2)$-parameterization such that its parameterization metric has a positive entry in the upper right corner. Therefore from Corollary 1 it follows that $(n-3,1) \notin \mathrm{NS}\big(M(t)\big)$. Since $(n-2,2) \prec (n-2,1)$, we can conclude from (3.42) and (3.43) that $(n-2,2)$ is the minimal element of $\mathrm{NS}\big(M(t)\big)$. In other words, $(n-2,2)$ is the characteristic pair of our classification problem. Also, it implies that the characteristic matrix $\mathcal{C}$ of the problem is equal to $\mathrm{Norm}(P)$, where P is as in Theorem 2. If $n=5$, then obviously $\mathcal{C}=\mathrm{Norm}(P)=P$. For $n>5$ one can calculate all nontrivial entries of $\mathrm{Norm}(P)$, using identities (2.21). This gives all nontrivial entries of the characteristic matrix $\mathcal{C}$:

(3.44)

$$\begin{aligned} &\mathcal{C}_{n-1,n-4} = n-2,\\ &\mathcal{C}_{n-2,n-4} = n(n-3),\\ &\mathcal{C}_{i,n-4} = n\left(1+\sum_{l=1}^{n-3-i}\tbinom{i+2l-3}{l}+\tbinom{2n-9-i}{n-3-i}+2\tbinom{2n-8-i}{n-2-i}+\tbinom{2n-7-i}{n-7-i}\right), \quad 2\leq i\leq n-3,\\ &\mathcal{C}_{1,n-4} = n,\\ &\mathcal{C}_{1j} = n\left(\tbinom{n-4+j}{j-1}+2\tbinom{n-5+j}{j-1}+2\tbinom{n-6+j}{j-1}+\tbinom{n-7+j}{j-1}-1-\sum_{l=1}^{j-1}\tbinom{n+2l-7-j}{l}\right),\\ &\quad 1\leq j\leq n-5. \end{aligned}$$

Recall that $\mathcal{C}_{ij}$ is the intrinsic number of the functional invariants of $i+1$ variables and the weight $j+2$. In order to obtain a characteristic parameterization from the parameterization by the tuple of the primary invariants, one can implement some series of rearrangement of the primary invariants according to the series of elementary transformations from the matrix P to $\mathrm{Norm}(P)$, as described in the previous section (see, for example, the formula (2.19) and the paragraph that follows it).

Remark 6. Two control systems $\dot{y}=\mathcal{F}(y,v)$ and $\dot{\tilde{y}}=\widetilde{\mathcal{F}}(\tilde{y},\tilde{v})$, with m-dimensional state-space S and one-dimensional control space V, are called *locally in state-input space state-feedback equivalent at the point* $(y_0,v_0)\in S\times V$ if there exists the state-feedback transformation

$$\begin{cases} \tilde{y}=\Phi_1(y),\\ \tilde{v}=\Phi_2(y,v),\\ y_0=\Phi_1(y_0), \quad v_0=\Phi_2(y_0,v_0) \end{cases}$$

such that in a neighborhood of (y_0, v_0) in $S \times V$ the following identity holds:

$$\widetilde{\mathcal{F}}\big(\Phi_1(y), \Phi_2(y,v)\big) = d\Phi_1 \mathcal{F}(y,v).$$

The affine $(1, m+1)$ control system (1.5) will be called *the affine extension* of the control system (1.4). It is not difficult to show that two control systems with scalar input are locally in state-input space state-feedback equivalent at some point $(y_0, v_0) \in S \times V$ if and only if their affine extensions are locally equivalent w.r.t. the state-feedback transformations of type (1.3) at the same point. Note also that the affine extensions of generic m-dimensional control systems with scalar input are generic in the set of all $(1, m+1)$-affine systems. Using this fact and Theorems 1 and 2, one obtains the local in state-input space state-feedback classification of generic germs of nonaffine m-dimensional control systems with scalar inputs, and $m \geq 3$ by the tuples of the primary invariants of their affine extensions (in the C^∞ category for $m = 3$ and the C^ω category for $m \geq 4$). Obviously, the Poincaré series, the characteristic pair, and the characteristic matrix of the local in state-input space state-feedback classification problem for m-dimensional control systems with scalar input are exactly the same as in the case of the state-feedback classification problem of $(1, m+1)$ control-affine systems. Besides, since in our method of normalization of $(1, n)$ control-affine systems with $n \geq 5$ we rectify the vector field f_1, in the case $m \geq 4$ the generic germ at (y_0, v_0) of an m-dimensional real analytic control system with the prescribed tuple (3.38) of the primary invariants of its affine extension has the following normal form w.r.t. the local in state-input space state-feedback equivalence:

$$\dot{\xi}_s = f_s(\xi_1, \dots, \xi_m, \nu), \quad 1 \leq s \leq m,$$

such that $f_s(x_2, x_3, \dots, x_{m+1}, x_1) = a_{s+1}(x_1, \dots, x_{m+1})$, where the tuple $\{a_k\}_{k=1}^{m+1}$ is the solution of the system of partial differential equations (3.34) with boundary conditions, which can be expressed by the primary invariants β_k, ψ_k, and ϕ_{kjl}, using (3.37) (the point (y_0, v_0) corresponds to the origin of the coordinates $(\xi_1, \dots, \xi_m, \nu)$ in the state-input space).

4. Reduction of control-affine systems with two-dimensional input to the scalar input case in dimensions 4 and 5. For $r = 2$ the system (1.1) has the following form:

$$\dot{q} = f_0(q) + u_1 f_1(q) + u_2 f_2(q), \quad q \in M, \quad u_1, u_2 \in \mathbb{R}. \tag{4.1}$$

Our aim is to assign to the system (4.1) in a canonical way an affine subsystem with scalar input.[2] It turns out that in the case $n = 4$ the original system can be recovered from it uniquely up to a feedback transformation, while in the case $n = 5$ such unique recovering is possible after introducing an additional invariant function of n variables (which is natural in view of the estimates for the number of functional parameters given in the introduction).

4.1. Preliminaries. Let us look on (4.1) as on the time optimal control problem and find its extremals. First we introduce some notation. Let T^*M be the cotangent bundle of M with canonical symplectic form σ. Denote by h_i, $0 \leq i \leq 2$, the following functions on T^*M:

$$h_i(\lambda) = p \cdot f_i(q), \quad \lambda = (p, q), \quad q \in M, \quad p \in T_q^* M. \tag{4.2}$$

[2]The meaning of the word "subsystem" is that at any point q the set of its admissible velocities is a subset of the set of the admissible velocities of the original system at q.

For a given function $G : T^*M \mapsto \mathbb{R}$ denote by $\vec{G}$ the corresponding Hamiltonian vector field defined by the relation $\sigma(\vec{G}, \cdot) = -d\,G(\cdot)$. For a given vector distribution D on M (i.e., a subbundle of the tangent bundle), define the lth power D^l by the recursive relation

$$D^l = D^{l-1} + [D, D^{l-1}], \quad D^1 = D,$$

and denote by $(D^l)^\perp \subset T^*M$ the annihilator of D^l, namely,

$$(D^l)^\perp = \{(p, q) \in T^*M : p \cdot v = 0 \ \forall v \in D^l(q)\}.$$

In the introduced notation the Hamiltonian of Pontryagin's maximum principle for the time optimal problem (4.1) can be written as follows:

$$H(\lambda, u_1, u_2) = h_0(\lambda) + u_1 h_1(\lambda) + u_2 h_2(\lambda), \quad \lambda \in T^*M, \quad u_1, u_2 \in \mathbb{R}. \tag{4.3}$$

Let $\gamma(\cdot)$ be an extremal of (4.1) with extremal control functions $\bar{u}_1(t)$ and $\bar{u}_2(t)$. Then

$$\dot{\gamma}(t) = \vec{h}_0\big(\gamma(t)\big) + \bar{u}_1(t)\vec{h}_1\big(\gamma(t)\big) + \bar{u}_2(t)\vec{h}_2\big(\gamma(t)\big), \tag{4.4}$$

and from the maximality condition for H it follows that

$$\gamma(\cdot) \subset \{\lambda \in T^*M : h_1(\lambda) = h_2(\lambda) = 0\}. \tag{4.5}$$

If we denote $D_2 = \operatorname{span}(f_1, f_2)$, then (4.5) is equivalent to $\gamma(\cdot) \subset (D_2)^\perp$. Combining (4.4) and (4.5), we obtain

$$d_{\gamma(t)} h_i\big(\dot{\gamma}(t)\big) = 0, \quad i = 1, 2. \tag{4.6}$$

Then from (4.4) and (4.6) it follows that

$$\begin{aligned} &\{h_0, h_1\}\big(\gamma(t)\big) + \bar{u}_2(t)\{h_2, h_1\}\big(\gamma(t)\big) = 0, \\ &\{h_0, h_2\}\big(\gamma(t)\big) + \bar{u}_1(t)\{h_1, h_2\}\big(\gamma(t)\big) = 0 \end{aligned} \tag{4.7}$$

(here $\{h_i, h_j\}$ are Poisson brackets of the Hamiltonians h_i and h_j: $\{h_i, h_j\} = dh_j(\vec{h}_i)$). Now suppose that

$$\dim D_2^2 = 3. \tag{4.8}$$

Then relations (4.7) imply that the extremals of (4.1), lying in $(D_2)^\perp \backslash (D_2^2)^\perp$, are exactly the integral curves of the vector field

$$\vec{X} = \vec{h}_0 + \frac{\{h_0, h_2\}}{\{h_2, h_1\}}\vec{h}_1 + \frac{\{h_1, h_0\}}{\{h_2, h_1\}}\vec{h}_2 \tag{4.9}$$

(which is the Hamiltonian vector field, corresponding to the Hamiltonian $X = h_0 + \frac{\{h_0,h_2\}}{\{h_2,h_1\}}h_1 + \frac{\{h_1,h_0\}}{\{h_2,h_1\}}h_2$). Denote by V the affine subbundle of TM, defined by system (4.1), and let $V(q)$ be the set of all admissible velocities of the system (4.1) at the point q,

$$V(q) = \{f_0(q) + u_1 f_1(q) + u_2 f_2(q) : u_1, u_2 \in \mathbb{R}\}.$$

Let $\pi : T^*M \mapsto M$ be the canonical projection. The set

$$\operatorname{Ext}(q) = \{\pi_* \vec{X}(\lambda) : \lambda \in T_q^*M \cap (D_2)^\perp \backslash (D_2^2)^\perp\}, \quad q \in M, \tag{4.10}$$

is the subset of $V(q)$, consisting of the velocities of all extremal trajectories starting at q and having a lift in $(D_2)^\perp \backslash (D_2^2)^\perp$.

Among all extremals on $(D_2)^\perp \backslash (D_2^2)^\perp$, one can distinguish so-called abnormal extremals, i.e., the extremals lying on the zero level set of the Hamiltonian X. Denote $D_3 = \operatorname{span}(f_0, f_1, f_2)$ and suppose that

$$\dim(D_2^2 + D_3) = 4. \tag{4.11}$$

The set

$$\operatorname{Abn}(q) = \{\pi_* \vec{X}(\lambda) : \lambda \in T_q^* M \cap (D_3)^\perp \backslash (D_2^2)^\perp\}, \quad q \in M, \tag{4.12}$$

is the subset of $\operatorname{Ext}(q)$, consisting of the velocities of all abnormal extremal trajectories starting at q and having a lift in $(D_2)^\perp \backslash (D_2^2)^\perp$. One can show that for generic germs of affine systems of type (4.1), $\operatorname{Ext}(q) = V(q)$ in the case $n \geq 5$ and $\operatorname{Abn}(q) = V(q)$ in the case $n \geq 6$. But in the cases $n = 4$ and $n = 5$ either $\operatorname{Ext}(q)$ or $\operatorname{Abn}(q)$ (or both of them) defines the proper subsystem of the original system (4.1). Moreover, it turns out that these subsystems are affine with scalar input, so one can apply the theory of the previous section. Now let us consider the cases $n = 4$ and $n = 5$ separately.

4.2. The case $n = 4$. Let

$$[V, D_2](q) = \{[X, Y](q) : X \in V,\ Y \in D_2, \text{ are vector fields}\}.$$

It is not difficult to show that $[V, D_2](q)$ is a linear space and

$$[V, D_2](q) = \operatorname{span}\big(f_1(q), f_2(q), [f_1, f_2](q), [f_0, f_1](q), [f_0, f_2](q)\big). \tag{4.13}$$

The crucial observation is formulated in the following proposition.

PROPOSITION 3. *The set* $\operatorname{Ext}(q)$ *is an affine line, provided that* (4.8) *holds and*

$$\dim [V, D_2](q) = 4. \tag{4.14}$$

Proof. Take some vector field f_3 such that the tuple (f_0, f_1, f_2, f_3) constitutes the frame on M. Denote by c_{ji}^k the structural functions of this frame, i.e., the functions satisfying

$$[f_i, f_j] = \sum_{k=0}^{3} c_{ji}^k f_k. \tag{4.15}$$

Using the well-known property of the Poisson brackets,

$$\{h_i, h_j\}(p, q) = p \cdot [f_i, f_j](q), \quad q \in M, \quad p \in T_q^* M, \tag{4.16}$$

and (4.9), one can easily obtain that

$$\operatorname{Ext}(q) = \Pi(q) \cap V(q), \tag{4.17}$$

where

$$\begin{aligned} \Pi(q) = \big\{ &\big(c_{12}^0(q)\nu + c_{12}^3(q)\mu\big) f_0(q) + \big(c_{20}^0(q)\nu + c_{20}^3(q)\mu\big) f_1(q) \\ &+ \big(c_{01}^0(q)\nu + c_{01}^3(q)\mu\big) f_2(q) : \mu, \nu \in \mathbb{R}\big\}. \end{aligned} \tag{4.18}$$

From assumption (4.14) and identity (4.13) it follows that $\Pi(q)$ is a plane. Assumption (4.8) implies that the plane $\Pi(q)$ is not parallel to the plane $V(q)$. Note also that both $\Pi(q)$ and $V(q)$ belong to $D_3(q)$. Hence by (4.17) it follows that the set $\mathrm{Ext}(q)$ is an affine line. □

Consider the control system such that $\mathrm{Ext}(q)$ is its set of admissible velocities at q. By Proposition 3 it is an affine system with scalar input. We call this system *the reduction of the four-dimensional control-affine system* (4.1). The following proposition gives another characterization of the reduction of the system (4.1).

PROPOSITION 4. *Assume that the four-dimensional control-affine system* (4.1) *satisfies the conditions* (4.8) *and* (4.14). *Then the subsystem*

$$\dot{q} = g_0 + ug_1 \tag{4.19}$$

of (4.1) *is its reduction if and only if*

$$[g_0, g_1] \in D_2. \tag{4.20}$$

Proof. By definition, the system (4.19) is the reduction of (4.1) if and only if

$$\mathrm{Ext}(q) = \{g_0(q) + tg_1(q) : t \in \mathbb{R}\}. \tag{4.21}$$

On the other hand, one can take from the beginning $f_0 = g_0$ and $f_1 = g_1$. Then comparing (4.21) with (4.17) and (4.18), we obtain that the system (4.19) is the reduction of (4.1) if and only if $c_{01}^0 = c_{01}^3 = 0$, which is equivalent to $[g_0, g_1] \in \mathrm{span}(g_1, f_2) = D_2$. □

COROLLARY 2. *Assume that a four-dimensional affine control system* (4.19) *satisfies*

$$\dim \mathrm{span}\big(g_1, [g_1, g_0], \big[g_1, [g_1, g_0]\big], \big[g_0, [g_1, g_0]\big]\big) = 4. \tag{4.22}$$

Then any four-dimensional control-affine system with two-dimensional input, having the system (4.19) *as its reduction, is feedback equivalent to the system*

$$\dot{q} = g_0 + u_1 g_1 + u_2 [g_0, g_1], \quad u_1, u_2 \in \mathbb{R}. \tag{4.23}$$

Proof. First, by assumption (4.22) and relation (4.13), the system (4.23) satisfies conditions (4.8) and (4.14) (where f_0, f_1, and f_2 are replaced by g_0, g_1, and $[g_0, g_1]$). Hence, by Proposition 3 the system (4.23) admits the reduction, and by Proposition 4 this reduction is the system (4.19). On the other hand, suppose that some system (4.1) has the reduction (4.19). Then from the previous proposition $[g_0, g_1] \in \mathrm{span}(g_1, f_2)$. According to (4.22), g_1 and $[g_0, g_1]$ are linearly independent. Hence the system (4.1) is feedback equivalent to (4.23). □

According to the previous proposition, a generic germ of a four-dimensional control-affine system with two-dimensional input can be uniquely—up to a feedback transformation—recovered from its reduction. Suppose that the reduction (4.19) of the system (4.1) satisfies (4.22) and

$$\dim \mathrm{span}\big(g_0, g_1, [g_1, g_0], \big[g_1, [g_1, g_0]\big]\big) = 4. \tag{4.24}$$

Then we can apply to the system (4.19) all constructions of section 2. In particular, one can construct the tuple of the primary invariants of (4.19) at a given point, which are also feedback invariants of the original system (4.1). Note that the set of germs of systems of type (4.1) having the reductions, which satisfies conditions (4.22) and

(4.24), is generic. Combining Theorem 1, Corollary 2, and normal form (3.32), we obtain the following classification of generic germs of systems of type (4.1) in terms of the tuple of the primary invariants of their reductions as follows.

THEOREM 3. *Given two arbitrary germs $\mathcal{I}_1$ and $\mathcal{I}_2$ of functions of four variables at 0, four arbitrary germs $\beta_1, \beta_2, \beta_3, \beta_4$ of functions of three variables at 0, and four arbitrary germs $\psi_1, \psi_2, \psi_3, \psi_4$ of functions of two variables at 0, there exists a unique, up to state-feedback transformation of type* (1.3)*, four-dimensional control-affine system with two-dimensional input such that its reduction satisfies genericity assumptions* (4.22) *and* (4.24) *and $(\mathcal{I}_1, \mathcal{I}_2, \beta_1, \beta_2, \beta_3, \beta_4, \psi_1, \psi_2, \psi_3, \psi_4)$ is the tuple of the primary invariants of the reduction at the given point q_0. This control system is state-feedback equivalent to the following system:*

$$\begin{cases} \dot{x}_1 = 1 + (N + B_1\rho_2)u_1 + \left(\dfrac{\partial N}{\partial x_1} + \beta_1 \dfrac{\partial \rho_2}{\partial x_1}\right) u_2, \\ \dot{x}_2 = (\rho_1 + B_2\rho_2)u_1 + \left(\dfrac{\partial \rho_1}{\partial x_1} + B_2 \dfrac{\partial \rho_2}{\partial x_1}\right) u_2, \qquad u_1, u_2 \in \mathbb{R}, \\ \dot{x}_i = B_i\rho_2 u_1 + \beta_i \dfrac{\partial \rho_2}{\partial x_1} u_2, \quad i = 3, 4, \end{cases} \tag{4.25}$$

where N is the solution of (3.29)*, ρ_i, $i = 1, 2$, are the solutions of* (3.30)*, and B_k, $1 \le k \le 4$, are as in* (3.31)*. The Poincaré series, the characteristic pair, and the characteristic matrix of the classification problem are exactly the same as in the case of* $(1, 4)$ *control-affine systems.*

Remark 7. It is easy to show that in the case $n = 4$ the set $\mathrm{Abn}(q)$ consists of one vector provided that (4.11) holds. Besides, if the system (4.19) is the reduction of the system (4.1) and it satisfies (4.22) and (4.24), then $\mathrm{Abn}(q)$ is exactly its canonical drift. □

Remark 8. Actually, there is another intrinsic way to assign to the system (4.1), satisfying (4.11), an affine subsystem with scalar input: As a drift one can take again $\mathrm{Abn}(q)$. It remains to define canonically the direction of the affine line of the reduction. For this note first that the distribution D_2 satisfies (4.8) because of assumption (4.11). Therefore through any point of M the unique (unparameterized) abnormal extremal trajectory of the rank 2 distribution D_2 passes: The line subdistribution L of D_2, tangent to the abnormal extremal trajectories at any point, is characterized by the relation $[L, D^2] \subseteq D^2$. The direction of the affine line of the reduction can be taken parallel to L. The direction of L is different in general from the direction of the affine line in the first reduction. But this new reduction is worse than the previous one, because the original system (4.1) is not uniquely recovered from it: If $(\bar{g}_0, \bar{g}_1)$ is the canonical pair of the new reduction (by construction and the previous remark $g_0 = \mathrm{Abn}$), then the field f_2 can be taken in the form $f_2 = \alpha\bar{g}_0 + [\bar{g}_1, \bar{g}_0]$, where the function α satisfies some second order ordinary differential equation along each integral curve of $\bar{g}_1$. Note that the direction $\bar{g}_1$ depends on the second jet of the original system (4.1), while the direction of the affine line in the first reduction depends only on the first jet. This could be the reason for the loss of some information about the original system during the reduction described in the present remark. □

4.3. The case $n = 5$. In this case by analogy with Proposition 3 we have the following proposition.

PROPOSITION 5. *The set* $\mathrm{Abn}(q)$ *is an affine line, provided that* (4.11) *holds and*

$$\dim D_3^2 = 5. \tag{4.26}$$

Proof. Take some vector fields f_3 and f_4 such that the tuple $(f_0, f_1, f_2, f_3, f_4)$ constitutes the frame on M. By analogy with (4.15), let c_{ji}^k, $0 \le i, j, k \le 4$, be the structural functions of this frame. From (4.9) and (4.12), using (4.16) and the fact that $h_0 = 0$ on $(D_3)^\perp$, one can easily obtain that

$$\mathrm{Abn}(q) = \Pi_1(q) \cap V(q), \tag{4.27}$$

where

$$\begin{aligned}\Pi_1(q) = \big\{ &\big(c_{12}^3(q)\nu + c_{12}^4(q)\mu\big) f_0(q) + \big(c_{20}^3(q)\nu + c_{20}^4(q)\mu\big) f_1(q) \\ &+ \big(c_{01}^3(q)\nu + c_{01}^4(q)\mu\big) f_2(q) : \mu, \nu \in \mathbb{R}\big\}.\end{aligned} \tag{4.28}$$

From assumption (4.26) and identity (4.13) it follows that $\Pi_1(q)$ is a plane. Assumption (4.11) implies that the plane $\Pi_1(q)$ is not parallel to the plane $V(q)$. Note also that both $\Pi_1(q)$ and $V(q)$ belong to $D_3(q)$. Hence by (4.27) the set $\mathrm{Abn}(q)$ is an affine line. □

Consider the control system such that $\mathrm{Abn}(q)$ is its set of admissible velocities at q. By Proposition 5 it is an affine system with scalar input. We call this system *the reduction of the five-dimensional control-affine system* (4.1). The following proposition gives another characterization of the reduction of the system (4.1).

PROPOSITION 6. *Assume that the five-dimensional control-affine system* (4.1) *satisfies the conditions* (4.26) *and* (4.11). *Then the subsystem*

$$\dot q = g_0 + u g_1 \tag{4.29}$$

of (4.1) *is its reduction if and only if*

$$[g_0, g_1] \in D_3. \tag{4.30}$$

Proof. By definition, the system (4.19) is the reduction of (4.1) if and only if

$$\mathrm{Abn}(q) = \{g_0(q) + t g_1(q) : t \in \mathbb{R}\}. \tag{4.31}$$

On the other hand, one can take from the beginning $f_0 = g_0$ and $f_1 = g_1$. Then comparing (4.31) with (4.27) and (4.28), we obtain that the system (4.29) is the reduction of (4.1) if and only if $c_{01}^3 = c_{01}^4 = 0$, which is equivalent to $[g_0, g_1] \in \mathrm{span}(g_0, g_1, f_2) = D_3$. □

COROLLARY 3. *Assume that a five-dimensional affine control system* (4.19) *satisfies*

$$\dim \mathrm{span}\big(g_0, g_1, [g_1, g_0], \big[g_1, [g_1, g_0]\big], \big[g_0, [g_1, g_0]\big]\big) = 5. \tag{4.32}$$

Then a five-dimensional control-affine system with two-dimensional input has the system (4.19) *as its reduction if and only if it is feedback equivalent to the system*

$$\dot q = g_0 + u_1 g_1 + u_2(\alpha g_0 + [g_0, g_1]), \quad u_1, u_2 \in \mathbb{R}, \tag{4.33}$$

where α is some function.

Proof. First by assumption (4.32) the system (4.33) satisfies conditions (4.26) and (4.11). Hence, by Proposition 5 the system (4.33) admits the reduction, and by Proposition 6 this reduction is the system (4.19). On the other hand, suppose that some system (4.1) has the reduction (4.19). Then from the previous proposition

$[g_0, g_1] \in \operatorname{span}(g_0, g_1, f_2)$. According to (4.22), g_0, g_1, and $[g_0, g_1]$ are linearly independent. Hence, $f_2 = \xi_0 g_0 + \xi_1 g_1 + \xi_2 [g_0, g_1]$, where $\xi_2 \neq 0$. Hence by a feedback transformation we can replace f_2 with $\alpha g_0 + [g_0, g_1]$. □

Thus, in contrast to the case $n = 4$, five-dimensional control-affine systems with two-dimensional input cannot be recovered from their reduction only. Suppose that the system (4.1) has the reduction (4.29) satisfying condition (4.32) and also

$$\dim \operatorname{span}\big(g_0, g_1, [g_1, g_0], \big[g_1, [g_1, g_0]\big], \big[g_1, \big[g_1, [g_1, g_0]\big]\big]\big) = 5, \tag{4.34}$$

$$\dim \operatorname{span}\big(g_1, \big[g_0, [g_0, g_1]\big], [g_1, g_0], \big[g_1, [g_1, g_0]\big], \big[g_1, \big[g_1, [g_1, g_0]\big]\big]\big) = 5. \tag{4.35}$$

Then we can apply to the system (4.29) all constructions of section 2. In particular, let (G_0, G_1) be the canonical pair of the system (4.29). As before, let V be the affine subbundle of TM, defined by system (4.1). Then by the same arguments, as in the proof of Corollary 3, there exists a unique vector field $G_2 \in V$ such that

$$G_2 = RG_0 + [G_0, G_1]. \tag{4.36}$$

By construction, the function R is a feedback invariant of system (4.1). Moreover, the system (4.1) can be uniquely—up to a feedback transformation—recovered from its reduction and the function R.

Fix some point q_0 in M. Let Φ_5 be as in (3.17). Denote

$$\mathcal{R} = R \circ \Phi_5. \tag{4.37}$$

Note that $\mathcal{R}$ is the germ of a function of five variables at 0. We call it *the recovering invariant of* (4.1) *at the point* q_0. By construction, it is an invariant of the weight 0. Note that the set of germs of systems of type (4.1) having the reductions, which satisfies conditions (4.32), (4.34), and (4.35), is generic. Using Theorem 2 in the case $n = 5$ and the definition of the recovering invariant, we obtain the following classification of generic real analytic germs of systems of type (4.1) in terms of the tuple of the primary invariants of their reductions and their recovering invariant.

THEOREM 4. *Given four germs $\mathcal{I}_1$, $\mathcal{I}_2$, $\mathcal{I}_3$, and $\mathcal{R}$ of real analytic functions of five variables at 0 such that $\mathcal{I}_3(0) \neq 0$, 10 germs $\{\beta_k\}_{k=1}^5$ and $\{\phi_{k22}\}_{k=1}^5$ of real analytic functions of four variables at 0, 10 germs $\{\psi_k\}_{k=1}^5$ and $\{\phi_{k23}\}_{k=1}^5$ of real analytic functions of three variables at 0, and five germs $\{\phi_{k24}\}_{k=1}^5$ of real analytic functions of two variables at 0, there exists a unique, up to state-feedback real analytic transformation of type* (1.3)*, five-dimensional real analytic control-affine system with two inputs such that first, its reduction satisfies genericity assumptions* (4.32), (4.34), (4.35)*, second,* $\big(\{\mathcal{I}_j\}_{j=1}^3, \{\beta_k\}_{k=1}^5, \{\psi_k\}_{k=1}^5, \{\phi_{k2l} : 1 \leq k \leq 5,\ 2 \leq l \leq 4\}\big)$ *is its tuple of the primary invariants, and finally, $\mathcal{R}$ is its recovering invariant at the given point q_0. In other words, the tuples of the primary invariants give the regular* $(3, 2)$*-parameterization of the considered classification problem in the real analytic category with the following* (4×4)*-parameterization matrix P:*

$$P = \begin{pmatrix} 0 & 0 & 0 & 5 \\ 0 & 0 & 0 & 10 \\ 0 & 0 & 0 & 10 \\ 1 & 0 & 0 & 3 \end{pmatrix}. \tag{4.38}$$

The Poincaré series $M(t)$ of the considered classification problem satisfies

$$M(t) = \frac{1}{(1-t)^6} + t^3 \left(\frac{5}{(1-t)^3} + \frac{10}{(1-t)^4} + \frac{10}{(1-t)^5} + \frac{3}{(1-t)^6} \right). \tag{4.39}$$

By the same arguments, as in the case of the $(1,n)$-affine control system with $n \geq 5$, treated in section 3, one can show that $(3,2)$ is the characteristic pair of the considered classification problem (just use formulas (2.24), (2.25), and Corollary 1). The characteristic matrix $\mathcal{C}$ of the problem is equal to $\mathrm{Norm}(P)$, which can be found easily by the series of elementary transformations. Namely,

$$\mathcal{C} = \mathrm{Norm}(P) = \begin{pmatrix} 1 & 3 & 6 & 5 \\ 0 & 0 & 0 & 16 \\ 0 & 0 & 0 & 13 \\ 0 & 0 & 0 & 4 \end{pmatrix}. \tag{4.40}$$

Conclusion 2. *The characteristic parameterization of* $(2,5)$ *control-affine systems, up to state-feedback transformations, consists of four functional invariants of five variables and the weight* 3*,* 13 *functional invariants of four variables and the weight* 3*,* 16 *functional invariants of three variables and the weight* 3*, one functional invariant of two variables and the weight* 0*, three functional invariants of two variables and the weight* 1*, six functional invariants of two variables and the weight* 2*, and five functional invariants of two variables and the weight* 3*.*

In order to obtain a characteristic parameterization from the parameterization by the tuple of the primary invariants and the recovering invariant, one can implement some series of rearrangement of the primary invariants according to the series of elementary transformations from the matrix P to $\mathrm{Norm}(P)$, as described in the previous section (see, for example, formula (2.19) and the paragraph that follows it).

REFERENCES

[1] A. A. Agrachev and R. V. Gamkrelidze, *Feedback-invariant optimal control theory.* I. *Regular extremals*, J. Dynam. Control Systems, 3 (1997), pp. 343–389.

[2] A. A. Agrachev, *Feedback-invariant optimal control theory.* II. *Jacobi curves for singular extremals*, J. Dynam. Control Systems, 4 (1998), pp. 583–604.

[3] A. Agrachev and I. Zelenko, *Geometry of Jacobi curves.* I, J. Dynam. Control Systems, 8 (2002), pp. 93–140.

[4] A. Agrachev and I. Zelenko, *Geometry of Jacobi curves.* II, J. Dynam. Control Systems, 8 (2002), pp. 167–215.

[5] V. I. Arnold, *Mathematical problems in classical physics*, in Trends and Perspectives in Applied Mathematics, Appl. Math. Sci. 100, F. John, J. E. Marsden, and L. Sirovich, eds., Springer, New York, 1994, pp. 1–20; in Selected Works of V. I. Arnold, Phasis, Moscow, 1997, pp. 553–575 (in Russian).

[6] B. Bonnard, *Feedback equivalence for nonlinear systems and the time optimal control problem*, SIAM J. Control Optim., 29 (1991), pp. 1300–1321.

[7] R. B. Gardner, *The geometry of nonlinear control systems*, in Differential Geometry: A Symposium in Honor of Manfredo do Carmo, Pitman Monogr. Surveys Pure Appl. Math. 52, B. Lawson and K. Tenenblatt, eds., Longman Scientific and Technical, Harlow, UK, 1991, pp. 179–198.

[8] B. Jakubczyk, *Equivalence and invariants of nonlinear control systems*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel Dekker, New York, Basel, 1990, pp. 177–218.

[9] B. Jakubczyk, *Critical Hamiltonians and feedback invariants*, in Geometry of Feedback and Optimal Control, B. Jakubczyk and W. Respondek, eds., Marcel Dekker, New York, Basel, 1998, pp. 219–256.

[10] B. Jakubczyk, *Feedback invariants, critical trajectories, and Hamiltonian formalism*, in Nonlinear Control in the Year 2000, Vol. 1, Lecture Notes in Control and Inform. Sci. 258, A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, eds., Springer, London, 2001, pp. 219–256.

[11] B. Jakubczyk and W. Respondek, *Feedback classification of analytic control systems in the plane*, in Analysis of Controlled Dynamical Systems, B. Bonnard et al., eds., Birkhäuser Boston, Boston, 1991, pp. 262–273.

[12] W. Kang, *Extended controller form and invariants of nonlinear control systems with a single input*, J. Math. Systems Estim. Control, 6 (1996), pp. 27–51.

[13] W. Kang and A. J. Krener, *Extended quadratic controller normal form and dynamic state feedback linearization of nonlinear systems*, SIAM J. Control Optim., 30 (1992), pp. 1319–1337.

[14] I. Kupka, *On feedback equivalence*, in Differential Geometry, Global Analysis, and Topology, CMS Conf. Proc. 12, AMS, Providence, RI, 1991, pp. 105–117.

[15] W. Respondek, *Feedback classification of nonlinear control systems in* $\mathbb{R}^2$ *and* $\mathbb{R}^3$, in Geometry of Feedback and Optimal Control, B. Jakubczyk and W. Respondek, eds., Marcel Dekker, New York, 1998, pp. 347–382.

[16] U. Serres, *On the curvature of two-dimensional optimal control systems and Zermelo's navigation problem*, J. Math. Sci., 135 (2006), pp. 3224–3243.

[17] I. A. Tall and W. Respondek, *Feedback classification of nonlinear single-input control systems with controllable linearization: Normal forms, canonical forms, and invariants*, SIAM J. Control Optim., 41 (2003), pp. 1498–1531.

[18] G. R. Wilkens, *Centro-affine geometry in the plane and feedback invariants of two-state scalar control systems*, in Differential Geometry and Control (Boulder, CO, 1997), Proc. Sympos. Pure Math. 64, AMS, Providence, RI, 1999, pp. 319–333.

[19] M. Zhitomirskii, *Typical Singularities of Differential* 1*-Forms and Pfaffian Equations*, Transl. Math. Monogr. 113, AMS, Providence, RI, 1992.

[20] M. Zhitomirskii and W. Respondek, *Simple germs of corank one affine distributions*, in Singularities Symposium-Lojasiewicz 70, Banach Center Publ. 44, B. Jakubczyk, W. Pawlucki, and J. Stasica, eds., Polish Acad. Sci., Warsaw, 1998, pp. 269–276.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1461–1482

OPTIMAL TERMINAL WEALTH UNDER PARTIAL INFORMATION: BOTH THE DRIFT AND THE VOLATILITY DRIVEN BY A DISCRETE-TIME MARKOV CHAIN*

MICHAEL TAKSAR† AND XUDONG ZENG†

Abstract. We consider a multistock market model. The stock price process satisfies a stochastic differential equation where both the drift and the volatility are driven by a discrete-time Markov chain of finite states. Not only the underlying Brownian motion but also the Markov chain in the stochastic differential equation are assumed to be unobservable. Investors can observe the stock price process only. The main result of this paper is that we derive the approximation of the optimal trading strategy and the corresponding optimal expected utility function from the terminal wealth for the CRRA utility function.

Key words. partial information, optimal terminal wealth, CRRA utility, dynamic programming

AMS subject classifications. 93E10, 49L20, 91B28

DOI. 10.1137/050639351

1. Introduction. In this paper, we consider a model of an incomplete market in which stocks are driven by an m-dimensional geometric Brownian motion the same as in the Black–Scholes model. However, the drift and the diffusion coefficients of this process depend on a discrete-time Markov chain whose states are not observable. Moreover, we do not observe the driving diffusion process—rather only the values of the price process.

There are quite a few papers devoted to studies of the problem of maximizing the expected utility function from the terminal wealth under partial information. Pham and Quenez [15] considered a stochastic volatility model. They solved the portfolio optimization problem under partial information by stochastic filtering techniques and adapting martingale duality methods. For more literature on partial information and stochastic volatility problems, we refer the reader to Lakner [12], [13], Frey [8], Runggaldier [16], Frey and Runggaldier [9], Cvitanić, Lipster, and Rozovskii [1], Elliott and Rishel [6], Elliott [7], and Henderson and Hobson [10].

Sass and Haussman [18] considered a multistock market model in continuous–time. The drift is a continuous–time, finite state Markov chain, and the volatility matrix is constant and nonsingular. They used Malliavin calculus and hidden Markov chain theory to derive an explicit expression for the optimal portfolio selection. However, their method cannot be extended to the case in which the volatility is driven by a Markov chain, because the EM algorithm they used to estimate the drift does not work for the volatility due to the fact that the measures involved in their method are not equivalent if the volatility is driven by a Markov chain.

In this paper, we consider a discrete-time multistock market model where both the drift and the volatility are driven by a Markov chain. In our paper, we use the method of estimating volatility studied by Elliott [4], Elliott, Hunter, and Jamieson

*Received by the editors August 31, 2005; accepted for publication (in revised form) March 13, 2007; published electronically September 28, 2007. This research was supported by National Science Foundation grant NSF DMS 0505435.

http://www.siam.org/journals/sicon/46-4/63935.html.

†Department of Mathematics, University of Missouri, Columbia, MO 65203 (taksar@math.missouri.edu, zeng@math.missouri.edu).

[5], developed for the discrete-time setting. The algorithm enables us to estimate the states of the Markov chain and its transition matrix. We solve the problem of optimizing the expected utility from the terminal wealth problem, and using dynamic programming we construct the optimal strategy in terms of the filter of the price process.

The rest of this paper is structured as follows. In section 2, we present the discrete-time model from the general continuous model and describe the relationship between the two models. We prove some preliminary results, which will be used in subsequent calculation. In section 3, we introduce some definitions and state the optimization problem. Section 4 is devoted to an important lemma related to an approximation of the optimal trading strategy for the CRRA utility function. In section 5 we show the results of simulation to illustrate our results.

2. Models and preliminary results.

2.1. Regime switching model: Continuous time. Consider the m-dimensional stock price process whose dynamics is given by a geometric Brownian motion equation:

$$dS(t) = \mathrm{diag}(S_t)(\mu(Y(t))dt + \hat{\sigma}(Y(t))dW(t)), \quad 0 \le t \le T. \tag{2.1}$$

Here $S_t = (S_t^{(1)}, S_t^{(2)}, \ldots, S_t^{(m)})'$, and the column vector W_t is an m-dimensional standard Brownian motion. $Y(t)$ is a finite state, homogeneous Markov chain with a generator $Q = (q_{ij})_{d\times d}$, independent of $W(t)$. The distribution of $Y(0)$ is known. $Y(t)$ has a state space $\mathcal{M} = \{e_1, \ldots, e_d\}$, where $e_i, i = 1, 2, \ldots, d$, is the unit vector in $\mathcal{R}^d$.

$$Y(t) \in \mathcal{M} := \{e_1, \ldots, e_d\}.$$

There are different values for the drift and different matrices for the volatility corresponding to states of the Markov chain $Y(t)$. Thus $\mu(\cdot)$ (resp., $\sigma(\cdot)$) is a mapping of $\mathcal{M}$ (resp., $\mathcal{N} := \{B := (b_{i,j})_{1\le i,j\le m}$ is invertible $|b_{i,j} \in \mathcal{R}^+\}$) into R^m (resp., into $R^{m\times m}$).

Suppose r is a constant interest rate. Then (2.1) may be written as follows.

$$d\ \log(e^{-rt}S(t)) = (\mu(Y(t)) - r1_m - \mathrm{diag}(\hat{\sigma}_{n-1}\hat{\sigma}'_{n-1}))dt + \hat{\sigma}(Y(t))dW(t), \tag{2.2}$$

where $0 \le t \le T, 1_m = (1, 1, \ldots, 1)' \in \mathcal{R}^{m\times 1}$, where we use the following convention:

$$\log((x_1, x_2, \ldots, x_n)') = (\log(x_1), \ldots, \log(x_n))'.$$

2.2. Regime switching model: Discrete time. In this paper, we will consider a discrete approximation to the continuous-time model (2.1).

Let $\Delta t = \frac{T}{N}$, $Y_n = Y(n\Delta t)$,

$$\mu_n = \mu(Y(n\Delta t)), \hat{\sigma}_n = \hat{\sigma}(Y(n\Delta t)), S_n = S(n\Delta t),$$

where $n = 0, 1, \ldots, N$. Then (2.2) becomes

$$\begin{aligned} &\log(S_n e^{-r\Delta t}) - \log(S_{n-1}) \\ &= (\mu_{n-1} - r1_m - \mathrm{diag}(\hat{\sigma}_{n-1}\hat{\sigma}'_{n-1})/2)\Delta t + \hat{\sigma}_{n-1}(W_n - W_{n-1}) \\ &:= y_n \end{aligned} \tag{2.3}$$

for $n = 1, 2, 3, \ldots, N$.

Let

$$\begin{cases} g_n : & = (\mu_n - r1_m - \operatorname{diag}(\hat{\sigma}_n \hat{\sigma}_n')/2)\Delta t \\ & = (\mu(Y_n) - r - \operatorname{diag}(\hat{\sigma}(Y_n)\hat{\sigma}(Y_n)')/2)\Delta t, \\ \sigma_n : & = \hat{\sigma}_n \sqrt{\Delta t} = \hat{\sigma}(Y_n)\sqrt{\Delta t}. \end{cases} \tag{2.4}$$

Then

$$y_n = g_{n-1} + \sigma_{n-1} Z_n, \quad n = 1, 2, \ldots, N, \tag{2.5}$$

where $Z_n = (W_n - W_{n-1})/\sqrt{\Delta t}, n = 1, 2, \ldots, N$, is a sequence of standard normal independently and identically distributed random variables.

Note that g_n and σ_n are functions of Y_n and can be written as $G(Y_n)$ and $H(Y_n)$, respectively, which obviously satisfy

$$\begin{array}{l} \frac{G(e_i)}{\Delta t} = \mu(e_i) - r1_m - \operatorname{diag}(\hat{\sigma}(e_i)\hat{\sigma}(e_i)')/2, \\ \frac{H(e_i)}{\sqrt{\Delta t}} = \hat{\sigma}(e_i). \end{array} \tag{2.6}$$

In this paper, we assume only the price of stock S_n or y_n can be observed. Denote the filtration generated by S_n by $\{\mathcal{F}_n\}$. We will study an optimization of the utility function of the terminal wealth in the discrete-time model (2.5).

2.3. Preliminary results. We present some preliminary results which will be used in the proofs in what follows.

By the definition of y_n (2.3), we know that for each $i = 1, 2, \ldots, m$,

$$S_n^{(i)} = S_{n-1}^{(i)} e^{y_n^{(i)}} e^{r\Delta t}.$$

For $k = 1, 2, \ldots, d$, denote

$$b_k := G(e_k) \in \mathcal{R}^{m\times 1}, f_k := H(e_k) \in \mathcal{R}^{m\times m}.$$

Let $b_k(i)$ stand for the ith component of b_k, and let $f_k(i)$ stands for the ith row of f_k. Then we have the following lemma.

LEMMA 2.1.

$$Pr(y_n^{(i)} \le t|\mathcal{F}_{n-1}) = \sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1}) \int_{-\infty}^{t-b_k(i)} \phi_{ik}(x)dx,$$

where $\phi_{ik}(x) = \frac{1}{\sqrt{2\pi(f_k(i)\cdot f_k(i)')}} e^{-\frac{x^2}{2f_k(i)\cdot f_k(i)'}}$, $i = 1, 2, \ldots, m$.

Proof.

$$\begin{aligned} Pr(y_n^{(i)} \le t|\mathcal{F}_{n-1}) &= Pr(g_{n-1}^{(i)} + \sigma_{n-1}^{(i)} Z_n \le t|\mathcal{F}_{n-1}) \\ &= \sum_{k=1}^{d} Pr(b_k(i) + f_k(i)Z_n \le t, Y_{n-1} = e_k|\mathcal{F}_{n-1}) \\ &= \sum_{k=1}^{d} Pr(b_k(i) + f_k(i)Z_n \le t) Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1}) \\ &= \sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1}) \int_{-\infty}^{t-b_k(i)} \phi_{ik}(x)dx. \end{aligned}$$

□

Remark 1. Similarly, for the multidimensional case, $x \in R^{m\times 1}$, we have

$$\begin{aligned} &Pr(y_n \le x|\mathcal{F}_{n-1}) \\ &= \textstyle\sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1}) \int_{-\infty}^{x_1-b_k(1)} \dots \int_{-\infty}^{x_m-b_k(m)} \phi_k(z)dz, \end{aligned} \tag{2.7}$$

where $\phi_k(x) = (2\pi|f_k f_k'|)^{\frac{-1}{2}} e^{-x'(f_k f_k')^{-1}x/2}, x \in R^{m\times 1}$.

LEMMA 2.2. *We have a recursive filter:*

$$Pr(Y_n = e_k|\mathcal{F}_n) = \frac{\sum_{i=1}^{d} Pr(Y_{n-1} = e_i|\mathcal{F}_{n-1})\phi_i(y_n - b_i)p_{ki}}{\sum_{i=1}^{d} Pr(Y_{n-1} = e_i|\mathcal{F}_{n-1})\phi_i(y_n - b_i)}, \tag{2.8}$$

where p_{ki} is the (k,i) entry of the transition matrix P.

Proof. This is Theorem 3.1 of Elliott [4]. □

In what follows, we will use the notation $E_n[\zeta] := E[\zeta|\mathcal{F}_n]$.

LEMMA 2.3. *$\alpha < 1$. For $i = 1, 2, \dots m$, we have*

(i) $|E_{n-1}[e^{y_n^{(i)}} - 1]| = O(\Delta t)$, (ii) $|E_{n-1}[(e^{y_n^{(i)}} - 1)^2]| = O(\Delta t)$,
(iii) $|E_{n-1}[(e^{y_n^{(i)}} - 1)^3]| = O(\Delta t)^2$, (iv) $|E_{n-1}[(1 - e^{-y_n^{(i)}})^3]| = O(\Delta t)^2$,
(v) $|E_{n-1}[e^{\alpha y_n^{(i)}}(1 - e^{-y_n^{(i)}})^3]| = O(\Delta t)^2$.

Proof. The proof of this lemma is in Appendix A. □

Remark 2. Using Lemmas 2.1 and 2.3, we have for $m = 1$ (i.e., when there is only one stock in the model)

$$\begin{aligned} &\tfrac{E_{n-1}[\Delta S_n]}{E_{n-1}[(\Delta S_n)^2]} \\ &= \frac{\sum_{k=1}^{d} Pr(Y_{n-1}=e_k|\mathcal{F}_{n-1})e^{(f_k^2/2+b_k)}-1}{e^{r\Delta t}S_{n-1}(\sum_{k=1}^{d} Pr(Y_{n-1}=e_k|\mathcal{F}_{n-1})e^{(2f_k^2+2b_k)}-2\sum_{k=1}^{d} Pr(Y_{n-1}=e_k|\mathcal{F}_{n-1})e^{(f_k^2/2+b_k)}+1)} \\ &= \frac{\sum_{k=1}^{d} Pr(Y_{n-1}=e_k|\mathcal{F}_{n-1})b_k}{e^{r\Delta t}S_{n-1}\sum_{k=1}^{d} Pr(Y_{n-1}=e_k|\mathcal{F}_{n-1})f_k^2} + O(\Delta t). \end{aligned} \tag{2.9}$$

When $d = 1$, denote the unique state of μ_n by μ and the unique state of $\hat{\sigma}_n$ by σ. Then (2.9) is reduced to

$$\begin{aligned} \tfrac{(E_{n-1}[\Delta S_n])}{E_{n-1}[(\Delta S_n)^2]} &= \tfrac{1}{e^{r\Delta t}S_{n-1}} \tfrac{e^{\sigma^2/2+g}-1}{e^{2\sigma^2+2g}-2e^{\sigma^2/2+g}+1} \\ &= \tfrac{1}{e^{r\Delta t}S_{n-1}} \tfrac{\mu-r}{\sigma^2} + O(\Delta t). \end{aligned} \tag{2.10}$$

3. Definition and optimization problem.

3.1. Wealth process and admissible strategies. In this section we describe a discrete-time optimization model which approximates the original continuous-time model presented at the beginning of this paper.

DEFINITION 3.1. *$h_{n-1} \in \mathcal{F}_{n-1}, n = 1, \dots, N$, are column vectors $\in R^{m\times 1}$. A wealth process $\{X_n^{h_{n-1}}\}_{n=1,2,\dots,N}, X_0 = x_0$ is defined as*

$$X_n^{h_{n-1}} = X_{n-1}^{h_{n-2}} e^{r\Delta t}\left(1 - \sum_{i=1}^{m} h_{n-1}^{(i)}\right) + X_{n-1}^{h_{n-2}} \sum_{i=1}^{m} h_{n-1}^{(i)} \frac{S_n^{(i)}}{S_{n-1}^{(i)}} + o(\sqrt{\Delta t}),$$

where $h_{n-1}^{(i)}$ or $S_{n-1}^{(i)}$ denotes the ith component of the vector h_{n-1} or S_{n-1}.

Using the notation defined in section 1, the wealth process has a simpler expression:

$$X_n^{h_{n-1}} = X_{n-1}^{h_{n-2}} e^{r\Delta t}(1 + h_{n-1} \cdot (e^{y_n} - 1)),$$

where "·" stands for the inner product in R^m. Generally, we have

$$X_n^{h_{n-1}} = X_0 e^{rn\Delta t}(1 + h_0 \cdot (e^{y_1} - 1)) \dots (1 + h_{n-1} \cdot (e^{y_n} - 1)), \quad n = 1, \dots, N.$$

DEFINITION 3.2. *A vector sequence* $h = \{h_i\}_{i=0}^{N-1}$ *is an* admissible *strategy if* $Pr(X_n^{h_{n-1}} > 0$ *for all* $n = 1, 2, \dots, N) = 1$.

We use $\mathcal{H}$ to denote the set of all admissible strategies.

Remark 3. If h is admissible, then

$$Pr(X_n^{h_{n-1}} > 0 | X_{n-1}^{h_{n-2}} > 0) = \frac{Pr(X_n^{h_{n-1}} > 0, X_{n-1}^{h_{n-2}} > 0)}{Pr(X_{n-1}^{h_{n-2}} > 0)} = 1.$$

Then we have

$$Pr(X_{n-1}^{h_{n-2}} e^{r\Delta t}(1 + h_{n-1} \cdot (e^{y_n} - 1)) > 0 | X_{n-1}^{h_{n-2}} > 0) = 1.$$

So if h is admissible, then

$$Pr(1 + h_{n-1} \cdot (e^{y_n} - 1) > 0) = 1. \tag{3.1}$$

By (2.5), we have $y_n^{(i)} \in (-\infty, \infty), e^{y_n^{(i)}} - 1 \in (-1, \infty)$. Hence the equality (3.1) implies

$$\|h_{n-1}\|_1 \le 1, \quad 1 \ge h_{n-1}^{(i)} \ge 0, \tag{3.2}$$

for each $i \in \{1, 2, \dots, m\}$.

The inequalities above imply no stock shorting as well as no money borrowing in our model. Rogers also mentioned such a restriction on the portfolio in his h-investor model (see [17]).

3.2. HARA utility functions.

DEFINITION 3.3. *A function* $u : (x_u, \infty) \to \mathcal{R}, x_u \in \mathcal{R}$, *is called a* utility function *if* u *is strictly increasing, strictly concave, and twice continuously differentiable on* (x_u, ∞) *and satisfies* $\lim_{x\to\infty} u'(x) = 0$ *and* $\lim_{x\to x_u^+} u'(x) = \infty$.

Next we define the coefficient of absolute risk aversion,

$$R_a(x) = -\frac{u''(x)}{u'(x)}.$$

DEFINITION 3.4. *If* $R_a^{-1}(x)$ *is a linear function, i.e.,* $R_a^{-1}(x) = a + bx$, *then we say that* $u(x)$ *is of the hyperbolic absolute risk aversion (HARA) class.*

It should be mentioned that the solution to the differential equation

$$\frac{1}{a + bx} = -\frac{u''(x)}{u'(x)}$$

is a power, logarithmic, or exponential function, i.e.,

$$u(x) = \begin{cases} \frac{c}{b-1}(a + bx)^{1-\frac{1}{b}} + d, x > \frac{-a}{b} & \text{if } b > 0, b \ne 1, \\ c\ \ln(a + x) + e,\ x > -a & \text{if } b = 1, \\ \frac{-f}{a} e^{-x/a} + g, x \in \mathcal{R} & \text{if } b = 0, \end{cases}$$

where c, d, e, f, g are some constants.

Hence, a HARA utility function is thirdly continuously differentiable and

$$\frac{u'''(x)}{u''(x)} = -\frac{1+b}{a+bx}.$$

The most popular HARA utility functions are

- constant relative risk aversion (CRRA): $u(x) = x^\beta/\beta$, where $\beta < 1$, and
- constant absolute risk aversion (CARA): $u(x) = -e^{-\beta x}/\beta$, where $\beta > 0$.

3.3. Optimization problem. Let $u(x)$ be a utility function, and $\{X_n^h\}$ be the wealth process. The objective is to calculate

$$(*) \qquad V^* = \sup_{h\in\mathcal{H}}\{E[u(X_N^h)]\}$$

and find an admissible trading strategy h^* s.t. $E[u(X_N^{h^*})] = V^*$.

4. Dynamic programming. Define

$$(4.1) \qquad U_n(x) = \sup_{h\in\mathcal{H}} E[u(X_N^h)|\mathcal{F}_n],\ n = 0, 1, \ldots, N.$$

From this definition, $U_0 = V^*$, as defined in the optimization problem of section 3.3. The dynamic programming equation for the sequence $u_0, u_1, \ldots, u_N$ is

$$(4.2) \qquad \begin{cases} U_n(x) = \sup_{h_n} E[U_{n+1}(xe^{r\Delta t} + xe^{r\Delta t}h_n \cdot (e^{y_{n+1}} - 1))|\mathcal{F}_n], \\ U_N(x) = u(x). \end{cases}$$

In what follows we derive a numerical scheme for the solution of these dynamic programming equations and approximations for the optimal strategies for the original optimization problem.

For each $\eta_n \in \mathcal{F}_n, n = 0, 2, \ldots, N-1$, define

$$A_{n-1,\eta_n} = E_{n-1}\left[\eta_n \begin{pmatrix} (e^{y_n^{(1)}} - 1)^2 & . & . & (e^{y_n^{(m)}} - 1)(e^{y_n^{(1)}} - 1) \\ (e^{y_n^{(2)}} - 1)(e^{y_n^{(1)}} - 1) & . & . & (e^{y_n^{(m)}} - 1)(e^{y_n^{(2)}} - 1) \\ . & . & . & . \\ (e^{y_n^{(m)}} - 1)(e^{y_n^{(1)}} - 1) & . & . & (e^{y_n^{(m)}} - 1)^2 \end{pmatrix}\right].$$

PROPOSITION 4.1. *Let* $u(x) = x^\alpha/\alpha, 0 \neq \alpha < 1$. *Let*

$$\lambda_N := 1,\ \eta_n := E_n[\textstyle\prod_{k=n}^N \lambda_k],$$
$$\lambda_{n-1} := (1 + \tfrac{1}{1-\alpha}(e^{y_n} - 1)' A_{n-1,\eta_n}^{-1} E_{n-1}[(e^{y_n} - 1)\eta_n])^\alpha,$$

$n = 1, \ldots, N$. *Then the following hold:*

(i) $\eta_n \in \mathcal{F}_n$ *is bounded.*

(ii) A_{n-1,η_n} *is invertible and there exists a constant* C *s.t.* $\Delta t\|A_{n-1,\eta_n}^{-1}\| < C$ *for all* $n = 1, 2, \ldots, N$.

(iii) $\|E_{n-1}[\eta_n(e^{y_n} - 1)]\| = o(\sqrt{\Delta t})$.

The proof of this proposition follows from Theorem B.3 (iii), (ii) and Theorem B.4.

For each $\eta_n \in \mathcal{F}_n, n = 0, 2, \ldots, N-1$, define

$$W_n(x_n, h_n) := E[\eta_{n+1} u(x_n^{h_n}) | \mathcal{F}_n] \tag{4.3}$$
$$= E[\eta_{n+1} u(x_n e^{r\Delta t} + x_n e^{r\Delta t} h_n \cdot (e^{y_{n+1}} - 1)) | \mathcal{F}_n],$$

$$V_n(x_n, h_n) \tag{4.4}$$
$$:= E\Big[\eta_{n+1}(u(x_n e^{r\Delta t}) + u'(x_n e^{r\Delta t}) x_n e^{r\Delta t} h_n \cdot (e^{y_{n+1}} - 1)$$
$$+ \frac{1}{2} u''(x_n e^{r\Delta t})(x_n e^{r\Delta t} h_n \cdot (e^{y_{n+1}} - 1))^2) | \mathcal{F}_n\Big].$$

Assume for any $x_{n-1} \geq 0$ that there exist $h_{n-1}^*, h_{n-1}^{**} \in [0,1]^m \subset \mathcal{R}^m$ s.t.

$$W_{n-1}(x_{n-1}, h_{n-1}^*) = \sup_{h_{n-1}} W_{n-1}(x_{n-1}, h_{n-1}) \tag{4.5}$$

and

$$V_{n-1}(x_{n-1}, h_{n-1}^{**}) = \sup_{h_{n-1}} V_{n-1}(x_{n-1}, h_{n-1}). \tag{4.6}$$

We use h_n^{**} defined in (4.6) to approximate h_n^* defined in (4.5).

As is seen from (4.4), (4.5) V_n is obtained from W_n by taking a Taylor expansion up to the second term. Thus the function V_n approximates W_n when Δt is small (and as a result y_n is close to 0).

LEMMA 4.2. *Let $u(x) = x^\alpha/\alpha, 0 \neq \alpha < 1$. Let η_n be defined as in Proposition 4.1. Then*

$$|W_{n-1}(x_{n-1}, h_{n-1}^{**}) - W_{n-1}(x_{n-1}, h_{n-1}^*)| = x^\alpha o(\Delta t), \tag{4.7}$$

where

$$h_{n-1}^{**} = \frac{1}{1-\alpha} A_{n-1,\eta_n}^{-1} \cdot E_{n-1}[\eta_n (e^{y_n} - 1)]. \tag{4.8}$$

The proof of this lemma is in Appendix B.

Remark 4. Using the same route we can prove that Lemma 4.2 holds for any η_n satisfying the three conditions in Proposition 4.1.

Remark 5. The motivation to define W_n and V_n can be seen as follows. Let $n = N-1$. Then in (4.2) we have

$$U_{N-1}(x) = \sup_{h_{N-1}} E[u(xe^{r\Delta t} + xe^{r\Delta t} h_{N-1} \cdot (e^{y_N} - 1)) | \mathcal{F}_{N-1}].$$

Clearly, $U_{N-1}(x_{N-1})$ coincides with $W_{N-1}(x_{N-1}, h_{N-1}^*)$ defined by (4.4), (4.5) with $\eta_N = 1$. Below, we write $\mathcal{E}_{n-1}(x_{n-1}) := W_{n-1}(x_{n-1}, h_{n-1}^*) - W_{n-1}(x_{n-1}, h_{n-1}^{**})$ for convenience.

In other words,

$$\begin{aligned} &U_{N-1}(x_{N-1}) = W_{N-1}(x_{N-1}, h_{N-1}^*) = W_{N-1}(x_{N-1}, h_{N-1}^{**}) + \mathcal{E}_{N-1} \\ &\approx W_{N-1}(x_{N-1}, h_{N-1}^{**}) = E_{N-1}[u(x_{N-1} e^{r\Delta t}(1 + h_{N-1}^{**} \cdot (e^{y_N} - 1)))], \end{aligned} \tag{4.9}$$

since $\mathcal{E}_{N-1}$ is small by Lemma 4.2.

As a result, one can find an approximation for the optimal strategy in a recursive way. Moreover, the expected utility of the terminal wealth associated with $h^{**} = \{h_k^{**}\}_{k=0}^{N-1}$ is an approximation for the value function V^* in the optimization problem. The next two theorems show that in a limit as Δt tends to zero the value of V^{**} converges to that of V^*. Thus h_n^{**} can serve as an approximation for the optimal strategy.

Hence Lemma 4.2 shows that the difference between the wealth associated with optimal portfolio h^* and the portfolio h^{**} is small when Δt is small.

THEOREM 4.3. *Let* $u(x) = \frac{x^\alpha}{\alpha}, 0 \neq \alpha < 1$. *Define*

$$\begin{aligned} &\lambda_N := 1, \eta_N = 1, \\ &h_{n-1}^{**} = \tfrac{1}{1-\alpha} A_{n-1,\eta_n}^{-1} \cdot E_{n-1}[(e^{y_n} - 1)\eta_n], \\ &\lambda_{n-1} := e^{r\Delta t\alpha}(1 + \tfrac{1}{1-\alpha} h_{n-1}^{**} \cdot (e^{y_n} - 1))^\alpha, \\ &\eta_n := E_n[\textstyle\prod_{i=n}^N \lambda_i], 1 \le n \le N. \end{aligned} \tag{4.10}$$

Then

$$|U_n^{**}(x_n) - U_n(x_n)| = u(x_n)o(1), \tag{4.11}$$

where

$$\begin{aligned} &U_n^{**}(x_n) := E_n[u(x_{n,N}^{**})], \\ &x_{n,k}^{**} := x_n e^{(k-n)r\Delta t} \cdot \textstyle\prod_{i=n}^{k-1}(1 + h_i^{**} \cdot (e^{y_{i+1}} - 1)). \end{aligned} \tag{4.12}$$

Proof. Define the following sequences:

$$\begin{aligned} &U_N^{**} := u(x), \\ &U_k^{**}(x_k) := W_k(x_k, h_k^{**}), \\ &U_k^*(x_k) := W_k(x_k, h_k^*). \end{aligned}$$

From the definition of η_n and (4.4), (4.8), we verify that

$$U_k^{**}(x_k) = u(x_k)\eta_k, \tag{4.13}$$

and using (4.13) above we can write

$$\begin{aligned} &U_k^*(x_k) = W_k(x_k, h_k^*) \\ &= \sup_{h_k} E_k[\eta_{k+1} u(x_k e^{r\Delta t}(1 + h_k \cdot (e^{y_{k+1}} - 1)))] \\ &= \sup_{h_k} E_k[U_{k+1}^{**}(x_k e^{r\Delta t}(1 + h_k \cdot (e^{y_{k+1}} - 1)))]. \end{aligned} \tag{4.14}$$

First, we prove that

$$|U_{N-1}^{**}(x_{N-1}) - U_{N-1}(x_{N-1})| \le u(x_{N-1})o(\Delta t).$$

Really $U_{N-1}^{**}(x_{N-1}) = u(x_{N-1})\eta_{N-1}$, and

$$\begin{aligned} &U_{N-1}^*(x_{N-1}) = \sup_{h_{N-1}} E_{N-1}[U_N^{**}(x_{N-1}e^{r\Delta t}(1 + h_{N-1} \cdot (e^{y_N} - 1)))] \\ &= \sup_{h_{N-1}} E_{N-1}[u(x_{N-1}e^{r\Delta t}(1 + h_{N-1} \cdot (e^{y_N} - 1)))] = U_{N-1}(x_{N-1}). \end{aligned}$$

Therefore,

$$\begin{aligned} |U_{N-1}^{**}(x_{N-1}) - U_{N-1}(x_{N-1})| &= |W_{N-1}(x_{N-1}, h_{N-1}^{**}) - U_{N-1}^*(x_{N-1})| \\ &= |W_{N-1}(x_{N-1}, h_{N-1}^{**}) - W_{N-1}(x_{N-1}, h_{N-1}^*)| \le u(x_{N-1})o(\Delta t). \end{aligned}$$

The last inequality is due to Lemma 4.2.

Let $C_1 = 1$. Suppose for $n \le N$ we have

$$|U_n^{**}(x_n) - U_n(x_n)| = C_{N-n} u(x_n) o(\Delta t), \tag{4.15}$$

where $\{C_k\}_{k=2}^N$ is a bounded sequence to be determined later. Then

$$\begin{aligned}
|U_{n-1}^{**} - U_{n-1}| &= |W_{n-1}(x_{n-1}, h_{n-1}^{**}) - U_{n-1}| \\
&\le |W_{n-1}(x_{n-1}, h_{n-1}^{**}) - W_{n-1}(x_{n-1}, h_{n-1}^{*})| \\
&\quad + |W_{n-1}(x_{n-1}, h_{n-1}^{*}) - U_{n-1}| \\
&\le o(\Delta t) u(x_{n-1}) + | \sup_{h_{n-1}} E_{n-1}[U_n^{**}(x_{n-1} e^{r\Delta t}(1 + h_{n-1} \cdot (e^{y_n} - 1)))] \\
&\quad - \sup_{h_{n-1}} E_{n-1}[U_n(x_{n-1} e^{r\Delta t}(1 + h_{n-1} \cdot (e^{y_n} - 1)))]|.
\end{aligned}$$

Then, using (4.14), (4.15), and Lemma 4.2, we get

$$\begin{aligned}
&|U_{n-1}^{**} - U_{n-1}| \\
&\le o(\Delta t) u(x_{n-1}) + \sup_{h_{n-1}} E_{n-1}[|U_n^{**}(\gamma) - U_n(\gamma)|]|_{\gamma = x_{n-1} e^{r\Delta t}(1 + h_{n-1} \cdot (e^{y_n} - 1))} \\
&\le o(\Delta t) u(x_{n-1}) + \sup_{h_{n-1}} E_{n-1}[u(x_{n-1} e^{r\Delta t}(1 + h_{n-1} \cdot (e^{y_n} - 1))) C_{N-n} o(\Delta t)]
\end{aligned}$$

for all sufficiently small Δt (so as $o(\Delta t) < 1$).

Applying Remark 4 with $\eta_n = 1$, we have

$$\begin{aligned}
&\sup_{h_{n-1}} E_{n-1}[u(x_{n-1} e^{r\Delta t}(1 + h_{n-1} \cdot (e^{y_n} - 1)))] \\
&\le E_{n-1}[u(x_{n-1} e^{r\Delta t}(1 + h_{n-1}^{**} \cdot (e^{y_n} - 1)))] + o(\Delta t) u(x_{n-1}) \\
&= (\nu_{n-1} + o(\Delta t)) u(x_{n-1}),
\end{aligned}$$

where $\nu_{n-1} := e^{r\alpha\Delta t} E_{n-1}[(1 + h_{n-1}^{**} \cdot (e^{y_n} - 1))^{\alpha}]$. Therefore,

$$|U_{n-1}^{**} - U_{n-1}| \le (1 + C_{N-n}(\nu_{n-1} + o(\Delta t))) o(\Delta t) u(x_{n-1}).$$

Define $C_n = 1 + C_{N-n}(\nu_{N-n} + o(\Delta t))$, $C_1 = 1$. Then

$$|U_{n-1}^{**} - U_{n-1}| \le C_{N-n+1} o(\Delta t) u(x_{n-1}).$$

Continuing this process we get

$$|U_0^{**} - U_0| \le C_N u(x_0) o(\Delta t).$$

To complete the proof, we have to prove $C_N o(\Delta t) = o(1)$. In fact, ν_{N-k} has the limit 1 as $\Delta t \to 0$ by virtue of Theorem B.3. Moreover, there exists a constant $D_2 > 0$ s.t.

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} |\nu_{N-k} - 1| < D_2$$

for all $k = 1, \dots, N$. From the definition of C_k and the properties of ν_k for $k \ge 2$, we have

$$C_k \le 1 + C_{k-1}(1 + (D_2 + 1)\Delta t) \le \cdots \le (k-2) + \frac{1}{(D_2+1)\Delta t}((1 + (D_2+1)\Delta t)^k - 1).$$

Then $C_N\Delta t$ converges to a constant as $\Delta t \to 0$. Consequently $C_N o(\Delta t) \to 0$ as $\Delta t \to 0$. Hence we have proved that U_0^{**} has the same limit as U_0 (the optimal expected utility of the terminal wealth), as $\Delta t \to 0$, i.e.,

$$|U_0^{**} - U_0|u(x_0) = o(1).$$

By the definition of U_n^{**}, it is easy to see that

$$\begin{aligned}
&U_n^{**}(x_n) := W_n(x_n, h_n^{**}) = E_n[\eta_{n+1}u(x_n e^{r\Delta t}(1 + h_n^{**} \cdot (e^{y_{n+1}} - 1)))] \\
&= E_n[W_{n+1}(x_n e^{r\Delta t}(1 + h_n^{**} \cdot (e^{y_{n+1}} - 1)), h_{n+1}^{**})] \\
&= E_n[\eta_{n+2}u(x)|_{x=x_{n+1}e^{r\Delta t}(1+h_n^{**}(e^{y_{n+1}}-1))}] \\
&= E_n[W_{N-1}(x_{n,N-1}^{**}, h_{N-1}^{**})] = E_n[u(x_{n,N}^{**})],
\end{aligned}$$

where $x_{n,k}^{**} = x_n e^{(k-n)r\Delta t} \cdot \prod_{i=n}^{k-1}(1 + h_i^{**} \cdot (e^{y_{i+1}} - 1))$. □

The case of a logarithmic utility function can be treated as the same way as the power utility function. The same results hold, although with a higher rate of convergence.

LEMMA 4.4. *Let $u(x) = \log(x)$, and choose $\eta_n = 1, n = 1, 2, \ldots, N$. Then we have*

$$|W_{n-1}(x_{n-1}, h_{n-1}^{**}) - W_{n-1}(x_{n-1}, h_{n-1}^{*})| = O(\Delta t)^2,$$

where

$$h_{n-1}^{**} = A_{n-1,1}^{-1} \cdot E_{n-1}[e^{y_n} - 1]. \tag{4.16}$$

The proof is the same as that of Lemma 4.2. Moreover when we repeat the proof, we see that the resulting convergence rate is $O(\Delta t)^2$, which is higher than $O(\Delta t)$ obtained in the case of the power utility function.

THEOREM 4.5. *Let $u(x) = \log(x)$, and choose $\eta_n = 1, n = 1, 2, \ldots, N$. Then*

$$|U_n(x_n) - U_n^{**}(x_n)| = O(\Delta t),$$

where

$$\begin{aligned}
&U_n^{**}(x_n) := E_n[u(x_{n,N}^{**})], \\
&x_{n,k}^{**} := x_n e^{(k-n)r\Delta t} \cdot \textstyle\prod_{i=n}^{k-1}(1 + h_i^{**} \cdot (e^{y_{i+1}} - 1)),
\end{aligned}$$

and h_i^{**} is defined by (4.16).

Proof. This is the special case of Lemma 4.2 with $\alpha = 0$. However, the condition $\alpha = 0$ enables us to obtain a higher convergence rate of $O(\Delta t)$. □

Remark 6. There are particular cases depending on the structure of the transition matrix P, when we have the convergence rate $O(\Delta t)$ as above even for the power utility function. One of those cases is when the transition matrix has identical columns. We also have another case in Appendix D, Theorem B.5 (ii), when the convergence rate is of a higher rate of $O(\Delta t)$.

Remark 7. Let $u(x) = \frac{1}{\gamma}e^{-x\gamma}$. Define

$$\begin{aligned}
&\eta_N := 1, \\
&\eta_{n-1} := e^{-\gamma h_{n-1}^{**}(e^{y_n}-1)}, \\
&h_{n-1}^{**} = A_{n-1,\eta_n}^{-1} \cdot E_{n-1}[(e^{y_n} - 1)\eta_n](\tfrac{1}{\gamma x_{n-1}}e^{-r(N-n)\Delta t}), \\
&U_{n-1}^{**} = u(x_{n-1}e^{r(N-n+1)\Delta t}\textstyle\prod_{k=n-1}^{N-1}(1 + h_i^{**} \cdot (e^{y_{i+1}} - 1))), \quad n = N, \ldots, 1.
\end{aligned}$$

Then

$$|U_0(x_0) - U_0^{**}(x_0)| = o(1).$$

The proof follows the same route as that of Theorem 4.3.

Remark 8. From the definition (4.10) of η_n, we can see that η_n is the expected utility under strategy h^{**}, given $\mathcal{F}_n$, and given $x_n = 1$. That is,

$$\eta_n = E[u(X_N^{h^{**}})|\mathcal{F}_n, x_n = 1].$$

5. Simulations. Generally, as a result it is not easy to compute the approximate optimal strategy $\{h_n^{**}\}_{n=0}^{N-1}$ in (4.8) in the case of a power utility function. However, we can get an estimation using a simplified strategy:

$$\bar{h}_n^{**} := -\frac{1}{1-\alpha} A_{n-1,1}^{-1} \cdot E_{n-1}[(e^{y_n} - 1)], \quad n = 0, 1, \ldots, N-1.$$

From the definition of V^*, we see that the expected utility $E[u(x_N^{\bar{h}^{**}})]$ associated with $\bar{h}^{**}$ is a lower bound for V^*.

There are cases, however, where (4.8) is relatively easy to evaluate. For example, if the transition matrix has identical columns, then the conditional probability $Pr(Y_n = e_k|\mathcal{F}_n), n \geq 1$, would be a constant regardless of n, k. Thus η_n is a constant, and it can be excluded from the expression for h_{n-1}^{**} (4.8). Therefore, the strategy h^{**} is the same as $\bar{h}^{**}$.

We apply the results of section 2 to obtain an applicable representation of the strategy for the case of $m = 1$,

$$\bar{h}_{n-1}^{**} \overset{m=1}{=} \frac{\sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1})(\mu(e_k) - r)}{(1-\alpha)\sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1})\hat{\sigma}(e_k)^2} + O(\Delta t), n = 1, 2, \ldots, N. \tag{5.1}$$

The simulations in this section deal with (5.1). We use Lemma 2.2 to calculate $Pr(Y_n = e_k|\mathcal{F}_n)$ recursively.

Comparing the strategies (5.1) with Merton strategies, we can see that in our case the constant drift or the constant volatility in the expression for Merton strategies is replaced by linear combinations of the drifts or of the volatilities corresponding to different states of the Markov chain. The weights of the linear combinations are probabilities that the Markov chain is in those states. We divide the time interval [0,T] into N parts and assume the transition of the Markov chain occurs only at those points of time. Hence we have the Merton model on each interval. The consequence is that we might obtain a solution directly like (5.1) in the case of the logarithmic utility function. However, it is not true for the power or the exponential utility functions.

To illustrate this point, let us assume that the transition matrix does not have identical columns. Then in (4.8) we cannot choose $\eta_n = 1$ as in the case of a logarithmic utility function, nor can we simplify the expression by canceling η_n as in the case when all the columns of the transition matrix are identical. A representation as simple as (5.1) cannot not be obtained. We have to employ the Monte Carlo method to calculate the portfolio (4.8). However, we may use (5.1) to get a lower bound for V^*.

In our simulations, W_0 stands for the initial wealth. The default value of W_0 is 1. P denotes the transition matrix. The interest rate r is equal to 0.06. The time horizon T is 1, and it is divided into $N = 1000$ parts, i.e., $\Delta t = 10^{-3}$.

We compare our optimal strategy with the Merton strategy. Since the Markov chain has several states, we use the Merton strategy, replacing the drift and volatility in it with those obtained from taking the average of the drift and volatility, respectively, over different states of the Markov chain. The resulting Merton's strategy is

$$h_{n-1} = \frac{\bar{\mu} - r}{(1-\alpha)\bar{\sigma}^2}, \quad n = 1, \ldots, N, \tag{5.2}$$

where $\bar{\mu} = \sum_i \mu(e_i)/d, \bar{\sigma} = \sum_i \hat{\sigma}(e_i)/d$. One may think of using $\sum_i \sigma(e_i)^2/d$ instead of $\bar{\sigma}^2$ in the formula. However, our simulation shows that it does not provide a better result.

We also compare our optimal strategy with the buy-and-hold strategy, which is denoted as "b/h." The buy-and-hold strategy means to buy the stock using all cash available at the beginning and then hold the stock until the end. We generate the wealth process 1000 times and calculate the average of the utilities from the terminal wealth.

Table 1 lists the result for a Markov chain which has a transition matrix with identical columns. For the power utility function, we can obtain that our optimal strategy is a constant 0.1133 while the Merton strategy is also a constant 0.3636 different from ours as seen from (5.1). Table 1 shows that our optimal strategy on average gives better utilities with smaller standard deviations for both the logarithm and the power law utility functions. The last line of Table 1 shows the number of simulations in which our optimal strategy generates a better utility than the Merton strategy or the b/h strategy. Note that in the case of the power utility function, even though our optimal strategy generates only 487 better than the Merton strategy among 1000 simulations, the average (-0.2748) is still significantly higher than the one (-0.2974) generated by the Merton strategy, and the standard deviation (0.0380) is also significantly less than 0.1348.

In Table 2, we use the same parameters except the transition matrix is replaced by a matrix with nonidentical columns. In this case for the power utility function, the Merton strategy is still a constant ($h = 0.3636$). However, our optimal strategy h^{**} varies from 0.1294 to 0.3324 with a mean 0.1450, while it is a constant 0.1133 in the previous case. The result are similar to those of Table 1, though.

The average utility may vary slightly if more wealth processes are generated in the simulation. However, we always find that our optimal strategy generates on average the best utilities. The results in Table 3 for 5000 and 10000 simulations show that although the utilities vary slightly, our optimal strategy still has the best performance among the three strategies.

For one more example, we choose the same parameters as in Example 1 of Sass and Haussmann [18]. The results for this example are listed in Table 4. Table 5 is a copy of Table 2 of Sass and Haussmann [18]. One can see that our optimal strategy generates the average utilities (0.3969 for the logarithm, and -0.1128 for the power) very close to theirs (0.399 for the logarithm, and -0.121 for the power). It is not surprising because our model can be viewed as an extension of theirs in an approximate sense. Therefore, similar results are expected when the same parameters are employed.

Finally, we provide standard deviations in Tables 1 and 2. One can see that the standard deviation associated with the optimal strategy is always the smallest one.

TABLE 1
$\mu = [0.1, 0.9]', \hat{\sigma} = [0.4, 0.7]', P = [0.95, 0.95; 0.05, 0.05], \Delta t = 10^{-3}$. 1000 *simulations.*

u(x)	*log(x)*			$-x^{-3}/3$		
Strategy	*opt*	*Merton*	*b/h*	*opt*	*Merton*	*b/h*
av. u(x)	0.0772	0.0041	0.0498	−0.2748	− 0.2974	− 0.6352
med u(x)	0.0781	0.0084	0.0535	−0.2719	− 0.2687	− 0.2839
std u(x)	0.1871	0.5592	0.4039	0.0380	0.1348	0.9696
opt better		574	548		487	515

TABLE 2
$\mu = [0.1, 0.9]', \hat{\sigma} = [0.4, 0.7]', P = [0.95, 0.5; 0.05, 0.5], \Delta t = 10^{-3}$. 1000 *simulations.*

u(x)	*log(x)*			$-x^{-3}/3$		
Strategy	*opt*	*Merton*	*b/h*	*opt*	*Merton*	*b/h*
av. u(x)	0.0946	0.0360	0.0760	−0.2715	− 0.2877	− 0.6236
med u(x)	0.0930	0.0177	0.0644	−0.2664	− 0.2587	− 0.2747
std u(x)	0.2703	0.6012	0.4441	0.0557	0.1450	1.1093
opt better		565	552		470	512

Appendix A. Proof of Lemma 2.3.

Proof. For $k = 1, 2, \ldots, d$, denote

$$b_k := g_{n-1}|_{Y_{n-1}=e_k} \in \mathcal{R}^{m\times 1}, f_k := \sigma_{n-1}|_{Y_{n-1}=e_k} \in \mathcal{R}^{m\times m}.$$

(i) Using Lemma 2.1, we have

$$\begin{aligned}
&|E[y_n^{(i)} - 1|\mathcal{F}_{n-1}]| \\
&= \left| \int e^{y_n^{(i)}} \sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1}) \phi_{ik}(y_n^{(i)} - b_k(i)) dy_n^{(i)} - 1 \right| \\
&= \left| \sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1}) e^{f_k(i) f_k(i)'/2 + b_k(i)} - 1 \right| \\
&\leq \max_k \{|e^{f_k(i) f_k(i)'/2 + b_k(i)} - 1|\}.
\end{aligned}$$

By virtue of (2.6) the right-hand side of the above expression is of the order Δt. Therefore, there exists a constant $C_1 > 0$ s.t.

$$|E_{n-1}[e^{y_n^{(i)}} - 1]| \leq C_1 \Delta t.$$

(ii) Similarly,

$$\begin{aligned}
&E_{n-1}[(e^{y_n(i)} - 1)^2] \\
&= \sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1}) e^{2 f_k(i) f_k(i)' + 2 b_k(i)} \\
&- 2 \sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1}) e^{f_k(i) f_k(i)'/2 + b_k(i)} + 1 \\
&= \sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1}) (e^{2 f_k(i) f_k(i)' + 2 b_k(i)} - 2 e^{f_k(i) f_k(i)'/2 + b_k(i)} + 1).
\end{aligned}$$

TABLE 3
5000, 10000 *simulations.*

u(x)	*log(x)*			$-x^{-3}/3$		
Strategy	*opt*	*Merton*	*b/h*	*opt*	*Merton*	*b/h*
5000 sim.	0.0923	0.0338	0.0743	−0.2719	− 0.2883	− 0.6541
10000 sim.	0.0966	0.0422	0.0805	−0.2709	− 0.2857	− 0.6284

TABLE 4
$\mu = [0.8, -0.4]', \hat{\sigma} = [0.2, 0.2]', Q = [-30, 24; 30, -24], P = e^{Q/250}, \Delta t = 1/250$. 500 *simulations.*

u(x)	*log(x)*			$-x^{-5}/5$		
Strategy	*opt*	*Merton*	*b/h*	*opt*	*Merton*	*b/h*
av. u(x)	0.3969	0.1125	0.1205	−0.1128	− 0.1525	− 0.2180
med u(x)	0.2965	0.1033	0.1186	−0.0902	− 0.1218	− 0.1105
opt better		319	313		338	288

Note that we assumed $f_k(i,j)$ to be positive for any k, i, j, and

$$e^{2f_k(i)f_k(i)'+2b_k(i)} - 2e^{f_k(i)f_k(i)'/2+b_k(i)} + 1 = f_k(i)f_k(i)' + O(\Delta t^2).$$

Therefore, using (2.6) once more we conclude that there exist constants $C_2 > 0, C_3 > 0$ s.t.

$$C_2\Delta t < E_{n-1}[(e^{y_n^{(i)}} - 1)^2] \le C_3\Delta t.$$

(iii) Again, we have

$$\begin{aligned}
&|E_{n-1}[(e^{y_n^{(i)}} - 1)^3]| \\
&= \left|\sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1})(e^{9f_{ik}^2/2+3b_{k(i)}} - 3e^{2f_{ik}^2+2b_k(i)} + 3e^{f_{ik}^2/2+b_k(i)} - 1)\right| \\
&\le \max_k\{|e^{9f_{ik}^2/2+3b_{k(i)}} - 3e^{2f_{ik}^2+2b_k(i)} + 3e^{f_{ik}^2/2+b_k(i)} - 1|\} = O((\Delta t)^2).
\end{aligned}$$

Therefore,

$$|E[(e^{y_n^{(i)}} - 1)^3]|\mathcal{F}_{n-1}]| = O((\Delta t)^2).$$

(iv) Similarly,

$$\begin{aligned}
&|E[(1 - e^{-y_n^{(i)}})^3|\mathcal{F}_{n-1}]| \\
&\le \sum_{k=1}^{d} Pr(Y_{n-1} = e_k|\mathcal{F}_{n-1})|(-e^{9f_{ik}^2/2-3b_k(i)} + 3e^{2f_{ik}^2-2b_k(i)} - 3e^{f_{ik}^2/2-b_k(i)} + 1)| \\
&\le \max_k| - e^{9f_{ik}^2/2-3b_k(i)} + 3e^{2f_{ik}^2-2b_k(i)} - 3e^{f_{ik}^2/2-b_k(i)} + 1| \\
&= O(\Delta t)^2.
\end{aligned}$$

(v) Similarly, we have

$$E[e^{y_n\alpha}(1 - e^{-y_n})^3|\mathcal{F}_{n-1}] = O(\Delta t)^2. \qquad \square$$

TABLE 5
Sass and Haussmann's Table 2. Known parameters.

u(x)	*log(x)*			$-x^{-5}/5$		
Strategy	*opt*	*Merton*	*b/h*	*opt*	*Merton*	*b/h*
av. u(x)	0.399	0.136	0.118	−0.121	− 0.141	− 0.214
med u(x)	0.305	0.150	0.126	−0.091	− 0.131	− 0.107
opt better than		296	288		359	292

Appendix B. Proof of Lemma 4.2.

Proof. Using the definition (4.5), we have the first order condition:

$$\begin{aligned}
0 &= \left(\frac{\partial W_{n-1}}{\partial h_{n-1}}\right)|_{h_{n-1}=h^*_{n-1}} = \left(\frac{\partial W_{n-1}(x_{n-1},h_{n-1})}{\partial h_{n-1}}\right)|_{h_{n-1}=h^*_{n-1}} \\
&= E[\eta_n u'(x_{n-1}e^{r\Delta t} + x_{n-1}e^{r\Delta t}h^*_{n-1}\cdot(e^{y_n}-1))(e^{y_n}-1)x_{n-1}e^{r\Delta t}|\mathcal{F}_{n-1}] \\
&= E_{n-1}[\eta_n(e^{y_n}-1)]u'(x_{n-1}e^{r\Delta t})x_{n-1}e^{r\Delta t} + (x_{n-1}e^{r\Delta t})^2 E_{n-1}[\eta_n(h^*_{n-1} \\
&\quad \cdot (e^{y_n}-1))(e^{y_n}-1)u''(x_{n-1}e^{r\Delta t}) \\
&\quad + x^3_{n-1}e^{3r\Delta t}E_{n-1}[\eta_n(h^*_{n-1}\cdot(e^{y_n}-1))^2(e^{y_n}-1)u'''(\xi)]/2,
\end{aligned}$$

where $\xi = x_{n-1}e^{r\Delta t}(1+t\cdot(e^{y_n}-1))$ for some $t\in[0,1]^m \subset \mathcal{R}^{m\times 1}$. Solving h^*_{n-1} from the equations above, we obtain

(B.1)

$$\begin{aligned}
h^*_{n-1} = &- A^{-1}_{n-1,\eta_n}\cdot E_{n-1}[\eta_n(e^{y_n}-1)]\frac{u'(x_{n-1}e^{r\Delta t})}{u''(x_{n-1}e^{r\Delta t})x_{n-1}e^{r\Delta t}} \\
&- A^{-1}_{n-1,\eta_n}\cdot E_{n-1}[\eta_n(h^*_{n-1}\cdot(e^{y_n}-1))^2(e^{y_n}-1)u'''(\xi)]\frac{x_{n-1}e^{2r\Delta t}}{2u''(x_{n-1}e^{r\Delta t})},
\end{aligned}$$

where A_{n-1,η_n} is defined right before Proposition 4.1.

At the same time, according to the first order condition for the function V_{n-1}, we have

(B.2)
$$\begin{aligned}
h^{**}_{n-1} &= -A^{-1}_{n-1,\eta_n}\cdot E_{n-1}[\eta_n(e^{y_n}-1)]\frac{u'(x_{n-1}e^{r\Delta t})}{u''(x_{n-1}e^{r\Delta t})x_{n-1}e^{r\Delta t}} \\
&= \frac{1}{1-\alpha}A^{-1}_{n-1,\eta_n}\cdot E_{n-1}[\eta_n(e^{y_n}-1)].
\end{aligned}$$

Therefore, using (B.1) and (B.2), we have

$$\|h^*_{n-1}-h^{**}_{n-1}\|_1 = \left\|A^{-1}_{n-1,\eta_n}\cdot E_{n-1}[\eta_n(e^{y_n}-1)(h^*_{n-1}\cdot(e^{y_n}-1))^2u'''(\xi)]\frac{x_{n-1}e^{r\Delta t}}{2u''(x_{n-1}e^{r\Delta t})}\right\|_1.$$

Since each entry of A_{n-1,η_n} is of the order $O(\Delta t)$, it is easy to show that every entry of A^{-1}_{n-1,η_n} is of the order of $1/O(\Delta t)$. By Theorem B.2, we have

$$\|E_{n-1}[\eta_n(e^{y_n}-1)(h^*_{n-1}\cdot(e^{y_n}-1))^2u'''(\xi)]\| = u'''(x_{n-1}e^{r\Delta t})o(\Delta t)^{3/2}.$$

Therefore,

$$\begin{aligned}
\|h^*_{n-1}-h^{**}_{n-1}\|_1 &= \|A^{-1}_{n-1,\eta_n}\|_\infty u'''(x_{n-1}e^{r\Delta t})o(\Delta t^{3/2})\left|\frac{x_{n-1}e^{r\Delta t}}{u''(x_{n-1}e^{r\Delta t})}\right| \\
&= \frac{x_{n-1}u'''(x_{n-1}e^{r\Delta t})}{u''(x_{n-1}e^{r\Delta t})}\cdot o\Delta t^{1/2} = o(\Delta t)^{1/2}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}&|W_{n-1}(x_{n-1},h^{**}_{n-1})-W_{n-1}(x_{n-1},h^{*}_{n-1})|\\&=\|E_{n-1}[\eta_n u'(\zeta)(e^{y_n}-1)]\|_1\|h^{*}_{n-1}-h^{**}_{n-1}\|_1 x_{n-1}e^{r\Delta t}\\&=x_{n-1}o(\Delta t^{1/2})\|E_{n-1}[\eta_n(e^{y_n}-1)u'(\zeta)]\|_1,\end{aligned}$$

where $\zeta = x_{n-1}e^{r\Delta t}(1+s\cdot(e^{y_n}-1))$ for some s between h^{*}_{n-1} and h^{**}_{n-1}.

Note that since we assume $h^{*}_{n-1}, h^{**}_{n-1}\in[0,1]^m$, then $s\in[0,1]^m$. By the assumption (i), η_n is bounded, and we can apply the dominated convergence theorem to obtain

$$\begin{aligned}&\lim_{\Delta t\to 0}\|E_{n-1}[\eta_n(e^{y_n}-1)((1+s\cdot(e^{y_n}-1))^{\alpha-1}-1)]\|_1\frac{1}{\Delta t}\\&\le E_{n-1}\left[\lim_{\Delta t\to 0}\left\|\frac{\eta_n(e^{y_n}-1)}{\sqrt{\Delta t}}\right\|_1\ \lim_{\Delta t\to 0}\left|\frac{1}{\sqrt{\Delta t}}((1+s(e^{y_n}-1))^{\alpha-1}-1)\right|\right]\\&<\text{Constant}.\end{aligned}$$

Therefore,

$$\|E_{n-1}[\eta_n(e^{y_n}-1)((1+s\cdot(e^{y_n}-1))^{\alpha-1}-1)]\|_1 = O(\Delta t). \tag{B.3}$$

To complete our proof, we apply (B.3) and Theorem B.4 to obtain

$$\begin{aligned}&\|E_{n-1}[\eta_n(e^{y_n}-1)u'(\zeta)]\|_1\\&\le\|E_{n-1}[\eta_n(e^{y_n}-1)(u'(\zeta)-x_{n-1}^{\alpha-1}e^{(\alpha-1)r\Delta t})]\|_1+\|E_{n-1}[\eta_n(e^{y_n}-1)x_{n-1}^{\alpha-1}e^{(\alpha-1)r\Delta t}]\|_1\\&=x_{n-1}^{\alpha-1}e^{(\alpha-1)r\Delta t}E_{n-1}[\eta_n(e^{y_n}-1)((1+s(e^{y_n}-1))^{\alpha-1}-1)]+x_{n-1}^{\alpha-1}o(\Delta t^{1/2})\\&=x_{n-1}^{\alpha-1}o(\Delta t^{1/2}).\end{aligned}$$

As a result we have proved

$$|W_{n-1}(x_{n-1},h^{**}_{n-1})-W_{n-1}(x_{n-1},h^{*}_{n-1})| = x_{n-1}^{\alpha}o(\Delta t). \qquad \square$$

PROPOSITION B.1. *If $\xi=(1+s\cdot(e^y-1))^{\alpha-3}, \alpha<1, s\in[0,1]^m, \sum_{i=1}^m s_i\le 1$, $y=g\Delta t+\sigma\sqrt{\Delta t}Z, \sigma>0, Z=(Z_1,\dots,Z_m)'\in\mathcal{R}^m, Z_i\sim N(0,1)$, then there exists a constant C s.t. $|\xi-1|\le C(1+\|Z\|_1)e^{\|Z\|_1}\sqrt{\Delta t}$.*

Proof. First, we suppose $\sum_{i=1}^m s_i<1$. It is easy to see that

$$(1+s\cdot(e^y-1))^{\alpha-3}<\left(1-\sum_{i=1}^m s_i\right)^{\alpha-3}.$$

Then for small Δt, we have

$$\begin{aligned}&\frac{1}{\sqrt{\Delta t}}|(1+s\cdot(e^y-1))^{\alpha-3}-1|\\&=(3-\alpha)(1+s\cdot(e^{g\tau-\sigma\sqrt{\tau}Z}-1))^{\alpha-4}(s\cdot(\operatorname{diag}(2g^{(i)}\sqrt{\tau}+\sigma^{(i)}Z)e^{g\tau+\sigma\sqrt{\tau}Z}))|\\&\le C\left(1-\sum_i^m s_i\right)^{\alpha-4}\sum_{i=1}^m(2|g^{(i)}\tau|+\sigma^{(i)}|Z|)e^{g^{(i)}\tau+\sigma^{(i)}\sqrt{\tau}Z}\\&\le C(1+\|Z\|_1)e^{\|Z\|_1},\end{aligned}$$

where $\tau \in [0, \sqrt{\Delta t}]$. Therefore $|(1 + s \cdot (e^y - 1))^{\alpha-3} - 1| \le C(1 + \|Z\|_1)e^{\|Z\|_1}\sqrt{\Delta t}$.

In case $\sum_{i=1}^m s_i = 1$, we observe

$$\begin{aligned}(1 + s \cdot (e^y - 1))^{\alpha-3} &= \left(1 + \sum_{i=2}^m s_i(e^{y^{(i)}-y^{(1)}} - 1)\right)^{\alpha-3} e^{y^{(1)}(\alpha-3)} \\ &< \left(1 - \sum_{i=2}^m s_i\right)^{\alpha-3} e^{y^{(1)}(\alpha-3)}.\end{aligned}$$

We still can prove the proposition as above. □

THEOREM B.2. *$u(x) = x^\alpha/\alpha, \alpha < 1$. If $\alpha = 0$, define $u(x) = \log(x)$. $\xi = xe^{r\Delta t}(1 + s \cdot (e^{y_n} - 1))$ for some $s \in (0,1)^m \subset \mathcal{R}^m$. y_n is defined by (2.5). $h_{n-1} \in (0,1)^m \subset \mathcal{R}^m$ satisfies the constraints (3.2). $\eta_n \in \mathcal{F}_n$ is bounded by a constant δ. Then*

$$\|E_{n-1}[\eta_n(e^{y_n} - 1)(h_{n-1} \cdot (e^{y_n} - 1))^2 u'''(\xi)]\|_1 = u'''(x_{n-1})o(\Delta t^{3/2}).$$

Proof. Write $G_i = e^{y(i)} - 1, h_{n-1} = (h_{n-1}(1), \ldots, h_{n-1}(m)) \in R^M$. We have

$$\begin{aligned}&\|E_{n-1}[\eta_n(e^y - 1)(h_{n-1} \cdot (e^y - 1))^2 u'''(\xi)]\|_1 \\ &\le m \max_i \left| E_{n-1}\left[\eta_n \sum_{j,k} h_{n-1}(j)G_j h_{n-1}(k)G_k G_i u'''(\xi)\right]\right| \\ &\le m^3 \max_{i,j,k} |E_{n-1}[\eta_n h_{n-1}(j)G_j h_{n-1}(k)G_k G_i u'''(\xi)]| \\ &\le m^3 \max_{i,j,k} |E_{n-1}[\eta_n G_j G_k G_i u'''(\xi)]|.\end{aligned}$$

Recall (3.2); that is, $\|h\|_1 \le 1$ and $1 \ge h(i) \ge 0$ for $i = 1, \ldots, m$.

$$\begin{aligned}&|E_{n-1}[\eta_n G_j G_k G_i u'''(\xi)]| \\ &\le |E_{n-1}[\eta_n G_j G_k G_i (u'''(\xi) - u'''(x_{n-1}))]| + u'''(x_{n-1})|E_{n-1}[\eta_n G_j G_k G_i]| \\ &\le |E_{n-1}[\eta_n G_j G_k G_i (u'''(\xi) - u'''(x_{n-1}))]| + |E_{n-1}[\eta_n G_j G_k G_i]|u'''(x_{n-1}).\end{aligned} \tag{B.4}$$

The second term is $u'''(x_{n-1})o(\Delta t^{3/2})$ according to Theorem B.4.

Now consider the first term. We have

$$\begin{aligned}|u'''(\xi) - u'''(x_{n-1})| &= (\alpha - 1)(\alpha - 2)|((1 + s \cdot (e^{y_n} - 1))^{\alpha-3} - 1)x_{n-1}^{\alpha-3}| \\ &= u'''(x_{n-1})|(1 + s \cdot (e^{y_n} - 1))^{\alpha-3} - 1|.\end{aligned}$$

By Proposition B.1, we know

$$\lim_{\Delta t \to 0} \frac{1}{\sqrt{\Delta t}} |(1 + s \cdot (e^{y_n} - 1))^{\alpha-3} - 1| = \text{Constant, a.s.}$$

Proposition B.1 also guarantees that the dominated convergence theorem works. Hence we have

$$\begin{aligned}&\lim_{\Delta t \to 0} \frac{1}{u'''(x_{n-1})\Delta t^2} |E_{n-1}[\eta_n G_i G_j G_k (u'''(\xi) - u'''(x_{n-1}))]| \\ &\le \delta\ E_{n-1}\left[\lim_{\Delta t \to 0} \frac{|G_i G_j G_k|}{\Delta t^{3/2}} \frac{|u'''(\xi) - u'''(x_{n-1})|}{\sqrt{\Delta t} u'''(x_{n-1})}\right] = 0.\end{aligned}$$

So the first term of (B.4) is of order $O(\Delta t^2)$. Therefore,

$$\begin{aligned}&\|E_{n-1}[\eta_n(e^y-1)(h_{n-1}\cdot(e^y-1))^2u'''(\xi)]\|_1\\ &\le m^3\max_{i,j,k}|E_{n-1}[\eta_nG_jG_kG_iu'''(\xi)]| = u'''(x_{n-1})o(\Delta t)^{3/2}\\ &= u'''(x_{n-1})(O(\Delta t^2)+o(\Delta t^{3/2})) = u'''(x_{n-1})o(\Delta t^{3/2}).\end{aligned}$$

□

Define

$$\lambda_N := 1, \eta_n := E_n\left[\prod_{k=n}^{N}\lambda_k\right],$$

$$h_{n-1}^{**} = \frac{1}{1-\alpha}A_{n-1,\lambda_n}^{-1}E_{n-1}[(e^{y_n}-1)\eta_n],$$

$$\lambda_{n-1} := (1+h_{n-1}^{**}\cdot(e^{y_n}-1))^\alpha = \left(1+\frac{1}{1-\alpha}(e^{y_n}-1)'A_{n-1,\lambda_n}^{-1}E_{n-1}[(e^{y_n}-1)\eta_n]\right)^\alpha,$$

$$A_{n-1,\eta_n} := E_{n-1}\left[\eta_n\begin{pmatrix}(e^{y_n^{(1)}}-1)^2 & . & . & (e^{y_n^{(m)}}-1)(e^{y_n^{(1)}}-1)\\ (e^{y_n^{(2)}}-1)(e^{y_n^{(1)}}-1) & . & . & (e^{y_n^{(m)}}-1)(e^{y_n^{(2)}}-1)\\ . & . & . & .\\ (e^{y_n^{(m)}}-1)(e^{y_n^{(1)}}-1) & . & . & (e^{y_n^{(m)}}-1)^2\end{pmatrix}\right]$$

for $i = 1,\dots,N$.

THEOREM B.3. *Let $u(x) = \frac{x^\alpha}{\alpha}, \alpha < 1$. Assume that $h_{n-1}^{**}, n = 1,2,\dots,N$, defined above are admissible; then we have the following:*

(i) *For all n, there exist positive constants $D > \frac{1}{\Delta t}$ s.t.*

$$\frac{|E_n[\lambda_n]-1|}{\Delta t} < D.$$

(ii) *For all n, $\Delta t\|A_{n-1,\eta_n}^{-1}\|_\infty \le C_2$, where C_2 is a constant.*

(iii) *For all n,*

$$(1-D\Delta t)^{N-n} \le \eta_n \le (1+D\Delta t)^{N-n}.$$

Proof. We prove the theorem for the case $m = 1$. It becomes much more complex if $m > 1$; the idea of the proof is the same though.

Directly, we have

$$\begin{aligned}\lim_{\Delta t\to 0}\frac{|E_n[\lambda_n]-1|}{\Delta t} &= \lim_{\Delta t\to 0}\left|E_n\left[\frac{1}{\Delta t}|(1+(h_n^{**})'(e^{y_{n+1}}-1))^\alpha-1\right]\right|\\ &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}|E_n[\alpha(h_n^{**})'(e^{y_{n+1}}-1)]+O(\Delta t)|\\ &= \lim_{\Delta t\to 0}\left|\alpha(h_n^{**})'E_n\left[\frac{\alpha(e^{y_{n+1}}-1)}{\Delta t}\right]\right|+C_0\\ &\le \lim_{\Delta t\to 0}\left\|E_n\left[\frac{\alpha(e^{y_{n+1}}-1)}{\Delta t}\right]\right\|_1+C_0 = D.\end{aligned}$$

In the last inequality above, we use the assumption that h^{**} are admissible (and therefore bounded). (We may get the same conclusion by applying Proposition B.1.)

Therefore, λ_n satisfies (i).

To prove (iii), note that λ_n is $\mathcal{F}_{n+1}$ measurable, and so

$$\begin{aligned}\eta_n &= E_n[\lambda_n\eta_{n+1}] = E_n[\lambda_n\lambda_{n+1}\dots\lambda_{N-2}E_N[\lambda_{N-1}]]\\ &\le E_n[\lambda_n\dots\lambda_{N-2}](1+D\Delta t)\\ &\le \cdots \le (1+D\Delta t)^{N-n}.\end{aligned}$$

Similarly, we obtain $\eta_n > (1-D\Delta t)^{N-n}$. So we proved (iii). Applying (iii), we can know that each of the diagonal elements of A_{n-1,η_n} is of order $O(\Delta t)$, while nondiagonal elements are of order $O(\Delta t)^2$. Hence, $\|A_{n-1,\eta_n}\|_\infty$ are of order $1/O(\Delta t)$. □

THEOREM B.4. *η_n defined in* (4.10) *or in Proposition* 4.1 *satisfies*

$$E_{n-1}[\eta_n(e^{y_n}-1)] = o(\sqrt{\Delta t})$$

and

$$E_{n-1}[\eta_n(e^{y_n}-1)^3] = o(\Delta t^{3/2}).$$

Proof. Write $t=(n-1)\Delta t$. Fix t; then

$$E_{n-1}[\eta_n(e^{y_n}-1)^3] = E[\eta_n(e^{y_n}-1)^3|\mathcal{F}_t].$$

By Theorem B.3 (iii), η_n is bounded, so there exists an upper limit as $\Delta t\to 0$. Denote the limit by $\bar\eta_t$. Then

$$\begin{aligned}&\bar{\lim}_{\Delta t\to 0}\frac{1}{\sqrt{\Delta t}}E_{n-1}[\eta_n(e^{y_n}-1)]\\ &= \bar{\lim}_{\Delta t\to 0}\frac{1}{\sqrt{\Delta t}}E_{n-1}[(\eta_n-\bar\eta_t)(e^{y_n}-1)] + E_{n-1}[\bar\eta_t(e^{y_n}-1)]\\ &= E_{n-1}\left[\bar{\lim}_{\Delta t\to 0}(\eta_n-\bar\eta_t)\frac{1}{\sqrt{\Delta t}}(e^{y_n}-1)\right] + \bar\eta_t E_{n-1}\left[\frac{1}{\sqrt{\Delta t}}(e^{y_n}-1)\right]\\ &= 0.\end{aligned}$$

In the last step, we use the fact

$$\frac{e^{y_n^{(i)}}-1}{\sqrt{\Delta t}} \le Ce^{\|Z\|_1}(1+\|Z\|_1)$$

for each $i=1,2,\dots,m$, so that the dominated convergence theorem can be applied.

Therefore, we proved the first equality. Similarly, we can prove the second one. □

THEOREM B.5. *For either of the two following cases:*

(i) *the transition matrix P has identical columns;*

(ii) *all $b_i = (\mu(e_i) - r1_m - \operatorname{diag}(\hat\sigma(e_i)\hat\sigma(e_i)')/2)\Delta t$, $i=1,\dots,d$, are the same,*

we have

$$\begin{aligned}E_{n-1}[\eta_n(e^{y_n}-1)] &= O(\Delta t),\\ E_{n-1}[\eta_n(e^{y_n}-1)^3] &= O(\Delta t)^2.\end{aligned}$$

Proof. Again, we suppose $m=1$ (there is only one stock).

Note that by the definition, η_n is the expected utility of terminal wealth given $x_n=1$ and $\mathcal{F}_n$, while the wealth process is associated with the strategy $\{h_n^{**}\}$.

Intuitively, the expected utility of the terminal wealth should depend only on the initial wealth and the initial state of the Markov chain. Since the initial wealth is given as 1, η_n should be a functional of $E_n[Y_n]$—the expected state of Markov chain Y_n at time $n * \Delta t$. To see this point, one may use the definition of η_n, λ_n and Lemma 2.2 to calculate η_n directly. For example,

$$\begin{aligned}\eta_{N-1} &= E_{N-1}[\lambda_{N-1}] \\ &= E_{N-1}[(1 + h_{N-1}^{**}{}'(e^{y_N} - 1))^{\alpha}] \\ &\overset{(2.1)}{=} \int_{\mathcal{R}} (1 + h_{N-1}^{**}{}'(e^{y_N} - 1))^{\alpha} \sum_{k=1}^{d} Pr(Y_{N-1} = e_k | \mathcal{F}_{N-1}) \phi_k(y_N - b_k) dy_N.\end{aligned}$$

We can see that η_{N-1} is a function of $E_{N-1}[Y_{N-1}]$. Recursively, we may conclude that each η_n is a function of $E_n[Y_n]$.

Recalling $E_n[Y_n] \in [0,1]^d \subset R^d$, we can rewrite (2.2) in a vector form,

$$H(y_n) := E_n[Y_n] = \frac{P\Phi(y_n)E_{n-1}[Y_{n-1}]}{\|\Phi(y_n)E_{n-1}[Y_{n-1}]\|_1},$$

where

$$\begin{aligned}\Phi(y_n) &= \operatorname{diag}\{(2\pi f_i^2)^{-\frac{1}{2}} e^{-\frac{(y_n - b_i)^2}{f_i^2}}, i = 1, \ldots, d\}, \\ f_i &= \mu(e_i)\sqrt{\Delta t}, b_i = (\mu(e_i) - r - \hat{\sigma}(e_i)^2/2)\Delta t.\end{aligned}$$

Using the density function of y_n in Lemma 2.1, we have

$$\begin{aligned}&E_{n-1}[\eta_n(H(y_n))y_n] \\ &= \int_{\mathcal{R}} \eta_n(H(y_n))\ y_n\ \sum_{k=1}^{d} Pr(Y_{n-1} = e_k | \mathcal{F}_{n-1}) \phi_k(y_n - b_k) dy_n \\ &= \sum_{k=1}^{d} Pr(Y_{n-1} = e_k | \mathcal{F}_{n-1}) \int_{\mathcal{R}} \eta_n(H(b_k + f_k Z_n)) \\ &\quad \cdot (b_k + f_k Z_n)\ \phi_k(b_k + f_k Z_n - b_k) d(f_k Z_n)\ \text{(change variable)}.\end{aligned}$$

Note $b_i/\Delta t = \text{Constant}$, $f_i/\sqrt{\Delta t} = \text{Constant}$ for $i = 1, \ldots, d$. Therefore, we have

$$\begin{aligned}&E_{n-1}[\eta_n(H(y_n))] \\ &= \sum_{k=1}^{d} Pr(Y_{n-1} = e_k | \mathcal{F}_{n-1}) f_k^2 \int_{\mathcal{R}} \eta_n(H(b_k + f_k Z_n))\ Z_n\ \phi_k(f_k Z_n) dZ_n + O(\Delta t) \\ &= \sum_{k=1}^{d} Pr(Y_{n-1} = e_k | \mathcal{F}_{n-1}) \int_{\mathcal{R}} \eta_n(H(b_k + f_k Z_n))\ Z_n\ \frac{f_k}{\sqrt{2\pi}} e^{-Z_n^2/2} dZ_n + O(\Delta t).\end{aligned}$$

One can see that if $H(b_k + f_k Z_n)$ is an even function of Z_n, then

$$\int_{\mathcal{R}} \eta_n(H(b_k + f_k Z_n))\ Z_n\ \frac{f_k}{\sqrt{2\pi}} e^{-Z_n^2/2} dZ_n = 0;$$

therefore,

$$E_{n-1}[\eta_n(H(y_n))] = O(\Delta t). \tag{B.5}$$

For the case (i) given in the condition of the theorem, one can easily check that the value of H depends only on P—the transition matrix. In fact, H equals one column of P which is constant, so the theorem is true for this case.

For the case (ii), we can see that $H(b_k + f_k Z_n)$ is even by applying the condition $b_k = b_j, k \neq j$, in the definition of $H(y_n)$.

Therefore,

$$E_{n-1}[\eta_n y_n] = O(\Delta t),$$

and

$$E_{n-1}[\eta_n(e^{y_n} - 1)] = E_{n-1}[\eta_n y_n] + O(\Delta t) = O(\Delta t).$$

Similarly, we can prove

$$E_{n-1}[\eta_n(e^{y_n} - 1)^3] = O(\Delta t)^2. \qquad \square$$

Acknowledgments. The authors are grateful to the Associate Editor and two anonymous referees for their constructive suggestions, especially the comments on the structure of the paper and the model setting. The authors are also indebted to J. Zhang and J. Cvitanić for constructive conversations.

REFERENCES

[1] J. Cvitanić, R. Lipster, and B. Rozovskii, *A filtering approach to tracking volatility from prices observed at random times*, Ann. Appl. Probab., 16 (2005), pp. 1633–1652.

[2] J. Cvitanić and I. Karatzas, *Convex duality in constrained portfolio optimization*, Ann. Appl. Probab., 2 (1992), pp. 767–818.

[3] J. Cvitanić and I. Karatzas, *Hedging contingent claims with constrained portfolios*, Ann. Appl. Probab., 3 (1993), pp. 652–681.

[4] R. J. Elliott, *Exact adaptive filter for Markov chains observed in Gaussian noise*, Automatica J. IFAC, 20 (1994), pp. 1399–1408.

[5] R. J. Elliott, W. C. Hunter, and B. M. Jamieson, *Drift and volatility estimation in discrete time*, J. Econom. Dynam. Control, 22 (1998), pp. 209–218.

[6] R. J. Elliott and R. W. Rishel, *Estimating the implicit interest rate of a risky asset*, Stochastic Process. Appl., 49 (1994), pp. 199–206.

[7] R. J. Elliott, *New finite-dimensional filters and smoothers for noisily observed Markov chains*, IEEE Trans. Inform. Theory, 39 (1993), pp. 265–271.

[8] R. Frey, *Derivative asset analysis in models with level-dependent and stochastic volatility*, CWI Quarterly, 10 (1997), pp. 1–34.

[9] R. Frey and W. J. Runggaldier, *A nonlinear filtering approach to volatility estimation with a view towards high frequency data*, Internat. J. Theoret. Appl. Finance, 4 (2001), pp. 199–210.

[10] V. Henderson and D. Hobson, *Utility indifference pricing—an overview*, in Volume on Indifference Pricing, R. Carmona, ed., Princeton University Press, Princeton, NJ, 2007, to appear.

[11] M. R. James, V. Krishnamurthy, and F. L. Gland , *Time discretization of continuous-time filters and smoothers for HMM parameter estimation*, IEEE Trans. Inform. Theory, 42 (1996), pp. 593–605.

[12] P. Lakner, *Optimal trading strategy for an investor: The case of partial information*, Stochastic Process. Appl., 76 (1998), pp. 77–97.

[13] P. Lakner, *Utility maximization with partial information*, Stochastic Process. Appl., 56 (1995), pp. 247–273.

[14] H. Nagai and S. Peng, *Risk-senstive dynamic portfolio optimization with partial information on infinite time horizon*, Ann. Appl. Probab., 12 (2002), pp. 173–195.

[15] H. Pham and M. C. Quenez, *Optimal portfolio in partially observed stochastic volatility models*, Ann. Appl. Probab., 11 (2001), pp. 210–238.

[16] W. J. Runggaldier, *Estimation via stochastic filtering in financial market models*, in Mathematics of Finance, G. Yin and Q. Zhang, eds., Contemp. Math. 351, AMS, Providence, RI, 2004, pp. 309–318.
[17] L. C. G. Rogers, *The relaxed investor and parameter uncertainty*, Finance Stoch., 5 (2001), pp. 131–154.
[18] J. Sass and U. G. Haussmann, *Optimizing the terminal wealth under partial information: The drift process as a continuous time Markov chain*, Finance Stoch., 8 (2004), pp. 553–577.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1483–1517

A SMALL-GAIN THEOREM FOR A WIDE CLASS OF FEEDBACK SYSTEMS WITH CONTROL APPLICATIONS*

IASSON KARAFYLLIS† AND ZHONG-PING JIANG‡

Abstract. A small-gain theorem, which can be applied to a wide class of systems that includes systems satisfying the weak semigroup property, is presented in the present work. The result generalizes all existing results in the literature and exploits notions of weighted, uniform, and nonuniform input-to-output stability properties. Applications to partial state feedback stabilization problems with sampled-data feedback applied with zero order hold and positive sampling rate are also presented.

Key words. input-to-output stability, control systems, small-gain theorem

AMS subject classifications. 37C75, 37N35, 93A10, 93B52, 93C10, 93C25, 93D25

DOI. 10.1137/060669310

1. Introduction.

A common feature of stability analysis for complex interconnected systems is the application of small-gain results. Small-gain theorems for continuous-time finite-dimensional systems expressed in terms of "nonlinear gain functions" have a long history (see [14, 32, 51, 52] and the references therein). A nonlinear small-gain result was presented in [14], which allowed numerous applications to feedback stabilization problems. The methodology presented in [14] was followed by many researchers (see [15, 16, 22, 24, 47, 50]). A common characteristic of current research on nonlinear small-gain results in mathematical systems theory is the use of the notion of uniform input-to-state stability (ISS), introduced by Sontag in [44] for systems described by ordinary differential equations, or the notion of uniform input-to-output stability (IOS), introduced by Sontag and Wang in [46] (also see [14]) and extended in [10]. Small-gain theorems for discrete-time systems can be found in [17, 18, 19].

Extensions of small-gain results were presented recently in the literature. In [12, 13] less conservative small-gain conditions were presented for finite-dimensional systems. In [2, 3] matrix gain functions were used for the study of large scale finite-dimensional systems. A nonuniform in time small-gain theorem for continuous-time finite-dimensional systems was presented in [22]. Moreover, in [24, 47] small-gain results for wide classes of systems were provided. The classes of systems considered in [24, 47] satisfy the classical semigroup property (see [24, 25, 26, 45]). Small-gain results for hybrid systems satisfying the classical semigroup property were recently presented in [30].

An important feature of certain hybrid systems is that they do not satisfy the classical semigroup property: For example, the solution $x(t)$ of a system Σ with initial condition $x(t_0) = x_0$ does not coincide (in general) for $t \geq t_1 > t_0$ with the solution $\tilde{x}(t)$ of Σ with initial condition $\tilde{x}(t_1) = x(t_1)$. Such systems arise when sampled-data feedback laws are applied to finite-dimensional control systems or when numerical

*Received by the editors September 7, 2006; accepted for publication (in revised form) June 22, 2007; published electronically September 28, 2007. This work is supported partly by the U.S. NSF under grants ECS-0093176, OISE-0408925, and DMS-0504462.

http://www.siam.org/journals/sicon/46-4/66931.html

†Department of Environmental Engineering, Technical University of Crete, 73100, Chania, Greece (ikarafyl@enveng.tuc.gr).

‡Corresponding author. Department of Electrical and Computer Engineering, Polytechnic University, Six Metrotech Center, Brooklyn, NY 11201 (zjiang@poly.edu).

discretization schemes are applied for the numerical solution of a system of ordinary differential equations. However, from a mathematical point of view, these structures cannot be considered as "systems" in the sense given in [20, 24, 45]. This feature has important consequences, since the researcher cannot use the tools developed for systems theory and mathematical control theory. In [25, 26] the notion of a system was relaxed so that the semigroup property does not hold in a strict sense. Moreover, the modification introduced allows the results obtained in [24] to hold. Thus we are in a position to develop a complete stability theory, which covers the systems that satisfy the so-called "weak semigroup property."

The purpose of the present work is to present a small-gain result (Theorem 3.1 and Corollary 3.4), which

- ∗ can be applied to a very general class of systems (including systems that do not satisfy the classical semigroup property),
- ∗ unifies all existing results, which make use of uniform or nonuniform and weighted notions of ISS or IOS,
- ∗ can be used directly for the solution of sampled-data feedback stabilization problems or problems of numerical stability of discretization schemes, and
- ∗ can be applied to uncertain time-varying systems with vanishing or nonvanishing perturbations.

We believe that the main result of the present work is a valuable tool for establishing stability and will be used frequently in future research. However, we would like to emphasize the theoretical significance of our main result: It is a method for establishing qualitative properties expressed in a very general framework that unifies works from various fields as well as different stability notions. The results presented in the paper can be extended without much difficulty to the case of local stability notions.

The contents of this paper are presented as follows. In section 2 we provide the definitions of the notions used and several examples of systems that have the "boundedness-implies-continuation" (BIC) property. In section 3 the main result is stated and proved. In section 4, it is shown how the main result of the present work can be applied to an ISS stabilization problem of a certain class of systems with partial-state sampled-data feedback. It should be emphasized that sampled-data control systems cannot be handled with small-gain results that have appeared so far in the literature, since sampled-data control systems do not satisfy the classical semigroup property. Finally, section 5 contains the conclusions of the paper. The proofs of some basic results are given in the appendix.

Notations. Throughout this paper we adopt the following notations:

- ∗ We denote by K^+ the class of positive, continuous functions defined on $\Re^+$. We say that a function $\rho : \Re^+ \to \Re^+$ is of class $\mathcal{N}$, if ρ is continuous, nondecreasing, with $\rho(0) = 0$. By K we denote the set of positive definite, increasing and continuous functions. We say that a positive definite, increasing and continuous function $\rho : \Re^+ \to \Re^+$ is of class K_∞ if $\lim_{s\to+\infty} \rho(s) = +\infty$. By KL we denote the set of all continuous functions $\sigma = \sigma(s,t) : \Re^+ \times \Re^+ \to \Re^+$ with the properties: (i) for each $t \geq 0$ the mapping $\sigma(\cdot\,, t)$ is of class K; (ii) for each $s \geq 0$, the mapping $\sigma(s, \cdot\,)$ is nonincreasing with $\lim_{t\to+\infty} \sigma(s,t) = 0$.
- ∗ By $\|\ \|_{\mathcal{X}}$, we denote the norm of the normed linear space $\mathcal{X}$. By $|\ |$ we denote the Euclidean norm of $\Re^n$. Let $U \subseteq \mathcal{X}$, with $0 \in U$. By $B_U[0,r] := \{\, u \in U\ ;\ \|u\|_{\mathcal{X}} \leq r \,\}$ we denote the intersection of $U \subseteq \mathcal{X}$ with the closed sphere of radius $r \geq 0$, centered at $0 \in U$.

∗ Let a set U be a subset of a normed linear space $\mathcal{U}$, with $0 \in U$. By $\mathcal{M}(U)$ we denote the set of all locally bounded functions $u : \Re^+ \to U$. By u_0 we denote the identically zero input, i.e., the input that satisfies $u_0(t) = 0 \in U$ for all $t \geq 0$. If $U \subseteq \Re^n$, then $L^\infty_{loc}(\Re^+; U)$ denotes the space of measurable, locally bounded functions $u : \Re^+ \to U$.

The following convention will be adopted throughout the paper: The Cartesian product of two normed linear spaces $C := \mathcal{X} \times \mathcal{Y}$ will be considered to be endowed with the norm $\|(x, y)\|_C := \sqrt{\|x\|^2_{\mathcal{X}} + \|y\|^2_{\mathcal{Y}}}$, unless stated otherwise. Furthermore, our results can be extended to the case of measurable and locally essentially bounded inputs (where the "sup" operator is to be understood as "essential supermum").

2. Input-to-output stability in a system-theoretic framework. In this section we first give the notion of a control system with outputs. We emphasize that we consider control systems which do not necessarily satisfy the classical semigroup property (see [20, 24, 45]).

DEFINITION 2.1. *A control system* $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ *with outputs consists of*

(i) *a set* U *(control set) which is a subset of a normed linear space* $\mathcal{U}$*, with* $0 \in U$*, and a set* $M_U \subseteq \mathcal{M}(U)$ *(allowable control inputs) which contains at least the identically zero input* u_0*,*

(ii) *a set* D *(disturbance set) and a set* $M_D \subseteq \mathcal{M}(D)$*, which is called the "set of allowable disturbances,"*

(iii) *a pair of normed linear spaces* $\mathcal{X}, \mathcal{Y}$ *called the "state space" and the "output space," respectively,*

(iv) *a continuous map* $H : \Re^+ \times \mathcal{X} \times U \to \mathcal{Y}$ *that maps bounded sets of* $\Re^+ \times \mathcal{X} \times U$ *into bounded sets of* $\mathcal{Y}$*, called the "output map,"*

(v) *a set-valued map* $\Re^+ \times \mathcal{X} \times M_U \times M_D \ni (t_0, x_0, u, d) \to \pi(t_0, x_0, u, d) \subseteq [t_0, +\infty)$*, with* $t_0 \in \pi(t_0, x_0, u, d)$ *for all* $(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$*, called the set of "sampling times," and*

(vi) *the map* $\phi : A_\phi \to \mathcal{X}$*, where* $A_\phi \subseteq \Re^+ \times \Re^+ \times \mathcal{X} \times M_U \times M_D$*, called the "transition map," which has the following properties:*

(1) *Existence: For each* $(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$*, there exists* $t > t_0$ *such that* $[t_0, t] \times \{(t_0, x_0, u, d)\} \subseteq A_\phi$*.*

(2) *Identity property: For each* $(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$*, it holds that* $\phi(t_0, t_0, x_0, u, d) = x_0$*.*

(3) *Causality: For each* $(t, t_0, x_0, u, d) \in A_\phi$*, with* $t > t_0$*, and for each* $(\tilde{u}, \tilde{d}) \in M_U \times M_D$*, with* $(\tilde{u}(\tau), \tilde{d}(\tau)) = (u(\tau), d(\tau))$ *for all* $\tau \in [t_0, t]$*, it holds that* $(t, t_0, x_0, \tilde{u}, \tilde{d}) \in A_\phi$*, with* $\phi(t, t_0, x_0, u, d) = \phi(t, t_0, x_0, \tilde{u}, \tilde{d})$*.*

(4) *Weak semigroup property: There exists a constant* $r > 0$ *such that for each* $t \geq t_0$ *with* $(t, t_0, x_0, u, d) \in A_\phi$*:*

(a) $(\tau, t_0, x_0, u, d) \in A_\phi$ *for all* $\tau \in [t_0, t]$*;*

(b) $\phi(t, \tau, \phi(\tau, t_0, x_0, u, d), u, d) = \phi(t, t_0, x_0, u, d)$ *for all* $\tau \in [t_0, t] \cap \pi(t_0, x_0, u, d)$*;*

(c) *if* $(t+r, t_0, x_0, u, d) \in A_\phi$*, then it holds that* $\pi(t_0, x_0, u, d) \cap [t, t+r] \neq \emptyset$*;*

(d) *for all* $\tau \in \pi(t_0, x_0, u, d)$*, with* $(\tau, t_0, x_0, u, d) \in A_\phi$*, we have* $\pi(\tau, \phi(\tau, t_0, x_0, u, d), u, d) = \pi(t_0, x_0, u, d) \cap [\tau, +\infty)$*.*

In order to develop stability notions for a control system with outputs we need to clarify the notions of an equilibrium point as well as certain other important notions

and classes of systems that characterize the dynamic behavior of the system (see [24, 25]).

DEFINITION 2.2. *Let $T > 0$. A control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with outputs is called T-periodic, if:*

(a) *$H(t+T, x, u) = H(t, x, u)$ for all $(t, x, u) \in \Re^+ \times \mathcal{X} \times U$,*

(b) *for every $(u, d) \in M_U \times M_D$ and integer k there exist inputs $P_{kT}u \in M_U$, $P_{kT}d \in M_D$, with $(P_{kT}u)(t) = u(t + kT)$ and $(P_{kT}d)(t) = d(t + kT)$, for all $t + kT \geq 0$,*

(c) *for each $(t, t_0, x_0, u, d) \in A_\phi$, with $t \geq t_0$, and for each integer k, with $t_0 - kT \geq 0$ it follows that $(t - kT, t_0 - kT, x_0, P_{kT}u, P_{kT}d) \in A_\phi$ and $\pi(t_0 - kT, x_0, P_{kT}u, P_{kT}d) = \cup_{\tau \in \pi(t_0, x_0, u, d)}\{\tau - kT\}$, with $\phi(t, t_0, x_0, u, d) = \phi(t - kT, t_0 - kT, x_0, P_{kT}u, P_{kT}d)$.*

DEFINITION 2.3. *A control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with outputs is called time-invariant or autonomous, if it is T-periodic for all $T > 0$.*

DEFINITION 2.4. *Consider a control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with outputs. We say that system Σ*

(i) *has the BIC property if for each $(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$, there exists a maximal existence time, i.e., there exists $t_{\max} := t_{\max}(t_0, x_0, u, d) \in (t_0, +\infty]$, such that $A_\phi = \cup_{(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D}[t_0, t_{\max}) \times \{(t_0, x_0, u, d)\}$. In addition, if $t_{\max} < +\infty$, then for every $M > 0$ there exists $t \in [t_0, t_{\max})$, with $\|\phi(t, t_0, x_0, u, d)\|_{\mathcal{X}} > M$; and*

(ii) *is robustly forward complete (RFC) from the input $u \in M_U$ if it has the BIC property and for every $r \geq 0, T \geq 0$, it holds that*

$$\sup\{\|\phi(t_0 + s, t_0, x_0, u, d)\|_{\mathcal{X}}; \\ u \in \mathcal{M}(B_U[0, r]) \cap M_U, s \in [0, T], \|x_0\|_{\mathcal{X}} \leq r, t_0 \in [0, T], d \in M_D\} < +\infty.$$

DEFINITION 2.5. *Consider a control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$, and suppose that $H(t, 0, 0) = 0$ for all $t \geq 0$. We say that $0 \in \mathcal{X}$ is a robust equilibrium point from the input $u \in M_U$ for Σ if*

(i) *for every $(t, t_0, d) \in \Re^+ \times \Re^+ \times M_D$, with $t \geq t_0$, it holds that $\phi(t, t_0, 0, u_0, d) = 0$; and*

(ii) *for every $\varepsilon > 0$, $T, h \in \Re^+$ there exists $\delta := \delta(\varepsilon, T, h) > 0$ such that for all $(t_0, x, u) \in [0, T] \times \mathcal{X} \times M_U$, $\tau \in [t_0, t_0 + h]$, with $\|x\|_{\mathcal{X}} + \sup_{t \geq 0}\|u(t)\|_{\mathcal{U}} < \delta$, it holds that $(\tau, t_0, x, u, d) \in A_\phi$ for all $d \in M_D$ and*

$$\sup\{\|\phi(\tau, t_0, x, u, d)\|_{\mathcal{X}}; d \in M_D, \tau \in [t_0, t_0 + h], t_0 \in [0, T]\} < \varepsilon.$$

Remark 2.6. Consider a control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with the BIC property. It follows that Σ satisfies the (classical) semigroup property (see [24, 45]) if the weak semigroup property holds with $\pi(t_0, x_0, u, d) = [t_0, t_{\max})$, where $t_{\max} \in (t_0, +\infty]$ is the maximal existence time of the transition map for Σ that corresponds to $(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$, i.e.,

"for each $t \in [t_0, t_{\max})$ it holds that
$\phi(t, \tau, \phi(\tau, t_0, x_0, u, d), u, d) = \phi(t, t_0, x_0, u, d)$ for all $\tau \in [t_0, t]$"
(classical semigroup property).

The following example shows the difference between the classical semigroup property and the weak semigroup property for simple systems.

Example 2.7. Consider the following system:

$$\begin{array}{l} \dot{x}(t) = -x(\tau_i)\,,\ t \in [\tau_i, \tau_{i+1}), \\ \tau_{i+1} = \tau_i + 1, \\ x(t) \in \Re, \end{array} \tag{2.1}$$

with initial condition $x(t_0) = x_0 \in \Re$ and $\tau_0 = t_0 \geq 0$. Such systems will be characterized as hybrid systems with a sampling partition generated by the system (see Example 2.11), and they satisfy the BIC property. In this case we can determine analytically the transition map for all $t \geq t_0$ (u, d in this example are irrelevant):

$$\phi(t, t_0, x_0) = \begin{cases} (1 - t + t_0)x_0 & \text{for } t \in [t_0, t_0 + 1), \\ 0 & \text{for } t \geq t_0 + 1. \end{cases}$$

It is clear that the state space is $\Re$ and that the classical semigroup property does not hold for this system. On the other hand, the weak semigroup property holds for this system with $\pi(t_0, x_0) = \{t_0, t_0 + 1, t_0 + 2, \ldots\}$. Notice that the set of sampling times (the sampling partition) $\pi(t_0, x_0) = \{t_0, t_0 + 1, t_0 + 2, \ldots\}$ is generated by the system itself and depends on the initial condition. Furthermore, according to Definition 2.3, system (2.1) is autonomous.

Next, consider the following system:

$$\begin{array}{l} \dot{x}(t) = -x(\tau_i)\,,\ t \in [\tau_i, \tau_{i+1}), \\ x(t) \in \Re\,,\ \pi = \{\tau_i\}_{i=0}^{\infty} = \{0, 1, 2, \ldots\}. \end{array} \tag{2.2}$$

Such systems will be characterized as hybrid systems with impulses at fixed times (see Example 2.12), and they satisfy the BIC property. Notice that if the initial time t_0 is not a member of the partition $\pi = \{\tau_i\}_{i=0}^{\infty} = \{0, 1, 2, \ldots\}$, then it is not possible to determine the solution of (2.2) based only on the initial condition $x(t_0) = x_0 \in \Re$ and the transition map is not well-defined. In order to be able to determine the solution of (2.2), we need to know $x(t_0), x([t_0]) = x_0 = (x_{1,0}, x_{2,0}) \in \Re^2$ (where $[t_0]$ denotes the integer part of t_0). Indeed, we have

$$x(t) = \begin{cases} x_{1,0} - (t - t_0)x_{2,0}, & t_0 \leq t < [t_0] + 1, \\ (2 - t + [t_0])\,(x_{1,0} - ([t_0] + 1 - t_0)x_{2,0})\,, & [t_0] + 1 \leq t < [t_0] + 2, \text{ if } t_0 \notin \pi, \\ 0, & t \geq [t_0] + 2, \end{cases}$$

$$x(t) = \begin{cases} (1 - t + t_0)x_{1,0} & \text{for } t \in [t_0, t_0 + 1), \\ 0 & \text{for } t \geq t_0 + 1, \end{cases} \quad \text{if } t_0 \in \pi.$$

In this case the state space is $\Re^2$, and the state of (2.2) at time $t \geq t_0$ is $\phi(t, t_0, x_0) = (x(t), x([t])) \in \Re^2$. Furthermore, notice that the classical semigroup property holds and that the partition $\pi = \{\tau_i\}_{i=0}^{\infty} = \{0, 1, 2, \ldots\}$ is fixed and does not depend on the initial condition. Finally, according to Definitions 2.2 and 2.3, system (2.2) is T-periodic, with $T = 1$, but it is not autonomous.

It should be emphasized that there are systems which do not satisfy the weak semigroup property (e.g., systems described by integrodifferential equations studied in [29], such as $\dot{x}(t) = -x(t) + \int_{t_0}^{t} \sin(tx(s))\, ds$, $x(t) \in \Re$, with initial condition $x(t_0) = x_0 \in \Re$). However, many classes of systems used in physics and engineering satisfy the weak semigroup property and the BIC property and have a robust equilibrium point. The following examples provide classes of control systems which satisfy the weak semigroup property and the BIC property and possess a robust equilibrium point.

The examples help the reader to understand that the notions defined by Definitions 2.4–2.5 are typical for a wide class of systems under minimal assumptions.

Example 2.8 (finite-dimensional control systems described by ordinary differential equations (ODEs)). Consider the class of systems described by ODEs of the form

$$\begin{aligned} &\dot{x}(t) = f(t, x(t), u(t), d(t)), \\ &Y(t) = H(t, x(t), u(t)), \\ &x(t) \in \Re^n\,,\ u(t) \in U\,,\ d(t) \in D\,,\ t \geq t_0, \end{aligned} \tag{2.3}$$

where $U \subseteq \Re^m$, $D \subseteq \Re^l$, with $0 \in U$, and $f : \Re^+ \times \Re^n \times U \times D \to \Re^n$, $H : \Re^+ \times \Re^n \times U \to \Re^k$ are two locally bounded mappings, with $H(t, 0, 0) = 0$, $f(t, 0, 0, d) = 0$ for all $(t, d) \in \Re^+ \times D$, that satisfy the following hypotheses.

(A1) The mapping $(x, u, d) \to f(t, x, u, d)$ is continuous for each fixed $t \geq 0$, measurable with respect to $t \geq 0$ for each fixed $(x, u, d) \in \Re^n \times U \times D$ and such that for every pair of bounded sets $I \subseteq \Re^+$, $S \subset \Re^n \times U$, there exists a constant $L \geq 0$ such that

$$\begin{aligned} &(x - y)'\,(f(t, x, u, d) - f(t, y, u, d)) \leq L\,|x - y|^2 \\ &\forall t \in I\,,\ \forall (x, u, y, u) \in S \times S\,,\ \forall d \in D. \end{aligned}$$

(A2) The mapping $H : \Re^+ \times \Re^n \times U \to \Re^k$ is continuous.

(A3) There exist functions $\gamma \in K^+$, $a \in K_\infty$ such that $|f(t, x, u, d)| \leq \gamma(t)a(|x| + |u|)$ for all $(t, x, u, d) \in \Re^+ \times \Re^n \times U \times D$.

The theory of ordinary differential equations guarantees that, under hypotheses (A1)–(A3), for each $(t_0, x_0) \in \Re^+ \times \Re^n$ and for each pair of measurable and locally bounded inputs $(u, d) \in \mathcal{M}(U) \times \mathcal{M}(D)$, there exists a unique absolutely continuous mapping $x(t)$ that satisfies a.e. the differential equation (2.3) with initial condition $x(t_0) = x_0 \in \Re^n$. Moreover, certain results from the theory of ordinary differential equations guarantee that (2.3) is a control system $\Sigma := (\Re^n, \Re^k, M_U, M_D, \phi, \pi, H)$ with outputs that satisfies the BIC property with M_U, M_D the sets of all measurable and locally bounded mappings $u : \Re^+ \to U$, $d : \Re^+ \to D$, respectively. Furthermore, the classical semigroup property is satisfied for this system; i.e., we have $\pi(t_0, x_0, u, d) = [t_0, t_{\max})$, where $t_{\max} > t_0$ is the maximal existence time of the solution. Finally, hypotheses (A1)–(A3) guarantee that $0 \in \Re^n$ is a robust equilibrium point from the input $u \in M_U$ for Σ.

The following example presents a class of neutral functional equations described by continuous-time difference equations. Such systems were recently studied in [26, 42]. The importance of functional difference equations in applications is explained in [42].

Example 2.9 (control systems described by functional difference equations (FDEs)). Consider the class of systems described by FDEs of the form

$$\begin{aligned} &x(t) = f(t, T_{r-\tau(t)}(t - \tau(t))x, u(t), d(t)), \\ &Y(t) = H(t, T_r(t)x, u(t)), \\ &x(t) \in \Re^n\,,\ Y(t) \in Y\,,\ u(t) \in U\,,\ d(t) \in D\,,\ t \geq t_0, \end{aligned} \tag{2.4}$$

where $r > 0$ is a constant, $\tau : \Re^+ \to (0, +\infty)$ is a positive continuous function, with $\sup_{t \geq 0} \tau(t) \leq r$, $D \subset \Re^l$, $U \subseteq \Re^m$, with $0 \in U$, are nonempty sets, $T_{r-\tau(t)}(t - \tau(t))x := x(t - \tau(t) + \theta); \theta \in [-r + \tau(t), 0]$, $T_r(t)x := x(t + \theta); \theta \in [-r, 0]$ and H, $f : \Omega \times U \times D \to \Re^n$, where $\Omega = \cup_{t \geq 0}\{t\} \times \mathcal{F}_t$ and $\mathcal{F}_t$ denotes the set of bounded functions $x : [-r + \tau(t), 0] \to \Re^n$, are locally bounded mappings which satisfy the following hypotheses.

(R1) There exist functions $\gamma \in K^+$, $a \in K_\infty$ such that $|f(t, T_{r-\tau(t)}(-\tau(t))x, u, d)| \leq \gamma(t)a(\sup_{\theta\in[-r,-\tau(t)]} |x(\theta)| + |u|)$ for all $(t, x, u, d) \in \Re^+ \times \mathcal{X} \times U \times D$, where $\mathcal{X}$ is the normed linear space of the bounded functions $x : [-r, 0] \to \Re^n$, with $\|x\|_{\mathcal{X}} := \sup_{\theta\in[-r,0]} |x(\theta)|$.

(R2) The output map $H : \Re^+ \times \mathcal{X} \times U \to \mathcal{Y}$, where $\mathcal{Y}$ is a normed linear space, is a continuous mapping that maps bounded sets of $\Re^+ \times \mathcal{X} \times U$ into bounded sets of $\mathcal{Y}$ with $H(t, 0, 0) = 0$ for all $t \geq 0$.

It should be clear that, under the hypotheses stated above, for each $(t_0, x_0) \in \Re^+ \times \mathcal{X}$ and for each pair of locally bounded functions $u : \Re^+ \to U$, $d : \Re^+ \to D$, there exists a unique locally bounded mapping $x(t)$ that satisfies the difference equations (2.4) with initial condition $x(t_0+\theta) = x_0(\theta); \theta \in [-r, 0]$. Consequently, (2.4) describes a control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U M_D, \phi, \pi, H)$, with outputs and evolution map ϕ defined by $\phi(t, t_0, x_0, u, d) = x(t + \theta); \theta \in [-r, 0]$, where $\mathcal{U} := \Re^m$, M_U the set of all locally bounded functions $u : \Re^+ \to U$ and M_D the set of all functions $d : \Re^+ \to D$.

Systems described by functional difference evolution equations of the form (2.4) are considered in [6, 26, 42]. Working exactly in the same way as in [26], it can be shown that system (2.4) is RFC from the input $u \in M_U$ and that $0 \in \mathcal{X}$ is a robust equilibrium point from the input $u \in M_U$ for system (2.4).

Notice that a major advantage of allowing the output to take values in abstract normed linear spaces is that we are in a position to consider:

- outputs with no delays, e.g., $Y(t) = h(t, x(t), u(t))$, with $\mathcal{Y} = \Re^k$,
- outputs with discrete or distributed delay, e.g., $Y(t) = h(t, x(t), x(t-r), u(t))$ or $Y(t) = \sup_{\theta\in[t-r,t]} h(t, \theta, x(\theta), u(t))$, with $\mathcal{Y} = \Re^k$, and
- functional outputs with memory, e.g., $Y(t) = h(t, \theta, x(t + \theta)); \theta \in [-r, 0]$ or the identity output $Y(t) = T_r(t)x = x(t + \theta); \theta \in [-r, 0]$, with $\mathcal{Y} = \mathcal{X}$.

Finally, notice that the classical semigroup property is satisfied for this system; i.e., we have $\pi(t_0, x_0, u, d) = [t_0, +\infty)$.

The following example is an immediate consequence of Theorems 2.2 and 3.2 in [6], concerning continuous dependence on initial conditions and continuation of solutions of retarded functional differential equations, respectively.

Example 2.10 (control systems described by retarded functional differential equations (RFDEs)). Consider the class of systems described by RFDEs of the form

$$\begin{aligned} &\dot{x}(t) = f(t, T_r(t)x, u(t), d(t)), \\ &Y(t) = H(t, T_r(t)x, u(t)), \\ &x(t) \in \Re^n\,,\ u(t) \in U\,,\ d(t) \in D\,,\ t \geq t_0, \end{aligned} \tag{2.5}$$

where $T_r(t)x := x(t+\theta); \theta \in [-r, 0]$, $D \subseteq \Re^l$ is a nonempty set, $U \subseteq \Re^m$ is a nonempty set, with $0 \in U$, $f : \Re^+ \times C^0([-r, 0]; \Re^n) \times U \times D \to \Re^n$, $H : \Re^+ \times C^0([-r, 0]; \Re^n) \times U \to \mathcal{Y}$ ($\mathcal{Y}$ is a normed linear space) are locally bounded mappings, with $f(t, 0, 0, d) = 0$, $H(t, 0, 0) = 0$ for all $(t, d) \in \Re^+ \times D$, that satisfy the following hypotheses.

(S1) The mapping $(x, u, d) \to f(t, x, u, d)$ is continuous for each fixed $t \geq 0$ and such that, for every bounded $I \subseteq \Re^+$ and for every bounded $S \subset C^0([-r, 0]; \Re^n) \times U$, there exists a constant $L \geq 0$ such that

$$\begin{aligned} &(x(0) - y(0))' \left(f(t, x, u, d) - f(t, y, u, d)\right) \leq L \max_{\tau\in[-r,0]} |x(\tau) - y(\tau)|^2 \\ &\forall t \in I, \forall (x, u, y, u) \in S \times S, \forall d \in D. \end{aligned}$$

(S2) There exist functions $\gamma \in K^+$, $a \in K_\infty$ such that $|f(t, x, u, d)| \leq \gamma(t)a(\|x\|_r + |u|)$ for all $(t, x, u, d) \in \Re^+ \times C^0([-r, 0]; \Re^n) \times U \times D$, where $\|x\|_r$ denotes the sup-norm of the space $C^0([-r, 0]; \Re^n)$, i.e., $\|x\|_r := \max_{\theta\in[-r,0]} |x(\theta)|$.

(S3) There exists a countable set $A \subset \Re^+$, which is either finite or $A = \{t_k; k = 1, \dots, \infty\}$, with $t_{k+1} > t_k > 0$ for all $k = 1, 2, \dots$ and $\lim t_k = +\infty$, such that the mapping $(t, x, u, d) \in (\Re^+ \backslash A) \times C^0([-r, 0]; \Re^n) \times U \times D \to f(t, x, u, d)$ is continuous. Moreover, for each fixed $(t_0, x, u, d) \in \Re^+ \times C^0([-r, 0]; \Re^n) \times U \times D$, we have $\lim_{t \to t_0^+} f(t, x, u, d) = f(t_0, x, u, d)$.

(S4) The mapping $H : \Re^+ \times C^0([-r, 0]; \Re^n) \times U \to \mathcal{Y}$ is a continuous mapping that maps bounded sets of $\Re^+ \times C^0([-r, 0]; \Re^n) \times U$ into bounded sets of $\mathcal{Y}$.

The theory of retarded functional differential equations guarantees that under hypotheses (S1)–(S4), for each $(t_0, x_0) \in \Re^+ \times C^0([-r, 0]; \Re^n)$ and for each pair of measurable and locally bounded inputs $(u, d) \in \mathcal{M}(U) \times \mathcal{M}(D)$, there exists a unique absolutely continuous mapping $x(t)$ that satisfies a.e. the differential equation (2.5) with initial condition $x(t_0) = x_0 \in C^0([-r, 0]; \Re^n)$. Moreover, certain results from the theory of retarded functional differential equations (Theorems 2.2 and 3.2 in [6]) guarantee that (2.5) is a control system $\Sigma := (C^0([-r, 0]; \Re^n), \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with outputs that satisfies the BIC property with M_U, M_D the sets of all measurable and locally bounded mappings $u : \Re^+ \to U$, $d : \Re^+ \to D$, respectively. Furthermore, the classical semigroup property is satisfied for this system; i.e., we have $\pi(t_0, x_0, u, d) = [t_0, t_{\max})$, where $t_{\max} > t_0$ is the maximal existence time of the solution. Finally, hypotheses (S1)–(S4) guarantee that $0 \in C^0([-r, 0]; \Re^n)$ is a robust equilibrium point from the input $u \in M_U$ for Σ. Again notice that a major advantage of allowing the output to take values in abstract normed linear spaces is that we are in a position to consider various output cases (see previous example).

The following example presents an important class of systems that does not satisfy the classical semigroup property.

Example 2.11 (hybrid systems with sampling partition generated by the system). Consider the class of systems described by impulsive differential equations of the form

$$\begin{aligned} &\dot{x}(t) = f(t, \tau_i, x(t), x(\tau_i), u(t), u(\tau_i), d(t), d(\tau_i)), \quad t \in [\tau_i, \tau_{i+1}), \\ &\tau_0 = t_0,\ \tau_{i+1} = \tau_i + h(\tau_i, x(\tau_i), u(\tau_i), d(\tau_i))\,,\ i = 0, 1, \dots, \\ &x(\tau_{i+1}) = R\left(\tau_i, \lim_{t \to \tau_{i+1}^-} x(t), x(\tau_i), u(\tau_{i+1}), u(\tau_i), d(\tau_{i+1}), d(\tau_i)\right), \\ &Y(t) = H(t, x(t), u(t)), \end{aligned} \tag{2.6}$$

where $D \subseteq \Re^l$, $U \subseteq \Re^m$ is a closed set, with $0 \in U$, $h : \Re^+ \times \Re^n \times U \times D \to (0, r]$ is a positive function which is bounded by a certain constant $r > 0$, $f : \Re^+ \times \Re^+ \times \Re^n \times \Re^n \times U \times U \times D \times D \to \Re^n$, $H : \Re^+ \times \Re^n \times U \to \Re^p$, and $R : \Re^+ \times \Re^n \times \Re^n \times U \times U \times D \times D \to \Re^n$ is a triplet of vector fields that satisfy the following hypotheses.

(P1) $f(t, \tau, x, x_0, u, u_0, d, d_0)$ is measurable with respect to $t \geq 0$, continuous with respect to $(x, d, u) \in \Re^n \times D \times U$, and such that for every bounded $S \subset \Re^+ \times \Re^+ \times \Re^n \times \Re^n \times U \times U$ there exists a constant $L \geq 0$ such that

$$\begin{aligned} &(x - y)' \left(f(t, \tau, x, x_0, u, u_0, d, d_0) - f(t, \tau, y, x_0, u, u_0, d, d_0)\right) \leq L\, |x - y|^2 \\ &\forall (t, \tau, x, x_0, u, u_0, d, d_0) \in S \times D \times D\ ,\ \forall (t, \tau, y, x_0, u, u_0, d, d_0) \in S \times D \times D. \end{aligned}$$

(P2) There exist functions $\gamma \in K^+$, $a \in K_\infty$ such that

$$\begin{aligned} &|f(t, \tau, x, x_0, u, u_0, d, d_0)| \leq \gamma(t)\, a\left(|x| + |x_0| + |u| + |u_0|\right) \\ &\qquad \forall (\tau, u, u_0, d, d_0, x, x_0) \in \Re^+ \times U \times U \times D \times D \times \Re^n \times \Re^n, \quad \forall t \geq \tau, \\ &|R(t, x, x_0, u, u_0, d, d_0)| \leq \gamma(t)\, a\left(|x| + |x_0| + |u| + |u_0|\right) \\ &\qquad \forall (t, u, u_0, d, d_0, x, x_0) \in \Re^+ \times U \times U \times D \times D \times \Re^n \times \Re^n. \end{aligned}$$

(P3) $H : \Re^+ \times \Re^n \times U \to \Re^p$ is a continuous map, with $H(t,0,0) = 0$ for all $t \geq 0$.

(P4) There exist a positive, continuous, and bounded function $h_l : \Re^+ \times \Re^n \times U \to (0, r]$ and a partition $\pi = \{T_i\}_{i=0}^{\infty}$ of $\Re^+$, i.e., an increasing sequence of times with $T_0 = 0$ and $T_i \to +\infty$, such that

$$h(t,x,u,d) \geq \min\{p_\pi(t) - t\,,\, h_l(t,x,u)\} \quad \forall (t,x,u,d) \in \Re^+ \times \Re^n \times U \times D,$$

where $p_\pi(t) := \min\{T \in \pi; t < T\}$.

Hybrid systems of the form (2.6) under hypotheses (P1)–(P4) are considered in [25, 26], where it is shown that, for each $(t_0, x_0) \in \Re^+ \times \Re^n$ and for each pair of measurable and locally bounded inputs $u : \Re^+ \to U$ and $d : \Re^+ \to D$, there exists a unique piecewise absolutely continuous function $t \to x(t) \in \Re^n$ with initial condition $x(t_0) = x_0$, which is produced by the following algorithm:

Step i:

1. Given τ_i and $x(\tau_i)$, calculate τ_{i+1} using the equation $\tau_{i+1} = \tau_i + h(\tau_i, x(\tau_i), u(\tau_i), d(\tau_i))$.
2. Compute the state trajectory $x(t)$, $t \in [\tau_i, \tau_{i+1})$, as the solution of the differential equation $\dot{x}(t) = f(t, \tau_i, x(t), x(\tau_i), u(t), u(\tau_i), d(t), d(\tau_i))$.
3. Calculate $x(\tau_{i+1})$ using the equation $x(\tau_{i+1}) = R(\tau_i, \lim_{t \to \tau_{i+1}^-} x(t), x(\tau_i), u(\tau_{i+1}), u(\tau_i), d(\tau_{i+1}), d(\tau_i))$.
4. Compute the output trajectory $Y(t)$, $t \in [\tau_i, \tau_{i+1}]$, using the equation $Y(t) = H(t, x(t), u(t))$.

For $i = 0$ we take $\tau_0 = t_0$ and $x(\tau_0) = x_0$ (initial condition).

In [25] it is shown that system (2.6) under hypotheses (P1)–(P4) is a control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with outputs with the BIC property for which $0 \in \Re^n$ is a robust equilibrium point from the input $u \in M_U$. Particularly, we have $\mathcal{X} = \Re^n$, $\mathcal{Y} = \Re^p$, $\mathcal{U} = \Re^m$, and M_U, M_D the sets of measurable and locally bounded inputs $u : \Re^+ \to U$ and $d : \Re^+ \to D$, respectively. The set $\pi(t_0, x_0, u, d) \subseteq [t_0, +\infty)$ involved in the weak semigroup property consists of the sequence $\pi = \{\tau_0, \tau_1, \ldots\}$ generated by the recursive relation $\tau_{i+1} = \tau_i + h(\tau_i, x(\tau_i), u(\tau_i), d(\tau_i))$, $i = 0, 1, \ldots$, with $\tau_0 = t_0$. Notice that the control system (2.6) fails to satisfy the classical semigroup property.

If $h(\tau+T, x, u, d) = h(\tau, x, u, d)$, $f(t+T, \tau+T, x, x_0, u, u_0, d, d_0) = f(t, \tau, x, x_0, u, u_0, d, d_0)$, $R(\tau+T, x, x_0, u, u_0, d, d_0) = R(\tau, x, x_0, u, u_0, d, d_0)$, and $H(t+T, x, u) = H(t, x, u)$ for certain $T > 0$ and for $(t, \tau, u, u_0, d, d_0, x, x_0) \in \Re^+ \times \Re^+ \times U \times U \times D \times D \times \Re^n \times \Re^n$, with $t \geq \tau$, then system (2.6) is *T-periodic*. Moreover, if $h(\tau, x, u, d) = h(x, u, d)$, $f(t, \tau, x, x_0, u, u_0, d, d_0) = f(t-\tau, x, x_0, u, u_0, d, d_0)$, $R(\tau, x, x_0, u, u_0, d, d_0) = R(x, x_0, u, u_0, d, d_0)$, and $H(t, x, u) = H(x, u)$ for $(t, \tau, u, u_0, d, d_0, x, x_0) \in \Re^+ \times \Re^+ \times U \times U \times D \times D \times \Re^n \times \Re^n$, with $t \geq \tau$, then system (2.6) is autonomous.

Systems of the form (2.6) under hypotheses (P1)–(P4) arise frequently in certain applications in mathematical control theory and numerical analysis. Specifically, they arise when

(i) a (not necessarily continuous) sampled-data feedback law (with a possibly variable sampling rate) is applied to a finite-dimensional control system. For example, state-dependent sampling rates were related in [4] with the classical work on discontinuous stabilizability in [1], while feedback stabilization problems with zero order hold and a constant positive sampling rate were considered in [33, 34, 35, 36, 37, 38, 40] and time-varying sampling rates were considered in [8, 9],

(ii) a synchronous controller switching strategy is applied to a finite-dimensional control system (see [31, 43]), and

(iii) a numerical discretization method (with possibly variable integration step sizes) is applied in order to obtain the numerical solution of a given system of ordinary differential equations; see [5, 48] for the case of constant integration step sizes and [23, 25] for the case of variable integration step sizes.

For a unified description of the above problems, see [25, 26].

In contrast with the previous example, it should be noted that hybrid systems with impulses at fixed times satisfy the classical semigroup property. The following example illustrates this case.

Example 2.12 (hybrid systems with impulses at fixed times). Consider the class of systems described by impulsive differential equations of the form

$$\begin{aligned} &\dot{x}(t) = f\left(t, d(t), d\left(\tau_i\right), x(t), x\left(\tau_i\right), u(t), u\left(\tau_i\right)\right), \qquad \tau_i \le t < \tau_{i+1}, \\ &x(\tau_{i+1}) = R\left(\tau_i, \lim_{t \to \tau_{i+1}^-} x(t), x(\tau_i), u(\tau_{i+1}), u(\tau_i), d(\tau_{i+1}), d(\tau_i)\right), \\ &Y(t) = H(t, x(t), u(t)), \\ &x(t) \in \Re^n \ , \ Y(t) \in \Re^k \ , \ u(t) \in V \subseteq \Re^m \ , \ t \ge 0 \ , \ d(t) \in D, \end{aligned} \tag{2.7}$$

where $D \subseteq \Re^l$, $V \subseteq \Re^m$ is a closed set, with $0 \in V$, $\pi = \{\tau_i\}_{i=0}^{\infty}$ is a partition of $\Re^+$with diameter $r > 0$, i.e., an increasing sequence of times with $\tau_0 = 0$, $\sup\{\tau_{i+1} - \tau_i; i = 0, 1, 2, \ldots\} = r$, and $\tau_i \to +\infty$, $d(t)$ represents the disturbance vector or the vector of time-varying uncertainties taking values in the set $D \subset \Re^l$, $Y(t)$ represents the output of the system, and $u(t) \in V$ represents the input vector. A wide class of systems described by impulsive differential equations with impulses at fixed times, as well as hybrid systems of the form:

$$\begin{aligned} &\dot{x}(t) = f(t, x(t), u(t), w(i)), \qquad \tau_i \le t < \tau_{i+1}, \\ &w(i) = g(i, x(\tau_i), u(\tau_i)), \end{aligned} \tag{2.8}$$

where $\pi = \{\tau_i\}_{i=0}^{\infty}$ is a partition of $\Re^+$ of diameter $r > 0$, can be represented by the time-varying case (2.7). Fundamental properties of the solutions of systems of the form (2.8) are studied in [28, 29].

Consider system (2.7) under the following assumptions.

(Q1) $\pi = \{\tau_i\}_{i=0}^{\infty}$ is a partition of $\Re^+$ with finite diameter $r > 0$, i.e., an increasing sequence of times with $\tau_0 = 0$, $\sup\{\tau_{i+1} - \tau_i; i = 0, 1, 2, \ldots\} = r$, and $\tau_i \to +\infty$.

(Q2) $H : \Re^+ \times \Re^n \times V \to \Re^k$ is continuous, with $H(t, 0, 0) = 0$, for all $t \ge 0$.

(Q3) $f(t, d, d_0, x, x_0, u, u_0)$ is measurable with respect to $t \ge 0$, continuous with respect to $(x, d, u) \in \Re^n \times D \times V$, and such that, for every compact $S \subset \Re^n \times \Re^n \times V \times V$ and for every compact $I \subset \Re^+$, there exists a constant $L \ge 0$ such that

$$\begin{aligned} &(x - y)' \left(f(t, d, d_0, x, x_0, u, u_0) - f(t, d, d_0, y, x_0, u, u_0)\right) \le L \left|x - y\right|^2 \\ &\forall t \in I \ , \ \forall (d, d_0) \in D \times D \ , \ \forall (x, x_0, u, u_0) \in S \ , \ \forall (y, x_0, u, u_0) \in S. \end{aligned}$$

(Q4) There exist functions $\gamma \in K^+$, $a \in K_\infty$ such that $|f(t, d, d_0, x, x_0, u, u_0)| \le \gamma(t) a(|x_0| + |x| + |u| + |u_0|)$, $|R(t, x, x_0, u, u_0, d, d_0)| \le \gamma(t) a(|x| + |x_0| + |u| + |u_0|)$ for all $(t, d, d_0, x, x_0, u, u_0) \in \Re^+ \times D \times D \times \Re^n \times \Re^n \times V \times V$.

Systems of the form (2.7) with $R(t, x, x_0, u, u_0, d, d_0) \equiv x$ (impulse free case) were considered in [27]. Special classes of impulsive systems of the form (2.7) were studied in [7]. Using the method of steps on consecutive intervals, it is clear that system (2.7)

under hypotheses (Q1)–(Q4) defines a control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with outputs and the BIC property, with state space $\mathcal{X} = \Re^n \times \Re^n$, output space $\mathcal{Y} = \Re^k$, the set of structured uncertainties M_D being the set of mappings $t \in \Re^+ \to d(t) = \{\tilde{d}(t+\theta); \theta \in [-r, 0]\}$, where $\tilde{d} : \Re \to D$ is any measurable and locally bounded function, input space $\mathcal{U}$ the normed linear space of measurable and bounded functions on $[-r, 0]$ taking values in $\Re^m$ endowed with the sup-norm, $U \subseteq \mathcal{U}$ the set of measurable and bounded functions on $[-r, 0]$ taking values in $V \subseteq \Re^m$, and the set of external inputs M_U being the set of mappings $t \in \Re^+ \to u(t) = \{\tilde{u}(t+\theta); \theta \in [-r, 0]\} \in U$, where $\tilde{u} : \Re \to \Re^m$ is a measurable and locally bounded function. The reader may be surprised by the complicated definition of M_D and M_U, but it should be emphasized that this definition guarantees that the causality property of the control system (2.7) holds. Notice that the classical semigroup property is satisfied for this system; i.e., we have $\pi(t_0, x_0, u, d) = [t_0, t_{\max})$, where $t_{\max} > t_0$ is the maximal existence time of the solution. However, notice that if the vector fields f and R are independent of $d(\tau_i), x(\tau_i), u(\tau_i)$ (this is the case studied in [7]), then system (2.7) under hypotheses (Q1)–(Q4) defines a control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with outputs and the BIC property, with state space $\mathcal{X} = \Re^n$, output space $\mathcal{Y} = \Re^k$, the set of structured uncertainties M_D being the set of measurable and locally bounded functions $d : \Re \to D$, input space $\mathcal{U} = \Re^m$, and M_U being the set of measurable and locally bounded functions $u : \Re \to U$.

Let $q_\pi(t) = \max\{\tau_i; \tau_i \in \pi, \tau_i \leq t\}$. For all $(t_0, x_0, x_1, d, u) \in \Re^+ \times \Re^n \times \Re^n \times M_D \times M_U$, we denote by $x(t) = \phi(t, t_0, x_0, x_1; d, u) \in \Re^n$ the solution of (2.7) at time $t \geq t_0$ with initial condition $x(t_0) = x_0$ and the additional condition $x(q_\pi(t_0)) = x_1$, which holds only for the case $t_0 \notin \pi$, corresponding to inputs $(d, u) \in M_D \times M_U$ (this solution is unique by virtue of property (Q3)). Notice that the actual state of system (2.7) at time $t \geq t_0$ is given by $\tilde{\phi}(t, t_0, x_0, x_1; d, u) = (\phi(t, t_0, x_0, x_1; d, u), \phi(q_\pi(t), t_0, x_0, x_1; d, u)) \in \Re^n \times \Re^n$.

Hypotheses (Q3)–(Q4) can be used in order to show that $0 \in \Re^n \times \Re^n$ is a robust equilibrium point from the input $u \in M_U$, exactly in the same way with the proof of the analogous result in [27]. Notice that if $f(t+T, x, x_0, u, u_0, d, d_0) = f(t, x, x_0, u, u_0, d, d_0)$, $R(t+T, x, x_0, u, u_0, d, d_0) = R(t, x, x_0, u, u_0, d, d_0)$, $H(t+T, x, u) = H(t, x, u)$, and $\pi = \{iT\}_{i=0}^{\infty}$ for certain $T > 0$ and for all $(t, u, u_0, d, d_0, x, x_0) \in \Re^+ \times V \times V \times D \times D \times \Re^n \times \Re^n$, then system (2.7) is *T-periodic*. Moreover, it should be noted that system (2.7) fails to be autonomous for every possible selection of the sets D, V, vector fields f, R, H, and partition π.

For control systems with the BIC property the following lemma provides a useful characterization of the RFC property. Its proof is provided in the appendix.

LEMMA 2.13. *System $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ is RFC from the input $u \in M_U$ if and only if for every $\beta \in K^+$ there exist functions $\mu, c \in K^+, a, p \in K_\infty$ (depending only on $\beta \in K^+$) such that the following estimate holds for all $(t_0, x_0, d, u) \in \Re^+ \times \mathcal{X} \times M_D \times M_U$:*

$$\begin{aligned} &\beta(t)\, \|\phi(t, t_0, x_0, u, d)\|_{\mathcal{X}} \\ &\quad \leq \max\left\{ \mu(t - t_0)\,,\ c(t_0)\,,\ a\left(\|x_0\|_{\mathcal{X}}\right),\ \sup_{t_0 \leq \tau \leq t} p\left(\|u(\tau)\|_{\mathcal{U}}\right) \right\} \quad \forall t \geq t_0. \end{aligned} \tag{2.9}$$

Next we present the IOS property for the class of systems described by Definition 2.1.

DEFINITION 2.14. *Consider a control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with outputs and the BIC property and for which $0 \in \mathcal{X}$ is a robust equilibrium point from the input $u \in M_U$. Suppose that Σ is RFC from the input $u \in M_U$.*

- *If there exist functions $\sigma \in KL$, β, $\delta \in K^+$, $\gamma \in \mathcal{N}$ such that the following estimate holds for all $u \in M_U$, $(t_0, x_0, d) \in \Re^+ \times X \times M_D$, and $t \geq t_0$:*

$$\begin{aligned} &\|H(t, \phi(t, t_0, x_0, u, d), u(t))\|_{\mathcal{Y}} \\ &\quad \leq \sigma\left(\beta(t_0)\|x_0\|_{\mathcal{X}}, t - t_0\right) + \sup_{t_0 \leq \tau \leq t} \gamma\left(\delta(\tau)\|u(\tau)\|_{\mathcal{U}}\right), \end{aligned} \tag{2.10}$$

  then we say that Σ satisfies the weighted input-to-output stability (WIOS) property from the input $u \in M_U$, with gain $\gamma \in \mathcal{N}$ and weight $\delta \in K^+$. Moreover, if $\beta(t) \equiv 1$, then we say that Σ satisfies the uniform weighted input-to-output stability (UWIOS) property from the input $u \in M_U$, with gain $\gamma \in \mathcal{N}$ and weight $\delta \in K^+$.
- *If there exist functions $\sigma \in KL$, $\beta \in K^+$, $\gamma \in \mathcal{N}$ such that the following estimate holds for all $u \in M_U$, $(t_0, x_0, d) \in \Re^+ \times \mathcal{X} \times M_D$, and $t \geq t_0$:*

$$\|H(t, \phi(t, t_0, x_0, u, d), u(t))\|_{\mathcal{Y}} \leq \sigma\left(\beta(t_0)\|x_0\|_{\mathcal{X}}, t - t_0\right) + \sup_{t_0 \leq \tau \leq t} \gamma\left(\|u(\tau)\|_{\mathcal{U}}\right), \tag{2.11}$$

  then we say that Σ satisfies the IOS property from the input $u \in M_U$, with gain $\gamma \in \mathcal{N}$. Moreover, if $\beta(t) \equiv 1$, then we say that Σ satisfies the uniform input-to-output stability (UIOS) property from the input $u \in M_U$, with gain $\gamma \in \mathcal{N}$.

Finally, for the special case of the identity output mapping, i.e., $H(t, x, u) := x$, the (uniform) (weighted) input-to-output stability property from the input $u \in M_U$ is called the (uniform) (weighted) input-to-state stability ((U)(W)ISS) property from the input $u \in M_U$.

Remark 2.15. Using the inequalities $\max\{a, b\} \leq a + b \leq \max\left\{a + \rho(a), b + \rho^{-1}(b)\right\}$ (which hold for all $\rho \in K_\infty$ and $a, b \geq 0$), it should be clear that the WIOS property for $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ can be defined by using an estimate of the form

$$\begin{aligned} &\|H(t, \phi(t, t_0, x_0, u, d), u(t))\|_{\mathcal{Y}} \\ &\quad \leq \max\left\{\sigma\left(\beta(t_0)\|x_0\|_{\mathcal{X}}, t - t_0\right), \sup_{t_0 \leq \tau \leq t} \gamma\left(\delta(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right\} \end{aligned} \tag{2.10$'$}$$

instead of (2.10). Similarly, the IOS property for $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ can be defined by using an estimate of the form

$$\|H(t, \phi(t, t_0, x_0, u, d), u(t))\|_{\mathcal{Y}} \leq \max\left\{\sigma\left(\beta(t_0)\|x_0\|_{\mathcal{X}}, t - t_0\right), \sup_{t_0 \leq \tau \leq t} \gamma\left(\|u(\tau)\|_{\mathcal{U}}\right)\right\} \tag{2.11$'$}$$

instead of (2.11).

The following lemmas provide $\varepsilon - \delta$ characterizations of the WIOS and UWIOS properties, which are going to be used in the following section of the paper. Their proofs are provided in the appendix.

LEMMA 2.16. *Consider a control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with outputs and the BIC property and for which $0 \in \mathcal{X}$ is a robust equilibrium point from the input $u \in M_U$. Suppose that Σ is RFC from the input $u \in M_U$. Furthermore, suppose that there exist functions $V : \Re^+ \times \mathcal{X} \times U \to \Re^+$, with $V(t, 0, 0) = 0$, for all $t \geq 0$, $\gamma \in \mathcal{N}$, and $\delta \in K^+$ such that the following properties hold:*

P1. *For every $s \geq 0$, $T \geq 0$, it holds that*

$$\sup\left\{V(t,\phi(t,t_0,x_0,u,d),u(t)) - \sup_{t_0\leq\tau\leq t}\gamma\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right);\right.$$
$$\left. t\geq t_0, \left\|x_0\right\|_{\mathcal{X}} \leq s, t_0 \in [0,T], d\in M_D, u\in M_U\right\} < +\infty.$$

P2. *For every $\varepsilon > 0$ and $T \geq 0$, there exists a $\rho := \rho(\varepsilon, T) > 0$ such that*

$$\sup\left\{V(t,\phi(t,t_0,x_0,u,d),u(t)) - \sup_{t_0\leq\tau\leq t}\gamma\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right);\right.$$
$$\left. t\geq t_0, \left\|x_0\right\|_{\mathcal{X}} \leq \rho, t_0 \in [0,T], d\in M_D, u\in M_U\right\} \leq \varepsilon.$$

P3. *For every $\varepsilon > 0$, $T \geq 0$, and $R \geq 0$, there exists a $\tau := \tau(\varepsilon, T, R) \geq 0$ such that*

$$\sup\left\{V(t,\phi(t,t_0,x_0,u,d),u(t)) - \sup_{t_0\leq\tau\leq t}\gamma\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right);\right.$$
$$\left. t\geq t_0+\tau, \left\|x_0\right\|_{\mathcal{X}} \leq R, t_0 \in [0,T], d\in M_D, u\in M_U\right\} \leq \varepsilon.$$

Then there exist functions $\sigma \in KL$ and $\beta \in K^+$ such that the following estimate holds for all $u \in M_U$, $(t_0, x_0, d) \in \Re^+ \times \mathcal{X} \times M_D$, and $t \geq t_0$:

$$V(t,\phi(t,t_0,x_0,u,d),u(t)) \leq \sigma\left(\beta(t_0)\left\|x_0\right\|_{\mathcal{X}}, t-t_0\right) + \sup_{t_0\leq\tau\leq t}\gamma\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right). \tag{2.12}$$

Moreover, if there exists $a \in \mathcal{N}$ such that $\left\|H(t,x,u)\right\|_{\mathcal{Y}} \leq a(V(t,x,u))$ for all $(t,x,u) \in \Re^+ \times \mathcal{X} \times U$, then for every $\rho \in K_\infty$, Σ satisfies the WIOS property from the input $u \in M_U$, with gain $\tilde{\gamma} \in \mathcal{N}$ and weight $\delta \in K^+$, where $\tilde{\gamma}(s) := a\left(\gamma(s) + \rho\left(\gamma(s)\right)\right)$.

LEMMA 2.17. *Consider a control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with outputs and the BIC property and for which $0 \in \mathcal{X}$ is a robust equilibrium point from the input $u \in M_U$. Suppose that Σ is RFC from the input $u \in M_U$. Furthermore, suppose that there exist functions $V : \Re^+ \times \mathcal{X} \times U \to \Re^+$, with $V(t,0,0) = 0$, for all $t \geq 0$, $\gamma \in \mathcal{N}$, and $\delta \in K^+$ such that the following properties hold:*

P1. *For every $s \geq 0$, it holds that*

$$\sup\left\{V(t,\phi(t,t_0,x_0,u,d),u(t)) - \sup_{t_0\leq\tau\leq t}\gamma\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right);\right.$$
$$\left. t\geq t_0, \left\|x_0\right\|_{\mathcal{X}} \leq s, t_0 \geq 0, d\in M_D, u\in M_U\right\} < +\infty.$$

P2. *For every $\varepsilon > 0$, there exists a $\rho := \rho(\varepsilon) > 0$ such that*

$$\sup\left\{V(t,\phi(t,t_0,x_0,u,d),u(t)) - \sup_{t_0\leq\tau\leq t}\gamma\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right);\right.$$
$$\left. t\geq t_0, \left\|x_0\right\|_{\mathcal{X}} \leq \rho, t_0 \geq 0, d\in M_D, u\in M_U\right\} \leq \varepsilon.$$

P3. *For every $\varepsilon > 0$ and $R \geq 0$, there exists a $\tau := \tau(\varepsilon, R) \geq 0$ such that*

$$\sup\left\{V(t,\phi(t,t_0,x_0,u,d),u(t)) - \sup_{t_0\leq\tau\leq t}\gamma\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right);\right.$$
$$\left. t\geq t_0+\tau, \left\|x_0\right\|_{\mathcal{X}} \leq R, t_0 \geq 0, d\in M_D, u\in M_U\right\} \leq \varepsilon.$$

Then there exists a function $\sigma \in KL$ such that estimate (2.12) *holds for all $u \in M_U$, $(t_0, x_0, d) \in \Re^+ \times \mathcal{X} \times M_D$, and $t \geq t_0$, with $\beta(t) \equiv 1$. Moreover, if there exists $a \in \mathcal{N}$ such that $\|H(t,x,u)\|_{\mathcal{Y}} \leq a(V(t,x,u))$ for all $(t,x,u) \in \Re^+ \times \mathcal{X} \times U$, then for every $\rho \in K_\infty$, Σ satisfies the UWIOS property from the input $u \in M_U$, with gain $\tilde{\gamma} \in \mathcal{N}$ and weight $\delta \in K^+$, where $\tilde{\gamma}(s) := a\left(\gamma(s) + \rho\left(\gamma(s)\right)\right)$.*

Remark 2.18. Notice that Lemmas 2.16 and 2.17 can be very useful for the demonstration of the (U)WIOS property, because in practice we show properties (P1)–(P3) for some Lyapunov functional V and not necessarily for the norm of the output map. Moreover, notice that it is not required that V is continuous. If $V : \Re^+ \times \mathcal{X} \times U \to \Re^+$ is a continuous functional that maps bounded sets of $\Re^+ \times \mathcal{X} \times U$ into bounded sets of $\Re^+$, then Lemmas 2.16 and 2.17 guarantee that Σ satisfies the WIOS and the UWIOS properties with V as output, respectively, from the input $u \in M_U$, with gain $\gamma \in \mathcal{N}$ and weight $\delta \in K^+$.

Finally, we end this section with some useful observations for *T-periodic* control systems. It turns out that periodicity guarantees uniformity with respect to the initial times. The following lemmas should be compared with Lemma 1.1, p. 131 in [6]. Their proofs are provided in the appendix.

LEMMA 2.19. *Suppose that $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ is T-periodic. If Σ satisfies the WIOS property from the input $u \in M_U$, then Σ satisfies the UWIOS property from the input $u \in M_U$.*

LEMMA 2.20. *Suppose that $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ is T-periodic. If Σ satisfies the IOS property from the input $u \in M_U$, then Σ satisfies the UIOS property from the input $u \in M_U$.*

3. A small-gain theorem for a wide class of systems. The main result of the present work is stated next.

THEOREM 3.1. *Consider the system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ with the BIC property for which $0 \in \mathcal{X}$ is a robust equilibrium point from the input $u \in M_U$, and suppose that there exist maps $V_1 : \Re^+ \times \mathcal{X} \times U \to \Re^+$, $V_2 : \Re^+ \times \mathcal{X} \times U \to \Re^+$, with $V_i(t,0,0) = 0$ for all $t \geq 0$ $(i = 1, 2)$ such that the following hypotheses hold.*

(H1) *There exist functions $\sigma_1 \in KL$, β_1, μ_1, c_1, δ_1, δ_1^u, $q_1^u \in K^+$, γ_1, γ_1^u, a_1, p_1, $p_1^u \in \mathcal{N}$, $L_1 : \Re^+ \times \mathcal{X} \to \Re^+$, with $L_i(t,0) = 0$ for all $t \geq 0$, such that for every $(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$ the mapping $t \to V_1(t, \phi(t, t_0, x_0, u, d), u(t))$ is locally bounded on $[t_0, t_{\max})$, and the following estimates hold for all $t \in [t_0, t_{\max})$:*

$$V_1(t, \phi(t, t_0, x_0, u, d), u(t)) \leq \sigma_1\left(\beta_1(t_0) L_1(t_0, x_0), t - t_0\right) + \sup_{t_0 \leq \tau \leq t} \gamma_1\left(\delta_1(\tau) V_2(\tau)\right) + \sup_{t_0 \leq \tau \leq t} \gamma_1^u\left(\delta_1^u(\tau) \|u(\tau)\|_{\mathcal{U}}\right), \tag{3.1}$$

$$\beta_1(t) L_1(t, \phi(t, t_0, x_0, u, d)) \leq \max\left\{\mu_1(t - t_0), c_1(t_0), a_1\left(\|x_0\|_{\mathcal{X}}\right), \sup_{t_0 \leq \tau \leq t} p_1\left(V_2(\tau)\right), \sup_{t_0 \leq \tau \leq t} p_1^u\left(q_1^u(\tau) \|u(\tau)\|_{\mathcal{U}}\right)\right\}, \tag{3.2}$$

where $V_2(t) = V_2\left(t, \phi(t, t_0, x_0, u, d), u(t)\right)$ and $t_{\max}$ is the maximal existence time of the transition map of Σ.

(H2) *There exist functions $\sigma_2 \in KL$, β_2, μ_2, c_2, δ_2, δ_2^u, $q_2^u \in K^+$, γ_2, γ_2^u, a_2, p_2, $p_2^u \in \mathcal{N}$, $L_2 : \Re^+ \times \mathcal{X} \to \Re^+$, with $L_2(t,0) = 0$ for all $t \geq 0$, such that for every $(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$ the mapping $t \to V_2(t, \phi(t, t_0, x_0, u, d), u(t))$ is*

locally bounded on $[t_0, t_{\max})$*, and the following estimates hold for all* $t \in [t_0, t_{\max})$*:*

$$V_2(t, \phi(t, t_0, x_0, u, d), u(t)) \le \sigma_2\left(\beta_2(t_0) L_2(t_0, x_0), t - t_0\right) + \sup_{t_0 \le \tau \le t} \gamma_2\left(\delta_2(\tau) V_1(\tau)\right) + \sup_{t_0 \le \tau \le t} \gamma_2^u\left(\delta_2^u(\tau) \left\|u(\tau)\right\|_{\mathcal{U}}\right), \tag{3.3}$$

$$\beta_2(t) L_2(t, \phi(t, t_0, x_0, u, d)) \le \max\left\{\mu_2(t - t_0), c_2(t_0), a_2\left(\left\|x_0\right\|_{\mathcal{X}}\right), \sup_{t_0 \le \tau \le t} p_2\left(V_1(\tau)\right), \sup_{t_0 \le \tau \le t} p_2^u\left(q_2^u(\tau) \left\|u(\tau)\right\|_{\mathcal{U}}\right)\right\}, \tag{3.4}$$

where $V_1(t) = V_1\left(t, \phi(t, t_0, x_0, u, d), u(t)\right)$ *and* $t_{\max}$ *is the maximal existence time of the transition map of* Σ*.*

(H3) *There exist a function* $\rho \in K_\infty$ *and a constant* $M > 0$ *such that*

$$\delta_1(t) \le M \quad \forall t \ge 0, \tag{3.5}$$

$$g_1\left(\delta_1(t) g_2\left(\delta_2(\tau) s\right)\right) \le s \quad \forall t, s \ge 0 \text{ and } \tau \in [0, t], \tag{3.6}$$

where $g_i(s) := \gamma_i(s) + \rho\left(\gamma_i(s)\right)$*,* $i = 1, 2$*.*

(H4) *There exists a function* $a \in \mathcal{N}$ *such that the following inequality holds for all* $(t, x, u) \in \Re^+ \times \mathcal{X} \times U$*:*

$$\left\|H(t, x, u)\right\|_{\mathcal{Y}} \le a\left(V_1(t, x, u) + \gamma_1\left(\delta_1(t) V_2(t, x, u)\right)\right). \tag{3.7}$$

(H5) *There exist functions* $b \in \mathcal{N}$*,* $\mu \in K^+$ *such that the following inequalities hold for all* $(t, x) \in \Re^+ \times \mathcal{X}$*:*

$$\mu(t) \left\|x\right\|_{\mathcal{X}} \le b\left(L_1(t, x) + L_2(t, x)\right); \quad \max\left(L_1(t, x), L_2(t, x)\right) \le b\left(\left\|x\right\|_{\mathcal{X}}\right). \tag{3.8}$$

Then there exists a function $\gamma \in \mathcal{N}$ *such that system* Σ *satisfies the WIOS property from the input* $u \in M_U$*, with gain* $\gamma \in \mathcal{N}$ *and weight* $\delta \in K^+$*, where*

$$\delta(t) := \max\{\delta_1^u(t), \delta_2^u(t), q_1^u(t), q_2^u(t)\}. \tag{3.9}$$

Moreover, if $\beta_1, \beta_2, c_1, c_2, \delta_2 \in K^+$ *are bounded, then system* Σ *satisfies the UWIOS property from the input* $u \in M_U$*, with gain* $\gamma \in \mathcal{N}$ *and weight* $\delta \in K^+$*.*

Remark 3.2.

(a) It should be clear that Theorem 3.1 takes into account all possible cases (weights, nonuniformity with respect to initial times) and thus is applicable to a very wide class of systems.

(b) When $\gamma_1 \in \mathcal{N}$ (or $\gamma_2 \in \mathcal{N}$) is identically zero, it follows that (3.6) is automatically satisfied. This is the case of systems in cascade (see [14]). On the other hand, if $\gamma_i(s) = K_i s$ for certain constants $K_i \ge 0$ $(i = 1, 2)$, then inequality (3.6) is satisfied if $K_1 K_2 \sup_{t \ge 0}\left(\delta_1(t) \max_{\tau \in [0,t]} \delta_2(\tau)\right) < 1$. Moreover, if $\gamma_i(s) = K_i s$ for certain constants $K_i \ge 0$ $(i = 1, 2)$ and $\delta_1(t) \equiv \delta_2(t) \equiv 1$, then hypothesis (H3) is satisfied if $K_1 K_2 < 1$. This is the case of the classical small-gain theorem.

(c) If, instead of hypothesis (H4), there exists a function $a \in \mathcal{N}$ such that $\left\|H(t, x, u)\right\|_{\mathcal{Y}} \le a\left(V_2(t, x, u) + \gamma_2\left(\delta_2(t) V_1(t, x, u)\right)\right)$ holds for all $(t, x, u) \in \Re^+ \times \mathcal{X} \times U$, then indices 1 and 2 must be changed in hypothesis (H3). Furthermore, in this case system Σ satisfies the UWIOS property from the input $u \in M_U$, with gain $\gamma \in \mathcal{N}$ and weight $\delta \in K^+$, if functions $\beta_1, \beta_2, c_1, c_2, \delta_1 \in K^+$ are bounded.

(d) In previous nonlinear small-gain theorems (e.g., [14]), the functions $L_1 : \Re^+ \times \mathcal{X} \to \Re^+$ and $L_2 : \Re^+ \times \mathcal{X} \to \Re^+$ take the form $L_1(t,x) = x_1$ and $L_2(t,x) = x_2$, respectively, where $x = (x_1, x_2)$. It follows that hypothesis (H5) automatically holds with $b(s) := s$ and $\mu(t) \equiv 1$. This is the case in Corollary 3.4 below.

(e) Hypothesis (H4) in conjunction with (3.5) guarantees that: (i) if the mappings $t \to V_1(t, \phi(t, t_0, x_0, u, d), u(t))$, $t \to V_2(t, \phi(t, t_0, x_0, u, d), u(t))$ are bounded, then the mapping $t \to \|H(t, \phi(t, t_0, x_0, u, d), u(t))\|_{\mathcal{Y}}$ is bounded as well, and (ii) if $V_1(t, \phi(t, t_0, x_0, u, d), u(t)) \to 0$, $V_2(t, \phi(t, t_0, x_0, u, d), u(t)) \to 0$, then $\|H(t, \phi(t, t_0, x_0, u, d), u(t))\|_{\mathcal{Y}} \to 0$. In other words, hypothesis (H4) guarantees that the behavior of the output of system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ can be determined by studying the behavior of the functionals $V_1 : \Re^+ \times \mathcal{X} \times U \to \Re^+, V_2 : \Re^+ \times \mathcal{X} \times U \to \Re^+$.

Since small-gain results are frequently applied to feedback interconnections of control systems, we need to clarify the notion of the feedback interconnection of two control systems. However, the fact that we do not require the classical semigroup property for each of the interconnected subsystems creates technical difficulties: For example, the determination of the set of sampling times for the composite system is not trivial. In order to guarantee the existence of a set of sampling times for the composite system, we assume that the sampling times of the composite system are the common sampling times of the interconnected subsystems. The details are given in the following definition.

DEFINITION 3.3. *Consider a pair of control systems* $\Sigma_1 = (\mathcal{X}_1, \mathcal{Y}_1, M_{S_2 \times U}, M_D, \tilde{\phi}_1, \pi_1, H_1)$, $\Sigma_2 = (\mathcal{X}_2, \mathcal{Y}_2, M_{S_1 \times U}, M_D, \tilde{\phi}_2, \pi_2, H_2)$ *with outputs* $H_1 : \Re^+ \times \mathcal{X}_1 \times \mathcal{Y}_2 \times U \to S_1 \subseteq Y_1, H_2 : \Re^+ \times \mathcal{X}_2 \times \mathcal{Y}_1 \times U \to S_2 \subseteq \mathcal{Y}_2$ *and the BIC property and for which* $0 \in \mathcal{X}_i$, $i = 1, 2$, *are robust equilibrium points from the inputs* $(v_2, u) \in M_{S_2 \times U}, (v_1, u) \in M_{S_1 \times U}$, *respectively. Suppose that there exists a* unique *pair of a map* $\phi = (\phi_1, \phi_2) : A_\phi \to \mathcal{X}$ *and a set-valued map* $\Re^+ \times \mathcal{X} \times M_U \times M_D \ni (t_0, x_0, u, d) \to \pi(t_0, x_0, u, d) \subseteq [t_0, +\infty)$, *where* $A_\phi \subseteq \Re^+ \times \Re^+ \times \mathcal{X} \times M_U \times M_D$, $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, *such that for every* $(t, t_0, x_0, u, d) \in A_\phi$, *with* $t \geq t_0$, $x_0 = (x_{1,0}, x_{2,0}) \in \mathcal{X}_1 \times \mathcal{X}_2$, *it holds that:*

> *"there exists a pair of external inputs* $v_i \subseteq \mathcal{M}(S_i), i = 1, 2$, *with* $v_1(\tau) = H_1(\tau, \phi_1(\tau, t_0, x_0, u, d), v_2(\tau), u(\tau)), v_2(\tau) = H_2(\tau, \phi_2(\tau, t_0, x_0, u, d), v_1(\tau), u(\tau))$ *for all* $\tau \in [t_0, t], (v_i, u) \in M_{S_i \times U}, i = 1, 2, \pi(t_0, x_0, u, d) = \pi_1(t_0, x_{1,0}, (v_2, u), d) \cap \pi_2(t_0, x_{2,0}, (v_1, u), d)$, *and* $\phi_1(\tau, t_0, x_0, u, d) = \tilde{\phi}_1(\tau, t_0, x_{1,0}, (v_2, u), d), \phi_2(\tau, t_0, x_0, u, d) = \tilde{\phi}_2(\tau, t_0, x_{2,0}, (v_1, u), d)$ *for all* $\tau \in [t_0, t]$*."*

Moreover, let $\mathcal{Y}$ *be a normed linear space and* $H : \Re^+ \times \mathcal{X} \times U \to \mathcal{Y}$ *a continuous map that maps bounded sets of* $\Re^+ \times \mathcal{X} \times U$ *into bounded sets of* $\mathcal{Y}$, *with* $H(t, 0, 0) = 0$ *for all* $t \geq 0$, *and suppose that* $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ *is a control system with outputs and the BIC property, for which* $0 \in \mathcal{X}$ *is a robust equilibrium point from the input* $u \in M_U$. *Then system* Σ *is said to be the* feedback connection *or the* interconnection *of systems* Σ_1 *and* Σ_2.

It should be emphasized that the feedback interconnection of two systems may create a system which has different qualitative properties from each of the interconnected subsystems. For example, if we interconnect a subsystem described by RFDEs (see Example 2.10) with a hybrid subsystem with impulses at fixed times (see Example 2.12), then the overall system will be a system with both "memory" and impulses (discontinuous systems described by RFDEs—see [49]).

We are now in a position to state our main result for feedback interconnections of control systems. It is a direct consequence of Theorem 3.1, and its proof is omitted.

COROLLARY 3.4. *Suppose that* $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ *is the feedback connection of systems* $\Sigma_1 = (\mathcal{X}_1, \mathcal{Y}_1, M_{S_2\times U}, M_D, \tilde{\phi}_1, \pi_1, H_1)$ *and* $\Sigma_2 = (\mathcal{X}_2, \mathcal{Y}_2, M_{S_1\times U}, M_D, \tilde{\phi}_2, \pi_2, H_2)$ *with outputs* $H_1 : \Re^+ \times \mathcal{X}_1 \times \mathcal{Y}_2 \times U \to S_1 \subseteq \mathcal{Y}_1, H_2 : \Re^+ \times \mathcal{X}_2 \times \mathcal{Y}_1 \times U \to S_2 \subseteq \mathcal{Y}_2$. *We assume the following.*

(H1′) *Subsystem* Σ_1 *satisfies the WIOS property from the inputs* $v_2 \in \mathcal{M}(S_2)$ *and* $u \in M_U$. *Particularly, there exist functions* $\sigma_1 \in KL$, $\beta_1, \mu_1, c_1, \delta_1, \delta_1^u, q_1^u \in K^+, \gamma_1, \gamma_1^u, a_1, p_1, p_1^u \in \mathcal{N}$ *such that the following estimate holds for all* $(t_0, x_1, (v_2, u_0), d) \in \Re^+ \times \mathcal{X}_1 \times M_{S_2\times U} \times M_D$ *and* $t \geq t_0$:

$$\begin{aligned}\left\|H_1(t, \tilde{\phi}_1(t, t_0, x_1, (v_2, u), d), v_2(t), u(t))\right\|_{\mathcal{Y}_1} &\leq \sigma_1\left(\beta_1(t_0)\left\|x_1\right\|_{\mathcal{X}_1}, t - t_0\right) \\ &+ \sup_{t_0\leq\tau\leq t} \gamma_1\left(\delta_1(\tau)\left\|v_2(\tau)\right\|_{\mathcal{Y}_2}\right) + \sup_{t_0\leq\tau\leq t} \gamma_1^u\left(\delta_1^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right),\end{aligned} \tag{3.10}$$

$$\begin{aligned}\beta_1(t)\left\|\tilde{\phi}_1(t, t_0, x_1, (v_2, u), d)\right\|_{\mathcal{X}_1} \leq \max\Bigg\{&\mu_1(t - t_0), c_1(t_0), a_1\left(\left\|x_1\right\|_{\mathcal{X}_1}\right), \\ &\sup_{t_0\leq\tau\leq t} p_1\left(\left\|v_2(\tau)\right\|_{\mathcal{Y}_2}\right), \sup_{t_0\leq\tau\leq t} p_1^u\left(q_1^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right)\Bigg\}.\end{aligned} \tag{3.11}$$

(H2′) *Subsystem* Σ_2 *satisfies the WIOS property from the inputs* $v_1 \in \mathcal{M}(S_1)$ *and* $u \in M_U$. *Particularly, there exist functions* $\sigma_2 \in KL, \beta_2, \mu_2, c_2, \delta_2, \delta_2^u, q_2^u \in K^+, \gamma_2, \gamma_2^u, a_2, p_2, p_2^u \in \mathcal{N}$ *such that the following estimate holds for all* $(t_0, x_2, (v_1, u_0), d) \in \Re^+ \times \mathcal{X}_2 \times M_{S_1\times U} \times M_D$ *and* $t \geq t_0$:

$$\begin{aligned}&\left\|H_2(t, \tilde{\phi}_2(t, t_0, x_2, (v_1, u), d), v_1(t), u(t))\right\|_{\mathcal{Y}_2} \\ &\leq \sigma_2\left(\beta_2(t_0)\left\|x_2\right\|_{\mathcal{X}_2}, t - t_0\right) + \sup_{t_0\leq\tau\leq t} \gamma_2\left(\delta_2(\tau)\left\|v_1(\tau)\right\|_{\mathcal{Y}_1}\right) + \sup_{t_0\leq\tau\leq t} \gamma_2^u\left(\delta_2^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right),\end{aligned} \tag{3.12}$$

$$\begin{aligned}\beta_2(t)\left\|\tilde{\phi}_2(t, t_0, x_2, (v_1, u), d)\right\|_{\mathcal{X}_2} \leq \max\Bigg\{&\mu_2(t - t_0), c_2(t_0), a_2\left(\left\|x_2\right\|_{X_2}\right), \\ &\sup_{t_0\leq\tau\leq t} p_2\left(\left\|v_1(\tau)\right\|_{\mathcal{Y}_1}\right), \sup_{t_0\leq\tau\leq t} p_2^u\left(q_2^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right)\Bigg\}.\end{aligned} \tag{3.13}$$

Moreover, assume that hypothesis (H3) *of Theorem* 3.1 *holds and there exists a function* $a \in \mathcal{N}$ *such that the following inequality holds for all* $(t, x, u, Y_1, Y_2) \in \Re^+ \times \mathcal{X} \times U \times \mathcal{Y}_1 \times \mathcal{Y}_2$, *with* $x = (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2, Y_1 = H_1(t, x_1, Y_2, u), Y_2 = H_2(t, x_2, Y_1, u)$:

$$\left\|H(t, x, u)\right\|_{\mathcal{Y}} \leq a\left(\left\|Y_1\right\|_{\mathcal{Y}_1} + \gamma_1\left(\delta_1(t)\left\|Y_2\right\|_{\mathcal{Y}_2}\right)\right). \tag{3.14}$$

Then there exists a function $\gamma \in \mathcal{N}$ *such that system* Σ *satisfies the WIOS property from the input* $u \in M_U$, *with gain* $\gamma \in \mathcal{N}$ *and weight* $\delta \in K^+$, *where* $\delta \in K^+$ *is defined by* (3.9). *Moreover, if* $\beta_1, \beta_2, c_1, c_2, \delta_2 \in K^+$ *are bounded, then system* Σ *satisfies the UWIOS property from the input* $u \in M_U$, *with gain* $\gamma \in \mathcal{N}$ *and weight* $\delta \in K^+$.

Remark 3.5.

(a) When $\delta_1(t) \equiv \delta_2(t) \equiv 1$, $H_1 : \Re^+ \times \mathcal{X}_1 \times U \to S_1 \subseteq \mathcal{Y}_1$, and $\gamma_1 \in K_\infty$, then the result of Corollary 3.4 guarantees the WIOS property from the input $u \in M_U$ for the output $H(t, x, u) := (H_1(t, x_1, u), H_2(t, x_2, H_1(t, x_1, u), u))$, i.e., for the output that combines the outputs of each individual subsystem. Moreover, if in addition the functions $\delta_i^u(t)$, $q_i^u(t)$ $(i = 1, 2)$ are bounded, then the result of Corollary 3.4 guarantees the IOS from the input $u \in M_U$ for the output $H(t, x, u) := (H_1(t, x_1, u), H_2(t, x_2, H_1(t, x_1, u), u))$. Finally,

if in addition the functions $\beta_1, \beta_2, c_1, c_2, \delta_2 \in K^+$ are bounded, then the result of Corollary 3.4 guarantees the UIOS from the input $u \in M_U$ for the output $H(t,x,u) := (H_1(t,x_1,u), H_2(t,x_2,H_1(t,x_1,u),u))$. This particular case coincides with the prior result of the nonlinear ISS small-gain theorem presented in [14] for control systems described by ODEs.

(b) Conditions (3.11) and (3.13) hold automatically, when each one of the subsystems Σ_1 and Σ_2 satisfy the WISS property.

(c) If, instead of hypothesis (3.14), there exists a function $a \in \mathcal{N}$ such that $\|H(t,x,u)\|_{\mathcal{Y}} \leq a\left(\|Y_2\|_{\mathcal{Y}_2} + \gamma_2\left(\delta_2(t)\|Y_1\|_{\mathcal{Y}_1}\right)\right)$ holds for all $(t,x,u,Y_1,Y_2) \in \Re^+ \times \mathcal{X} \times U \times \mathcal{Y}_1 \times \mathcal{Y}_2$, with $x = (x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$, $Y_1 = H_1(t,x_1,Y_2,u)$, $Y_2 = H_2(t,x_2,Y_1,u)$, then indices 1 and 2 must be changed in hypothesis (H3). Furthermore, in this case system Σ satisfies the UWIOS property from the input $u \in M_U$, with gain $\gamma \in \mathcal{N}$ and weight $\delta \in K^+$, if functions $\beta_1, \beta_2, c_1, c_2, \delta_1 \in K^+$ are bounded.

Proof of Theorem 3.1. The proof consists of three steps:

Step 1. We show that Σ is RFC from the input $u \in M_U$.

Step 2. Let $\tilde{\varphi}(s) := s + \frac{1}{2}\rho(s)$, where $\rho \in K_\infty$ is the function involved in hypothesis (H3). We show that properties P1 and P2 of Lemma 2.16 hold for system Σ with $V = V_1$ or $V = \tilde{\varphi}\left(\gamma_1\left(\delta_1(t)V_2\right)\right)$, for appropriate $\tilde{\gamma} \in \mathcal{N}$ and $\delta \in K^+$ as defined by (3.9). Moreover, if $\beta_1, \beta_2 \in K^+$ are bounded, we show that properties P1 and P2 of Lemma 2.17 hold for system Σ with $V = V_1$ or $V = \tilde{\varphi}\left(\gamma_1\left(\delta_1(t)V_2\right)\right)$, for appropriate $\tilde{\gamma} \in \mathcal{N}$ and $\delta \in K^+$ as defined by (3.9).

Step 3. We show that property P3 of Lemma 2.16 holds for system Σ with $V = V_1$ or $V = \tilde{\varphi}\left(\gamma_1\left(\delta_1(t)V_2\right)\right)$, for appropriate $\tilde{\gamma} \in \mathcal{N}$ and $\delta \in K^+$ as defined by (3.9). Moreover, if $\beta_1, \beta_2, c_1, c_2, \delta_2 \in K^+$ are bounded, we show that property P3 of Lemma 2.17 holds for system Σ with $V = V_1$ or $V = \tilde{\varphi}\left(\gamma_1\left(\delta_1(t)V_2\right)\right)$, for appropriate $\tilde{\gamma} \in \mathcal{N}$ and $\delta \in K^+$ as defined by (3.9).

It then follows from Lemma 2.16 that there exist functions $\sigma \in KL$ and $\beta \in K^+$ such that the following estimate holds for all $u \in M_U$, $(t_0,x_0,d) \in \Re^+ \times \mathcal{X} \times M_D$, and $t \geq t_0$:

$$V_1(t,\phi(t,t_0,x_0,u,d),u(t)) \leq \sigma\left(\beta(t_0)\|x_0\|_{\mathcal{X}}, t-t_0\right) + \sup_{t_0 \leq \tau \leq t} \gamma\left(\delta(\tau)\|u(\tau)\|_{\mathcal{U}}\right), \tag{3.15a}$$

$$\begin{aligned} &\tilde{\varphi}\left(\gamma_1\left(\delta_1(t)V_2(t,\phi(t,t_0,x_0,u,d),u(t))\right)\right) \\ &\quad \leq \sigma\left(\beta(t_0)\|x_0\|_{\mathcal{X}}, t-t_0\right) + \sup_{t_0 \leq \tau \leq t} \gamma\left(\delta(\tau)\|u(\tau)\|_{\mathcal{U}}\right). \end{aligned} \tag{3.15b}$$

Moreover, if $\beta_1, \beta_2, c_1, c_2, \delta_2 \in K^+$ are bounded, it follows from Lemma 2.17 that estimates (3.15a)–(3.15b) hold with $\beta(t) \equiv 1$.

Thus, using (3.7) and the fact that $\tilde{\varphi}(s) \geq s$ for all $s \geq 0$, we conclude that Σ satisfies the WIOS property from the input $u \in M_U$, with gain $\gamma(s) := a\left(4\tilde{\gamma}(s)\right) \in \mathcal{N}$ and weight $\delta \in K^+$. Moreover, if $\beta_1, \beta_2, c_1, c_2, \delta_2 \in K^+$ are bounded, we conclude that Σ satisfies the UWIOS property from the input $u \in M_U$, with gain $\gamma(s) := a\left(4\tilde{\gamma}(s)\right) \in \mathcal{N}$ and weight $\delta \in K^+$.

Step 1. Let arbitrary $(t,t_0,x_0,u,d) \in A_\phi$, with $t \geq t_0$, $x_0 \in \mathcal{X}$, and let $t_{\max} \in (t_0, +\infty]$ the maximal existence time of the transition map ϕ of $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ that corresponds to $(t_0,x_0,u,d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$. Notice that, by virtue of the BIC property, if $t_{\max} < +\infty$, then for every $M > 0$ there exists $t \in [t_0, t_{\max})$, with $\|\phi(t,t_0,x_0,u,d)\|_{\mathcal{X}} > M$. We define $V_1(\tau) = V_1(\tau,\phi(\tau,t_0,x_0,u,d),u(\tau))$, $L_1(\tau) = L_1(\tau,\phi(\tau,t_0,x_0,u,d))$ and $V_2(\tau) = V_2(\tau,\phi(\tau,t_0,x_0,u,d),u(\tau))$, $L_2(\tau) = L_2(\tau,\phi(\tau,t_0,x_0,u,d))$ for all $\tau \in [t_0,t]$.

The previous definitions in conjunction with (3.1), (3.2), (3.3), (3.4) imply the following inequalities for all $t \in [t_0, t_{\max})$:

$$V_1(t) \leq \sigma_1\left(\beta_1(t_0)L_1(t_0), t-t_0\right) + \sup_{t_0\leq\tau\leq t} \gamma_1\left(\delta_1(\tau)V_2(\tau)\right) + \sup_{t_0\leq\tau\leq t} \gamma_1^u\left(\delta_1^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right), \tag{3.16}$$

$$\begin{aligned}\beta_1(t)L_1(t) \leq \max\Bigg\{&\mu_1(t-t_0), c_1(t_0), a_1\left(\left\|x_0\right\|_{\mathcal{X}}\right), \sup_{t_0\leq\tau\leq t} p_1\left(V_2(\tau)\right), \\ &\sup_{t_0\leq\tau\leq t} p_1^u\left(q_1^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right)\Bigg\},\end{aligned} \tag{3.17}$$

$$\begin{aligned}V_2(t) \leq{}& \sigma_2\left(\beta_2(t_0)L_2(t_0), t-t_0\right) + \sup_{t_0\leq\tau\leq t} \gamma_2\left(\delta_2(\tau)V_1(\tau)\right) \\ &+ \sup_{t_0\leq\tau\leq t} \gamma_2^u\left(\delta_2^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right),\end{aligned} \tag{3.18}$$

$$\begin{aligned}\beta_2(t)L_2(t) \leq \max\Bigg\{&\mu_2(t-t_0), c_2(t_0), a_2\left(\left\|x_0\right\|_{\mathcal{X}}\right), \sup_{t_0\leq\tau\leq t} p_2\left(V_1(\tau)\right), \\ &\sup_{t_0\leq\tau\leq t} p_2^u\left(q_2^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right)\Bigg\}.\end{aligned} \tag{3.19}$$

Let $\rho \in K_\infty$ the function involved in hypothesis (H3), and define $\kappa(s) := s + \rho^{-1}(s)$, $\varphi(s) := s + \rho(s)$. Using the inequality $r + s \leq \max\{\kappa(r); \varphi(s)\}$ (which holds for all $r, s \geq 0$) as well as the equality $g_2(s) = \varphi\left(\gamma_2(s)\right)$, we obtain from (3.18) for all $t \in [t_0, t_{\max})$:

$$\begin{aligned}V_2(t) \leq \max\Bigg\{&\kappa\left(\sigma_2\left(\beta_2(t_0)L_2(t_0), t-t_0\right) + \sup_{t_0\leq\tau\leq t} \gamma_2^u\left(\delta_2^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right)\right); \\ &\sup_{t_0\leq\tau\leq t} g_2\left(\delta_2(\tau)V_1(\tau)\right)\Bigg\}.\end{aligned} \tag{3.20}$$

Notice that inequality (3.6) implies that $\gamma_1\left(\delta_1(t)g_2\left(\delta_2(\tau)s\right)\right) \leq \varphi^{-1}(s)\ \forall t, s \geq 0$ and $\tau \in [0, t]$. Thus (3.20) in conjunction with (3.5) and the previous observation implies the following estimate which holds for all $t \in [t_0, t_{\max})$:

$$\begin{aligned}\gamma_1\left(\delta_1(t)V_2(t)\right) \leq \max\Bigg\{&\gamma_1\Bigg(M\kappa\Bigg(\sigma_2\left(\beta_2(t_0)L_2(t_0), t-t_0\right) \\ &+ \sup_{t_0\leq\tau\leq t} \gamma_2^u\left(\delta_2^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right)\Bigg)\Bigg); \varphi^{-1}\left(\sup_{t_0\leq\tau\leq t} V_1(\tau)\right)\Bigg\}.\end{aligned} \tag{3.21}$$

Combining estimate (3.16) with (3.21), we obtain

$$\begin{aligned}\sup_{t_0\leq\tau\leq t} V_1(\tau) \leq{}& \sigma_1\left(\beta_1(t_0)L_1(t_0), 0\right) + \sup_{t_0\leq\tau\leq t} \gamma_1^u\left(\delta_1^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right) \\ &+ \max\Bigg\{\gamma_1\Bigg(M\kappa\Bigg(\sigma_2\left(\beta_2(t_0)L_2(t_0), 0\right) + \sup_{t_0\leq\tau\leq t} \gamma_2^u\left(\delta_2^u(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right)\Bigg)\Bigg); \\ &\varphi^{-1}\left(\sup_{t_0\leq\tau\leq t} V_1(\tau)\right)\Bigg\}.\end{aligned} \tag{3.22}$$

Distinguishing the cases $\gamma_1(M\kappa(\sigma_2(\beta_2(t_0)L_2(t_0), 0) + \sup_{t_0\leq\tau\leq t} \gamma_2^u(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}))) \geq \varphi^{-1}(\sup_{t_0\leq\tau\leq t} V_1(\tau))$, $\gamma_1(M\kappa(\sigma_2(\beta_2(t_0)L_2(t_0), 0) + \sup_{t_0\leq\tau\leq t} \gamma_2^u(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}))) \leq \varphi^{-1}(\sup_{t_0\leq\tau\leq t} V_1(\tau))$, using the identity $s - \varphi^{-1}(s) = \kappa^{-1}(s)$ and the fact that $\kappa(s) \geq$

s in conjunction with (3.22) and (3.8) (which implies $L_i(t_0) \leq b\left(\|x_0\|_{\mathcal{X}}\right)$, $i = 1, 2$) gives the following estimate which holds for all $t \in [t_0, t_{\max})$:

$$\sup_{t_0 \leq \tau \leq t} V_1(\tau) \leq \max\left\{\kappa\left(\sigma_1\left(\beta_1(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right) + \sup_{t_0 \leq \tau \leq t} \gamma_1^u\left(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right); W\right\} \tag{3.23a}$$

where

$$\begin{aligned} W := {} & \sigma_1\left(\beta_1(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right) + \sup_{t_0 \leq \tau \leq t} \gamma_1^u\left(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right) \\ & + \gamma_1\left(M\kappa\left(\sigma_2\left(\beta_2(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right) + \sup_{t_0 \leq \tau \leq t} \gamma_2^u\left(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right) \end{aligned} \tag{3.23b}$$

We show next that Σ is RFC from the input $u \in M_U$ by contradiction. Suppose that $t_{\max} < +\infty$. Then by virtue of the BIC property for every $M > 0$ there exists $t \in [t_0, t_{\max})$ with $\|\phi(t, t_0, x_0, u, d)\|_{\mathcal{X}} > M$. On the other hand, estimate (3.23a) in conjunction with the hypothesis $t_{\max} < +\infty$ shows that there exists $M_1 \geq 0$ such that $\sup_{t_0 \leq \tau < t_{\max}} V_1(\tau) \leq M_1$. The fact that $V_1(t)$ is bounded in conjunction with estimates (3.18) and (3.19) implies that there exist constants $M_2, M_3 \geq 0$ such that $\sup_{t_0 \leq \tau < t_{\max}} V_2(\tau) \leq M_2$ and $\sup_{t_0 \leq \tau < t_{\max}} L_2(\tau) \leq M_3$. Finally, the fact that $V_2(t)$ is bounded in conjunction with estimate (3.17) implies that there exists a constant $M_4 \geq 0$ such that $\sup_{t_0 \leq \tau < t_{\max}} L_1(\tau) \leq M_4$. It follows from (3.8) and inequality $\mu(t)\|\phi(t, t_0, x_0, u, d)\|_{\mathcal{X}} \leq b\left(L_1(t) + L_2(t)\right)$ that the transition map of Σ, i.e., $\phi(t, t_0, x_0, u, d)$, is bounded on $[t_0, t_{\max})$, and this contradicts the requirement that for every $M > 0$ there exists $t \in [t_0, t_{\max})$ with $\|\phi(t, t_0, x_0, u, d)\|_{\mathcal{X}} > M$. Hence, we must have $t_{\max} = +\infty$.

Let arbitrary $R \geq 0, T \geq 0$. For every $u \in M(B_U[0, R]) \cap \mathcal{M}_U, s \in [0, T], \|x_0\|_{\mathcal{X}} \leq R, t_0 \in [0, T], d \in M_D$ estimate (3.23a) shows that there exists $M_1(T, R) \geq 0$ such that $V_1(t_0 + s) \leq M_1(T, R)$ for all $s \in [0, T]$. The previous observation in conjunction with estimates (3.18), (3.19), and (3.8) (which gives $L_i(t_0) \leq b\left(\|x_0\|_X\right)$, $i = 1, 2$) implies that there exist $M_2(T, R), M_3(T, R) \geq 0$ such that for every $u \in \mathcal{M}(B_U[0, R]) \cap M_U, s \in [0, T], \|x_0\|_{\mathcal{X}} \leq R, t_0 \in [0, T], d \in M_D$ we have $V_2(t_0 + s) \leq M_2(T, R)$ and $L_2(t_0 + s) \leq M_3(T, R)$ for all $s \in [0, T]$. Finally, inequality $V_2(t_0 + s) \leq M_2(T, R)$ in conjunction with estimate (3.17) implies that there exists a constant $M_4(T, R) \geq 0$ such that for every $u \in \mathcal{M}(B_U[0, R]) \cap M_U, s \in [0, T], \|x_0\|_{\mathcal{X}} \leq R, t_0 \in [0, T], d \in M_D$ we have $L_1(t_0 + s) \leq M_4(T, R)$ for all $s \in [0, T]$. It follows from (3.8) and inequality $\mu(t)\|\phi(t, t_0, x_0, u, d)\|_{\mathcal{X}} \leq b\left(L_1(t) + L_2(t)\right)$ that for every $u \in \mathcal{M}(B_U[0, R]) \cap M_U, s \in [0, T], \|x_0\|_{\mathcal{X}} \leq R, t_0 \in [0, T], d \in M_D$ the transition map of Σ, i.e., $\phi(t, t_0, x_0, u, d)$, satisfies $\|\phi(t, t_0, x_0, u, d)\|_{\mathcal{X}} \leq \frac{b(M_3(T,R) + M_4(T,R))}{\min\{\mu(t): t \in [0, 2T]\}} < +\infty$, and this according to Definition 2.2 implies that Σ is RFC from the input $u \in M_U$.

Step 2. Using (3.23a) in conjunction with the inequality $q(r + s) \leq q\left(\kappa(r)\right) + q\left(\varphi(s)\right)$ (which holds for all $r, s \geq 0$ and $q \in \mathcal{N}$) gives the following estimate, which holds for all $t \geq t_0$:

$$\begin{aligned} V_1(t) \leq {} & \kappa\left(\kappa\left(\sigma_1\left(\beta_1(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right)\right)\right) + \gamma_1\left(M\kappa\left(\kappa\left(\sigma_2\left(\beta_2(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right)\right)\right)\right) \\ & + \sup_{t_0 \leq \tau \leq t} \kappa\left(\varphi\left(\gamma_1^u\left(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right) + \sup_{t_0 \leq \tau \leq t} \gamma_1\left(M\kappa\left(\varphi\left(\gamma_2^u\left(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right)\right). \end{aligned} \tag{3.24}$$

Moreover, combining estimates (3.21) and (3.23a) and using the equalities $\varphi^{-1}\left(\kappa(s)\right) = \rho(s)$ and $\varphi(s) := s + \rho\left(s\right)$ as well as the inequalities $\varphi^{-1}(s) \leq s$ and (3.8) (which gives

$L_i(t_0) \leq b(\|x_0\|_{\mathcal{X}})$, $i = 1, 2)$ gives the following estimate, which holds for all $t \geq t_0$:

$$\begin{aligned}\gamma_1(\delta_1(t)V_2(t)) \leq \gamma_1\left(M\kappa\left(\sigma_2(\beta_2(t_0)b(\|x_0\|_{\mathcal{X}}),0) + \sup_{t_0\leq\tau\leq t}\gamma_2^u(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}})\right)\right)\\ + \varphi\left(\sigma_1(\beta_1(t_0)b(\|x_0\|_{\mathcal{X}}),0) + \sup_{t_0\leq\tau\leq t}\gamma_1^u(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}})\right).\end{aligned} \tag{3.25}$$

Using (3.25) in conjunction with the inequality $q(r+s) \leq q(\kappa(r)) + q(\varphi(s))$ (which holds for all $r, s \geq 0$ and $q \in \mathcal{N}$) gives the following estimate, which holds for all $t \geq t_0$:

$$\begin{aligned}\gamma_1(\delta_1(t)V_2(t)) \leq \gamma_1(&M\kappa(\kappa(\sigma_2(\beta_2(t_0)b(\|x_0\|_{\mathcal{X}}),0))))\\ &+ \sup_{t_0\leq\tau\leq t}\gamma_1(M\kappa(\varphi(\gamma_2^u(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}))))\\ &+ \varphi(\kappa(\sigma_1(\beta_1(t_0)b(\|x_0\|_{\mathcal{X}}),0)))\\ &+ \sup_{t_0\leq\tau\leq t}\varphi(\varphi(\gamma_1^u(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}))).\end{aligned} \tag{3.26}$$

Let $\tilde{\varphi}(s) := s + \frac{1}{2}\rho(s)$. Using (3.26) in conjunction with the inequality $q(r+s) \leq q(\kappa(r)) + q(\varphi(s))$ (which holds for all $r, s \geq 0$ and $q \in \mathcal{N}$), we obtain the following estimate, which holds for all $t \geq t_0$:

$$\begin{aligned}\tilde{\varphi}(\gamma_1(\delta_1(t)V_2(t))) \leq \tilde{\varphi}(&\kappa(\gamma_1(M\kappa(\kappa(\sigma_2(\beta_2(t_0)b(\|x_0\|_{\mathcal{X}}),0))))\\ &+ \varphi(\kappa(\sigma_1(\beta_1(t_0)b(\|x_0\|_{\mathcal{X}}),0)))))\\ &+ \sup_{t_0\leq\tau\leq t}\tilde{\varphi}(\varphi(\varphi(\gamma_1^u(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}))))\\ &+ \gamma_1(M\kappa(\varphi(\gamma_2^u(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}})))))).\end{aligned} \tag{3.27}$$

Estimates (3.24), (3.27) show that properties P1 and P2 of Lemma 2.16 hold for system Σ with $V = V_1$ or $V = \tilde{\varphi}(\gamma_1(\delta_1(t)V_2))$, for appropriate $\tilde{\gamma} \in \mathcal{N}$ and $\delta \in K^+$ as defined by (3.9). Moreover, if $\beta_1, \beta_2 \in K^+$ are bounded, then estimates (3.24), (3.27) show that properties P1 and P2 of Lemma 2.17 hold for system Σ with $V = V_1$ or $V = \tilde{\varphi}(\gamma_1(\delta_1(t)V_2))$, for appropriate $\tilde{\gamma} \in \mathcal{N}$ and $\delta \in K^+$ as defined by (3.9). Particularly, $\tilde{\gamma} \in \mathcal{N}$ satisfies

$$\begin{aligned}&\tilde{\gamma}(s) \geq \tilde{\varphi}(\varphi(\varphi(\gamma_1^u(s))) + \gamma_1(M\kappa(\varphi(\gamma_2^u(s))))))\\ &\text{and } \tilde{\gamma}(s) \geq \kappa(\varphi(\gamma_1^u(s))) + \gamma_1(M\kappa(\varphi(\gamma_2^u(s)))) \text{ for all } s \geq 0.\end{aligned} \tag{3.28}$$

Step 3. Let $\tilde{\varphi}(s) := s + \frac{1}{2}\rho(s)$, $\tilde{\kappa}(s) := s + \rho^{-1}(2s)$. Exploiting estimates (3.16), (3.21) in conjunction with the inequality $r + s \leq \max\{\tilde{\kappa}(r); \tilde{\varphi}(s)\}$ (which holds for all $r, s \geq 0$), we obtain:

$$\begin{aligned}V_1(t) \leq \max\Bigg\{&\tilde{\kappa}\left(\sigma_1(\beta_1(t_0)L_1(t_0), t-t_0) + \sup_{t_0\leq\tau\leq t}\gamma_1^u(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}})\right);\\ &\sup_{t_0\leq\tau\leq t}\tilde{\varphi}(\gamma_1(\delta_1(\tau)V_2(\tau)))\Bigg\},\end{aligned} \tag{3.29}$$

$$\begin{aligned}\tilde{\varphi}(\gamma_1(\delta_1(t)V_2(t))) \leq \max\Bigg\{&\tilde{\varphi}\Bigg(\gamma_1(M\kappa(\sigma_2(\beta_2(t_0)L_2(t_0), t-t_0)\\ &+ \sup_{t_0\leq\tau\leq t}\gamma_2^u(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}})))); \tilde{\varphi}\left(\varphi^{-1}\left(\sup_{t_0\leq\tau\leq t}V_1(\tau)\right)\right)\Bigg\}.\end{aligned} \tag{3.30}$$

Let arbitrary $\xi \in \pi(t_0, x_0, u, d)$ and $t \geq \xi$. Estimates (3.29), (3.30) in conjunction with estimates (3.17), (3.19) and the weak semigroup property imply

$$V_1(t) \leq \max \left\{ \begin{array}{l} \tilde{\kappa}\left(\kappa\left(\sigma_1\left(a_1\left(\|x_0\|_{\mathcal{X}}\right) + c_1(t_0) + \mu_1(\xi - t_0), t - \xi\right)\right)\right), \\ \sup\limits_{t_0 \leq \tau \leq \xi} \tilde{\kappa}\left(\kappa\left(\sigma_1\left(p_1^u\left(q_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right), 0\right)\right)\right), \sup\limits_{t_0 \leq \tau \leq \xi} \tilde{\kappa}\left(\kappa\left(\sigma_1\left(p_1\left(V_2(\tau)\right), t - \xi\right)\right)\right) \\ \sup\limits_{\xi \leq \tau \leq t} \tilde{\varphi}\left(\gamma_1\left(\delta_1(\tau) V_2(\tau)\right)\right), \sup\limits_{\xi \leq \tau \leq t} \tilde{\kappa}\left(\varphi\left(\gamma_1^u\left(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right) \end{array} \right\},$$

(3.31)

$$\tilde{\varphi}\left(\gamma_1\left(\delta_1(t) V_2(t)\right)\right) \leq \max \left\{ \begin{array}{l} \tilde{\varphi}\left(\gamma_1\left(M\kappa\left(\kappa\left(\sigma_2\left(a_2\left(\|x_0\|_{\mathcal{X}}\right) + c_2(t_0) + \mu_2(\xi - t_0), t - \xi\right)\right)\right)\right)\right), \\ \sup\limits_{\xi \leq \tau \leq t} \tilde{\varphi}\left(\gamma_1\left(M\kappa\left(\varphi\left(\gamma_2^u\left(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right)\right)\right), \\ \sup\limits_{t_0 \leq \tau \leq \xi} \tilde{\varphi}\left(\gamma_1\left(M\kappa\left(\kappa\left(\sigma_2\left(p_2^u\left(q_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right), 0\right)\right)\right)\right)\right), \\ \tilde{\varphi}\left(\varphi^{-1}\left(\sup\limits_{\xi \leq \tau \leq t} V_1(\tau)\right)\right), \sup\limits_{t_0 \leq \tau \leq \xi} \tilde{\varphi}\left(\gamma_1\left(M\kappa\left(\kappa\left(\sigma_2\left(p_2\left(V_1(\tau)\right), t - \xi\right)\right)\right)\right)\right) \end{array} \right\}.$$

(3.32)

Estimate (3.32) combined with estimate (3.24) gives

$$\tilde{\varphi}\left(\gamma_1\left(\delta_1(t) V_2(t)\right)\right) \leq \max \left\{ \begin{array}{l} \tilde{\varphi}\left(\gamma_1\left(M\kappa\left(\kappa\left(\sigma_2\left(a_2\left(\|x_0\|_{\mathcal{X}}\right) + c_2(t_0) + \mu_2(\xi - t_0) \right.\right.\right.\right.\right. \\ \qquad \left.\left.\left.\left.\left. + p_2(\kappa(A)), t - \xi\right)\right)\right)\right)\right), \\ \sup\limits_{\xi \leq \tau \leq t} \tilde{\varphi}\left(\gamma_1\left(M\kappa\left(\varphi\left(\gamma_2^u\left(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right)\right)\right), \\ \sup\limits_{t_0 \leq \tau \leq \xi} \tilde{\varphi}\left(\gamma_1\left(M\kappa\left(\kappa\left(\sigma_2\left(p_2^u\left(q_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right), 0\right)\right)\right)\right)\right), \\ \tilde{\varphi}\left(\varphi^{-1}\left(\sup\limits_{\xi \leq \tau \leq t} V_1(\tau)\right)\right), \\ \sup\limits_{t_0 \leq \tau \leq t} \tilde{\varphi}\left(\gamma_1\left(M\kappa\left(\kappa\left(\sigma_2\left(p_2\left(\varphi\left(B\right)\right), 0\right)\right)\right)\right)\right) \end{array} \right\},$$

(3.33)

where

$$\begin{aligned} A &= \kappa\left(\kappa\left(\sigma_1\left(\beta_1(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right)\right)\right) + \gamma_1\left(M\kappa\left(\kappa\left(\sigma_2\left(\beta_2(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right)\right)\right)\right), \\ B &= \kappa\left(\varphi\left(\gamma_1^u\left(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right) + \gamma_1\left(M\kappa\left(\varphi\left(\gamma_2^u\left(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right)\right). \end{aligned}$$

Similarly, estimate (3.18) combined with estimate (3.24) and inequality (3.8) (which implies $L_i(t_0) \leq b\left(\|x_0\|_{\mathcal{X}}\right)$, $i = 1, 2$) gives

$$\begin{aligned} V_2(t) &\leq \sigma_2\left(\beta_2(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right) + \sup_{t_0 \leq \tau \leq t} \gamma_2^u\left(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right) \\ &+ \gamma_2\left(\tilde{\delta}_2(t)\kappa\left(\kappa\left(\kappa\left(\sigma_1\left(\beta_1(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right)\right)\right) + \gamma_1\left(M\kappa\left(\kappa\left(\sigma_2\left(\beta_2(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right)\right)\right)\right)\right)\right) \\ &+ \sup_{t_0 \leq \tau \leq t} \gamma_2\left(\tilde{\delta}_2(t)\varphi\left(\kappa\left(\varphi\left(\gamma_1^u\left(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right) + \gamma_1\left(M\kappa\left(\varphi\left(\gamma_2^u\left(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right)\right)\right)\right), \end{aligned}$$

(3.34)

where

$$\tilde{\delta}_2(t) := \max_{0 \leq \tau \leq t} \delta_2(\tau).$$

Consequently, by combining estimates (3.31) and (3.34) we obtain

$$V_1(t) \leq \max \left\{ \begin{array}{l} \tilde{\kappa}\left(\kappa\left(\sigma_1\left(a_1\left(\|x_0\|_{\mathcal{X}}\right) + c_1(t_0) + \mu_1(\xi - t_0), t - \xi\right)\right)\right), \tilde{\kappa}\left(\kappa\left(\sigma_1\left(p_1\left(\kappa(C)\right), t - \xi\right)\right)\right), \\ \sup\limits_{t_0 \leq \tau \leq \xi} \tilde{\kappa}\left(\kappa\left(\sigma_1\left(p_1^u\left(q_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right), 0\right)\right)\right), \sup\limits_{t_0 \leq \tau \leq t} \tilde{\kappa}\left(\kappa\left(\sigma_1\left(p_1\left(\varphi(D)\right), 0\right)\right)\right), \\ \sup\limits_{\xi \leq \tau \leq t} \tilde{\varphi}\left(\gamma_1\left(\delta_1(\tau) V_2(\tau)\right)\right), \sup\limits_{\xi \leq \tau \leq t} \tilde{\kappa}\left(\varphi\left(\gamma_1^u\left(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right), \tilde{\kappa}\left(\kappa\left(\sigma_1\left(p_1\left(E\right), t - \xi\right)\right)\right) \end{array} \right\},$$

(3.35)

where

$$C := \sigma_2\left(\beta_2(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right) + \gamma_2\Big(\tilde{\delta}_2(\xi)\kappa\left(\kappa\left(\kappa\left(\sigma_1\left(\beta_1(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right)\right)\right)\right.$$
$$\left. + \gamma_1\ \left(M\kappa\left(\kappa\left(\sigma_2\left(\beta_2(t_0) b\left(\|x_0\|_{\mathcal{X}}\right), 0\right)\right)\right)\right)\right),$$

$$D := \gamma_2^u\left(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right) + \gamma_2\left(\varphi\left(\frac{1}{2}\varphi^2\left(\kappa\left(\varphi\left(\gamma_1^u\left(\delta_1^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right)\right)\right.\right.$$
$$\left.\left. + \gamma_1\ \left(M\kappa\left(\varphi\left(\gamma_2^u\left(\delta_2^u(\tau)\|u(\tau)\|_{\mathcal{U}}\right)\right)\right)\right)\right)\right),$$

$$E := \gamma_2\left(\kappa\left(\frac{1}{2}\tilde{\delta}_2^2(\xi)\right)\right).$$

From (3.33) and (3.35) we conclude that there exist functions $S_1, S_2 \in KL$, continuous functions $M_1, M_2 : (\Re^+)^3 \to \Re^+$, and $\tilde{\gamma} \in \mathcal{N}$ such that the following estimates hold for all $\xi \in \pi(t_0, x_0, u, d)$ and $t \geq \xi$:

$$\tilde{\varphi}\left(\gamma_1\left(\delta_1(t)V_2(t)\right)\right) \leq \max\left\{\begin{array}{l} S_2\left(M_2\left(t_0, \xi - t_0, \|x_0\|_{\mathcal{X}}\right), t - \xi\right), \\ \tilde{\varphi}\left(\varphi^{-1}\left(\sup\limits_{\xi\leq\tau\leq t} V_1(\tau)\right)\right), \sup\limits_{t_0\leq\tau\leq t}\tilde{\gamma}\left(\delta(\tau)\|u(\tau)\|_{\mathcal{U}}\right) \end{array}\right\},$$

(3.36)

$$V_1(t) \leq \max\left\{\begin{array}{l} S_1\left(M_1\left(t_0, \xi - t_0, \|x_0\|_{\mathcal{X}}\right), t - \xi\right), \\ \sup\limits_{\xi\leq\tau\leq t}\tilde{\varphi}\left(\gamma_1\left(\delta_1(\tau)V_2(\tau)\right)\right), \sup\limits_{t_0\leq\tau\leq t}\tilde{\gamma}\left(\delta(\tau)\|u(\tau)\|_{\mathcal{U}}\right) \end{array}\right\}, \quad (3.37)$$

where $\delta \in K^+$ is defined by (3.9). Notice that if $\beta_1, \beta_2, c_1, c_2, \delta_2 \in K^+$ are bounded, then the functions $M_1, M_2 : (\Re^+)^3 \to \Re^+$ are independent of $t_0 \in \Re^+$ (but still depend on $\xi - t_0$). Moreover, the function $\tilde{\gamma} \in \mathcal{N}$ in addition to (3.28) satisfies for all $s \geq 0$:

$$\tilde{\gamma}(s) \geq \max\left\{\tilde{\kappa}\left(\kappa\left(\sigma_1\left(p_1^u(s), 0\right)\right)\right), \tilde{\kappa}\left(\kappa\left(\sigma_1\left(p_1\left(\varphi(D(s))\right), 0\right)\right)\right), \tilde{\kappa}\left(\varphi\left(\gamma_1^u(s)\right)\right)\right\}, \quad (3.38a)$$

$$\tilde{\gamma}(s) \geq \max\left\{\tilde{\varphi}\left(\gamma_1\left(M\kappa\left(\varphi\left(\gamma_2^u(s)\right)\right)\right)\right), \tilde{\varphi}\left(\gamma_1\left(M\kappa\left(\kappa\left(\sigma_2\left(p_2^u(s), 0\right)\right)\right)\right)\right),\right.$$
$$\left.\tilde{\varphi}\left(\gamma_1\left(M\kappa\left(\kappa\left(\sigma_2\left(p_2\left(\varphi\left(B(s)\right)\right), 0\right)\right)\right)\right)\right)\right\}, \quad (3.38b)$$

where $D(s) := \gamma_2^u(s) + \gamma_2\left(\varphi\left(\frac{1}{2}\varphi^2\left(\kappa\left(\varphi\left(\gamma_1^u(s)\right)\right) + \gamma_1\left(M\kappa\left(\varphi\left(\gamma_2^u(s)\right)\right)\right)\right)\right)\right)$ and $B(s) := \kappa\left(\varphi\left(\gamma_1^u(s)\right)\right) + \gamma_1\left(M\kappa\left(\varphi\left(\gamma_2^u(s)\right)\right)\right)$ are functions of class $\mathcal{N}$.

We define:

$$a_1(h, T, R) := \sup\left\{V_1(t_0 + h) - \sup_{t_0\leq\tau\leq t_0+h}\tilde{\gamma}\left(\delta(\tau)\|u(\tau)\|_{\mathcal{U}}\right);\right.$$
$$\left.\|x_0\|_{\mathcal{X}} \leq R, t_0 \in [0, T], d \in M_D, u \in M_U\right\}, \quad (3.39)$$

$$a_2(h, T, R) := \sup\left\{\tilde{\varphi}\left(\gamma_1\left(\delta_1(t_0 + h)V_2(t_0 + h)\right)\right) - \sup_{t_0\leq\tau\leq t_0+h}\tilde{\gamma}\left(\delta(\tau)\|u(\tau)\|_{\mathcal{U}}\right);\right.$$
$$\left.\|x_0\|_{\mathcal{X}} \leq R, t_0 \in [0, T], d \in M_D, u \in M_U\right\}, \quad (3.40)$$

$$l_1 := \limsup_{h\to+\infty} a_1(h, T, R), \quad l_2 := \limsup_{h\to+\infty} a_2(h, T, R). \quad (3.41)$$

By virtue of (3.24) and (3.27), the limits defined in (3.41) exist and are finite. Definition (3.41) implies that, for every $\varepsilon > 0$, $T \geq 0$, and $R \geq 0$, there exists a $\tau := \tau(\varepsilon, T, R) \geq 0$ such that

$$a_1(h, T, R) \leq l_1 + \varepsilon, \quad a_2(h, T, R) \leq l_2 + \varepsilon \quad \forall h \geq \tau. \quad (3.42)$$

By virtue of the weak semigroup property for system Σ, there exists a constant $r > 0$ such that for each $(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$ we have $\pi(t_0, x_0, u, d) \cap [t_0 + \tau, t_0 + \tau + r] \neq \emptyset$. Let $\xi \in \pi(t_0, x_0, u, d) \cap [t_0 + \tau, t_0 + \tau + r]$. Estimates (3.36), (3.37) in conjunction with definitions (3.39), (3.40) and inequalities (3.42) give

$$V_1(t) - \sup_{t_0 \leq \tau \leq t} \tilde{\gamma}\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right) \leq \max\left\{S_1\left(M_1\left(t_0, \xi - t_0, \left\|x_0\right\|_{\mathcal{X}}\right), t - \xi\right); l_2 + \varepsilon\right\}, \tag{3.43}$$

$$\begin{aligned}&\tilde{\varphi}\left(\gamma_1\left(\delta_1(t)V_2(t)\right)\right) - \sup_{t_0 \leq \tau \leq t} \tilde{\gamma}\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right)\\ &\leq \max\left\{\begin{array}{l} S_2\left(M_2\left(t_0, \xi - t_0, \left\|x_0\right\|_{\mathcal{X}}\right), t - \xi\right),\\ \tilde{\varphi}\left(\varphi^{-1}\left(l_1 + \varepsilon + \sup\limits_{t_0 \leq \tau \leq t} \tilde{\gamma}\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right)\right)\right) - \sup\limits_{t_0 \leq \tau \leq t} \tilde{\gamma}\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right)\end{array}\right\}.\end{aligned} \tag{3.44}$$

Using the identity $\tilde{\varphi}\left(\varphi^{-1}(s)\right) = s - \frac{1}{2}\rho\left(\varphi^{-1}(s)\right)$ and inequality (3.44), we get

$$\begin{aligned}&\tilde{\varphi}\left(\gamma_1\left(\delta_1(t)V_2(t)\right)\right) - \sup_{t_0 \leq \tau \leq t} \tilde{\gamma}\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right)\\ &\qquad \leq \max\left\{S_2\left(M_2\left(t_0, \xi - t_0, \left\|x_0\right\|_{\mathcal{X}}\right), t - \xi\right); l_1 + \varepsilon - \frac{1}{2}\rho\left(\varphi^{-1}\left(l_1\right)\right)\right\}.\end{aligned} \tag{3.45}$$

The properties of the KL functions in conjunction with estimates (3.43), (3.45), the fact that $\xi \in [t_0 + \tau, t_0 + \tau + r]$, and definitions (3.39), (3.40), (3.41) give for all $\varepsilon > 0$:

$$l_1 \leq l_2 + \varepsilon; \quad l_2 \leq l_1 + \varepsilon - \frac{1}{2}\rho\left(\varphi^{-1}\left(l_1\right)\right).$$

From the first inequality we obtain $l_1 \leq l_2$. The second inequality implies $\rho(\varphi^{-1}(l_1)) \leq 2\varepsilon$ for all $\varepsilon > 0$, which directly gives $l_1 = l_2 = 0$.

Definitions (3.39), (3.40), and (3.41) imply that P3 of Lemma 2.16 holds for system Σ with $V = V_1$ or $V = \tilde{\varphi}\left(\gamma_1\left(\delta_1(t)V_2\right)\right)$, for appropriate $\tilde{\gamma} \in \mathcal{N}$ (which satisfies (3.28) and (3.39)) and $\delta \in K^+$ as defined by (3.9).

Notice that if $\beta_1, \beta_2, c_1, c_2, \delta_2 \in K^+$ are bounded, then definitions (3.39), (3.40) are modified as follows:

$$\begin{aligned}a_1(h, R) := \sup\Big\{&V_1(t_0 + h) - \sup_{t_0 \leq \tau \leq t_0 + h} \tilde{\gamma}\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right);\\ &\left\|x_0\right\|_{\mathcal{X}} \leq R, t_0 \geq 0, d \in M_D, u \in M_U\Big\},\end{aligned} \tag{3.46}$$

$$\begin{aligned}a_2(h, R) := \sup\Big\{&\varphi\left(\gamma_1\left(\delta_1(t_0 + h)V_2(t_0 + h)\right)\right) - \sup_{t_0 \leq \tau \leq t_0 + h} \tilde{\gamma}\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right);\\ &\left\|x_0\right\|_X \leq R, t_0 \geq 0, d \in M_D, u \in M_U\Big\}.\end{aligned} \tag{3.47}$$

Similar arguments as above show that property P3 of Lemma 2.17 holds for system Σ with $V = V_1$ or $V = \tilde{\varphi}\left(\gamma_1\left(\delta_1(t)V_2\right)\right)$, for appropriate $\tilde{\gamma} \in \mathcal{N}$ (which satisfies (3.28) and (3.38)) and $\delta \in K^+$ as defined by (3.9). The proof is complete. □

Remark 3.6. If the functions $\gamma_1^u, p_1^u, \gamma_2^u, p_2^u \in \mathcal{N}$ are all identically zero, then it follows that the gain function $\gamma \in \mathcal{N}$ is identically zero. Indeed, the reader should notice that $\gamma \in \mathcal{N}$ may be selected as $\gamma(s) := a\left(4\tilde{\gamma}(s)\right) \in \mathcal{N}$, where $a \in \mathcal{N}$ is the function involved in hypothesis (H4) and $\tilde{\gamma} \in \mathcal{N}$ is the function that satisfies (3.28), (3.38a)–(3.38b). Moreover, notice that for the input free case Theorem 3.1 and Corollary 3.4 imply (uniform) robust global asymptotic output stability (RGAOS)

for the corresponding system. The following example shows the applicability of this particular remark to systems with impulses at fixed times.

Example 3.7. Consider the following system:

$$\begin{aligned} \dot{z}(t) &= Az(t) + g(x(t)), \\ \dot{x}(t) &= f(x(t)), t \notin \pi, \end{aligned} \tag{3.48a}$$

$$x(\tau_i) = h\left(\lim_{t \to \tau_i^-} x(t)\right), \tag{3.48b}$$

$$z(t) \in \Re^k, x(t) \in \Re^n,$$

where $A \in \Re^{k \times k}$ is a Hurwitz matrix, $\pi = \{\tau_i\}_{i=0}^{\infty}$ is a partition of $\Re^+$ with diameter $r > 0$, $f : \Re^n \to \Re^n$, $g : \Re^n \to \Re^k$, $h : \Re^n \to \Re^n$ are continuous vector fields, $f(x)$ being locally Lipschitz with respect to $x \in \Re^n$, with $f(0) = 0$, $g(0) = 0$, $h(0) = 0$. Notice that subsystem (3.48a) is a system described by ODEs which satisfies hypotheses (A1)–(A3) of Example 2.8. Hence, subsystem (3.48a) satisfies the BIC property, and $0 \in \Re^k$ is a robust equilibrium point from the input x. Moreover, subsystem (3.48b) is a hybrid system with impulses at fixed times, which satisfies hypotheses (Q1)–(Q4) of Example 2.12. Hence, subsystem (3.48b) satisfies the BIC property, and $0 \in \Re^n$ is a robust equilibrium point (from the zero input). We remark that both subsystems (3.48a)–(3.48b) satisfy the classical semigroup property, and consequently the composite system (3.48) can be regarded as the feedback interconnection of subsystems (3.48a)–(3.48b).

Since $A \in \Re^{k \times k}$ is Hurwitz, it follows that subsystem (3.48a) satisfies the UISS property from the input x. Moreover, if there exists a C^1 positive definite and radially unbounded function $V : \Re^n \to \Re^+$ and constants $c_1, c_2 \in \Re$, with $c_2 \neq 0$, $\mu, \lambda > 0$, such that

$$\nabla V(x) f(x) \leq -c_1 V(x) \quad \forall x \in \Re^n, \tag{3.49a}$$

$$V(h(x)) \leq \exp(-c_2) V(x) \quad \forall x \in \Re^n, \tag{3.49b}$$

$$-c_2 card\,(\pi \cap [s, s+t)) \leq \mu + (c_1 - \lambda) t \forall s, t \in \Re^+. \tag{3.49c}$$

where $card(S)$ denotes the cardinal number of the set S, then Theorem 1 in [7] implies that $0 \in \Re^n$ is uniformly globally asymptotically stable for subsystem (3.48b). Taking into account Remarks 3.2(b) and 3.6, we conclude that $0 \in \Re^k \times \Re^n$ is uniformly globally asymptotically stable for the composite system (3.48) under the hypotheses stated above.

4. Application to partial-state sampled-data control. In this section we present applications of the small-gain results (Theorem 3.1 and Corollary 3.4) to partial-state sampled-data control problems. It should be emphasized that sampled-data control systems cannot be handled with small-gain results that have appeared so far in the literature, since sampled-data control systems do not satisfy the classical semigroup property (see Example 2.11).

Consider the following control system described by ODEs:

$$\begin{aligned} \dot{z} &= f(t, d, z, x, u), \\ & z \in \Re^k\,,\ d \in D\,,\ u \in U\,,\ t \geq 0, \end{aligned} \tag{4.1a}$$

$$\begin{aligned} \dot{x} &= Ax + Bv + Bg(t, d, z, u), \\ & x \in \Re^n\,,\ v \in \Re\,,\ d \in D\,,\ u \in U\,,\ t \geq 0, \end{aligned} \tag{4.1b}$$

where (A, B) is a controllable pair of matrices, $D \subset \Re^l$ is a compact set, $U \subseteq \Re^p$ is nonempty, with $0 \in U$, and the mappings $f : \Re^+ \times D \times \Re^k \times \Re^n \times U \to \Re^k$, $g : \Re^+ \times D \times \Re^k \times U \to \Re$ are continuous, locally Lipschitz in (z, x), uniformly in $d \in D$, with $f(t, d, 0, 0, 0) = 0$, $g(t, d, 0, 0) = 0$ for all $(t, d) \in \Re^+ \times D$. The problem we consider is the (W)ISS stabilization problem for (4.1) with sampled-data feedback applied with zero order hold and depending only on $x \in \Re^n$; i.e., we want to find a function $k : \Re^n \to \Re$ with $k(0) = 0$ and a constant $r > 0$ such that system (4.1a) with

$$\begin{aligned} &\dot{x}(t) = Ax(t) + Bk(x(\tau_i)) + Bg(t, d(t), z(t), u(t)), t \in [\tau_i, \tau_{i+1}), \\ &\tau_{i+1} = \tau_i + \exp(-w(\tau_i))\, r, \quad w(t) \in \Re^+ \end{aligned} \tag{4.2}$$

satisfies the WISS property from the inputs (u, w). Notice that the input w has been introduced in order to quantify the uncertainty in sampling times; i.e., we have to guarantee stability properties for the closed-loop system (4.1a)–(4.2) for all sampling schedules of diameter less than or equal to $r > 0$. To this purpose we make the following assumptions.

(A1) System (4.1a) satisfies the WISS property from the inputs x and u. Specifically, there exist functions $\sigma \in KL$, $\beta, \delta_1^u \in K^+$, γ_1, $\gamma_1^u \in \mathcal{N}$ such that for all $(t_0, z_0, d, x, u) \in \Re^+ \times \Re^k \times \mathcal{L}^\infty_{loc}(\Re^+; D) \times \mathcal{L}^\infty_{loc}(\Re^+; \Re^n) \times \mathcal{L}^\infty_{loc}(\Re^+; U)$ the solution of (4.1) with initial condition $z(t_0) = z_0$ corresponding to inputs $(d, x, u) \in \mathcal{L}^\infty_{loc}(\Re^+; D) \times \mathcal{L}^\infty_{loc}(\Re^+; \Re^n) \times \mathcal{L}^\infty_{loc}(\Re^+; U)$ satisfies the following estimate for all $t \geq t_0$:

$$|z(t)| \leq \sigma\left(\beta(t_0)\,|z_0|, t - t_0\right) + \sup_{t_0 \leq \tau \leq t} \gamma_1\left(|x(\tau)|\right) + \sup_{t_0 \leq \tau \leq t} \gamma_1^u\left(\delta_1^u(\tau)\,|u(\tau)|\right). \tag{4.3}$$

(A2) There exist functions $\delta_2^u \in K^+$, γ_2, $\gamma_2^u \in \mathcal{N}$ such that the following inequality holds for all $(t, z, d, u) \in \Re^+ \times \Re^k \times D \times U$:

$$|g(t, d, z, u)| \leq \gamma_2(|z|) + \gamma_2^u\left(\delta_2^u(t)\,|u|\right). \tag{4.4}$$

(A3) There exist a function $\rho \in K_\infty$and a constant $R \geq 1$ such that

$$\gamma_1\left(R^{-1}\gamma_2(s) + \rho\left(R^{-1}\gamma_2(s)\right)\right) + \rho\left(\gamma_1\left(R^{-1}\gamma_2(s) + \rho\left(R^{-1}\gamma_2(s)\right)\right)\right) \leq s \ \forall s \geq 0. \tag{4.5}$$

For example, hypothesis (A3) holds if $\gamma_i(s) = K_i s$, where $K_i \geq 0$ $(i = 1, 2)$, i.e., if the gain functions are linear.

Next we show that the problem of WISS stabilization problem for (4.1) with sampled-data feedback applied with zero order hold and depending only on $x \in \Re^n$ is solvable under hypotheses (A1)–(A3) by linear feedback. The proof of this result will be made by making use of Corollary 3.4.

Notice that since (A, B) is a controllable pair of matrices, it follows that for every $\mu > 0$, $R \geq 1$ there exist a symmetric positive definite matrix $P \in \Re^{n \times n}$, a vector $k \in \Re^n$, and constants $Q_1, Q_2 > 0$ such that the following inequalities hold for all $(x, u) \in \Re^n \times \Re$:

$$\begin{aligned} Q_1\,|x|^2 \leq x'Px \leq Q_2\,|x|^2, \\ 2x'P(A + Bk')x + 2x'PBu \leq -4\mu x'Px + \frac{Q_1}{4\mu R^2}u^2. \end{aligned} \tag{4.6}$$

Next we show the following claim.

Claim. *For every $\mu > 0$, $R \geq 1$ there exists a vector $k \in \Re^n$ and constants $M, r > 0$ such that for all $(t_0, x_0, d, z, u, w) \in \Re^+ \times \Re^n \times \mathcal{L}^\infty_{loc}(\Re^+; D) \times \mathcal{L}^\infty_{loc}(\Re^+; \Re^k) \times$*

$\mathcal{L}^\infty_{loc}(\Re^+;U)\times\mathcal{L}^\infty_{loc}(\Re^+;\Re^+)$ *the solution of the hybrid system*

$$\begin{aligned}\dot{x}(t) &= Ax(t)+Bk'x(\tau_i)+Bg(t,d(t),z(t),u(t)), t\in[\tau_i,\tau_{i+1}),\\ \tau_{i+1} &= \tau_i+\exp\left(-w(\tau_i)\right)r, \quad w(t)\in\Re^+,\end{aligned} \tag{4.7}$$

with initial condition $x(t_0)=x_0$ *corresponding to inputs* $(d,z,u,w)\in\mathcal{L}^\infty_{loc}(\Re^+;D)\times\mathcal{L}^\infty_{loc}(\Re^+;\Re^k)\times\mathcal{L}^\infty_{loc}(\Re^+;U)\times\mathcal{L}^\infty_{loc}(\Re^+;\Re^+)$*, satisfies the following estimate for all* $t\geq t_0$*:*

$$|x(t)|\leq M\exp\left(-\mu(t-t_0)\right)|x_0|+\sup_{t_0\leq\tau\leq t}R^{-1}\gamma_2\left(|z(\tau)|\right)+\sup_{t_0\leq\tau\leq t}R^{-1}\gamma_2^u\left(\delta_2^u(\tau)\,|u(\tau)|\right), \tag{4.8}$$

where $\delta_2^u\in K^+$*,* γ_2*,* $\gamma_2^u\in\mathcal{N}$ *are the functions involved in* (4.4).

Proof of Claim. Let arbitrary $\mu>0, R>1$. Since (A,B) is a controllable pair of matrices, it follows that there exists a symmetric positive definite matrix $P\in\Re^{n\times n}$, a vector $k\in\Re^n$, and constants $Q_1,Q_2>0$ such that inequalities (4.6) hold for all $(x,u)\in\Re^n\times\Re$. Let arbitrary $(t_0,x_0,d,z,u,w)\in\Re^+\times\Re^n\times\mathcal{L}^\infty_{loc}(\Re^+;D)\times\mathcal{L}^\infty_{loc}(\Re^+;\Re^k)\times\mathcal{L}^\infty_{loc}(\Re^+;U)\times\mathcal{L}^\infty_{loc}(\Re^+;\Re^+)$, and consider the solution $x(t)$ of (4.7) with initial condition $x(t_0)=x_0$ corresponding to inputs $(d,z,u,w)\in\mathcal{L}^\infty_{loc}(\Re^+;D)\times\mathcal{L}^\infty_{loc}(\Re^+;\Re^k)\times\mathcal{L}^\infty_{loc}(\Re^+;U)\times\mathcal{L}^\infty_{loc}(\Re^+;\Re^+)$ (the solution exists for all $t\geq t_0$). Finally, consider the function $V(t):=x'(t)Px(t)$, which is absolutely continuous on $[t_0,+\infty)$. By virtue of (4.6) the derivative of $V(t)$ satisfies a.e. on the interval $[\tau_i,\tau_{i+1})$:

$$\dot{V}(t)\leq-4\mu V(t)+2x'PBk'(x(\tau_i)-x(t))+\frac{Q_1}{4\mu R^2}\left|g(t,d(t),z(t),u(t))\right|^2. \tag{4.9}$$

Let $r>0$ be a constant that satisfies

$$\begin{gathered}r\leq\frac{2\mu}{M\,|Bk'|\,|A+Bk'|\,(2\,|A+Bk'|+|B|)+2\mu\,|A|+2\mu\,|A+Bk'|};\\ r\leq\frac{1}{4\mu R^2M\,|B|\,|Bk'|+|A|+|A+Bk'|},\end{gathered} \tag{4.10}$$

where $M:=\frac{Q_2}{Q_1}\geq1$.

It follows from (4.7) that $|x(t)-x(\tau_i)|\leq r|A|\sup_{\tau_i\leq s\leq t}|x(s)-x(\tau_i)|+r|A+Bk'||x(\tau_i)|+r|B|\sup_{\tau_i\leq s\leq t}|g(s,d(s),z(s),u(s))|$, which directly implies $|x(t)-x(\tau_i)|\leq\frac{r|A+Bk'|}{1-r|A}|x(\tau_i)|+\frac{r|B|}{1-r|A|}\sup_{\tau_i\leq s\leq t}|g(s,d(s),z(s),u(s))|$, for all $t\in[\tau_i,\tau_{i+1})$. Moreover, the previous inequality in conjunction with the triangle inequality $|x(\tau_i)|\leq|x(t)-x(\tau_i)|+|x(t)|$ implies the estimate $|x(t)-x(\tau_i)|\leq\frac{r|A+Bk'|}{1-r|A|-r|A+Bk'|}|x(t)|+\frac{r|B|}{1-r|A|-r|A+Bk'|}\sup_{\tau_i\leq s\leq t}|g(s,d(s),z(s),u(s))|$ for all $t\in[\tau_i,\tau_{i+1})$. Using the previous inequality in conjunction with (4.9) and completing the squares, we obtain for almost all $t\in[\tau_i,\tau_{i+1})$:

$$\begin{aligned}\dot{V}(t)\leq&-\left(4\mu-\frac{rM\,|Bk'|\;|A+Bk'|\;(2\,|A+Bk'|+|B|)}{1-r\,(|A|+|A+Bk'|)}\right)V(t)\\ &+\left(\frac{Q_1}{4\mu\,R^2}+\frac{r\,|B|\;Q_2\,|Bk'|}{1-r\,(|A|+|A+Bk'|)}\right)\sup_{\tau_i\leq s\leq t}|g(s,d(s),z(s),u(s))|^2.\end{aligned} \tag{4.11}$$

It follows from inequalities (4.10), (4.11) that the following estimate holds for the derivative of $V(t)$ a.e. on the interval $[\tau_i,\tau_{i+1})$:

$$\dot{V}(t)\leq-2\mu V(t)+\frac{Q_1}{2\mu R^2}\sup_{t_0\leq s\leq t}|g(s,d(s),z(s),u(s))|^2. \tag{4.12}$$

Notice that, since estimate (4.12) does not depend on the particular interval $[\tau_i, \tau_{i+1})$, we may conclude that estimate (4.12) holds a.e. for $t \geq t_0$. Estimate (4.12) implies directly that $V(t) \leq \exp(-2\mu(t-t_0))V(t_0) + \frac{Q_1}{R^2}\sup_{t_0\leq s\leq t}|g(s,d(s),z(s),u(s))|^2$ for all $t \geq t_0$. Finally, estimate (4.8) is an immediate consequence of the previous inequality, definitions $V(t) := x'(t)Px(t)$ and $M := \frac{Q_2}{Q_1} \geq 1$, as well as inequalities (4.4) and (4.6). The proof of the claim is complete. □

By making use of Corollary 3.4 and specifically Remark 3.5(b), we may conclude that the closed-loop system (4.1a) with (4.7) satisfies the WISS property from the inputs u and w, when $R > 1$ is chosen to be greater than or equal to the constant involved in hypothesis (A3). Moreover, the gain function for the input w is identically zero. Furthermore, if the functions $\delta_1^u, \delta_2^u \in K^+$ are bounded, then the closed-loop system (4.1a) with (4.7) satisfies the ISS property from the inputs u and w. Finally, if in addition $\beta \in K^+$ is bounded, then the closed-loop system (4.1a) with (4.7) satisfies the UISS property from the inputs u and w.

Example 4.1. The following planar system described by ODEs:

$$\begin{aligned} &\dot{z} = -z^3 + zx, \\ &\dot{x} = d(t)z^2 + u + v, \\ &(z,x)' \in \Re^2, u, v \in \Re \end{aligned} \tag{4.13}$$

is studied in [11], where it is shown that if $d(t) \equiv d$ with $|d| < \frac{1}{2}$, then the feedback law $v(t) = -x(t)$ guarantees the UISS property for the closed-loop system from the input $u \in \Re$. The proof of this fact is made by using a slightly modified version of the small-gain theorem presented in [14]. Here we study the possibility of robustly globally stabilizing the origin for system (4.13), using the following feedback law with zero order hold and a positive sampling rate:

$$\begin{aligned} &v(t) = -x(\tau_i), t \in [\tau_i, \tau_{i+1}), \\ &\tau_{i+1} = \tau_i + \exp(-w(\tau_i))r, w(t) \in \Re^+ \end{aligned} \tag{4.14}$$

for time-varying disturbances $d(t) \in D := [-\delta, \delta]$, with $\delta \in (0,1)$.

First notice that system (4.13) is a system of the form (4.1) with $f(t,d,z,x,u) = -z^3 + zx$, $g(t,d,z,u) = dz^2 + u$, $B = [1]$, $A = [0]$. Moreover, hypothesis (A2) holds with $\gamma_2(s) = \delta s^2$, $\gamma_2^u(s) := s$, and $\delta_2^u(t) \equiv 1$.

Working exactly as in [11] it may be shown that for every $\varepsilon \in (0,1)$ the subsystem Σ_1:

$$\dot{z} = -z^3 + zx \tag{4.15}$$

satisfies the UISS property from the input x with gain function $\gamma_1(s) := \sqrt{\frac{s}{1-\varepsilon}}$. Thus the subsystem Σ_1 satisfies hypothesis (A1) with $\gamma_1(s) := \sqrt{\frac{s}{1-\varepsilon}}$, $\beta(t) \equiv 1$, $\gamma_2^u(s) := s$, and appropriate $\sigma \in KL$.

For every $\delta \in (0,1)$ there exist $\varepsilon \in (0,1)$ and $L > 0$ such that

$$(1+L)\sqrt{\frac{(1+L)\delta}{1-\varepsilon}} \leq 1. \tag{4.16}$$

Selecting $\rho(s) := Ls$, with $L > 0$, we conclude from (4.16) that (4.5) holds with $R = 1$. Finally, since (4.6) holds with $Q_1 = Q_2 = 1$, $P = [1]$, $\mu = \frac{1}{4}$, and $k = -1$, we conclude that (4.13) with (4.14) satisfies the UISS property from the inputs u and w. Moreover, the gain function for the input w is identically zero. The maximum allowable sampling period (r) may be determined by inequalities (4.10), which give $r = 1/7$.

5. Conclusions. A small-gain theorem, which can be applied to a wide class of systems that includes systems that satisfy the weak semigroup property, is presented in the present work. The result generalizes all existing results in the literature and exploits notions of weighted, uniform, and nonuniform IOS property. Moreover, the small-gain theorem of the present work is a method for establishing qualitative properties expressed in a very general framework unifying works from various fields as well as different stability notions. The results presented in the paper can be extended without much difficulty to the case of local stability notions.

Applications to partial-state feedback stabilization problems with sampled-data feedback applied with zero order hold and a positive sampling rate are also presented. It should be emphasized that sampled-data control systems cannot be handled with small-gain results that have appeared so far in the literature, since sampled-data control systems do not satisfy the classical semigroup property. The results are illustrated by examples which show the usefulness of the main result for the stability analysis of interconnected systems. Other promising applications of the new generalized small-gain theorem include the popular topics of hybrid systems and networked control systems. Some initial, interesting results can be found in [30, 39, 41].

Appendix A.

Proof of Lemma 2.13. Lemma 3.5 in [24] guarantees that the control system $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ which has the BIC property is RFC from the input $u \in \mathcal{M}(U)$ if and only if there exist functions $q \in K^+$, $a \in K_\infty$ and a constant $R \geq 0$ such that the following estimate holds for all $(t_0, x_0, d, u) \in \Re^+ \times \mathcal{X} \times M_D \times M_U$ and $t \geq t_0$:

$$\|\phi(t, t_0, x_0, u, d)\|_{\mathcal{X}} \leq q(t) a \left(R + \|x_0\|_{\mathcal{X}} + \sup_{t_0 \leq \tau \leq t} \|u(\tau)\|_{\mathcal{U}} \right). \tag{A.1}$$

It should be emphasized that all results in [24] were proved under the assumption of the classical semigroup property for the control system. However, the proof of Lemma 3.5 does not depend on the semigroup property and consequently may be repeated as it stands for a system which satisfies the weak semigroup property. Let $\beta \in K^+$ arbitrary. Using (A.1) we obtain

$$\begin{aligned}
&\beta(t) \|\phi(t, t_0, x_0, u, d)\|_{\mathcal{X}} \\
&\quad \leq \frac{1}{2} q^2(t) \beta^2(t) + \frac{1}{2} \max \left\{ a^2(3R); a^2\left(3 \|x_0\|_{\mathcal{X}}\right); \sup_{t_0 \leq \tau \leq t} a^2\left(3 \|u(\tau)\|_{\mathcal{U}}\right) \right\} \\
&\quad \leq \max \left\{ q^2(t)\beta^2(t)\,;\, a^2(3R)\,;\, a^2\left(3 \|x_0\|_{\mathcal{X}}\right)\,;\, \sup_{t_0 \leq \tau \leq t} a^2\left(3 \|u(\tau)\|_{\mathcal{U}}\right) \right\} \\
&\quad \leq \max \left\{ \gamma(t)\,;\, a^2\left(3 \|x_0\|_{\mathcal{X}}\right)\,;\, \sup_{t_0 \leq \tau \leq t} a^2\left(3 \|u(\tau)\|_{\mathcal{U}}\right) \right\},
\end{aligned} \tag{A.2}$$

where $\gamma(t) = q^2(t)\beta^2(t) + a^2(3R)$. Define

$$a(T, s) := \max \{ \gamma(t_0 + h) - \gamma(t_0) : h \in [0, s], t_0 \in [0, T] \}. \tag{A.3}$$

Clearly, definition (A.3) implies that, for each fixed $s \geq 0$, $a(\cdot, s)$ is nondecreasing and, for each fixed $T \geq 0$, $a(T, \cdot)$ is nondecreasing. Furthermore, continuity of γ guarantees that, for every $T \geq 0, \lim_{s \to 0^+} a(T, s) = a(T, 0) = 0$. It turns out from Lemma 2.3 in [21] that there exist functions $\zeta \in K_\infty$ and $\kappa \in K^+$ such that

$$a(T, s) \leq \zeta(\kappa(T)s) \quad \forall (T, s) \in \left(\Re^+\right)^2. \tag{A.4}$$

Combining definition (A.3) with inequality (A.4), we conclude that, for all $t_0 \geq 0$ and $t \geq t_0$, it holds that

$$\begin{aligned}\gamma(t) &\leq \gamma(t_0) + \zeta\left(\kappa(t_0)\,(t-t_0)\right) \leq \gamma(t_0) + \zeta\left(\tfrac{1}{2}\kappa^2(t_0) + \tfrac{1}{2}\,(t-t_0)^2\right)\\ &\leq \gamma(t_0) + \zeta\left(\kappa^2(t_0)\right) + \zeta\left((t-t_0)^2\right) \leq \max\left\{2\gamma(t_0) + 2\zeta\left(\kappa^2(t_0)\right); 2\zeta\left((t-t_0)^2\right)\right\}.\end{aligned}$$

The above inequality in conjunction with (A.2) implies that (2.9) holds for all $(t_0, x_0, d, u) \in \Re^+ \times \mathcal{X} \times M_D \times M_U$ and $t \geq t_0$ with $\mu(t) := 2\zeta\left(t^2+1\right)$, $c(t) := 2\gamma(t) + 2\zeta\left(\kappa^2(t)\right)$, $a(s) := a^2(3s)$, and $p(s) := a^2(3s)$. The proof is complete. □

Proof of Lemma 2.16. As in the proof of Proposition 2.2 in [21], let $T, h \geq 0$, $s \geq 0$, and define

$$\begin{aligned}a(T,s) := \sup\Big\{ & V(t, \phi(t, t_0, x_0, u, d), u(t)) - \sup_{t_0 \leq \tau \leq t} \gamma\left(\delta(\tau)\,\|u(\tau)\|_{\mathcal{U}}\right);\\ & \|x_0\|_{\mathcal{X}} \leq s, t \geq t_0 \in [0,T], d \in M_D, u \in M_U\Big\},\end{aligned} \tag{A.5}$$

$$\begin{aligned}M(h,T,s) := \sup\Big\{ & V(t_0+h, \phi(t_0+h, t_0, x_0, u, d), u(t)) - \sup_{t_0 \leq \tau \leq t_0+h} \gamma\left(\delta(\tau)\,\|u(\tau)\|_{\mathcal{U}}\right);\\ & \|x_0\|_X \leq s, t_0 \in [0,T], d \in M_D, u \in M_U\Big\}.\end{aligned} \tag{A.6}$$

First notice that by virtue of property P1 it holds that $a(T,s) < +\infty$ for all $T \geq 0$, $s \geq 0$. Moreover, notice that, since $0 \in \mathcal{X}$ is a robust equilibrium point from the input $u \in M_U$ and $V(t,0,0) = 0$ for all $t \geq 0$, we have $a(T,s) \geq 0$ for all $T \geq 0$, $s \geq 0$. Furthermore, notice that M is well-defined, since by definitions (A.5), (A.6) the following inequality is satisfied for all $T, h \geq 0$ and $s \geq 0$:

$$0 \leq M(h,T,s) \leq a(T,s). \tag{A.7}$$

Clearly, definition (A.5) implies that, for each fixed $s \geq 0$, $a(\cdot, s)$ is nondecreasing and, for each fixed $T \geq 0$, $a(T,\cdot)$ is nondecreasing. Furthermore, property P2 asserts that, for every $T \geq 0$, $\lim_{s \to 0^+} a(T,s) = 0$. Hence, the inequality $a(T,0) \geq 0$ for all $T \geq 0$, in conjunction with $\lim_{s\to 0^+} a(T,s) = 0$ and the fact that $a(T,\cdot)$ is nondecreasing, implies $a(\cdot,0) = 0$. It turns out from Lemma 2.3 in [21] that there exist functions $\zeta \in K_\infty$ and $q \in K^+$ such that

$$a(T,s) \leq \zeta(q(T)s) \quad \forall (T,s) \in \left(\Re^+\right)^2. \tag{A.8}$$

Without loss of generality we may assume that $q \in K^+$ is nondecreasing. Moreover, property P3 guarantees that, for every $\varepsilon > 0$, $T \geq 0$, and $R \geq 0$, there exists a $\tau = \tau(\varepsilon, T, R) \geq 0$ such that

$$M(h,T,s) \leq \varepsilon \;\forall\; h \geq \tau(\varepsilon,T,R) \text{ and } 0 \leq s \leq R. \tag{A.9}$$

Let

$$g(s) := \sqrt{s} + s^2, \tag{A.10}$$

and let p be a nondecreasing function of class K^+, with $p(0) = 1$ and

$$\lim_{t \to +\infty} p(t) = +\infty. \tag{A.11}$$

Define

$$\mu(h) := \sup\left\{\frac{M(h,T,s)}{p(T)g\left(\zeta\left(q(T)s\right)\right)};\quad T\geq 0, s>0\right\}. \tag{A.12}$$

Obviously, by virtue of (A.7), (A.8), and (A.10) the function $\mu : \Re^+ \to \Re^+$ is well-defined and satisfies $\mu(\cdot)\leq 1$. We show that $\lim_{h\to+\infty}\mu(h)=0$; equivalently, we establish that for any given $\varepsilon>0$ there exists a $\delta=\delta(\varepsilon)\geq 0$ such that

$$\mu\left(h\right)\leq\varepsilon \text{ for } h\geq\delta\left(\varepsilon\right). \tag{A.13}$$

Notice first that for any given $\varepsilon>0$ there exist constants $a:=a(\varepsilon)$ and $b:=b(\varepsilon)$, with $0<a<b$, such that

$$x\notin(a,b)\Rightarrow\frac{x}{\sqrt{x}+x^2}\leq\varepsilon. \tag{A.14}$$

We next recall (A.11), which asserts that, for the above ε for which (A.14) holds, there exists a $c:=c(\varepsilon)\geq 0$ such that $p(T)\geq\frac{1}{\varepsilon}$ for all $T\geq c$. This by virtue of (A.7), (A.8), (A.10), and (A.14) yields

$$\frac{M(h,T,s)}{p(T)g(\zeta(q(T)s))}\leq\varepsilon\quad\forall h\geq 0,\ \text{when } T\geq c \text{ or } \zeta(q(T)s)\notin(a,b). \tag{A.15}$$

Hence, in order to establish (A.13), it remains to consider the case:

$$a\leq\zeta(q(T)s)\leq b \text{ and } 0\leq T\leq c. \tag{A.16}$$

Since, for each fixed $(h,s)\in\left(\Re^+\right)^2$, the mappings $M(h,\cdot,s)$, $M(h,s,\cdot)$, $q(\cdot)$, and $p(\cdot)$ are nondecreasing, we have that

$$\frac{M(h,T,s)}{p(T)g\left(\zeta\left(q(T)s\right)\right)}\leq\frac{M\left(h,c,\frac{\zeta^{-1}(b)}{q(0)}\right)}{g\left(a\right)}, \tag{A.17}$$

provided that (A.16) holds. By using (A.9) and (A.17) with

$$\varepsilon:=\varepsilon g(a), T:=c, R:=\frac{\zeta^{-1}(b)}{q(0)},$$

it follows

$$M\left(h,c,\frac{\zeta^{-1}(b)}{q(0)}\right)\leq\varepsilon g\left(a\right)\ \text{ for } h\geq\delta\left(\varepsilon\right):=\tau\left(\varepsilon\, g(a),c,\frac{\zeta^{-1}(b)}{q(0)}\right). \tag{A.18}$$

By taking into account (A.15), (A.16), (A.17), (A.18), and definition (A.12) of $\mu\left(\cdot\right)$, it follows that (A.13) holds with $\delta=\delta\left(\varepsilon\right)$ as selected in (A.18). Since $\varepsilon>0$ was arbitrary we conclude that $\lim_{h\to+\infty}\mu(h)=0$. Consequently, there exists a continuous strictly decreasing function $\bar{\mu}:\Re^+\to(0,+\infty)$ such that $\bar{\mu}(h)\geq\mu(h)$ for all $h\geq 0$ and $\lim_{h\to+\infty}\bar{\mu}(h)=0$. Thus, by recalling definition (A12) we obtain

$$M(h,T,s)\leq\bar{\mu}(h)\theta(T,s)\quad\forall(T,s)\in\left(\Re^+\right)^2,\quad\forall h\geq 0, \tag{A.19}$$

where $\theta(T,s):=p(T)g(\zeta(q(T)s))$. Clearly, θ satisfies all hypotheses of Lemma 2.3 in [21], and therefore there exist $\zeta_2\in K_\infty$ and $\beta\in K^+$ such that

$$\theta(T,s)\leq\zeta_2(\beta(T)s)\quad\forall(T,s)\in\left(\Re^+\right)^2. \tag{A.20}$$

Thus definition (A.6) implies that the following estimate holds for all $u \in M_U$, $(t_0, x_0, d) \in \Re^+ \times \mathcal{X} \times M_D$, and $t \geq t_0$:

$$V(t, \phi(t, t_0, x_0, u, d), u(t)) \leq \bar{\mu}(t - t_0)\zeta_2\left(\beta(t_0)\left\|x_0\right\|_X\right) + \sup_{t_0 \leq \tau \leq t} \gamma\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right). \tag{A.21}$$

Estimate (A.21) implies (2.12) with $\sigma(s,t) := \bar{\mu}(t)\zeta_2(s)$. □

Proof of Lemma 2.17. As in the proof of Lemma 2.16, let $h \geq 0$, $s \geq 0$, and define

$$a(s) := \sup\left\{V(t, \phi(t, t_0, x_0, u, d), u(t)) - \sup_{t_0 \leq \tau \leq t} \gamma\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right); \right. \\ \left. \left\|x_0\right\|_{\mathcal{X}} \leq s, t \geq t_0 \geq 0, d \in M_D, u \in M_U\right\}, \tag{A.22}$$

$$M(h,s) := \sup\left\{V(t_0 + h, \phi(t_0 + h, t_0, x_0, u, d), u(t)) - \sup_{t_0 \leq \tau \leq t_0 + h} \gamma\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right); \right. \\ \left. \left\|x_0\right\|_{\mathcal{X}} \leq s, t_0 \geq 0, d \in M_D, u \in M_U\right\}. \tag{A.23}$$

First notice that by virtue of property P1 it holds that $a(s) < +\infty$ for all $s \geq 0$. Moreover, notice that, since $0 \in \mathcal{X}$ is a robust equilibrium point from the input $u \in M_U$ and $V(t, 0, 0) = 0$ for all $t \geq 0$, we have $a(s) \geq 0$ for all $s \geq 0$. Furthermore, notice that M is well-defined, since by definitions (A.22), (A.23) the following inequality is satisfied for all $h \geq 0$ and $s \geq 0$:

$$0 \leq M(h,s) \leq a(s). \tag{A.24}$$

Clearly, definition (A.22) implies that $a(\cdot)$ is nondecreasing. Furthermore, property P2 asserts that $\lim_{s \to 0^+} a(s) = 0$. Hence, the inequality $a(0) \geq 0$, in conjunction with $\lim_{s \to 0^+} a(s) = 0$ and the fact that $a(\cdot)$ is nondecreasing, implies $a(0) = 0$. It turns out that a can be bounded from above by the K_∞ function $\tilde{a}$ defined by $\tilde{a}(s) := s + \frac{1}{s}\int_s^{2s} a(w)dw$ for $s > 0$ and $\tilde{a}(0) = 0$. Define

$$\mu(h) := \sup\left\{\frac{M(h,s)}{g\left(\tilde{a}\left(s\right)\right)}; s > 0\right\}, \tag{A.25}$$

where g is defined by (A.10). Working exactly as in the proof of Lemma 2.16 we can show that the function $\mu : \Re^+ \to \Re^+$ is well-defined and satisfies $\mu(\cdot) \leq 1$, $\lim_{h \to +\infty} \mu(h) = 0$. Consequently, there exists a continuous strictly decreasing function $\bar{\mu} : \Re^+ \to (0, +\infty)$ such that $\bar{\mu}(h) \geq \mu(h)$ for all $h \geq 0$ and $\lim_{h \to +\infty} \bar{\mu}(h) = 0$. Thus, by recalling definition (A.25) we obtain

$$M(h,s) \leq \bar{\mu}(h) g(\tilde{a}(s)) \quad \forall h, s \geq 0. \tag{A.26}$$

Hence definition (A.23) implies that the following estimate holds for all $u \in M_U$, $(t_0, x_0, d) \in \Re^+ \times \mathcal{X} \times M_D$, and $t \geq t_0$:

$$V(t, \phi(t, t_0, x_0, u, d), u(t)) \leq \bar{\mu}(t - t_0) g\left(\tilde{a}(\left\|x_0\right\|_X)\right) + \sup_{t_0 \leq \tau \leq t} \gamma\left(\delta(\tau)\left\|u(\tau)\right\|_{\mathcal{U}}\right). \tag{A.27}$$

Estimate (A.27) implies (2.12) with $\beta(t) \equiv 1$ and $\sigma(s,t) := \bar{\mu}(t) g(\tilde{a}(s))$. □

Proof of Lemmas 2.19–2.20. The proof is based on the following observation: If $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ is *T-periodic*, then for all $(t_0, x_0, u, d) \in \Re^+ \times$

$\mathcal{X} \times M_U \times M_D$ it holds that $\phi(t, t_0, x_0, u, d) = \phi(t - kT, t_0 - kT, x_0, P_{kT}u, P_{kT}d)$ and $H(t, \phi(t, t_0, x_0, u, d), u(t)) = H(t - kT, \phi(t - kT, t_0 - kT, x_0, P_{kT}u, P_{kT}d), (P_{kT}u)(t - kT))$, where $k := [t_0/T]$ denotes the integer part of t_0/T and the inputs $P_{kT}u \in M_U, P_{kT}d \in M_D$ are defined in Definition 2.2.

Since $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ satisfies the WIOS property from the input $u \in M_U$, there exist functions $\sigma \in KL, \beta, \delta \in K^+, \gamma \in \mathcal{N}$ such that (2.10) holds for all $(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$ and $t \geq t_0$. Consequently, it follows that the following estimate holds for all $(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$ and $t \geq t_0$:

$$\begin{aligned}&\|H(t, \phi(t, t_0, x_0, u, d), u(t))\|_{\mathcal{Y}} \\ &\quad \leq \sigma\left(\beta\left(t_0 - kT\right)\|x_0\|_{\mathcal{X}}, t - t_0\right) + \sup_{\tau \in [t_0 - kT, t - kT]} \gamma\left(\delta(\tau)\left\|\left(P_{kT}u\right)(\tau)\right\|_{\mathcal{U}}\right).\end{aligned}$$

Setting $\tau = s - kT$ and since $0 \leq t_0 - \left[\frac{t_0}{T}\right] T < T$ for all $t_0 \geq 0$, we obtain

$$\begin{aligned}&\|H(t, \phi(t, t_0, x_0, u, d), u(t))\|_{\mathcal{Y}} \\ &\quad \leq \tilde{\sigma}\left(\|x_0\|_{\mathcal{X}}, t - t_0\right) + \sup_{s \in [t_0, t]} \gamma\left(\delta(s - kT)\left\|\left(P_{kT}u\right)(s - kT)\right\|_{\mathcal{U}}\right),\end{aligned} \tag{A.28}$$

where $\tilde{\sigma}(s, t) := \sigma(rs, t)$ and $r := \max\{\beta(t); 0 \leq t \leq T\}$. Estimate (A.28) and the identity $\left(P_{kT}u\right)(s - kT) = u(s)$ for all $s \geq 0$ imply that the following estimate holds for all $(t_0, x_0, u, d) \in \Re^+ \times \mathcal{X} \times M_U \times M_D$ and $t \geq t_0$:

$$\|H(t, \phi(t, t_0, x_0, u, d), u(t))\|_{\mathcal{Y}} \leq \tilde{\sigma}\left(\|x_0\|_{\mathcal{X}}, t - t_0\right) + \sup_{s \in [t_0, t]} \gamma\left(\tilde{\delta}(s)\|u(s)\|_{\mathcal{U}}\right), \tag{A.29}$$

where $\tilde{\delta}(t) := \max\{\delta(s); s \in [0, t]\}$.

In the case that $\Sigma := (\mathcal{X}, \mathcal{Y}, M_U, M_D, \phi, \pi, H)$ satisfies the IOS property from the input $u \in M_U$, then all arguments above may be repeated with $\delta(t) \equiv 1$. Thus we conclude that (A.29) holds for all $(t_0, x_0, u, d) \in \Re^+ \times X \times M_U \times M_D$ and $t \geq t_0$ with $\tilde{\delta}(t) \equiv 1$. The proof is complete. ∎

REFERENCES

[1] F. H. Clarke, Y. S. Ledyaev, E. D. Sontag, and A. I. Subbotin, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[2] S. Dashkovskiy, B. S. Ruffer, and F. R. Wirth, *A small-gain type stability criterion for large scale networks of ISS systems*, in Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference 2005, Seville, Spain, 2005, pp. 5633–5638.

[3] S. Dashkovskiy, B. S. Ruffer, and F. R. Wirth, *An ISS small gain theorem for general networks*, Math. Control Signals Syst., 19 (2007), pp. 93–122.

[4] L. Grune, *Stabilization by sampled and discrete feedback with positive sampling rate*, in Stability and Stabilization of Nonlinear Systems, D. Aeyels, F. Lamnabhi-Lagarrigue, and A. van der Schaft, eds., Springer-Verlag, London, 1999, pp. 165–182.

[5] L. Grune, *Asymptotic Behavior of Dynamical and Control Systems under Perturbation and Discretization*, Springer-Verlag, Berlin, 2002.

[6] J. K. Hale and S. M. V. Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[7] J. P. Hespanha, D. Liberzon, and A. R. Teel, *On input-to-state stability of impulsive systems*, in Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference 2005, Seville, Spain, 2005, pp. 3992–3997.

[8] B. Hu and A. N. Michel, *Stability analysis of digital control systems with time-varying sampling periods*, Automatica J. IFAC, 36 (2000), pp. 897–905.

[9] B. Hu and A. N. Michel, *Robustness analysis of digital control systems with time-varying sampling periods*, J. Franklin Insti., 337 (2000), pp. 117–130.

[10] B. Ingalls and Y. Wang, *On input-to-output stability for systems not uniformly bounded*, in Proceedings of NOLCOS 2001, St. Petersburg, Russia, Elsevier, New York, 2002.
[11] A. Isidori, *Nonlinear Control Systems* II, Springer-Verlag, London, 1999.
[12] H. Ito, *State-dependent scaling problems and stability of interconnected iISS and ISS systems*, IEEE Trans. Automat. Control, 51 (2006), pp. 1626–1643.
[13] H. Ito and Z.-P. Jiang, *Nonlinear small-gain condition covering iISS systems: Necessity and sufficiency from a Lyapunov perspective*, in Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference 2005, Seville, Spain, 2005, pp. 355–360.
[14] Z. P. Jiang, A. Teel, and L. Praly, *Small-gain theorem for ISS systems and applications*, Math. Control Signals Systems, 7 (1994), pp. 95–120.
[15] Z. P. Jiang, I. M. Y. Mareels, and Y. Wang, *A Lyapunov formulation of the nonlinear small-gain theorem for interconnected systems*, Automatica J. IFAC, 32 (1996), pp. 1211–1215.
[16] Z. P. Jiang and I. M. Y. Mareels, *A small-gain control method for nonlinear cascaded systems with dynamic uncertainties*, IEEE Trans. Automat. Control, 42 (1997), pp. 292–308.
[17] Z. P. Jiang, Y. Lin, and Y. Wang, *A local nonlinear small-gain theorem for discrete-time feedback systems and its applications*, in Proceedings of 3rd Asian Control Conference, Shanghai, 2000, pp. 1227–1232.
[18] Z. P. Jiang, Y. Lin, and Y. Wang, *Nonlinear small-gain theorems for discrete-time feedback systems and applications*, Automatica J. IFAC, 40 (2004), pp. 2129–2136.
[19] Z. P. Jiang and Y. Wang, *Input-to-state stability for discrete-time nonlinear systems*, Automatica J. IFAC, 37 (2001), pp. 857–869.
[20] R. E. Kalman, *Mathematical description of linear dynamical systems*, SIAM J. Control Optim., 1 (1963), pp. 152–192.
[21] I. Karafyllis and J. Tsinias, *A converse Lyapunov theorem for nonuniform in time global asymptotic stability and its application to feedback stabilization*, SIAM J. Control Optim., 42 (2003), pp. 936–965.
[22] I. Karafyllis and J. Tsinias, *Non-uniform in time ISS and the small-gain theorem*, IEEE Trans. Automat. Control, 49 (2004), pp. 196–216.
[23] I. Karafyllis, *Non-uniform robust global asymptotic stability for discrete-time systems and applications to numerical analysis*, IMA J. Math. Control Inform., 23 (2006), pp. 11–41.
[24] I. Karafyllis, *The non-uniform in time small-gain theorem for a wide class of control systems with outputs*, Eur. J. Control, 10 (2004), pp. 307–323.
[25] I. Karafyllis, *A system-theoretic framework for a wide class of systems* I: *Applications to numerical analysis*, J. Math. Anal. Appl., 328 (2007), pp. 876–899.
[26] I. Karafyllis, *A system-theoretic framework for a wide class of systems* II: *Input-to-output stability*, J. Math. Anal. Appl., 328 (2007), pp. 466–486.
[27] I. Karafyllis, *Stabilization by means of time-varying hybrid feedback*, Math. Control Signals Systems, 18 (2006), pp. 236–259.
[28] V. Lakshmikantham, D. D. Bainov, and P. S. Simeonov, *Theory of Impulsive Differential Equations*, World Scientific, Singapore, 1989.
[29] V. Lakshmikantham and S. G. Deo, *Method of Variation of Parameters for Dynamic Systems-Volume* 1, Gordon and Breach Science Publishers, New Delhi, 1998.
[30] D. Liberzon and D. Nesic, *Stability analysis of hybrid systems via small-gain theorems*, in Proceedings of the 9th International Workshop on Hybrid Systems: Computation and Control, Santa Barbara, 2006, Lecture Notes in Comput. Sci. 3927, J. Hespanha and A. Tiwari, eds., Springer, Berlin, 2006, pp. 421–435.
[31] D. Liberzon, *Switching in Systems and Control*, Birkhauser, Boston, 2003.
[32] I. M. Y. Mareels and D. J. Hill, *Monotone stability of nonlinear feedback systems*, J. Math. Systems, Estimation and Control, 2 (1992), pp. 275–291.
[33] D. Nesic, A. R. Teel, and E. D. Sontag, *Formulas relating KL stability estimates of discrete-time and sampled-data nonlinear systems*, Sys. Control Lett., 38 (1999), pp. 49–60.
[34] D. Nesic, A. R. Teel, and P. V. Kokotovic, *Sufficient conditions for stabilization of sampled-data nonlinear systems via discrete-time approximations*, Syst. Control Lett., 38 (1999), pp. 259–270.
[35] D. Nesic and A. R. Teel, *Sampled-data control of nonlinear systems: An overview of recent results*, Perspectives on Robust Control, R. S. O. Moheimani, ed., Springer-Verlag, New York, 2001, pp. 221–239.
[36] D. Nesic and D. S. Laila, *A note on input-to-state stabilization for nonlinear sampled-data systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 1153–1158.

[37] D. Nesic and D. Angeli, *Integral versions of ISS for sampled-data nonlinear systems via their approximate discrete-time models*, IEEE Trans. Automat. Control, 47 (2002), pp. 2033–2038.
[38] D. Nesic and A. Teel, *A framework for stabilization of nonlinear sampled-data systems based on their approximate discrete-time models*, IEEE Trans. Automat. Control, 49 (2004), pp. 1103–1122.
[39] D. Nesic and A. R. Teel, *Input-output stability properties of networked control systems*, IEEE Trans. Automat. Control, 49 (2004), pp. 1650–1667.
[40] D. Nesic and L. Grune, *Lyapunov-based continuous-time nonlinear controller redesign for sampled-data implementation*, Automatica J. IFAC, 41 (2005), pp. 1143–1156.
[41] D. Nesic and D. Liberzon, *A Unified Approach to Controller Design for Systems with Quantization and Time Scheduling*, preprint, 2007.
[42] P. Pepe, *The Liapunov's second method for continuous time difference equations*, Internat. J. Robust Nonlinear Control, 13 (2003), pp. 1389–1405.
[43] A. V. Savkin and R. J. Evans, *Hybrid Dynamical Systems*, Birkhauser, Boston, 2002.
[44] E. D. Sontag, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.
[45] E. D. Sontag, *Mathematical Control Theory*, 2nd ed., Springer-Verlag, New York, 1998.
[46] E. D. Sontag and Y. Wang, *Notions of input to output stability*, Syst. Control Lett., 38 (1999), pp. 235–248.
[47] E. D. Sontag and B. Ingalls, *A small-gain theorem with applications to input/output systems, incremental stability, detectability, and interconnections*, J. Franklin Inst., 339 (2002), pp. 211–229.
[48] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, 1998.
[49] Y. Sun, A. N. Michel, and G. Zhai, *Stability of discontinuous retarded functional differential equations with applications*, IEEE Trans. Automat. Control, 50 (2005), pp. 1090–1105.
[50] A. Teel, *A nonlinear small gain theorem for the analysis of control systems with saturations*, IEEE Trans. Automat. Control, 41, pp. 1256–1270.
[51] G. Zames, *On the input-output stability of time-varying nonlinear feedback systems. Part* I: *Conditions using concepts of loop gain, conicity and positivity*, IEEE Trans. Automat. Control, 11 (1966), pp. 228–238.
[52] G. Zames, *On the input-output stability of time-varying nonlinear feedback systems. Part* II: *Conditions involving circles in the frequency plane and sector nonlinearities*, IEEE Trans. Automat. Control, 11 (1966), pp. 465–476.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 4, pp. 1518–1537

BILATERAL OBSTACLE CONTROL PROBLEM OF PARABOLIC VARIATIONAL INEQUALITIES*

QIHONG CHEN†, DELIN CHU‡, AND ROGER C. E. TAN‡

Abstract. In this paper, the optimality system as well as the existence theorem for an obstacle optimal control problem are established, in which the governing system is a parabolic bilateral variational inequality and the input control is the pair of upper and lower obstacles.

Key words. bilateral obstacle control problem, parabolic variational inequality, optimality system, monotonicity inequality

AMS subject classifications. 49J24, 49J35, 49K24, 49K35

DOI. 10.1137/050638047

1. Introduction.

Throughout this paper, the following notation is used:

- $\Omega \subset \mathcal{R}^n$ is a bounded domain with a $C^{1,1}$ boundary $\partial\Omega$;
- $Q = \Omega \times (0,T)$, $\Sigma = \partial\Omega \times (0,T)$, $\partial_p Q = \Sigma \bigcup \bar{\Omega} \times \{0\}$;
- $L^2(0,T;H_0^1(\Omega)) = \{z\colon (0,T) \to H_0^1(\Omega) |\ \int_0^T \|z(\cdot,t)\|^2_{H_0^1(\Omega)} dt < \infty\}$;
- $W = \{y \in L^2(0,T;H_0^1(\Omega)) |\ y_t \in L^2(0,T;H^{-1}(\Omega)\}$;
- $W_p^{2,1}(Q) = \{z \in L^p(Q) |\ z_t, z_x, z_{xx} \in L^p(Q)\}$;
- $\dot{W}_p^{2,1}(Q) = \{z \in W_p^{2,1}(Q) |\ z|_{\partial_p Q} = 0\}$.

This paper deals with an optimal obstacle control problem in which the state y is governed by a semilinear parabolic bilateral variational inequality

$$\begin{cases} y \in W_2^{2,1}(Q) \cap L^2(0,T;H_0^1(\Omega)), \quad y|_{t=0} = y_0 & \text{in } \Omega, \\ \varphi \le y \le \psi & \text{in } Q, \\ (y_t - \Delta y - f(x,t,y))(y-\varphi) \le 0 & \text{in } Q, \\ (y_t - \Delta y - f(x,t,y))(y-\psi) \le 0 & \text{in } Q, \end{cases} \tag{1.1}$$

and the input control (φ, ψ) is a pair of upper and lower obstacles.

We assume that

(i)

$$y_0 \in C_0^\alpha(\bar{\Omega}) \cap W^{2-\frac{1}{p},p}(\Omega)$$

for some $\alpha \in (0,1)$ and any $p > 1$.

(ii) the function $f\colon \Omega \times [0,T] \times \mathcal{R} \to \mathcal{R}$ has the following properties:

- $f(\cdot,\cdot,y)$ is measurable on $\Omega \times [0,T]$;
- $f(x,t,\cdot)$ is a decreasing function in $C^1(\mathcal{R})$, so it is monotone;

*Received by the editors August 12, 2005; accepted for publication (in revised form) February 19, 2007; published electronically October 5, 2007.

http://www.siam.org/journals/sicon/46-4/63804.html

†Department of Applied Mathematices, Shanghai University of Finance and Economics, 200433 Shanghai, China (chenqih@yahoo.com). The work of this author was partly supported by FANEDD grant 200218; NSFC grants 10171059 and 10571030; and NUS grants R-146-000-047-112 and R-146-000-087-112.

‡Department of Mathematics, National University of Singapore, Kent Ridge, Republic of Singapore (matchudl@nus.edu.sg, scitance@nus.edu.sg). The work of these authors was supported by NUS grants R-146-000-047-112 and R-146-000-087-112.

- $f(x,t,\cdot)$ is a Lipschitz function, and thus, there exists a constant $K>0$, such that

$$|f(x,t,y_1)-f(x,t,y_2)| \le K\ |y_1-y_2| \quad \forall\ (x,t)\in\Omega\times[0,T];\ y_1,\ y_2\in\mathcal{R}; \tag{1.2}$$

-

$$|f(x,t,0)|\le K \qquad \forall\ (x,t)\in\Omega\times[0,T]. \tag{1.3}$$

Denote

$$\mathcal{U}=\{(\varphi,\psi)\in W_p^{2,1}(Q)\times W_p^{2,1}(Q)|\varphi\le\psi \text{ in } Q,\ \varphi=0=\psi \text{ on } \Sigma,\ \varphi|_{t=0}\le y_0\le\psi|_{t=0} \text{ in } \Omega\}.$$

For any given $(\varphi,\psi)\in\mathcal{U}$, the bilateral variational inequality (1.1) is uniquely solvable (see Proposition 2.4 below). We will denote by $y=\mathcal{S}(\varphi,\psi)$ the unique solution of (1.1) corresponding to (φ,ψ).

We shall take the pair of upper and lower obstacles (φ,ψ) as the control so that the corresponding state y becomes close to a desired target profile $z_d\in L^2(Q)$. This leads to the following problem.

Problem (C). *Find a control pair* $(\bar\varphi,\bar\psi)\in\mathcal{U}$ *such that*

$$J(\bar\varphi,\bar\psi)=\inf_{(\varphi,\psi)\in\mathcal{U}} J(\varphi,\psi),$$

where the objective functional $J(\varphi,\psi)$ *is defined by*

$$J(\varphi,\psi)=\int_Q\left\{\frac{1}{2}(\mathcal{S}(\varphi,\psi)-z_d)^2+\frac{1}{p}[|\varphi_t|^p+|\Delta\varphi|^p+|\psi_t|^p+|\Delta\psi|^p]\right\}dxdt. \tag{1.4}$$

The variational inequalities and related optimal control problems have been studied extensively; see, e.g., [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 20, 21, 22, 23, 24, 27, 29] and the references therein. When the governing system is an obstacle variational inequality and the obstacle is considered as the control, the resulting problem is referred to as an optimal obstacle control problem. This type of problem usually appears in shape optimization (cf. [26]). It may concern, for example, the design of the shape for a string in the one-dimensional obstacle problem, or the optimal shape for a dam, in which case the obstacle gives the form to be designed such that the pressure of the fluid inside the dam is close to a desired value.

The motivations of our present work are as follows:

(a) In Problem (C), the state system is a parabolic bilateral variational inequality, which has played an important role in the following recent research on mathematical finance: (1) the pricing model for American contingent claims (cf. [28]) is formulated as a parabolic variational inequality with the obstacle being the "payoff" function, and in that particular case, controlling the obstacle is equivalent to the design of the "payoff"; (2) in [11] the optimal investment problem with finite-horizon and transaction costs for a constant relative risk aversion investor is linked with a parabolic bilateral problem involving two free boundaries which correspond to the optimal buying and selling boundaries, respectively; and (3) in [12] the pricing model of a callable convertible security is formulated as a parabolic bilateral variational inequality in order

to study the optimal policies of the holder's conversion and issuer's calling for the callable American warrants and callable convertible bonds. It should be acknowledged that there is a nontrivial gap between the results of the present paper and that of the problem arising from mathematical finance, since Black–Scholes with American options has singular coefficients and an unbounded domain. However, in practical computation, we always truncate the unbounded domain, then solve the related partial differential equations on a bounded domain. In fact, if information on the interface between buy and sell regions is available, then this can be used as a sophisticated way of "cutting off" the domain; on this point we refer the reader to [30]. We also note that there may be other discussions of this point, e.g., in the stochastic community. Thus, it is possible, in some sense, to apply our results to mathematical finance. This is a research topic worthy of further study.

(b) The existence and uniqueness as well as characterizations of the optimal pair for an optimal obstacle control problem of an elliptic variational inequality have been given in [1]. The optimal obstacle control problem for more general systems has also been studied in [2, 3, 6, 8, 9, 10, 21, 22]. A problem, which is similar to Problem (C) with $p = 2$, has been considered in [2]. However, it seems that the method used in [2] cannot be applied to Problem (C) with $p > 2$ in the sense that when $p > 2$ a nonlinear leading differential operator, which lacks weak continuity, appears in the approximate optimality system (3.4) in section 3, but this operator is linear in the case $p = 2$.

(c) In general, when the penalty on control is stronger, the proof of the existence of optimal control becomes easier, whereas the derivation of necessary conditions will become harder, since the corresponding duality relation has to be established in a bigger space. Moreover, when the state equation is a variational inequality, one has to prove the convergence of the approximate adjoint equation, which is difficult due to the fact that the approximate adjoint equation is defined in a very weak sense, and then some measure term will occur in the limit. As far as the optimality system is concerned, it is incomplete as long as the measure which intervenes in the adjoint equation has not been precisely described. Hence, Problem (C) with $p > 2$ is much more difficult than its version with $p = 2$ in the sense that it is hard to derive the desired optimality system.

(d) Generally the existing results for elliptic variational inequalities cannot be extended to Problem (C) for parabolic variational inequalities in a trivial manner. In addition, some existing ideas, for example, the idea in [6], cannot be used to solve Problem (C) with $p > 2$ because of the difficulties described below. Furthermore, it is not clear if some existing results for the optimal control of elliptic variational inequalities are still true for the case of parabolic variational inequality; for instance, it is interesting to investigate whether the optimality conditions in [17] and [23] can still be true for Problem (C).

In this paper, we mainly aim at establishing an optimality system for Problem (C). Our approach is based on the penalty method and Barbu's treatment (cf. [4, 5]) as a penalty parameter approaching zero. In order to derive a more complete optimality system for Problem (C) than the one presented in [2] for a similar problem with $p = 2$, we adopt the $W_p^{2,1}$ framework which, however, causes some further difficulties due to the lack of weak continuity of the leading differential operator in the approximate optimality systems. We have overcome the encountered difficulties by

making full use of the special structure of the approximate optimality systems, including the monotonicity of the leading differential operator. We will prove that for most optimality conditions in [17, Theorem 5.1] and [23, Theorem 3.2], we can find their counterparts in our parabolic bilateral obstacle control problem.

Our approach can also be applied to more general cases; for instance, the Laplacian in (1.1) may be replaced by a general second order uniform elliptic operator with smooth coefficients.

2. State analysis. This section is devoted to some necessary preliminaries for the developments in the next sections.

2.1. Weak formulation. Given $(\varphi, \psi) \in \mathcal{U}$, define

$$\mathbf{K}(\varphi, \psi) = \{w \in W \,|\, \varphi \le w \le \psi \text{ a.e. in } Q \text{ and } w|_{t=0} = y_0 \text{ a.e. in } \Omega\}.$$

Clearly, $\mathbf{K}(\varphi, \psi)$ is a nonempty convex and closed subset of W.

LEMMA 2.1. *Let $y \in W_2^{2,1}(Q) \cap L^2(0,T; H_0^1(\Omega))$. Then $y = \mathcal{S}(\varphi, \psi)$ if and only if*

(2.1)
$$\begin{cases} y \in \mathbf{K}(\varphi, \psi), \\ \int_Q [y_t(w-y) + \nabla y \cdot \nabla(w-y)]dxdt \ge \int_Q f(x,t,y)(w-y)dxdt \quad \forall w \in \mathbf{K}(\varphi, \psi). \end{cases}$$

Proof. If y solves (1.1), then

$$y \in \mathbf{K}(\varphi, \psi),$$

and, for any $w \in \mathbf{K}(\varphi, \psi)$, $(w-y)^+$ (resp., $(w-y)^-$) can differ from 0 only when $y - \psi < 0$ (resp., $y - \varphi > 0$) and therefore $y_t - \Delta y - f \ge 0$ (resp., $y_t - \Delta y - f \le 0$). Thus, by the divergence theorem, the following inequality holds:

$$\begin{aligned} &\int_Q [y_t(w-y) + \nabla y \cdot \nabla(w-y) - f(x,t,y)(w-y)dxdt \\ &\quad = \int_Q (y_t - \Delta y - f)(w-y)dxdt \\ &\quad = \int_Q (y_t - \Delta y - f)(w-y)^+ dxdt - \int_Q (y_t - \Delta y - f)(w-y)^- dxdt \\ &\quad \ge 0 \qquad \forall w \in \mathbf{K}(\varphi, \psi). \end{aligned} \tag{2.2}$$

On the other hand, any $y \in W_2^{2,1}(Q) \cap L^2(0,T; H_0^1(\Omega))$ satisfying (2.1) must be a solution of (1.1). In fact, for any fixed $D \subset Q$ and any sequence $\{\chi_n\}$ of functions in $C_c^\infty(Q)$ satisfying $0 \le \chi_n \le 1$ and $\chi_n \to \chi_D$ (characteristic function of D) a.e. in Q, we can insert $w = y + \chi_n(\varphi - y)$ and $w = y + \chi_n(\psi - y)$ into (2.2), and then get

$$\int_Q (y_t - \Delta y - f)\chi_n(\varphi - y)dxdt \ge 0, \quad \int_Q (y_t - \Delta y - f)\chi_n(\psi - y)dxdt \ge 0. \tag{2.3}$$

So, by letting $n \to \infty$ in (2.3) we obtain that

$$\int_D (y_t - \Delta y - f)(\varphi - y)dxdt \ge 0 \quad \text{and} \quad \int_D (y_t - \Delta y - f)(\psi - y)dxdt \ge 0.$$

Hence the last two inequalities in (1.1) follow directly from the arbitrariness of D. ∎

Clearly, (2.1) is a weak formulation of the bilateral variational problem (1.1).

2.2. Approximation to the state. The objective functional $J(\varphi,\psi)$ defined by (1.4) is not smooth, which will cause difficulties for deriving an optimal system for Problem (C), so we approximate it by a smooth functional. For this purpose, we define

$$\beta(s)=\begin{cases}0, & 0\le s<+\infty,\\ -s^2, & -\frac{1}{2}\le s<0,\\ s+\frac{1}{4}, & -\infty<s<-\frac{1}{2},\end{cases}$$

$$\gamma(s)=\begin{cases}0, & -\infty<s<0,\\ s^2, & 0\le s<\frac{1}{2},\\ s-\frac{1}{4}, & \frac{1}{2}\le s<+\infty\end{cases}$$

and introduce a family of approximations to the state equation (2.1):

$$\begin{cases} y_t^\epsilon-\Delta y^\epsilon+\frac{1}{\epsilon}[\beta(y^\epsilon-\varphi)+\gamma(y^\epsilon-\psi)]=f(x,t,y^\epsilon) & \text{in } Q,\\ y^\epsilon|_\Sigma=0,\ y^\epsilon|_{t=0}=y_0.\end{cases}\tag{2.4}$$

For any given $(\varphi,\psi)\in\mathcal{U}$ and $\epsilon>0$, (2.4) is uniquely solvable in W. This unique solution is denoted by $y^\epsilon=\mathcal{S}^\epsilon(\varphi,\psi)$.

We consider the $W_p^{2,1}$-estimation and the convergence for approximate state y^ϵ in the following lemmas.

LEMMA 2.2. *For any $(\varphi,\psi)\in\mathcal{U}$ and $\epsilon>0$, the following inequality always holds:*

$$\|y^\epsilon\|_{W_p^{2,1}(Q)}\le C\left(1+\|\varphi\|_{W_p^{2,1}(Q)}+\|\psi\|_{W_p^{2,1}(Q)}\right),\tag{2.5}$$

where C is a constant independent of $\epsilon>0$ and $(\varphi,\psi)\in\mathcal{U}$.

Proof. To obtain (2.5), it suffices to prove the two estimates

$$\|\beta(y^\epsilon-\varphi)\|_{L^p(Q)}\le\epsilon C\left(1+\|\varphi\|_{W_p^{2,1}(Q)}\right),\tag{2.6}$$

$$\|\gamma(y^\epsilon-\psi)\|_{L^p(Q)}\le\epsilon C\left(1+\|\psi\|_{W_p^{2,1}(Q)}\right),\tag{2.7}$$

since (2.5) follows immediately from (2.6)–(2.7) and the standard parabolic L^P-estimate (cf. [19]).

Define, for $s\in\mathcal{R}$, $B(s)=|\beta(s)|^{p-2}\beta(s)$ and $\Gamma(s)=|\gamma(s)|^{p-2}\gamma(s)$. Then we have

$$B(s)\le 0\quad\text{and}\quad\Gamma(s)\ge 0\qquad\forall s\in\mathcal{R},\tag{2.8}$$

$$B(s)=0\ \forall s\ge 0\quad\text{and}\quad\Gamma(s)=0\ \forall s\le 0,\tag{2.9}$$

$$B'(s)=(p-1)|\beta(s)|^{p-2}\beta'(s)\ge 0\quad\text{and}\quad\Gamma'(s)=(p-1)|\gamma(s)|^{p-2}\gamma'(s)\ge 0.\tag{2.10}$$

Let

$$\Phi(s)=\int_0^s B(\tau)d\tau.$$

By (2.8)–(2.10), we further have

$$\Phi(s) \geq 0, \quad \Phi'(s) = B(s) \leq 0 \quad \text{in } \mathcal{R}; \quad \Phi(s) = 0 \quad \text{in } \mathcal{R}^+.$$

Thus, we easily get

$$\begin{aligned} &\int_Q [(y^\epsilon - \varphi)_t B(y^\epsilon - \varphi) + \nabla(y^\epsilon - \varphi) \cdot \nabla B(y^\epsilon - \varphi)] dxdt \\ &\quad = \int_\Omega \Phi(y^\epsilon - \varphi)|_{t=T} dx + \int_Q B'(y^\epsilon - \varphi)|\nabla(y^\epsilon - \varphi)|^2 dxdt \\ &\quad \geq 0. \end{aligned} \tag{2.11}$$

Under assumption (ii) of the introduction, f is decreasing in y, so using (2.8)–(2.9) we have that

$$\int_Q f(x,t,y^\epsilon) B(y^\epsilon - \varphi) dxdt \leq \int_Q f(x,t,\varphi) B(y^\epsilon - \varphi) dxdt. \tag{2.12}$$

Note that (2.9) implies

$$B(y^\epsilon - \varphi)\gamma(y^\epsilon - \psi) = 0 \qquad \text{a.e. in } Q,$$

so we can multiply (2.4) by $\epsilon B(y^\epsilon - \varphi)$, integrate it over Q, and then get

$$\begin{aligned} &\epsilon \int_Q [y_t^\epsilon B(y^\epsilon - \varphi) + \nabla y^\epsilon \cdot \nabla B(y^\epsilon - \varphi)] dxdt + \int_Q |\beta(y^\epsilon - \varphi)|^p dxdt \\ &\qquad = \epsilon \int_Q f(x,t,y^\epsilon) B(y^\epsilon - \varphi) dxdt. \end{aligned} \tag{2.13}$$

Furthermore, from (2.13) using (2.11)–(2.12) and Hölder's inequality we can deduce that

$$\begin{aligned} &\|\beta(y^\epsilon - \varphi)\|^p_{L^p(Q)} \\ &\quad \leq \epsilon \int_Q \{f(x,t,\varphi) B(y^\epsilon - \varphi) - [\varphi_t B(y^\epsilon - \varphi) + \nabla\varphi \cdot \nabla B(y^\epsilon - \varphi)]\} dxdt \\ &\quad = \epsilon \int_Q [f(x,t,\varphi) - \varphi_t + \Delta\varphi] B(y^\epsilon - \varphi) dxdt \\ &\quad \leq \epsilon [\|f(\cdot,\cdot,\varphi(\cdot,\cdot))\|_{L^p(Q)} + \|\varphi\|_{W^{2,1}_p(Q)}] \, \|\beta(y^\epsilon - \varphi)\|^{p-1}_{L^p(Q)}. \end{aligned}$$

Finally, because (1.2) and (1.3) hold, we get

$$|f(x,t,\varphi(x,t))| \leq |f(x,t,0)| + |f(x,t,\varphi(x,t)) - f(x,t,0)| \leq K[1 + |\varphi(x,t)|] \tag{2.14}$$

and then

$$\|f(\cdot,\cdot,\varphi(\cdot,\cdot))\|_{L^p(Q)} \leq C \left(1 + \|\varphi\|_{W^{2,1}_p(Q)}\right) \tag{2.15}$$

with C being independent of $\epsilon > 0$ and φ. Thus, (2.6) follows.

The inequality (2.7) can be obtained similarly. □

LEMMA 2.3. *Given a sequence of obstacle pairs* $(\varphi^\epsilon, \psi^\epsilon) \in \mathcal{U}$, *let* $y^\epsilon = \mathcal{S}^\epsilon(\varphi^\epsilon, \psi^\epsilon)$. *If the sequence* $\{(\varphi^\epsilon, \psi^\epsilon)\}$ *is bounded in* $W_p^{2,1}(Q) \times W_p^{2,1}(Q)$ *with* $p > (n+2)/2$, *then for some subsequences (still denoted by themselves), as* $\epsilon \to 0^+$,

$$\left.\begin{array}{l} \varphi^\epsilon \to \varphi \\ \psi^\epsilon \to \psi \\ y^\epsilon \to y \end{array}\right\} \text{weakly in } W_p^{2,1}(Q) \text{ and strongly in } C^{\theta,\theta/2}(\bar{Q}) \cap L^2(0,T;H_0^1(\Omega))$$

for some $\theta \in (0,1)$, *where* $(\varphi, \psi) \in \mathcal{U}$ *and* $y = \mathcal{S}(\varphi, \psi)$.

Proof. First, by Simon's compactness lemma (cf. [25]), we know that, if $p > (n+2)/2$, any bounded subset of $W_p^{2,1}(Q)$ is relatively compact in $C^{\theta,\theta/2}(\bar{Q}) \cap L^2(0,T;H_0^1(\Omega))$ for some $\theta \in (0,1)$. Thus, we can extract a subsequence (still denoted by itself) by virtue of (2.5) such that

$$\left.\begin{array}{l} \varphi^\epsilon \to \varphi \\ \psi^\epsilon \to \psi \\ y^\epsilon \to y \end{array}\right\} \text{weakly in } W_p^{2,1}(Q) \text{ and strongly in } C^{\theta,\theta/2}(\bar{Q}) \cap L^2(0,T;H_0^1(\Omega)).$$

Obviously, $(\varphi, \psi) \in \mathcal{U}$.

Next, by using (2.6), (2.7), the strong convergence of $(\varphi^\epsilon, \psi^\epsilon, y^\epsilon)$ to (φ, ψ, y) in $C^{\theta,\theta/2}(\bar{Q})$, and the definition of $\beta(\cdot)$ and $\gamma(\cdot)$, we also have (as $\epsilon \to 0^+$)

$$\beta(y^\epsilon - \varphi^\epsilon) + \gamma(y^\epsilon - \psi^\epsilon) \to 0 \quad \text{in } L^p(Q),$$

$$\beta(y - \varphi) + \gamma(y - \psi) = 0 \qquad \text{a.e. in } Q,$$

and

$$\varphi(x,t) \le y(x,t) \le \psi(x,t) \qquad \text{a.e. in } Q.$$

Clearly, $y|_{t=0} = y_0$. Hence, $y \in \mathbf{K}(\varphi, \psi)$.

For any $w \in \mathbf{K}(\varphi, \psi)$, let $w^\epsilon = \sup(\varphi^\epsilon, \inf(\psi^\epsilon, w))$. Since $\beta(y^\epsilon - \varphi^\epsilon)$ can differ from 0 only when $y^\epsilon < \varphi^\epsilon \le w^\epsilon$, and $\gamma(y^\epsilon - \psi^\epsilon)$ can differ from 0 only when $y^\epsilon > \psi^\epsilon \ge w^\epsilon$, we deduce from (2.4) that

$$\begin{aligned} &\int_Q [y_t^\epsilon(w^\epsilon - y^\epsilon) + \nabla y^\epsilon \cdot \nabla(w^\epsilon - y^\epsilon)]dxdt \\ &= -\frac{1}{\epsilon}\int_Q [\beta(y^\epsilon - \varphi^\epsilon) + \gamma(y^\epsilon - \psi^\epsilon)](w^\epsilon - y^\epsilon)dxdt + \int_Q f(x,t,y^\epsilon)(w^\epsilon - y^\epsilon)dxdt \\ &\ge \int_Q f(x,t,y^\epsilon)(w^\epsilon - y^\epsilon)dxdt. \end{aligned}$$

Then letting $\epsilon \to 0^+$ in the above inequality, we obtain that $y = \lim_{\epsilon \to 0^+} y^\epsilon$ satisfies (2.1). Hence, $y = \mathcal{S}(\varphi, \psi)$. ∎

2.3. Unique solvability of the state system. As claimed in the introduction, the state system (1.1) is uniquely solvable, and we prove this point in this subsection.

PROPOSITION 2.4. *For any given* $(\varphi, \psi) \in \mathcal{U}$, *the state system* (1.1) *is uniquely solvable. Moreover, let* $y = \mathcal{S}(\varphi, \psi)$; *then*

$$\|y\|_{W_p^{2,1}(Q)} \le C\left(1 + \|\varphi\|_{W_p^{2,1}(Q)} + \|\psi\|_{W_p^{2,1}(Q)}\right), \tag{2.16}$$

where C is a constant independent of $(\varphi, \psi) \in \mathcal{U}$.

Proof. Choose $(\varphi^\epsilon, \psi^\epsilon) \equiv (\varphi, \psi)$ and let $y^\epsilon = \mathcal{S}^\epsilon(\varphi, \psi)$; then Lemma 2.3 gives

$$y^\epsilon \to y = \mathcal{S}(\varphi, \psi),$$

and thus (2.16) follows from Lemma 2.2.

To prove the uniqueness, let y_i $(i = 1, 2)$ be two weak solutions of the state system (1.1). Taking y_2 (resp., y_1) as a test function, substituting it into inequality (2.1) of y_1 (resp., y_2), and then adding, we get

$$\frac{1}{2}\int_\Omega (y_1 - y_2)^2|_{t=T} dx + \int_Q |\nabla(y_1 - y_2)|^2 dxdt \le \int_Q [f(x,t,y_1) - f(x,t,y_2)](y_1 - y_2) dxdt.$$

By assumption (ii) of the introduction, the integral on the right-hand side is nonpositive, so we can assert that

$$\|y_1 - y_2\|_{L^2(0,T;H_0^1(\Omega))} = 0,$$

i.e.,

$$y_1(x,t) = y_2(x,t) \qquad \text{a.e. in } Q. \qquad \square$$

3. Main results. In this section we establish our main results, which are the existence and the optimality system for Problem (C).

In our cost functional (1.4), there appears the Laplacian together with the t-derivative of the control (obstacle), which provides a certain compactness of the control. As a direct result, the existence of the optimal control is almost routine. Consequently, the following existence theorem for Problem (C) can be obtained by some standard ideas with some suitable variational inequality techniques.

THEOREM 3.1. *Problem* (C) *has at least one pair of optimal obstacles* $(\bar\varphi, \bar\psi) \in \mathcal{U}$.

Let $(\bar\varphi, \bar\psi)$ be an optimal pair for Problem (C) and let $\bar y = \mathcal{S}(\bar\varphi, \bar\psi)$. We first introduce a family of approximate control problems.

PROBLEM (C^ϵ). *Find a pair* $(\varphi^\epsilon, \psi^\epsilon) \in \mathcal{U}$ *such that*

$$J^\epsilon(\varphi^\epsilon, \psi^\epsilon) = \inf_{(\varphi,\psi)\in\mathcal{U}} J^\epsilon(\varphi, \psi),$$

where

$$J^\epsilon(\varphi,\psi) = \int_Q \Big\{ \frac{1}{2}(y^\epsilon - z_d)^2 + \frac{1}{p}[|\varphi_t|^p + |\Delta\varphi|^p + |\psi_t|^p + |\Delta\psi|^p + |\varphi - \bar\varphi|^p + |\psi - \bar\psi|^p] \Big\} dxdt$$

with $y^\epsilon = \mathcal{S}^\epsilon(\varphi, \psi)$ *being the approximate state solving* (2.4).

As an analogy to Problem (C), we have the following.

PROPOSITION 3.2. *There exists an optimal control pair* $(\varphi^\epsilon, \psi^\epsilon) \in \mathcal{U}$ *for Problem* (C^ϵ).

Before deriving the optimality system for Problem (C^ϵ) *we need a support result on the Gâteaux-differentiability of the approximate state operator* $\mathcal{S}^\epsilon$.

LEMMA 3.3. *For any fixed* $\epsilon > 0$, *the solution mapping* $\mathcal{S}^\epsilon : (\varphi, \psi) \mapsto y^\epsilon$ *of* (2.4) *is differentiable in the following sense:*

Given $(\varphi,\psi)\in\mathcal{U}$, *for any* $(u,\,v)\in\mathcal{U}$,

$$\frac{\mathcal{S}^\epsilon(\varphi+\delta(u-\varphi),\psi+\delta(v-\psi))-\mathcal{S}^\epsilon(\varphi,\psi)}{\delta}\to\xi^\epsilon \qquad \textit{weakly in } W$$

as $\delta\to 0^+$, *where* ξ^ϵ *satisfies*

$$\text{(3.1)}\qquad \begin{cases} \xi^\epsilon_t-\Delta\xi^\epsilon+\{\frac{1}{\epsilon}[\beta'(y^\epsilon-\varphi)+\gamma'(y^\epsilon-\psi)]-f_y(x,t,y^\epsilon)\}\xi^\epsilon \\ \qquad\qquad =\frac{1}{\epsilon}[\beta'(y^\epsilon-\varphi)(u-\varphi)+\gamma'(y^\epsilon-\psi)(v-\psi)] \qquad \textit{in } Q,\\ \xi^\epsilon|_{\partial_pQ}=0. \end{cases}$$

Proof. The proof is standard and tedious. We omit it here. □

PROPOSITION 3.4. *Assume* $(\varphi^\epsilon,\psi^\epsilon)$ *is an optimal solution to Problem* (C^ϵ) *and* $y^\epsilon=\mathcal{S}^\epsilon(\varphi^\epsilon,\psi^\epsilon)$. *Then, there exists* $p^\epsilon\in L^2(0,T;H_0^1(\Omega))$ *such that the following optimality system is satisfied:*

$$\text{(3.2)}\qquad \begin{cases} y^\epsilon_t-\Delta y^\epsilon+\frac{1}{\epsilon}[\beta(y^\epsilon-\varphi^\epsilon)+\gamma(y^\epsilon-\psi^\epsilon)]=f(x,t,y^\epsilon) \qquad \textit{in } Q,\\ y^\epsilon|_\Sigma=0,\ y^\epsilon|_{t=0}=y_0, \end{cases}$$

(3.3)

$$\begin{cases} -p^\epsilon_t-\Delta p^\epsilon+\{\frac{1}{\epsilon}[\beta'(y^\epsilon-\varphi^\epsilon)+\gamma'(y^\epsilon-\psi^\epsilon)]-f_y(x,t,y^\epsilon)\}p^\epsilon=y^\epsilon-z_d \quad \textit{in } Q,\\ p^\epsilon|_\Sigma=0,\\ p^\epsilon|_{t=T}=0, \end{cases}$$

and

(3.4)

$$\begin{aligned}\int_Q\Bigg\{&\left[\frac{1}{\epsilon}\beta'(y^\epsilon-\varphi^\epsilon)p^\epsilon+|\varphi^\epsilon-\bar\varphi|^{p-2}(\varphi^\epsilon-\bar\varphi)\right](\varphi-\varphi^\epsilon)\\ &+\left[\frac{1}{\epsilon}\gamma'(y^\epsilon-\psi^\epsilon)p^\epsilon+\ |\psi^\epsilon-\bar\psi|^{p-2}(\psi^\epsilon-\bar\psi)\right](\psi-\psi^\epsilon)\\ &+[|\varphi^\epsilon_t|^{p-2}\varphi^\epsilon_t(\varphi-\varphi^\epsilon)_t+|\Delta\varphi^\epsilon|^{p-2}\Delta\varphi^\epsilon\Delta(\varphi-\varphi^\epsilon)]\\ &+[|\psi^\epsilon_t|^{p-2}\psi^\epsilon_t(\psi-\psi^\epsilon)_t+|\Delta\psi^\epsilon|^{p-2}\Delta\psi^\epsilon\Delta(\psi-\psi^\epsilon)]\Bigg\}dxdt\ge 0\end{aligned}$$

$$\forall(\varphi,\psi)\in\mathcal{U}.$$

Proof. Let $(\varphi^\epsilon,\psi^\epsilon)\in\mathcal{U}$ be an optimal control pair for Problem (C^ϵ) and let $y^\epsilon=\mathcal{S}^\epsilon(\varphi^\epsilon,\psi^\epsilon)$. For any given $(\varphi,\psi)\in\mathcal{U}$, as $(\varphi^\epsilon,\psi^\epsilon)$ is optimal to Problem (C^ϵ), we have

$$\begin{aligned} 0 &\le \liminf_{\delta\to 0}\ \delta^{-1}[J^\epsilon(\varphi^\epsilon+\delta(\varphi-\varphi^\epsilon),\psi^\epsilon+\delta(\psi-\psi^\epsilon))-J^\epsilon(\varphi^\epsilon,\psi^\epsilon)]\\ &= \int_Q\{(y^\epsilon-z_d)\xi^\epsilon+[|\varphi^\epsilon_t|^{p-2}\varphi^\epsilon_t(\varphi-\varphi^\epsilon)_t+|\Delta\varphi^\epsilon|^{p-2}\Delta\varphi^\epsilon\Delta(\varphi-\varphi^\epsilon)]\\ &\qquad+[|\psi^\epsilon_t|^{p-2}\psi^\epsilon_t(\psi-\psi^\epsilon)_t+|\Delta\psi^\epsilon|^{p-2}\Delta\psi^\epsilon\Delta(\psi-\psi^\epsilon)]\\ &\qquad+[|\varphi^\epsilon-\bar\varphi|^{p-2}(\varphi^\epsilon-\bar\varphi)(\varphi-\varphi^\epsilon)+|\psi^\epsilon-\bar\psi|^{p-2}(\psi^\epsilon-\bar\psi)(\psi-\psi^\epsilon)]\}dxdt,\end{aligned}$$

where ξ^ϵ satisfies

$$\text{(3.5)}\qquad \begin{cases} \xi^\epsilon_t-\Delta\xi^\epsilon+\{\frac{1}{\epsilon}[\beta'(y^\epsilon-\varphi^\epsilon)+\gamma'(y^\epsilon-\psi^\epsilon)]-f_y(x,t,y^\epsilon)\}\xi^\epsilon\\ \qquad\qquad =\frac{1}{\epsilon}[\beta'(y^\epsilon-\varphi^\epsilon)(\varphi-\varphi^\epsilon)+\gamma'(y^\epsilon-\psi^\epsilon)(\psi-\psi^\epsilon)] \qquad \text{in } Q,\\ \xi^\epsilon|_{\partial_pQ}=0. \end{cases}$$

Let $p^\epsilon \in W$ be the unique solution of linear equation (3.3). Then, using (3.3) and (3.5) we get that

$$\int_Q (y^\epsilon - z_d)\xi^\epsilon dxdt = \int_Q \tfrac{1}{\epsilon} p^\epsilon [\beta'(y^\epsilon - \varphi^\epsilon)(\varphi - \varphi^\epsilon) + \gamma'(y^\epsilon - \psi^\epsilon)(\psi - \psi^\epsilon)]dxdt.$$

So (3.4) follows. □

With respect to the approximate optimality condition (3.4), we give the following remarks, which are crucial in deriving the optimality system for the original Problem (C).

Remark 1. Using the so-called monotonicity inequality (cf. [16]),

$$(|a|^{p-2}a - |b|^{p-2}b)(a-b) \geq 0 \qquad (a, b \in \mathcal{R}),$$

we can deduce from (3.4) that

$$\begin{aligned}(3.6)\qquad \int_Q \Big\{ & \left[\frac{1}{\epsilon}\beta'(y^\epsilon - \varphi^\epsilon)p^\epsilon + |\varphi^\epsilon - \bar\varphi|^{p-2}(\varphi^\epsilon - \bar\varphi)\right](\varphi - \varphi^\epsilon) \\ & + \left[\frac{1}{\epsilon}\gamma'(y^\epsilon - \psi^\epsilon)p^\epsilon + |\psi^\epsilon - \bar\psi|^{p-2}(\psi^\epsilon - \bar\psi)\right](\psi - \psi^\epsilon) \\ & + [|\varphi_t|^{p-2}\varphi_t(\varphi - \varphi^\epsilon)_t + |\Delta\varphi|^{p-2}\Delta\varphi\Delta(\varphi - \varphi^\epsilon)] \\ & + [|\psi_t|^{p-2}\psi_t(\psi - \psi^\epsilon)_t + |\Delta\psi|^{p-2}\Delta\psi\Delta(\psi - \psi^\epsilon)]\Big\}dxdt \geq 0 \\ & \forall (\varphi, \psi) \in \mathcal{U}.\end{aligned}$$

Remark 2. For any $w \in \dot W_p^{2,1}(Q)$, by inserting $\varphi = \varphi^\epsilon + w$ and $\psi = \psi^\epsilon + w$ into inequality (3.4), we get

$$\begin{aligned}(3.7)\qquad \int_Q \Big\{ & \left[\frac{1}{\epsilon}(\beta'(y^\epsilon - \varphi^\epsilon) + \gamma'(y^\epsilon - \psi^\epsilon))p^\epsilon + |\varphi^\epsilon - \bar\varphi|^{p-2}(\varphi^\epsilon - \bar\varphi) + |\psi^\epsilon - \bar\psi|^{p-2}(\psi^\epsilon - \bar\psi)\right] w \\ & + [|\varphi_t^\epsilon|^{p-2}\varphi_t^\epsilon + |\psi_t^\epsilon|^{p-2}\psi_t^\epsilon]w_t \\ & + [|\Delta\varphi^\epsilon|^{p-2}\Delta\varphi^\epsilon + |\Delta\psi^\epsilon|^{p-2}\Delta\psi^\epsilon]\Delta w\Big\}dxdt = 0 \\ & \forall w \in \dot W_p^{2,1}(Q).\end{aligned}$$

The equality in (3.7), instead of the inequality, is due to the arbitrariness of w.

The relationship between Problems (C) and (C$^\epsilon$) is revealed in the following proposition.

PROPOSITION 3.5. *Let $(\varphi^\epsilon, \psi^\epsilon) \in \mathcal{U}$ be an optimal control pair for Problem* (C$^\epsilon$) *and let $y^\epsilon = \mathcal{S}^\epsilon(\varphi^\epsilon, \psi^\epsilon)$. Then, for $p > (n+2)/2$,*

$$\left.\begin{array}{l}\varphi^\epsilon \to \bar\varphi \\ \psi^\epsilon \to \bar\psi \\ y^\epsilon \to \bar y\end{array}\right\} \text{ weakly in } W_p^{2,1}(Q) \text{ and strongly in } C^{\theta,\theta/2}(\bar Q) \cap L^2(0,T;H_0^1(\Omega)),$$

and

$$\lim_{\epsilon \to 0^+} J^\epsilon(\varphi^\epsilon, \psi^\epsilon) = J(\bar\varphi, \bar\psi),$$

where $\theta \in (0,1)$ and $(\bar{\varphi}, \bar{\psi}) \in \mathcal{U}$ is the optimal control pair for Problem (C) *given at the beginning of this subsection and $\bar{y} = \mathcal{S}(\bar{\varphi}, \bar{\psi})$ is the corresponding optimal state.*

Proof. First, Lemma 2.2 yields

$$J^\epsilon(\varphi^\epsilon, \psi^\epsilon) \le J^\epsilon(\bar{\varphi}, \bar{\psi}) \le C\left(1 + \|\bar{\varphi}\|_{W_p^{2,1}(Q)} + \|\bar{\psi}\|_{W_p^{2,1}(Q)}\right).$$

So, $(\varphi^\epsilon, \psi^\epsilon)$ is bounded in $W_p^{2,1}(Q) \times W_p^{2,1}(Q)$, due to the form of the functional J^ϵ.

Then, by Lemma 2.3, for some subsequences (still denoted by themselves) and some $\theta \in (0,1)$,

$$\left.\begin{array}{l} \varphi^\epsilon \to \varphi^* \\ \psi^\epsilon \to \psi^* \\ y^\epsilon \to y^* \end{array}\right\} \text{ weakly in } W_p^{2,1}(Q) \text{ and strongly in } C^{\theta,\theta/2}(\bar{Q}) \cap L^2(0,T; H_0^1(\Omega)),$$

where $(\varphi^*, \psi^*) \in \mathcal{U}$ and $y^* = \mathcal{S}(\varphi^*, \psi^*)$.

Next, from the weak lower semicontinuity of the L^p-norm, we have

(3.8)

$$\begin{aligned} &J(\varphi^*, \psi^*) \\ &\le J(\varphi^*, \psi^*) + \frac{1}{p}\int_Q \left\{|\varphi^* - \bar{\varphi}|^p + |\psi^* - \bar{\psi}|^p\right\} dxdt \\ &= \int_Q \left\{\frac{1}{2}(y^* - z_d)^2 + \frac{1}{p}[|\varphi_t^*|^p + |\Delta\varphi^*|^p + |\psi_t^*|^p + |\Delta\psi^*|^p + |\varphi^* - \bar{\varphi}|^p \right. \\ &\qquad \left. + |\psi^* - \bar{\psi}|^p]\right\} dxdt \\ &\le \liminf_{\epsilon\to 0} J^\epsilon(\varphi^\epsilon, \psi^\epsilon) \\ &\le \limsup_{\epsilon\to 0} J^\epsilon(\varphi^\epsilon, \psi^\epsilon) \\ &\le \lim_{\epsilon\to 0} J^\epsilon(\bar{\varphi}, \bar{\psi}) \\ &= J(\bar{\varphi}, \bar{\psi}). \end{aligned}$$

On the other hand, $(\bar{\varphi}, \bar{\psi})$ is optimal to Problem (C), and then $J(\bar{\varphi}, \bar{\psi}) \le J(\varphi^*, \psi^*)$. Thus, all the equalities in (3.9) must hold. This means that

$$\int_Q \left\{|\varphi^* - \bar{\varphi}|^p + |\psi^* - \bar{\psi}^p\right\} dxdt = 0;$$

i.e., $\varphi^* = \bar{\varphi}$ and $\psi^* = \bar{\psi}$. By the unique solvability of the state variational inequality (1.1), we get $y^* = \bar{y}$. In addition, we see that

$$\lim_{\epsilon\to 0^+} J^\epsilon(\varphi^\epsilon, \psi^\epsilon) = J(\bar{\varphi}, \bar{\psi}).$$

Finally, the uniqueness of limit point implies the convergence of the whole sequence of $(\varphi^\epsilon, \psi^\epsilon)$ and the whole sequence of y^ϵ as well. The proof is completed. ∎

We are now ready to present the optimality system for Problem (C).

THEOREM 3.6. *Let $(\bar{\varphi}, \bar{\psi})$ be an optimal control pair for Problem* (C) *and let $\bar{y} = \mathcal{S}(\bar{\varphi}, \bar{\psi})$. Then for $p > (n+2)/2$, there exist $\bar{p} \in L^2(0,T; H_0^1(\Omega))$ and $\bar{\mu} \in \mathcal{M}_0(\bar{Q})$*

satisfying[1]

$$\begin{cases} -\bar{p}_t - \Delta\bar{p} - f_y(x,t,\bar{y})\bar{p} = \bar{y} - z_d - \bar{\mu} & \text{in } Q, \\ \bar{p}|_\Sigma = 0, \\ \bar{p}|_{t=T} = 0, \end{cases} \tag{3.9}$$

with

$$\operatorname{supp}\bar{\mu} \subset \{(x,t) \in Q | \quad \bar{y}(x,t) = \bar{\varphi}(x,t) \quad \text{or} \quad \bar{y}(x,t) = \bar{\psi}(x,t)\}; \tag{3.10}$$

$$\begin{aligned} \operatorname{supp}[\bar{y}_t - \Delta\bar{y} - f(x,t,\bar{y})] \subset \{(x,t) \in Q | \quad & \bar{y}(x,t) = \bar{\varphi}(x,t) \quad \text{or} \\ & \bar{y}(x,t) = \bar{\psi}(x,t)\} \end{aligned} \tag{3.11}$$

such that

$$\int_Q [(|\bar{\varphi}_t|^{p-2}\bar{\varphi}_t + |\bar{\psi}_t|^{p-2}\bar{\psi}_t)w_t + (|\Delta\bar{\varphi}|^{p-2}\Delta\bar{\varphi} + |\Delta\bar{\psi}|^{p-2}\Delta\bar{\psi})\Delta w]dxdt + \langle\bar{\mu}, w\rangle = 0 \tag{3.12}$$

$$\forall w \in \dot{W}_p^{2,1}(Q)$$

and

$$\int_Q [\bar{y}_t - \Delta\bar{y} - f(x,t,\bar{y})]\bar{p}dxdt = 0. \tag{3.13}$$

Moreover,

$$-\bar{y}_t + \Delta\bar{y} + f(x,t,\bar{y}) = \bar{\lambda}_\varphi + \bar{\lambda}_\psi$$

with

$$\langle\bar{\lambda}_\varphi, \bar{y} - \bar{\varphi}\rangle = 0; \qquad \langle\bar{\lambda}_\psi, \bar{y} - \bar{\psi}\rangle = 0; \tag{3.14}$$

$$\operatorname{supp}\bar{\lambda}_\varphi \subset \{(x,t) \in Q | \, \bar{y}(x,t) = \bar{\varphi}(x,t)\}; \ \operatorname{supp}\bar{\lambda}_\psi \subset \{(x,t) \in Q | \, \bar{y}(x,t) = \bar{\psi}(x,t)\}; \tag{3.15}$$

and

$$\bar{\mu} = \bar{\mu}_\varphi + \bar{\mu}_\psi$$

with

$$\langle\bar{\mu}_\varphi, \bar{y} - \bar{\varphi}\rangle = 0; \qquad \langle\bar{\mu}_\psi, \bar{y} - \bar{\psi}\rangle = 0; \tag{3.16}$$

[1] According to [2], the function $\bar{p} \in L^2(0,T; H_0^1(\Omega))$ solves (3.9) in the sense that

$$\int_Q [\bar{p}\chi_t + \nabla\bar{p}\cdot\nabla\chi - f_y(x,t,\bar{y})\bar{p}\chi - (\bar{y} - z_d)\chi]\,dxdt + \langle\bar{\mu},\chi\rangle = 0 \quad \forall\chi \in \dot{W}_p^{2,1}(Q),$$

or equivalently,

$$\int_Q [\bar{p}\chi_t + \nabla\bar{p}\cdot\nabla\chi - f_y(x,t,\bar{y})\bar{p}\chi]\,dxdt = \int_Q (\bar{y} - z_d)\chi\,dxdt - \langle\bar{\mu},\chi\rangle \quad \forall\chi \in \dot{W}_p^{2,1}(Q).$$

(3.17)
$$\operatorname{supp}\bar{\mu}_\varphi \subset \{(x,t)\in Q|\, \bar{y}(x,t) = \bar{\varphi}(x,t)\};\ \operatorname{supp}\bar{\mu}_\psi \subset \{(x,t)\in Q|\, \bar{y}(x,t) = \bar{\psi}(x,t)\}.$$

Remark. In the above optimality system, (3.9) and (3.13) are referred to as the adjoint equation and the optimality condition, respectively. The condition (3.10) is understood as the following: For any $\eta \in C_0(\bar{Q}) = \{\eta \in C(\bar{Q})|\, \eta|_\Sigma = 0\}$ with $\operatorname{supp}\eta \subset Q' = \{(x,t)\in Q|\, \bar{\varphi}(x,t) < \bar{y}(x,t) < \bar{\psi}(x,t)\}$,

$$\langle \bar{\mu},\, \eta\rangle_{\mathcal{M}_0(\bar{Q}),\, C_0(\bar{Q})} = 0,$$

where $\mathcal{M}_0(\bar{Q}) = C_0(\bar{Q})^*$ is the set of all regular signed measures on $\bar{Q}$ with the support contained in $Q \cup (\Omega \times \{0,T\})$.

Proof. Let $(\bar{y},\bar{\varphi},\bar{\psi})$ be an optimal triple to Problem (C). Consider the approximate Problem (C^ϵ) related to $(\bar{y},\bar{\varphi},\bar{\psi})$. Let $(y^\epsilon,\varphi^\epsilon,\psi^\epsilon)$ be any optimal triple to Problem (C^ϵ). Then, by Proposition 3.4, (3.4) holds (furthermore, (3.6) and (3.7) also hold) for some $p^\epsilon \in L^2(0,T;H_0^1(\Omega))$ which satisfies (3.3). Note that by $\beta' \geq 0$, $\gamma' \geq 0$, and $f_y \leq 0$ (cf. assumption (ii) of the introduction), one can easily get the following estimate from (3.3):

$$\|p^\epsilon\|_{L^2(0,T;H_0^1(\Omega))} \leq C,$$

where C is independent of ϵ. This implies that for some subsequence,

$$p^\epsilon \rightharpoonup \bar{p} \qquad \text{weakly in } L^2(0,T;H_0^1(\Omega)). \tag{3.18}$$

Denote

$$\mu_\varphi^\epsilon = \frac{1}{\epsilon}\beta'(y^\epsilon-\varphi^\epsilon)p^\epsilon,\quad \mu_\psi^\epsilon = \frac{1}{\epsilon}\gamma'(y^\epsilon-\psi^\epsilon)p^\epsilon,\quad \mu^\epsilon = \mu_\varphi^\epsilon + \mu_\psi^\epsilon,$$

and let $S_\delta(\cdot)\in C^1(\mathcal{R})$ be a family of smooth approximations to the sign function and satisfy the following:

$$S_\delta'(r) \geq 0 \qquad \forall r\in\mathcal{R}$$

and

$$S_\delta(r) = \begin{cases} 1 & \text{if} \quad r>\delta,\\ 0 & \text{if} \quad r=0,\\ -1 & \text{if} \quad r<-\delta. \end{cases}$$

Then we can multiply (3.3) by $S_\delta(p^\epsilon)$ and integrate it over Q. As a result, we get

$$\int_Q \mu^\epsilon S_\delta(p^\epsilon)dxdt \leq C.$$

Moreover, by letting $\delta\to 0^+$ we have that

$$\|\mu^\epsilon\|_{L^1(Q)}\,(=\|\mu_\varphi^\epsilon\|_{L^1(Q)} + \|\mu_\psi^\epsilon\|_{L^1(Q)}) \leq C.$$

Thus, after extracting some subsequence if necessary, we can let

$$\left.\begin{array}{c} \mu_\varphi^\epsilon \rightharpoonup \bar{\mu}_\varphi \\ \mu_\psi^\epsilon \rightharpoonup \bar{\mu}_\psi \\ \mu^\epsilon \rightharpoonup \bar{\mu} \end{array}\right\} \text{weakly}^\star \text{ in } \mathcal{M}_0(\bar{Q}), \tag{3.19}$$

where $\bar{\mu} = \bar{\mu}_\varphi + \bar{\mu}_\psi$.

Multiplying (3.3) by any test function χ in $\dot{W}_p^{2,1}(Q)$, integrating it over Q, and then letting $\epsilon \to 0^+$, we get

$$\int_Q [\bar{p}\chi_t + \nabla\bar{p}\cdot\nabla\chi - f_y(x,t,\bar{y})\bar{p}\chi - (\bar{y}-z_d)\chi]\,dxdt + \langle\bar{\mu},\chi\rangle = 0 \quad \forall\chi \in \dot{W}_p^{2,1}(Q),$$

which shows that $\bar{p}$ solves (3.9).

Furthermore, by passing to the limit in (3.6) we have that

$$\begin{aligned}\langle\bar{\mu}_\varphi,\varphi-\bar{\varphi}\rangle &+ \langle\bar{\mu}_\psi,\psi-\bar{\psi}\rangle \\ &+\int_Q [|\varphi_t|^{p-2}\varphi_t(\varphi-\bar{\varphi})_t + |\Delta\varphi|^{p-2}\Delta\varphi\Delta(\varphi-\bar{\varphi}) \\ &\quad + |\psi_t|^{p-2}\psi_t(\psi-\bar{\psi})_t + |\Delta\psi|^{p-2}\Delta\psi\Delta(\psi-\bar{\psi})]dxdt \geq 0 \qquad \forall(\varphi,\psi)\in\mathcal{U}\end{aligned}$$

and consequently, by inserting $\varphi = \varphi^\epsilon + w$ and $\psi = \psi^\epsilon + w$ into the above inequality, we have that

(3.20)

$$\begin{aligned}&\langle\bar{\mu},w\rangle + \int_Q [|(\bar{\varphi}+w)_t|^{p-2}(\bar{\varphi}+w)_t + |(\bar{\psi}+w)_t|^{p-2}(\bar{\psi}+w)_t]w_t dxdt \\ &+\int_Q [|\Delta(\bar{\varphi}+w)|^{p-2}\Delta(\bar{\varphi}+w) + |\Delta(\bar{\psi}+w)|^{p-2}\Delta(\bar{\psi}+w)]\Delta w dxdt \geq 0 \ \forall w \in \dot{W}_p^{2,1}(Q).\end{aligned}$$

Note that for $p' = p/(p-1)$,

$$\||\varphi_t^\epsilon|^{p-2}\varphi_t^\epsilon + |\psi_t^\epsilon|^{p-2}\psi_t^\epsilon\|_{p'} \leq \|\varphi_t^\epsilon\|_p^{p-1} + \|\psi_t^\epsilon\|_p^{p-1} \leq C,$$

$$\||\Delta\varphi^\epsilon|^{p-2}\Delta\varphi^\epsilon + |\Delta\psi^\epsilon|^{p-2}\Delta\psi^\epsilon\|_{p'} \leq \|\Delta\varphi^\epsilon\|_p^{p-1} + \|\Delta\psi^\epsilon\|_p^{p-1} \leq C,$$

we may assume for some subsequence that

$$|\varphi_t^\epsilon|^{p-2}\varphi_t^\epsilon + |\psi_t^\epsilon|^{p-2}\psi_t^\epsilon \rightharpoonup F \quad \text{weakly in } L^{p'}(Q),$$

$$|\Delta\varphi^\epsilon|^{p-2}\Delta\varphi^\epsilon + |\Delta\psi^\epsilon|^{p-2}\Delta\psi^\epsilon \rightharpoonup G \quad \text{weakly in } L^{p'}(Q).$$

Taking the limit in (3.7), we obtain

$$\langle\bar{\mu},w\rangle + \int_Q (Fw_t + G\Delta w)dxdt = 0 \qquad \forall w \in \dot{W}_p^{2,1}(Q). \tag{3.21}$$

Now, to prove (3.13), we need only verify that

(3.22)

$$\begin{aligned}&\int_Q (F\chi_t + G\Delta\chi)dxdt \\ &= \int_Q [(|\bar{\varphi}_t|^{p-2}\bar{\varphi}_t + |\bar{\psi}_t|^{p-2}\bar{\psi}_t)\chi_t + (|\Delta\bar{\varphi}|^{p-2}\Delta\bar{\varphi} + |\Delta\bar{\psi}|^{p-2}\Delta\bar{\psi})\Delta\chi]dxdt \\ &\forall\chi \in \dot{W}_p^{2,1}(Q).\end{aligned}$$

In fact, by combining (3.21) and (3.21) we have that

(3.23)
$$\begin{aligned}&\int_Q (Fw_t + G\Delta w)dxdt\\ &\quad\le \int_Q [|(\bar\varphi + w)_t|^{p-2}(\bar\varphi + w)_t + |(\bar\psi + w)_t|^{p-2}(\bar\psi + w)_t]w_t dxdt\\ &\quad+\int_Q [|\Delta(\bar\varphi + w)|^{p-2}\Delta(\bar\varphi + w) + |\Delta(\bar\psi + w)|^{p-2}\Delta(\bar\psi + w)]\Delta w dxdt\\ &\quad\forall w\ \in \dot W_p^{2,1}(Q).\end{aligned}$$

For any given $\chi \in \dot W_p^{2,1}(Q)$, take $w = \kappa\chi$ ($\kappa \neq 0$) in (3.24), then divide it by κ. Finally, by letting $\kappa \to 0^+$ and $\kappa \to 0^-$, respectively, we get

$$\begin{aligned}&\int_Q (F\chi_t + G\Delta\chi)dxdt\\ &\quad\le \int_Q [(|\bar\varphi_t|^{p-2}\bar\varphi_t + |\bar\psi_t|^{p-2}\bar\psi_t)\chi_t + (|\Delta\bar\varphi|^{p-2}\Delta\bar\varphi + |\Delta\bar\psi|^{p-2}\Delta\bar\psi)\Delta\chi]dxdt\end{aligned}$$

and

$$\begin{aligned}&\int_Q (F\chi_t + G\Delta\chi)dxdt\\ &\quad\ge \int_Q [(|\bar\varphi_t|^{p-2}\bar\varphi_t + |\bar\psi_t|^{p-2}\bar\psi_t)\chi_t + (|\Delta\bar\varphi|^{p-2}\Delta\bar\varphi + |\Delta\bar\psi|^{p-2}\Delta\bar\psi)\Delta\chi]dxdt,\end{aligned}$$

respectively. Thus, (3.23) follows.

Denote

$$\lambda_\varphi^\epsilon = \frac{1}{\epsilon}\beta(y^\epsilon - \varphi^\epsilon), \quad \lambda_\psi^\epsilon = \frac{1}{\epsilon}\gamma(y^\epsilon - \psi^\epsilon), \quad \lambda^\epsilon = \lambda_\varphi^\epsilon + \lambda_\psi^\epsilon.$$

Recalling approximate equation (2.4), estimates (2.5)–(2.6), and Lemma 2.3, we know that

(3.24)
$$\left.\begin{array}{l}\lambda_\varphi^\epsilon \rightharpoonup \bar\lambda_\varphi\\ \lambda_\psi^\epsilon \rightharpoonup \bar\lambda_\psi\end{array}\right\} \quad \text{weakly in } L^p(Q),$$

$$\lambda^\epsilon = -y_t^\epsilon + \Delta y^\epsilon + f(x,t,y^\epsilon) \rightharpoonup -\bar y_t + \Delta\bar y + f(x,t,\bar y) \equiv \bar\lambda,$$

where $\bar\lambda = \bar\lambda_\varphi + \bar\lambda_\psi$.

Now, (3.13) is equivalent to

(3.25)
$$\int_Q \bar\lambda\,\bar p\,dxdt = 0,$$

which can be proved by the following steps:

(1) Prove

(3.26)
$$\langle\mu_\varphi^\epsilon,\, y^\epsilon - \varphi^\epsilon\rangle \to 0; \qquad \langle\mu_\psi^\epsilon,\, y^\epsilon - \psi^\epsilon\rangle \to 0$$

(which implies (3.16)).

As $\mu_\varphi^\epsilon = \frac{1}{\epsilon}\beta'(y^\epsilon - \varphi^\epsilon)p^\epsilon$ is different from zero only when $y^\epsilon - \varphi^\epsilon < 0$, we have

$$\langle \mu_\varphi^\epsilon,\, (y^\epsilon - \varphi^\epsilon)^+ \rangle = 0 \qquad \forall \epsilon > 0.$$

In addition, by (3.19) and Lemma 2.3, we have that

$$\langle \mu_\varphi^\epsilon,\, (y^\epsilon - \varphi^\epsilon)^- \rangle \to \langle \bar{\mu}_\varphi,\, (\bar{y} - \bar{\varphi})^- \rangle = \langle \bar{\mu}_\varphi,\, 0 \rangle = 0.$$

Hence,

$$\langle \bar{\mu}_\varphi, \bar{y} - \bar{\varphi} \rangle = \lim_{\epsilon \to 0^+} \langle \mu_\varphi^\epsilon,\, y^\epsilon - \varphi^\epsilon \rangle = 0.$$

The other assertion in (3.26) can be obtained similarly.

(2) Prove

$$\int_Q \lambda^\epsilon\, p^\epsilon\, dxdt \to 0. \tag{3.27}$$

By direct computation, we have

$$\int_Q \lambda_\psi^\epsilon\, p^\epsilon\, dxdt = \frac{1}{\epsilon}\int_Q \left[\left(y^\epsilon - \psi^\epsilon - \frac{1}{4} \right) \chi_{\{y^\epsilon - \psi^\epsilon \geq \frac{1}{2}\}} + (y^\epsilon - \psi^\epsilon)^2 \chi_{\{0 < y^\epsilon - \psi^\epsilon < \frac{1}{2}\}} \right] p^\epsilon\, dxdt$$

and

$$\langle \mu_\psi^\epsilon,\, y^\epsilon - \psi^\epsilon \rangle = \frac{1}{\epsilon}\int_Q [(y^\epsilon - \psi^\epsilon)\chi_{\{y^\epsilon - \psi^\epsilon \geq \frac{1}{2}\}} + 2(y^\epsilon - \psi^\epsilon)^2 \chi_{\{0 < y^\epsilon - \psi^\epsilon < \frac{1}{2}\}}]\, p^\epsilon\, dxdt.$$

Then

$$2\int_Q \lambda_\psi^\epsilon\, p^\epsilon\, dxdt - \langle \mu_\psi^\epsilon,\, y^\epsilon - \psi^\epsilon \rangle = \frac{1}{\epsilon}\int_Q \left(y^\epsilon - \psi^\epsilon - \frac{1}{2} \right) \chi_{\{y^\epsilon - \psi^\epsilon \geq \frac{1}{2}\}}\, p^\epsilon\, dxdt.$$

Noting the fact that

$$\frac{1}{4} \leq \left| y^\epsilon - \psi^\epsilon - \frac{1}{4} \right| \qquad \text{in } \left\{ y^\epsilon - \psi^\epsilon \geq \frac{1}{2} \right\},$$

we get by using (2.7) that

$$\frac{1}{4^p}\text{meas}\left\{ y^\epsilon - \psi^\epsilon \geq \frac{1}{2} \right\} \leq \int_Q \left| y^\epsilon - \psi^\epsilon - \frac{1}{4} \right|^p \chi_{\{y^\epsilon - \psi^\epsilon \geq \frac{1}{2}\}}\, dxdt \leq \|\gamma(y^\epsilon - \psi^\epsilon)\|_{L^p(Q)}^p \to 0.$$

Applying Hölder's inequality, we obtain

$$\begin{aligned}
&\left| 2\int_Q \lambda_\psi^\epsilon\, p^\epsilon\, dxdt - \langle \mu_\psi^\epsilon,\, y^\epsilon - \psi^\epsilon \rangle \right| \\
&\quad \leq \frac{1}{\epsilon}\left\| \left(y^\epsilon - \psi^\epsilon - \frac{1}{2} \right) \chi_{\{y^\epsilon - \psi^\epsilon \geq \frac{1}{2}\}} \right\|_{L^p(Q)} \|p^\epsilon \chi_{\{y^\epsilon - \psi^\epsilon \geq \frac{1}{2}\}}\|_{L^{p'}(Q)} \\
&\quad \leq \frac{2}{\epsilon}\left\| \left(y^\epsilon - \psi^\epsilon - \frac{1}{4} \right) \chi_{\{y^\epsilon - \psi^\epsilon \geq \frac{1}{2}\}} \right\|_{L^p(Q)} \|p^\epsilon\|_{L^2(Q)} \|\chi_{\{y^\epsilon - \psi^\epsilon \geq \frac{1}{2}\}}\|_{L^{\frac{2p'}{2-p'}}(Q)} \\
&\quad \leq \frac{2}{\epsilon}\|\gamma(y^\epsilon - \psi^\epsilon)\|_{L^p(Q)}\, \|p^\epsilon\|_{L^2(Q)}\, \left(\text{meas}\left\{ y^\epsilon - \psi^\epsilon \geq \frac{1}{2} \right\} \right)^{\frac{2-p'}{2p'}} \\
&\quad = 2\|\lambda_\psi^\epsilon\|_{L^p(Q)}\, \|p^\epsilon\|_{L^2(Q)}\, \left(\text{meas}\left\{ y^\epsilon - \psi^\epsilon \geq \frac{1}{2} \right\} \right)^{\frac{2-p'}{2p'}} \\
&\quad \to 0,
\end{aligned}$$

which, in view of (3.26), leads to

$$\int_Q \lambda_\psi^\epsilon \, p^\epsilon \, dxdt \to 0.$$

Similarly, we can prove that

$$\int_Q \lambda_\varphi^\epsilon \, p^\epsilon \, dxdt \to 0,$$

and hence (3.27) follows.

(3) Prove

$$\int_Q (\bar{\lambda} - \lambda^\epsilon) \, p^\epsilon \, dxdt \to 0. \tag{3.28}$$

Lemma 2.3 gives

$$y^\epsilon \to \bar{y} \qquad \text{strongly in } C_0(\bar{Q}),$$

and thus

$$\langle \mu^\epsilon, \, \bar{y} - y^\epsilon \rangle \to 0;$$

$$\|f(x,t,\bar{y}) - f(x,t,y^\epsilon)\|_{L^\infty(Q)} \to 0.$$

Then, it is not hard to get

$$\begin{aligned}
&\int_Q (\bar{\lambda} - \lambda^\epsilon) \, p^\epsilon \, dxdt \\
&\quad = \int_Q [-(\bar{y} - y^\epsilon)_t + \Delta(\bar{y} - y^\epsilon) + (f(x,t,\bar{y}) - f(x,t,y^\epsilon))] \, p^\epsilon \, dxdt \\
&\quad = \int_Q [(p_t^\epsilon + \Delta p^\epsilon)(\bar{y} - y^\epsilon) + (f(x,t,\bar{y}) - f(x,t,y^\epsilon))p^\epsilon] \, dxdt \\
&\quad = \int_Q [(\mu^\epsilon - f_y(x,t,y^\epsilon)p^\epsilon - (y^\epsilon - z_d))(\bar{y} - y^\epsilon) + (f(x,t,\bar{y}) - f(x,t,y^\epsilon))p^\epsilon] \, dxdt \\
&\quad \to 0.
\end{aligned}$$

By the weak convergency of p^ϵ to $\bar{p}$ in $L^2(Q)$,

$$\int_Q \bar{\lambda} \, \bar{p} \, dxdt = \lim_{\epsilon \to 0^+} \int_Q \bar{\lambda} \, p^\epsilon \, dxdt.$$

Thus, (3.27) combined with (3.28) implies (3.25). (Note that, since both $\{\lambda^\epsilon\}$ and $\{p^\epsilon\}$ are only weakly convergent, (3.27) does not suffice to imply (3.25) without (3.28).)

It remains to prove (3.10) and (3.11). As $p > (n+2)/2$, the $W_p^{2,1}$-bounded subset is relatively compact in $C^{\theta,\theta/2}(\bar{Q})$ for some $\theta \in (0,1)$. So, for any $\eta \in C_0(\bar{Q})$ with $\operatorname{supp} \eta \subset Q'$, the uniform convergence of the approximate optimal control and state (cf. Proposition 3.5), together with the compactness of $\operatorname{supp} \eta$, ensures that, for some $\epsilon_0 > 0$,

$$\varphi^\epsilon(x,t) < y^\epsilon(x,t) < \psi^\epsilon(x,t) \qquad \forall (x,t) \in \operatorname{supp} \eta, \quad 0 < \epsilon < \epsilon_0,$$

which yields

$$\begin{aligned}\langle \bar{\mu},\, \eta\rangle_{\mathcal{M}_0(\bar{Q}),\, C_0(\bar{Q})} &= \lim_{\epsilon\to 0}\int_Q \mu^\epsilon \eta\, dxdt \\ &= \lim_{\epsilon\to 0}\int_{\operatorname{supp}\eta} \frac{1}{\epsilon}[\beta'(y^\epsilon-\varphi^\epsilon)+\gamma'(y^\epsilon-\psi^\epsilon)]p^\epsilon\eta\, dxdt \\ &= 0.\end{aligned}$$

Thus (3.10) holds. Replacing μ^ϵ and $\bar{\mu}$ by λ^ϵ and $\bar{\lambda}$, respectively, (3.11) follows.

Alternatively, we can first prove (3.18) which also leads to (3.10).

In a similar way, (3.14) and (3.15) can be obtained.

The proof is completed. □

In Theorem 3.6, it is assumed that $p > (n+2)/2$, which excludes the case $p = 2$ when $n \geq 2$. Hence, before closing this section, we consider the case $p = 2$.

When $p = 2$, the objective functional $J(\varphi, \psi)$ and Problem (C) are reduced to

$$J_2(\varphi,\psi) \equiv \frac{1}{2}\int_Q \left\{[\mathcal{S}(\varphi,\psi)-z_d]^2 + |\varphi_t|^2 + |\Delta\varphi|^2 + |\psi_t|^2 + |\Delta\psi|^2\right\} dxdt$$

and the following.

PROBLEM (C_2). *Find a control pair* $(\bar{\varphi}, \bar{\psi}) \in \mathcal{U}_2$ *such that*

$$J_2(\bar{\varphi},\bar{\psi}) = \inf_{(\varphi,\psi)\in\mathcal{U}_2} J_2(\varphi,\psi),$$

respectively, where

$$\begin{aligned}\mathcal{U}_2 = \{(\varphi,\psi) \in W_2^{2,1}(Q)\times W_2^{2,1}(Q) | \varphi \leq \psi \text{ in } Q,\\ \varphi = 0 = \psi \quad \text{on } \Sigma,\ \varphi|_{t=0} \leq y_0 \leq \psi|_{t=0} \text{ in } \Omega\}.\end{aligned}$$

Accordingly, the approximate optimal control problem is to minimize the following approximate functional:

$$J_2^\epsilon(\varphi,\psi) \equiv \frac{1}{2}\int_Q [(y^\epsilon - z_d)^2 + |\varphi_t|^2 + |\Delta\varphi|^2 + |\psi_t|^2 + |\Delta\psi|^2 + |\varphi-\bar{\varphi}|^2 + |\psi-\bar{\psi}|^2] dxdt,$$

where $y^\epsilon = \mathcal{S}^\epsilon(\varphi,\psi)$ is the approximate state solving (2.4) and $(\bar{\varphi},\bar{\psi})$ is an optimal control pair of Problem (C_2).

The following result is an analogy to Proposition 3.5.

PROPOSITION 3.7. *Let* $(\bar{\varphi},\bar{\psi},\bar{y})$ *be an optimal triple for Problem* (C_2), *let* $(\varphi^\epsilon,\psi^\epsilon) \in \mathcal{U}_2$ *be approximate optimal control pairs, and let* $y^\epsilon = \mathcal{S}^\epsilon(\varphi^\epsilon,\psi^\epsilon)$. *Then*

$$\left.\begin{array}{l}\varphi^\epsilon \to \bar{\varphi}\\ \psi^\epsilon \to \bar{\psi}\\ y^\epsilon \to \bar{y}\end{array}\right\} \text{ weakly in } W_2^{2,1}(Q) \text{ and strongly in } L^2(0,T;H_0^1(\Omega)).$$

By taking limits in the corresponding approximate optimality condition, which can be obtained easily, owing to the quadratic cost functional and smooth state equation, we arrive at the following result for Problem (C_2).

THEOREM 3.8. *Let* $(\bar{\varphi},\bar{\psi})$ *be an optimal control pair for Problem* (C_2) *and let* $\bar{y} = \mathcal{S}(\bar{\varphi},\bar{\psi})$. *Then there exist* $\bar{p} \in L^2(0,T;H_0^1(\Omega))$ *and* $\bar{\mu} \in \mathcal{M}_0(\bar{Q})$ *such that* $\bar{p}$ *solves the following equation:*

$$\left\{\begin{array}{ll}-\bar{p}_t - \Delta\bar{p} - f_y(x,t,\bar{y})\bar{p} = \bar{y} - z_d - \bar{\mu} & \text{in } Q,\\ \bar{p}|_\Sigma = 0,\\ \bar{p}|_{t=T} = 0,\end{array}\right.$$

and furthermore,

$$\int_Q [(\bar{\varphi}_t + \bar{\psi}_t)w_t + (\Delta\bar{\varphi} + \Delta\bar{\psi})\Delta w]dxdt + \langle \bar{\mu}, w\rangle = 0 \qquad \forall\ w\ \in\ \dot{W}_2^{2,1}(Q).$$

In the present situation with $p = 2$, although the state analysis in section 2 is still valid for the set of approximate optimal controls and states, the compactness in Hölder spaces $C^{\theta,\theta/2}(\bar{Q})$ with $\theta \in (0,1)$ would not be expected to remain valid, since a bounded subset in $W_2^{2,1}(Q)$ is not necessarily compact in $C^{\theta,\theta/2}(\bar{Q})$ for any $\theta \in (0,1)$ when $n \geq 2$. As a result, if $p = 2$, the conditions (3.10), (3.11), *and* (3.13)–(3.17) *in Theorem* 3.6 *cannot be expected to remain valid; however, the results and analysis for the related Problem* (C_2) *and Theorem* 3.8 *would still be useful.*

Acknowledgments. We would like to thank the anonymous referees, the editor, and the associate editor for their valuable comments and constructive criticism, which helped us to improve the paper substantially.

REFERENCES

[1] D.R. Adams, S.M. Lenhart and J. Yong, *Optimal control of obstacle for elliptic variational inequality*, Appl. Math. Optim., 38 (1998), pp. 121–140.
[2] D.R. Adams and S.M. Lenhart, *Optimal control of the obstacle for a parabolic variational inequality*, J. Math. Anal. Appl., 268 (2002), pp. 602–614.
[3] D.R. Adams and S.M. Lenhart, *An obstacle control problem with a source term*, Appl. Math. Optim., 47 (2003), pp. 79–95.
[4] V. Barbu, *Optimal Control of Variational Inequalities*, Pitman, London, 1984.
[5] V. Barbu, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic Press, New York, 1993.
[6] M. Bergounioux and S. Lenhart, *Optimal control of bilateral obstacle problems*, SIAM J. Control Optim., 43 (2004), pp. 240–255.
[7] J.F. Bonnans and D. Tiba, *Pontryagin's principle in the control of semilinear elliptic variational inequalities*, Appl. Math. Optim., 23 (1991), pp. 299–312.
[8] Q. Chen, *Indirect obstacle control problem for semilinear elliptic variational inequalities*, SIAM J. Control Optim., 38 (1999), pp. 138–158.
[9] Q. Chen, *Indirect obstacle optimal control for evolutionary variational inequalities with state constraints*, Sci. China Ser. E, 43 (2000), pp. 653–669.
[10] Q. Chen, *Indirect obstacle minimax control for elliptic variational inequalities*, J. Optim. Theory Appl., 110 (2001), pp. 337–359.
[11] M. Dai and F. Yi, *Finite-Horizon Optimal Investment with Transaction Costs: A Parabolic Double Obstacle Problem*, working paper, 2006. Available online at http://www.math.nus.edu.sg/∼matdm/oitc.pdf
[12] M. Dai and Y.K. Kwok, *Optimal policies of call with notice period requirement for American warrants and convertible bonds*, Asia Pacific Financial Markets, 12 (2005), pp. 353–373.
[13] A. Friedman, *Variational Principles and Free-boundary Problems*, John Wiley & Sons, New York, 1982.
[14] A. Friedman, *Optimal control for variational inequalities*, SIAM J. Control Optim., 24 (1986), pp. 439–451.
[15] A. Friedman, *Optimal control for parabolic variational inequalities*, SIAM J. Control Optim., 25 (1987), pp. 482–497.
[16] M. Fuchs and N. Fusco, *Partial regularity results for vector valued functions which minimize certain functions having nonquadratic growth under smooth side conditions*, J. Reine Angew. Math., 390 (1988), pp. 67–78.
[17] K. Ito and K. Kunisch, *Optimal control of elliptic variational inequalities*, Appl. Math. Optim., 41 (2000), pp. 343–364.
[18] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980; SIAM Classics in Appl. Math. 31, SIAM, Philadelphia, 2000.
[19] O.A. Ladyzhenskaya, V.A. Solonnikov, and N.N. Ural'ceva, *Linear and Quasi-Linear Equations of Parabolic Type*, Amer. Math. Soc. Transl., American Mathematical Society, Providence, RI, 1968.

[20] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser Boston, Boston, MA, 1995.
[21] H. Lou, *On the regularity of an obstacle control problem*, J. Math. Anal. Appl., 258 (2001), pp. 32–51.
[22] H. Lou, *An optimal control problem governed by quasi-linear variational inequalities*, SIAM J. Control Optim., 41 (2002), pp. 1229–1253.
[23] F. Mignot and J.P. Puel, *Optimal control in some variational inequalities*, SIAM J. Control Optim., 22 (1984), pp. 466–476.
[24] J.F. Rodrigues, *Obstacle Problems in Mathematical Physics*, North–Holland Math. Stud. 134, North–Holland, Amsterdam, 1987.
[25] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
[26] J. Sokołowski and J.-P. Zolésio, *Introduction to Shape Optimization: Shape Sensitivity Analysis*, Springer-Verlag, New York, 1992.
[27] G.M. Troianiello, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, 1987.
[28] P. Wilmott, J. Dewynne, and S. Howison, *Option Pricing: Mathematical Models and Computation*, Oxford Financial Press, Oxford, UK, 1993.
[29] J. Yong, *Pontryagin maximum principle for semilinear second order elliptic partial differential equations and variational inequalities with state constraints*, Differential Integral Equations, 5 (1992), pp. 1307–1334.
[30] K. Ito and K. Kunisch, *Parabolic variational inequalities: The Lagrange multiplier approach*, J. Math. Pures Appl., 85 (2006), pp. 415–449.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1539–1561

DIET SELECTION AS A DIFFERENTIAL FORAGING GAME*

FRÉDÉRIC HAMELIN†, PIERRE BERNHARD†, A. J. SHAIJU‡, AND ÉRIC WAJNBERG§

Abstract. An important issue addressed by behavioral ecology is that of the evolutionary relevance of foraging strategies adopted by animals in quest of a patchily distributed resource, both in terms of diet selection and patch-leaving decisions under competition. We revisit the classical model of diet selection concerning an isolated—not subject to competition—forager; it yields a zero-one rule, i.e., a type of resource should be always accepted, or always rejected, that appears to be more the exception than the rule, as partial preferences are commonly observed in many species. Thus arises the question of the rule's robustness where there is an uncertainty on the time available to a forager to enjoy a patch, due to the possible occurrence of a perturbating event. We mean any event that would affect its gain with respect to what it would obtain by enjoying alone the patch as long as it wants. For instance, the sudden presence of a predator could force it to flee the patch or the arrival of a conspecific would deprive it of some good resources. By taking into account the potentially imminent arrival of a conspecific—but also any event that would suddenly shorten patch exploitation—we show that the classical policy of diet selection no longer holds, as it changes the qualitative aspect of the optimal foraging strategies. Qualitatively, the optimal strategy is close to, but less greedy than, the evolutionarily stable strategy that concerns foragers actually competing for resources. It consists in accepting only the most profitable resource until it is depleted down to a given level, after which time both resources are accepted. The underlying mathematical technique involves the solution of nonzero-sum differential games and synthesis techniques.

Key words. differential games, evolutionarily stable strategies, optimal foraging theory, behavioral ecology

AMS subject classification. 91A23, 91A22

DOI. 10.1137/06065814X

1. Introduction.

"Nothing in biology makes sense except in the light of evolution."[1] In this respect, behavioral ecology [18] interprets an animal's behavior through an evolutionary approach, via estimating its capacity to get through the natural selection process and thus maximize Darwinian *fitness* [22]—a notion analogous to that of "utility" in economics. Typically, in foraging theory [33], or the art of gathering resources in the environment, fitness is related to the amount of resource gathered. In many cases, the resource is patchily distributed and the utility function on each patch is strictly increasing, concave, and bounded with respect to time. As the *intake-rate* decreases with the quantity of resource available on the patch, it is likely advantageous for an animal to leave a patch not yet exhausted in order to find a new one, in spite of an uncertain *travel time*. Charnov's marginal value theorem [6] reveals that the optimal giving-up time is when the intake-rate is equal to the optimal long-term

*Received by the editors April 25, 2006; accepted for publication (in revised form) March 4, 2007; published electronically October 10, 2007. This work is part of GDR 2155 "Écologie Comportementale" (CNRS commission 29), INRA/Ecoger, and ESF/BEPAR scientific programmes.
http://www.siam.org/journals/sicon/46-5/65814.html

†Laboratory I3S, CNRS and Université de Nice–Sophia Antipolis, École Polytechnique Universitaire de Nice–Sophia Antipolis, 06903 Sophia Antipolis, France (hamelin@polytech.unice.fr, bernhard@polytech.unice.fr).

‡School of Information Technology and Electrical Engineering, UNSW@ADFA, Northcott Drive, Canberra ACT 2600, Australia (ajshaiju@gmail.com). At this time, this author was a postdoctoral fellow at Laboratory I3S, Université de Nice–Sophia Antipolis.

§INRA Sophia Antipolis, 400 Route des Chappes, BP 167, 06903 Sophia Antipolis, France (wajnberg@sophia.inra.fr).

[1] Theodosius Dobzhansky, geneticist, 1900–1975.

mean rate γ^* which, if achieved, gives the best fitness a forager can expect in its environment.

This famous theoretical model was originally designed for lone foragers in quest of a singular patchily distributed resource. In parallel, another branch of the theory grew by focusing on the optimal diet selection [5, 17, 31] when the environment offers a plural resource which varies both in profitability and abundance but spatially is regularly and homogeneously distributed. The authors of [23] have merged these two theories.

Naturally, the question arises of whether this theory holds for foragers competing for a common patchily distributed resource, i.e., whether this is an evolutionarily stable strategy [20]; for instance, it might have implications in terms of population dynamics [34, 37].

Concerning the singular resource case, Charnov's patch-leaving rule remains qualitatively unchanged under scramble competition, i.e., when the only competition between foragers is in sharing a common resource [32, 8]; γ^* is clearly affected by the number of potential competitors, but the patch-leaving rule is unchanged. However, if there is *interference*, i.e., a decline in intake-rate due to competition, the game results in a *war of attrition* [32, 9] or random patch-leaving strategies.

In the present paper, our aim is to determine the evolutionarily stable strategy that noninterfering foragers competing for a plural and depleting resource should adopt, both in terms of diet selection and patch-leaving decision [3, 33].

The remainder of the paper is organized as follows. In section 2, we reformulate the optimal diet selection policy for a lone forager free to leave the current patch of resources at any time. On our way, we solve the optimal diet selection problem for a single forager with a fixed end time; this is done in Appendix A. In section 3, we investigate the foraging game involving several foragers arriving simultaneously on a patch containing two distinct types of resources. Section 4 focuses on an asynchronous two-forager game, where the interarrival time is assumed deterministic. Finally, the game considered in section 5 lets the possible arrival of an opponent be a Poisson variable.[2]

2. Foraging alone. It is well known [5, 16] that a lone forager should accept a unit of resource i if its energy value e_i is worth the time required to retrieve it, i.e., the *handling-time* h_i. Indeed, Charnov's marginal value theorem [6, 22] prohibits the intake-rate from falling below a critical threshold γ^*. Hence the rule is to accept this resource if and only if $\gamma^* \leq e_i/h_i$. However, let us recall this result in order to introduce our modeling and solution approaches—the latter is close to that of [26]. We shall define *profitability* of resource i as the ratio e_i/h_i. We shall also let $\delta_i := e_i - \gamma^* h_i$.

Let x be the state vector containing the ratios $x_i \in [0,1]$ of each type of resource available in the patch. Let u be the control vector containing the controls $u_i \in [0,1]$ deciding the acceptance rate[3] of each type of resource available in the patch. Let $\dot{x} := \mathrm{d}x/\mathrm{d}t$, where t stands for the *residence time*.

Proceeding as in [14, 8] and most of the literature, an assumption of random

[2]A particular case of that game is that of a single player with an exponential random end time, such as the possible occurrence of a predator.

[3]Or equivalently the probability to accept a given type of resource when encountered.

probing on a patch yields the following dynamics:

$$(2.1)\quad \forall i \in \{1,\dots,N\}\,,\quad q\dot{x}_i = -\frac{u_i x_i}{\alpha + \sum_{j=1}^N u_j x_j h_j}\,,\quad x_i(0) = x_i^0\,,\quad \sum_{i=1}^N x_i^0 = 1\,,$$

where α is the time required to probe an area of the patch that could contain a unit of resource and q is the quality of the patch or the quantity of resources it initially contains.

Following [22], we want to maximize the criterion

$$(2.2)\qquad J = \int_0^{t^*} L(x,u)\mathrm{d}t \quad \text{with} \quad L(x,u) = -\sum_{j=1}^N e_j q\dot{x}_j - \gamma^*\,,$$

where t^* is a free final time.

We claim the following result.

THEOREM 2.1. *The optimal policy in the problem stated by* (2.1) *and* (2.2) *is given by*

- $\forall t \in [0,t^*]$, *take* $u_i = \begin{cases} 1 \text{ if } \gamma^* < e_i/h_i, \\ \text{arbitrary in } [0,1] \text{ if } \gamma^* = e_i/h_i, \\ 0 \text{ if } \gamma^* > e_i/h_i \end{cases}$
- *and leave as soon as* $\sum_{j=1}^N u_j x_j (e_j - \gamma^* h_j) - \gamma^* \alpha \le 0$.

Proof. Let s be such that $\mathrm{d}t = qD\mathrm{d}s$ with $D := \alpha + \sum_{j=1}^N h_j u_j x_j$. Let $\mathring{x} := \mathrm{d}x/\mathrm{d}s$ and $f(x,u) := \mathring{x}$. The dynamics become

$$\mathring{x_i} = -u_i x_i\,,\quad x_i(0) = x_i^0\,.$$

Our criterion can now be expressed as follows: Let

$$\mathcal{J} := J/q = \int_0^{s^*} \mathcal{L}(x,u)\mathrm{d}s \quad \text{with} \quad \mathcal{L}(x,u) = \sum_{j=1}^N u_j x_j e_j - \gamma^* D\,.$$

It directly yields that the optimal end time is such that $\mathcal{L}$ is zero on the optimal trajectories, since $\partial\mathcal{J}/\partial s^* = \mathcal{L}(x(s^*), u(s^*)) = 0$; this corresponds to Charnov's patch-leaving rule. Hence the claim of the theorem.

Let λ be the adjoint vector. It yields the Hamiltonian

$$\mathcal{H} = \mathcal{L}(x,u) + \langle \lambda, f(x,u)\rangle = \sum_{j=1}^N (\delta_j - \lambda_j) u_j x_j - \alpha\gamma^*\,.$$

According to Pontryagin's maximum principle [30], if a policy $u^*(s)$ generating a trajectory $x^*(s)$ is optimal, then there exists an adjoint trajectory $\lambda(s)$ such that

$$\begin{cases} \mathring{\lambda} = -\nabla_x \mathcal{H}(\lambda, u^*, x^*), \\ \lambda(s^*) = 0, \\ \mathcal{H}(s^*) = 0. \\ \quad \forall s \in [0,s^*], \text{ where } u^*(\cdot) \text{ is continuous}, \\ \quad \mathcal{H}(\lambda(s), u^*(s), x^*(s)) = \max_{u\in[0,1]^n} \mathcal{H}(\lambda(s), u, x^*(s)). \end{cases}$$

The last condition above translates into the switch-functions

$$\sigma_i := \partial\mathcal{H}/\partial u_i = (\delta_i - \lambda_i)x_i$$

and the *bang-bang* optimal policy

$$u_i^* = \begin{cases} 1 \text{ if } \sigma_i > 0, \\ 0 \text{ if } \sigma_i < 0. \end{cases}$$

The singular case $\sigma_i = 0$ allows the focal forager to either accept or reject the less profitable resource indifferently.

We also have

$$\mathring{\lambda}_i = -\partial\mathcal{H}/\partial x_i = -(\delta_i - \lambda_i)u_i\,, \quad \lambda_i(s^*) = 0\,.$$

It yields $\mathring{\sigma}_i = 0$. Hence the sign of σ_i never changes, and therefore the optimal policy is

$$\forall t \in [0, t^*]\,, \quad u_i = \begin{cases} 1 \text{ if } \gamma^* < e_i/h_i, \\ 0 \text{ if } \gamma^* > e_i/h_i. \end{cases}$$

As already mentioned in [23], the reason the author of [26] found another result is that he—consciously—considered, as a constraint, an arbitrarily predetermined residence time.[4] It has to be noticed that in this simple model, *partial preferences* [21] should occur only in the nongeneric case $\gamma^* = e_i/h_i$. □

Given the optimal policy as a function of γ^*, it is possible to compute both γ^* and the corresponding optimal diet, as done in [23], where the authors provide an algorithm that converges to the solution.

3. The synchronous foraging game. The authors of [15] argue that when "a large number" of foragers is competing for a plural and depleting resource, they should maximize their intake-rate. Thus the evolutionarily stable policy consists of being *selective* first and, after a while, being *opportunistic*.[5] The results of both [26, 35] are in agreement with [15], except that the authors of [26, 35] found "earlier" switch-times for a relatively low number of competitors. However, both approaches point to a convergence of the switch-time towards the intake-rate maximizing switch-time as the number of foragers increases.

Our aim now is to determine the evolutionarily stable policy via an approach similar to that of [26], except that we do not set any arbitrarily predetermined residence time or final patch state.

Following section 2, we now restrict the resource range to those resources which would be included in the diet of a lone forager: $\forall i \in \{1, 2\}\,, \delta_i \geq 0$—the resource types rejected by a lone forager should a fortiori be rejected under competition. We shall also let $\zeta := e_1h_2 - e_2h_1 \geq 0$ as $e_2/h_2 \leq e_1/h_1$ by hypothesis.

Proceeding as in [26], we look for the optimal policy against a strategy assumed commonly adopted by the opponents. If it leads to the latter, this is indeed an evolutionarily stable strategy—as this is a strict and symmetric Nash equilibrium [13, 9]. However, to be consistent, we need to assume a state feedback strategy for the opponents. Hence we must use a regular synthesis technique in order to recover

[4] This issue is addressed in Appendix A.

[5] The author of [10] also mentioned this "expanding-specialist" strategy under competition.

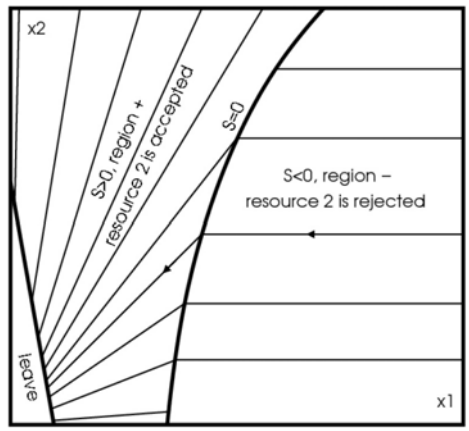


FIG. 3.1. *The Nash-optimal fields of trajectories in the state-space* (x_1, x_2).

the costate vector as a function of the current state and to construct a switch-manifold in the state space. This, in turn, induces discontinuities in the adjoint variables of the focal player and other difficulties that we must take into account. Throughout our reasoning, we shall refer to Figure 3.1, which represents the state-space (x_1, x_2).

Let n be the number of foragers on the patch. Let u be the decision variable of the focal forager, i.e., the acceptance rate of resource 2—as resource 1, the most profitable one, should, of course, always be accepted. Similarly, let v be the decision variable of its opponents.

Let $D(u) := \alpha + h_1 x_1 + u h_2 x_2$. The dynamics are now

$$\left\{ \begin{array}{ll} q\dot{x}_1 = -x_1/D(u) - (n-1)x_1/D(v)\,, & x_1(0) = x_1^0, \\ q\dot{x}_2 = -ux_2/D(u) - v(n-1)x_2/D(v)\,, & x_2(0) = x_2^0. \end{array} \right. \tag{3.1}$$

The criterion of the focal forager is

$$J = \int_0^{t^*} L(x,u)\mathrm{d}t \quad \text{with} \quad L(x,u) = e_1 x_1/D(u) + u e_2 x_2/D(u) - \gamma^*, \tag{3.2}$$

where t^* is a free final time.

We claim the following result (see Figure 3.1).

THEOREM 3.1. *The unique pure symmetric state feedback Nash equilibrium in the game stated by* (3.1) *and* (3.2) *corresponds to*

- *take* $\left\{ \begin{array}{l} u = 0 \text{ as long as } S(x) \le 0, \\ u = 1 \text{ as soon as } S(x) > 0, \end{array} \right.$ *where* $S(x)$ *is given by* (3.9)
- *and leave as soon as* $\delta_1 x_1 + \delta_2 x_2 - \gamma^* \alpha \le 0$.

Proof. Let s be such that $\mathrm{d}t =: qD(u)\mathrm{d}s$. Let $\mathring{x} := \mathrm{d}x/\mathrm{d}s$ and $f(x,u) := \mathring{x}$. The dynamics become

$$\left\{ \begin{array}{ll} \mathring{x}_1 = -x_1(1 + (n-1)D(u)/D(v))\,, & x_1(0) = x_1^0, \\ \mathring{x}_2 = -x_2(u + v(n-1)D(u)/D(v))\,, & x_2(0) = x_2^0. \end{array} \right.$$

Our criterion can now be expressed as follows: Let

$$\mathcal{J} := J/q = \int_0^{s^*} \mathcal{L}(x,u)\mathrm{d}s \quad \text{with} \quad \mathcal{L}(x,u) = e_1 x_1 + u e_2 x_2 - \gamma^* D(u)\,.$$

Clearly, $\partial\mathcal{J}/\partial s^* = \mathcal{L}(x(s^*), u(s^*)) = 0$. It directly yields that the optimal end time is such that $\mathcal{L}$—which does not depend on s—is zero on the optimal trajectories; this corresponds to Charnov's patch-leaving rule. The oblique line sloping to the left in Figure 3.1 represents this terminal manifold.

Let λ be the adjoint vector associated with the focal forager. It yields the Hamiltonian

$$\begin{aligned}\mathcal{H} = {} & e_1x_1 + ue_2x_2 - \gamma^* D(u) - \lambda_1 x_1(1 + (n-1)D(u)/D(v)) \\ & -\lambda_2 x_2(u + v(n-1)D(u)/D(v))\,.\end{aligned}$$

According to Pontryagin's maximum principle,[6] the optimal policy is thus bang-bang, according to the switch-function

$$\sigma = [\delta_2 - \lambda_1(n-1)h_2x_1/D(v) - \lambda_2(1 + v(n-1)h_2x_2/D(v))]x_2\,.$$

We have

$$\begin{aligned}\mathring{\lambda}_1 &= \lambda_1[(1 + (n-1)D(u)/D(v)) + (n-1)h_1x_1(1/D(v) - D(u)/D(v)^2)] \\ &\quad + \lambda_2 v(n-1)h_1x_2(1/D(v) - D(u)/D(v)^2) - \delta_1\,, \quad \lambda_1(s^*) = 0, \\ \mathring{\lambda}_2 &= \lambda_2[(u + v(n-1)D(u)/D(v)) + v(n-1)h_2x_2(u/D(v) - vD(u)/D(v)^2)] \\ &\quad + \lambda_1(n-1)h_2x_1(u/D(v) - vD(u)/D(v)^2) - u\delta_2\,, \quad \lambda_2(s^*) = 0\,.\end{aligned}$$

Clearly, $u^*(s^*) = 1$ as $\sigma(s^*) = \delta_2x_2 \geq 0$ by hypothesis.

Proceeding as in [26], we first assume that the opponents are opportunists. The optimal strategy is then given via integrating backward the above differential equations, with $v = 1$. As long—as from the end time—σ remains positive, being opportunistic is optimal. Thus if it remains so backward up to time zero, being opportunistic the whole time spent on the patch is the evolutionarily stable strategy. Otherwise, i.e., if the sign of σ changes in backward time, being selective is, at least locally, optimal before the switch-point. Thanks to the assumed symmetry among foragers, if such a switch-point appears, then it prevails for any competitor on the patch. Therefore, we shall assume in a second time that $v = 0$ from this possible switch-point down to $s = 0$. However, a prerequisite for reiterating a similar process backward in time is that being selective is optimal against selective opponents.

Let $(\hat{s}, \hat{x})$ be either the first—in backward time—switch-point, if there is one, or $(0, x(0))$ otherwise; i.e., beyond this point, being opportunistic remains optimal up to the end time. Let the superscript + denote the region of the state-space beyond the last switch-point; thus we postulate that in region + the Nash-optimal strategies are $u = v = 1$. For instance, let $D^+ := D(1) = \alpha + h_1x_1 + h_2x_2$. We have $\forall s \in (\hat{s}, s^*)$, $\forall i \in \{1, 2\}$,

$$\mathring{x}_i = -nx_i\,, \quad x_i(s^*) =: x_i^* \quad \text{and} \quad \mathring{\lambda}_i^+ = n\lambda_i^+ - \delta_i\,, \quad \lambda_i^+(s^*) = 0\,.$$

We also have

$$\sigma^+ = [\delta_2 - \lambda_1^+(n-1)h_2x_1/D^+ - \lambda_2^+\left(1 + (n-1)h_2x_2/D^+\right)]x_2 \tag{3.3}$$

and $\mathcal{H}^+ = e_1x_1 + e_2x_2 - \gamma^*D^+ - \lambda_1^+x_1n - \lambda_2^+x_2n$. It yields $\forall s \in (\hat{s}, s^*)$, $\forall i \in \{1, 2\}$,

$$x_i(s) = x_i^*\mathrm{e}^{n(s^*-s)} \quad \text{and} \quad \lambda_i^+(s) = \delta_i(1 - \mathrm{e}^{-n(s^*-s)})/n\,. \tag{3.4}$$

[6] As long as the opponent uses a Lipschitz continuous (here, constant) strategy w.r.t. x.

As a consequence

$$\lambda_1^+(s) = \delta_1(1 - x_1^*/x_1)/n \geq 0\,. \tag{3.5}$$

Moreover, one can notice from (3.4) that $\forall s \in (\hat{s}, s^*)$, $x_1^*/x_1 = x_2^*/x_2$ or, equivalently, that x_1/x_2 is invariant over $[\hat{s}, s^*]$—this results from our assumption of homogeneous probing on the patch, and that is why in Figure 3.1 the field of optimal trajectories is a radial one. Furthermore, as the Hamiltonian remains constant all along the optimal trajectory, it remains equal to zero here and this yields $\forall s \in (\hat{s}, s^*)$, $\forall i \in \{1, 2\}$,

$$x_i^*/x_i = \frac{\gamma^*\alpha}{x_1\delta_1 + x_2\delta_2}\,. \tag{3.6}$$

Hence the claim of the theorem.

The switch-function can also be rewritten as follows:

$$\sigma^+ = \left[(e_2 - \lambda_2^+)(\alpha + h_1x_1) - (e_1 - \lambda_1^+)h_2x_1\right] x_2/D^+\,.$$

Let us now assume that there is a switch-point by definition such that $\sigma^+(\hat{s}) = 0$. Our aim now is to verify that switching—in backward time—u to zero remains optimal if v switches to zero simultaneously at time $\hat{s}$. Let the superscript $-$ denote the region of the state-space where we conjecture that the Nash-optimal strategies are $u = v = 0$. For instance, let $D^- := D(0) = \alpha + h_1x_1$. We have

$$\sigma^-(\hat{s}) = [\delta_2 - \lambda_1^-(\hat{s})(n-1)h_2\hat{x}_1/D^-(\hat{s}) - \lambda_2^-(\hat{s})]\hat{x}_2\,.$$

One also has $\mathcal{H}^-(\hat{s}) = e_1\hat{x}_1 - \gamma^*D^-(\hat{s}) - \lambda_1^-(\hat{s})\hat{x}_1n$.

Notice that the time instant $\hat{s}$ depends on the trajectory considered, and thus $x(\hat{s})$ describes a switch-manifold $S(x) = 0$—the curve in Figure 3.1. Therefore, $\lambda_1^-(\hat{s})$ and $\lambda_2^-(\hat{s})$ must satisfy the system of equations below—the difference of the adjoint vectors is a normal to the manifold (see, e.g., [1]):

$$\begin{pmatrix} \lambda_1^-(\hat{s}) \\ \lambda_2^-(\hat{s}) \\ -\mathcal{H}^-(\hat{s}) \end{pmatrix} = \begin{pmatrix} \lambda_1^+(\hat{s}) \\ \lambda_2^+(\hat{s}) \\ -\mathcal{H}^+(\hat{s}) \end{pmatrix} + \kappa \begin{pmatrix} \partial S(\hat{s})/\partial x_1 \\ \partial S(\hat{s})/\partial x_2 \\ \partial S(\hat{s})/\partial s \end{pmatrix}, \tag{3.7}$$

where κ is a scalar that remains to be determined and S is any function that characterizes the manifold $\sigma^+ = 0$ in the plane (x_1, x_2). Indeed, (3.5) and (3.6) clearly show that σ^+ can be expressed as a function of x_1 and x_2 alone, and not s. Hence $\partial S/\partial s = 0$. Therefore, $\mathcal{H}^-(\hat{s}) = \mathcal{H}^+(\hat{s}) = 0$ and it yields $\lambda_1^-(\hat{s}) = [e_1\hat{x}_1 - \gamma^*D^-]/\hat{x}_1n$; thus $\lambda_1^-(\hat{s}) - \lambda_1^+(\hat{s}) = (\delta_1x_1^* - \gamma^*\alpha)/(\hat{x}_1n) \leq 0$ as $\mathcal{L}(s^*) = \delta_1x_1^* + \delta_2x_2^* - \gamma^*\alpha = 0$.

Moreover, using $\mathcal{H}^-(\hat{s}) = 0$, we can rewrite $\sigma^-(\hat{s})$ as follows:

$$\sigma^-(\hat{s}) = \left[(e_2 - \lambda_2^-(\hat{s}))(\alpha + h_1\hat{x}_1) - (e_1 - \lambda_1^-(\hat{s}))h_2\hat{x}_1\right] x_2/D^-(\hat{s})\,.$$

Using the fact that $\sigma^+(\hat{s}) = 0$ yields

$$\sigma^-(\hat{s}) = \kappa\left(\frac{\partial S(\hat{s})}{\partial x_1}h_2\hat{x}_1 - \frac{\partial S(\hat{s})}{\partial x_2}(\alpha + h_1\hat{x}_1)\right)\hat{x}_2/D^-(\hat{s}),$$

and we have

$$\sigma^-(\hat{s}) = [\lambda_1^-(\hat{s}) - \lambda_1^+(\hat{s})]\left(h_2\hat{x}_1 - (\alpha + h_1\hat{x}_1)\frac{\partial S(\hat{s})}{\partial x_2}\Big/\frac{\partial S(\hat{s})}{\partial x_1}\right)\hat{x}_2/D^-(\hat{s})\,. \tag{3.8}$$

Describing the switch manifold via an implicit function $\hat{x}_1 = \xi(\hat{x}_2)$, equation (3.8) also reads

$$\sigma^-(\hat{s}) = [\lambda_1^-(\hat{s}) - \lambda_1^+(\hat{s})] \left(h_2\hat{x}_1 + (\alpha + h_1\hat{x}_1)\frac{\mathrm{d}\xi(\hat{x}_2)}{\mathrm{d}x_2} \right) x_2/D^-(\hat{s})\,.$$

Choose $S(x_1, x_2) := n(x_1\delta_1 + x_2\delta_2)D^+\sigma^+/x_2$, and S can be expressed as follows:

(3.9)

$$\left| \begin{array}{l} S(x_1,x_2) = a(x_2)x_1^2 + b(x_2)x_1 + c(x_2) \text{ or } S(x_1,x_2) = d(x_1)x_2 + e(x_1), \text{ where} \\ \\ \left| \begin{array}{ll} a := -(n-1)\delta_1\zeta \le 0, & \left| \begin{array}{l} b := -fx_2 - g + h \text{ with} \\ f := -\delta_2 a/\delta_1 \ge 0, \\ g := \zeta\gamma^*\alpha \ge 0, \\ h := \alpha\delta_1(e_2 n - \delta_2) \ge 0. \end{array} \right. \\ c := x_2\delta_2 h/\delta_1 + \delta_2\gamma^*\alpha^2 \ge 0, & \\ d := -fx_1 + \delta_2 h/\delta_1 = \delta_2(ax_1 + h)/\delta_1, & \\ e := -ax_1^2 + (g+h)x_1 + \delta_2\gamma\alpha^2, & \end{array} \right. \end{array} \right.$$

Hence $\xi(x_2) = \left(-b - \sqrt{b^2 - 4ac}\right)/2a$ and one has

$$\frac{\mathrm{d}\xi(x_2)}{\mathrm{d}x_2} = -\frac{d}{2a\xi(x_2) + b} = \frac{\delta_2}{\delta_1}\frac{a\xi(x_2) + h}{2a\xi(x_2) + b} = -\frac{\delta_2}{2\delta_1}\left(1 + \frac{b - 2h}{\sqrt{b^2 - 4ac}}\right)\,.$$

As $b - 2h \le 0$, it yields

$$\frac{\mathrm{d}\xi(x_2)}{\mathrm{d}x_2} = -\frac{\delta_2}{2\delta_1}\left(1 - \sqrt{\frac{(b-2h)^2}{b^2 - 4ac}}\right) = -\frac{\delta_2}{2\delta_1}\left(1 - \sqrt{1 + \frac{\iota}{b^2 - 4ac}}\right) \ge 0$$

as $\iota = 4[ac - h(b-h)] = 4n\zeta{\gamma^*}^2\alpha^2\delta_1 h_2 \ge 0$. Hence $\forall x_2\,,\ \mathrm{d}\xi(x_2)/\mathrm{d}x_2 \ge 0$, which justifies the orientation of the curve in Figure 3.1. As a consequence, $\sigma^-(\hat{s}) \le 0$.

As long as u remains equal to zero while going backward in time from $\hat{s}$, one has

$$\left\{ \begin{array}{ll} \mathring{\hat{x}}_1 = -nx_1\,, & x_1(\hat{s}) =: \hat{x}_1, \\ \mathring{\hat{x}}_2 = 0\,, & x_2(\hat{s}) =: \hat{x}_2, \end{array} \right.$$

and $\sigma^- = \left[(e_2 - \lambda_2^-)(\alpha + h_1x_1) - (e_1 - \lambda_1^-)h_2x_1\right] x_2/D^-$, with

$$\left\{ \begin{array}{ll} \mathring{\lambda}_1^- = n\lambda_1^- - \delta_1\,, & \lambda_1^-(\hat{s}) = [e_1\hat{x}_1 - \gamma^* D^-(\hat{s})]/\hat{x}_1 n, \\ \mathring{\lambda}_2^- = 0\,, & \lambda_2^-(\hat{s}) = \cdot \end{array} \right.$$

Thus, still going backward in time from $\hat{s}$ with $u = v = 0$, one has

$$\left\{ \begin{array}{l} x_1(s) = \hat{x}_1 \mathrm{e}^{n(\hat{s}-s)}, \\ x_2(s) = \hat{x}_2 \end{array} \right. \quad \text{and} \quad \left\{ \begin{array}{l} \lambda_1^-(s) = (e_1x_1 - \gamma^* D^-)/x_1 n, \\ \lambda_2^-(s) = \lambda_2^-(\hat{s}). \end{array} \right.$$

Introducing $y(x_1) := e_1h_2(n-1)x_1/[n(\alpha + h_1x_1)]$ yields $\sigma^- - \sigma^-(\hat{s}) = y(\hat{x}_1) - y(x_1)$. It is easy to see that $y(x_1)$ is increasing. Thus $\forall s \in [0, \hat{s}]\,,\ \sigma^- \le 0$. Hence there is at most one switch-point.

Finally, it is also necessary to check that if the focal forager does not switch to the generalist strategy upon reaching the switch manifold the state nevertheless crosses the said manifold and enters the region where the optimal behavior for all players is to be opportunistic. This is the so-called "permeability condition" [1].[7]

[7]Yet, this is a nonzero sum game, and one cannot conclude as in [1] that the adjoint variables are continuous.

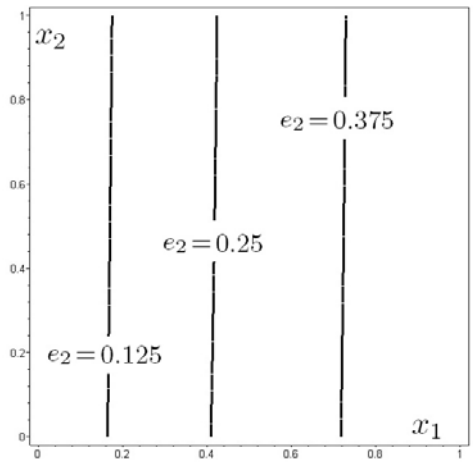


FIG. 3.2. *The switch-manifolds associated with several values of e_2 in the state-space (x_1, x_2). We took $n = 2$, $e_1 = 1$, $\alpha = 1$, $h_1 = h_2 = 1$, $\gamma^* = 0.1$, and from the left to the right, $e_2 = \{0.125, 0.25, 0.375\}$.*

To that aim, let $\nu = (-1, \mathrm{d}\xi/\mathrm{d}x_2)$ be a normal vector of the switch-manifold which points in the same direction as the outgoing trajectories. We have

$$\langle \nu, f(x, 1, 1) \rangle = \left(x_1 - x_2 \frac{\mathrm{d}\xi}{\mathrm{d}x_2} \right) > 0 \,.$$

We calculate

$$\langle \nu, f(x, 0, 1) \rangle = x_1 + (n-1) \frac{\alpha + h_1 x_1}{\alpha + h_1 x_1 + h_2 x_2} \left(x_1 - x_2 \frac{\mathrm{d}\xi}{\mathrm{d}x_2} \right) ,$$

which, taking the previous inequality into account, is clearly positive. Hence the permeability condition is satisfied.

Therefore, the evolutionarily stable strategy is indeed either opportunistic during the whole time spent on the patch, or it is selective first and, after a while, is opportunistic. □

Our aim now is to characterize this possible switch-point. Clearly, $\hat{x}_2 = 1 - x_1^0$ and thus $\hat{x}_1 = \xi(1 - x_1^0)$.

Figure 3.2 shows the switch-manifolds associated with several values of e_2 in the state-space (x_1, x_2). Interestingly, we see that the threshold $\hat{x}_1$ is almost independent from x_2. In other words, the curve in Figure 3.1 seems to be qualitatively very close to a straight line of constant x_1.

Figure 3.3 shows the mapping $n \mapsto \hat{x}_1$. Interestingly, the greater the number of foragers on the patch, the closer the evolutionarily stable strategy becomes to the intake-rate maximization.

3.1. Partial conclusion. Our results are in agreement with those of [26], obtained via a similar approach, although the author of [26] ignored the discontinuities on the adjoint variables; see (3.7). Our innovation lies in the fact that we do not consider any arbitrarily predetermined residence time or final patch state. It allows us to analyze the sensitivity of the switch-point to the initial conditions, and our

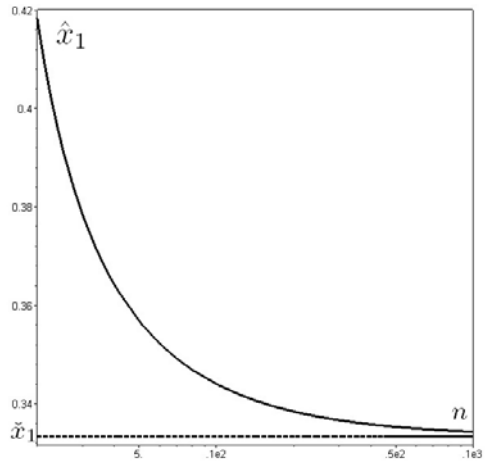


FIG. 3.3. *The mapping $n \mapsto \hat{x}_1$. We took $e_2 = 0.25$, $x_1^0 = 0.5$, $n \in [2, 100]$ and left the others parameters unchanged. The horizontal axis has a logarithmic scale. The dashed line represent the threshold $\check{x}_1$ that corresponds to intake-rate maximization; indeed, intake-rate maximization is selective, while x_1 remains larger than a given threshold $\check{x}_1 := e_2\alpha/\zeta$, independent of x_2.*

model reveals[8] that it seems almost independent from them. Qualitatively, the evolutionarily stable strategy is then close to intake-rate maximization, a policy that is selective until the best resource is depleted down to an optimal threshold—whatever the abundance of the less profitable resource. However, the intake-rate maximization threshold remains a lower bound; for instance, the larger the number of foragers on the patch, the closer the evolutionarily stable strategy to intake-rate maximization. Moreover, these results are also in agreement with those of [15, 35] obtained by quite different approaches.

As the diet selection policy of an isolated—not subject to competition—forager is really different from the evolutionarily stable strategy relevant in a situation of actual competition, the question that arises then is, What should a lone forager entering a patch do if the probability of facing a situation of competition is nonzero?

4. An asynchronous but deterministic foraging game. As a preliminary approach, this section focuses on an asynchronous two-forager game, where the arrival time t_a of the second one is assumed deterministic—this might be relevant in a case of group foraging with "information sharing" [7], assuming that the first forager on a patch has some time to take advantage of its discovery.

Once the second forager arrives, the evolutionarily stable policy depends only on the current patch-state x and is detailed in section 3. It thus remains to determine the optimal strategy before the intruder's arrival.

We claim the following result (see Figure 4.1).

THEOREM 4.1. *In the deterministic arrival time problem, the optimal strategy of the first forager before the arrival of the second one is as follows:*

- *Any admissible policy that leads to $S(x) = 0$ at $t = t_a$, provided that it is feasible; if so, there exists an infinite number of optimal trajectories.*

[8] Compared to [26], as the author of [35] also observed that the switch-point is "nearly independent of... the ratio of the prey types... initially present on the patch."

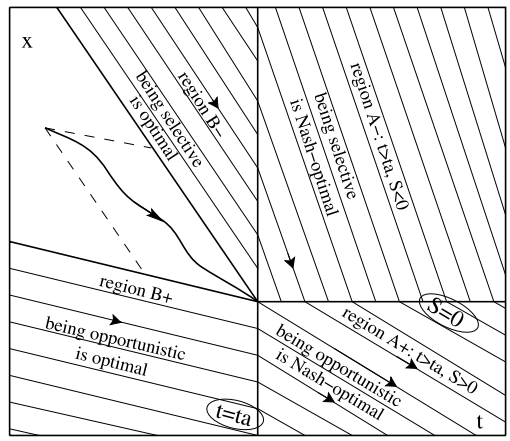


FIG. 4.1. *This graph represents the optimal fields of trajectories in the state-space* (t, x). *The regions* B *(before the opponent's arrival) and* A *(after) correspond, respectively, to* $t < t_a$ *and* $t \geq t_a$, *or equivalently, to* $n = 1$ *and* $n = 2$. *The signs* $-$ *and* $+$ *denote, respectively, that being strictly selective or strictly opportunistic is the (Nash) optimal policy in the region considered. In the latter case, optimal trajectories are, symbolically, linearly plotted. The vertical line represents the manifold* $t = t_a$. *The horizontal one in region* A *represents the switch-manifold given by* $S(x) = 0$. *The curvilinear trajectory starting in the region (not cross-hatched by "linear" trajectories) reflects the fact that any trajectory that remains in this region is optimal.*

- *Otherwise, take* $u = 0$, *respectively,* $u = 1$, *all along the trajectory if this leads to* $S(x) < 0$, *respectively,* $S(x) > 0$, *at time* $t = t_a$.

Proof. As we now need to consider the variable t explicitly, we let it be a state variable; thus the dynamics are extended as follows:

$$\begin{cases} \mathring{x}_1 = -x_1[1 + (n-1)D(u)/D(v)], & x_1(0) = x_1^0, \\ \mathring{x}_2 = -x_2[u + v(n-1)D(u)/D(v)], & x_2(0) = x_2^0 = 1 - x_1^0, \\ \mathring{t} = D(u), & t(0) = 0. \end{cases}$$

From now on, we shall refer to Figure 4.1 to support our reasoning; we stress that this is a symbolic sketch. In region A, the Nash-optimal fields of trajectories are perfectly known, thanks to section 3; i.e., the evolutionarily stable strategy depends only on the sign of $S(x)$.

It remains to determine the optimal fields of trajectories in region B.

For ease of notation, we let μ and H be, respectively, the adjoint vector and the Hamiltonian associated with the trajectories evolving in region B—the part of the game during which the forager is still alone. Connecting μ to λ is a matter of transversality conditions relative to the manifold $t = t_a$—or possibly only to its intersection with the switch-manifold given by $S(x) = 0$. The plane $t = t_a$, parallel to the x subspace, is consequently transparent for this patch-state variable. Thus the only possible discontinuity concerning the adjoint vector is on the costate variables associated with t, say, μ_3 and λ_3—except on the intersection of the two manifolds. Thus, apart from this particular 1-D curve, we have the following relation:

$$\begin{pmatrix} \mu_1(s_a) \\ \mu_2(s_a) \\ \mu_3(s_a) \\ -\mathsf{H}(s_a) \end{pmatrix} = \begin{pmatrix} \lambda_1(s_a) \\ \lambda_2(s_a) \\ \lambda_3(s_a) = 0 \\ -\mathcal{H}(s_a) = 0 \end{pmatrix} + \nu \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix},$$

where ν is a scalar that remains to be determined and s_a is such that $t(s_a) = t_a$. To be exhaustive, we show that the transversality condition associated with the curve given by the intersection of the two manifolds is the following, although we will not need it:

$$\begin{pmatrix} \mu_1(s_a) \\ \mu_2(s_a) \\ \mu_3(s_a) \\ -\mathsf{H}(s_a) \end{pmatrix} = \begin{pmatrix} \lambda_1(s_a) \\ \lambda_2(s_a) \\ \lambda_3(s_a) \\ -\mathcal{H}(s_a) \end{pmatrix} + \tilde{\kappa} \begin{pmatrix} \partial S(s_a)/\partial x_1 \\ \partial S(s_a)/\partial x_2 \\ \partial S(s_a)/\partial t = 0 \\ \partial S(s_a)/\partial s = 0 \end{pmatrix} + \tilde{\nu} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

As, in region B, $n = 1$, one has $\mathsf{H} = \mathcal{L} - \mu_1 x_1 - \mu_2 u x_2 + \mu_3 D(u)$.

Let $\varsigma = \partial \mathsf{H}/\partial u = (\delta_2 - \mu_2 + h_2\mu_3)x_2$. The fact that $\mathsf{H}(s_a) = \mathcal{H}(s_a) = 0$ yields

$$\mu_3(s_a) = \nu = -\frac{\mu_1 x_1(s_a) + \mu_2 u x_2(s_a)}{\alpha + h_1 x_1(s_a) + h_2 x_2(s_a)}.$$

Thus $\varsigma(s_a) = \sigma(s_a)$: the discontinuity on the costate variable associated with t precisely maintains the continuity of the switch-function. Moreover, one has

$$\begin{cases} \mathring{\mu}_1 = -\partial \mathsf{H}/\partial x_1 = \delta_1 - \mu_1 + h_1\mu_3\,, & \mu_1(s_a) = \lambda_1(s_a), \\ \mathring{\mu}_2 = -\partial \mathsf{H}/\partial x_2 = -u(\delta_2 - \mu_2 + h_2\mu_3)\,, & \mu_2(s_a) = \lambda_2(s_a), \\ \mathring{\mu}_3 = -\partial \mathsf{H}/\partial t = 0\,, & \mu_3(s_a) = \nu, \end{cases}$$

and it yields $\mathring{\varsigma} = 0$. We now investigate the possible geometry of the trajectory fields, referring the reader to the four sketches of Figure A.2. As it is clear that being opportunistic does not deplete the best resource as much as being selective during the same time, the fourth quadrant represents an impossible scenario. If being selective until the intruder's arrival yields $S(x(s_a)) < 0$ (second quadrant), then this is optimal. In a similar fashion, if being opportunistic yields $S(x(s_a)) > 0$ (first quadrant), then this is optimal. Otherwise (third quadrant), the optimal policy is such that $S(x(s_a)) = 0$. Moreover, as $\mathrm{d}\xi/\mathrm{d}x_2 \geq 0$, the scenario of the third quadrant, now considering the dashed line as the switch-manifold, cannot happen. Hence there is no state prior to t_a through which trajectories of the two extremal fields can pass. On the contrary, there is indeed a gap—a region uncovered by our extremal fields—between regions + and − in region B of Figure 4.1.

Appendix A provides a relation giving the time spent to move from a point of the state-space (x_1, x_2) to another—as it does not depend on the trajectory followed—via the same dynamics if taken with $n = 1$. It thus yields the locus of all points attainable in a time t_a from a given point (x_1^0, x_2^0)—the manifold represented by the dot-dashed line in the third quadrant of Figure 4.2 contains the said locus. More precisely, this manifold corresponds to the application

$$x_1 \mapsto x_2^0 - [t_a - \alpha \ln(x_1^0/x_1) - (x_1^0 - x_1)h_1]/h_2\,,$$

clearly monotonously decreasing. Hence for each initial condition, there is a unique point $(\hat{x}_1, \hat{x}_2)$ such that $S(\hat{x}_1, \hat{x}_2) = 0$ at time t_a—this is the intersection of the two manifolds. Therefore, any trajectory that remains in the gap has to reach this unique point, and therefore yields the same overall payoff. Further, since optimal trajectories cannot penetrate any of the two extremal fields, any trajectory remaining in the gap and reaching the switch-manifold at $t = t_a$ is optimal. □

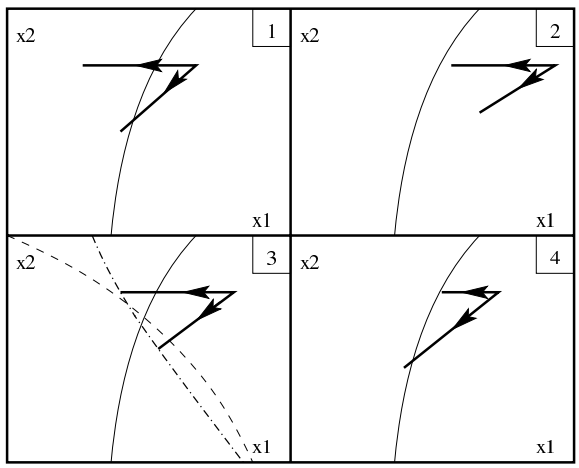


FIG. 4.2. *Each quadrant represents a state-space* (x_1, x_2). *The thick arrows are possible trajectories. The solid curves represent the switch-manifold* $S(x) = 0$. *The time horizon is the same for both trajectories plotted: "being selective" and "being opportunistic."*

5. An asynchronous stochastic foraging game. It remains to show how to forage optimally under the risk of competition, i.e., if the intruder's arrival is no longer deterministic.

As in the last section, once the possible intruder arrives, the optimal policy depends only on the current patch-state x and is as detailed in section 3. Therefore, the optimal total future reward $V_2(x)$ is known. It thus remains to determine the optimal strategy before a possible intruder's arrival.

Notice that taking $V_2 = 0$ addresses the question of the optimal diet selection when the end time is random; for instance, the sudden arrival of a predator could cause the forager to flee from the patch.

Let this possible perturbation be a Poisson variable of intensity π.

We still let the time have a cost—in terms of missed opportunities—of γ^* per unit, and thus the forager is nevertheless incited to leave the patch in order to avoid wasting its time.

Let t^* be the time the forager would remain on the patch if not interrupted and ϵ be the random event time, exponentially distributed with mean $1/\pi$.

Our dynamics are

$$\left\{ \begin{array}{ll} q\dot{x}_1 = -x_1/D(u)\,, & x_1(0) = x_1^0, \\ q\dot{x}_2 = -ux_2/D(u)\,, & x_2(0) = x_2^0. \end{array} \right. \tag{5.1}$$

Let our criterion be

$$G = \mathbb{E}\, J \quad \text{with} \quad J = \int_0^{\epsilon \wedge t^*} L(x,u)\mathrm{d}t + \left\{ \begin{array}{l} V_2(x(\epsilon \wedge t^*)) \text{ if } \epsilon < t^*, \\ 0 \text{ otherwise,} \end{array} \right. \tag{5.2}$$

and $L(x,u) = e_1 x_1/D(u) + u e_2 x_2/D(u) - \gamma^*$.

We claim the following result.

THEOREM 5.1. *The optimal policy in the stochastic arrival time problem stated by* (5.1) *and* (5.2) *is bang-bang with a single switch from* $u = 0$ *to* $u = 1$ *occurring before reaching the manifold* $S(x) = 0$. *The leaving policy is unchanged from Theorem* 3.1.

Proof. Using the fact that $P(\epsilon > t^*) = \mathrm{e}^{-\pi t^*}$, we have

$$\begin{aligned} G &= \mathbb{E}_{\epsilon<t^*}\left[\int_0^{\epsilon} L(x,u)\mathrm{d}t + V_2(x)\right] + \mathrm{e}^{-\pi t^*}\int_0^{t^*} L(x,u)\mathrm{d}t \\ &= \int_0^{t^*}\left[\left(\int_0^{\epsilon} L(x,u)\mathrm{d}t\right) + V_2(x)\right]\pi\mathrm{e}^{-\pi\epsilon}\mathrm{d}\epsilon + \mathrm{e}^{-\pi t^*}\int_0^{t^*} L(x,u)\mathrm{d}t\,, \\ &= \int_0^{t^*} L(x,u)\left(\int_t^{t^*}\pi\mathrm{e}^{-\pi\epsilon}\mathrm{d}\epsilon\right)\mathrm{d}t + \int_0^{t^*} V_2(x)\pi\mathrm{e}^{-\pi\epsilon}\mathrm{d}\epsilon + \mathrm{e}^{-\pi t^*}\int_0^{t^*} L(x,u)\mathrm{d}t\,, \\ &= \int_0^{t^*}\big[L(x,u) + \pi V_2(x)\big]\mathrm{e}^{-\pi t}\mathrm{d}t\,. \end{aligned}$$

Having in mind these equivalent dynamics $f(x,u)$,

$$\begin{cases} \mathring{x}_1 = -x_1\,, & x_1(0) = x_1^0, \\ \mathring{x}_2 = -ux_2\,, & x_2(0) = x_2^0, \end{cases}$$

our criterion can also be expressed as follows:

$$G = q\int_0^{s^*}\mathcal{L}(x,u)\mathrm{e}^{-\pi t(s)}\mathrm{d}s\,, \quad \mathcal{L}(x,u) = \delta_1 x_1 + u\delta_2 x_2 - \gamma^*\alpha + \pi V_2(x)D(u)\,,$$

with $t(s) := \alpha s + h_1(x_1^0 - x_1) + h_2(x_2^0 - x_2)$—see Appendix A—and $t(s^*) := t^*$.

However, let us consider an equivalent criterion:

$$\mathcal{G} = \mathrm{e}^{\pi\left(h_1x_1^0 + h_2x_2^0\right)}G/q = \int_0^{s^*}\mathcal{L}(x,u)\mathrm{e}^{-\pi(\alpha s - h_1x_1 - h_2x_2)}\mathrm{d}s\,.$$

Our stochastic end time optimization problem is thus equivalent to the above deterministic one. Formulated in such a fashion, it directly yields that the end time is such that $\mathcal{L}$ is zero. As V_2 is zero beyond the manifold $\delta_1x_1 + \delta_2x_2 - \gamma^*\alpha = 0$, the latter is the terminal manifold corresponding to this problem. Therefore, we introduce the *value function* V, or the optimal total future reward, which is the solution of the following Hamilton–Jacobi–Bellman equation:

$$\begin{cases} \forall x\,,\ V(x,s^*) = 0 \quad \text{and} \quad \forall(x, s < s^*)\,, \\ -\partial V(x,s)/\partial s = \max_u\left[\langle\nabla_x V(x,s), f(x,u)\rangle + \mathcal{L}(x,u)\mathrm{e}^{-\pi(\alpha s - h_1x_1 - h_2x_2)}\right]. \end{cases}$$

Let $V(x,s) =: \mathrm{e}^{-\pi(\alpha s - h_1x_1 - h_2x_2)}\mathcal{V}(x)\ \forall i \in \{1,2\}$, $\mu_i := \partial\mathcal{V}/\partial x_i$, and $\mathsf{V}(x) := \mathcal{V}(x) - V_2(x)$; V is thus the solution of the following stationary Hamilton–Jacobi–Bellman equation:

$$\begin{cases} \forall(x|\mathcal{L}\le 0)\,,\ \mathsf{V}(x) = 0, \quad \text{and} \quad \forall(x|\mathcal{L} > 0)\,, \\ \alpha\pi\mathsf{V}(x) = \max_u\left[x_1(\delta_1 - \pi h_1\mathsf{V}(x) - \mu_1) + ux_2(\delta_2 - \pi h_2\mathsf{V}(x) - \mu_2) - \gamma^*\alpha\right]. \end{cases}$$

Let us notice that $\mathcal{V}(x)$ is indeed the optimal value of our payoff $\mathcal{G}$. Hence $\mathsf{V}(x)$ is nonnegative.

As it is clear that the optimal control is bang-bang, let us introduce the switch-function $\sigma = \delta_2 - \pi h_2\mathsf{V}(x) - \mu_2$. We conjecture that there is at most one switch. Let the superscript + denote the region of the state-space beyond the switch-point, where we postulate that the optimal strategy is $u = 1$. For instance, let $D^+ := D(1) = \alpha + h_1x_1 + h_2x_2$.

We have $\forall i \in \{1,2\}\,,\ \mu_i^+(s^*) = 0$. It yields $\sigma^+(s^*) = \delta_2 \geq 0$ by hypothesis. Thus at the end time, the optimal policy is to be opportunistic, i.e., to take both resources.

The Hamilton–Jacobi–Bellman equation states that $\forall (x|\mathcal{L} > 0)$ in region $+$,

$$\pi D^+ \mathsf{V}^+(x) = x_1(\delta_1 - \mu_1^+) + x_2(\delta_2 - \mu_2^+) - \gamma^*\alpha\,.$$

Let $\forall i \in \{1,2\}\,,\ \nu_i := \partial \mathsf{V}/\partial x_i = \mu_i - \lambda_i$, as $\lambda_i = \partial V_2/\partial x_i$.

According to the classical theory of characteristics [4], $\forall i \in \{1,2\}$,

$$\mathring{\mu}_i^+ = \pi D^+ \nu_i^+ - (\delta_i - \pi h_i \mathsf{V}^+(x) - \mu_i^+)\,, \quad \mu_i^+(\hat{s}) = \hat{\mu}_i\,.$$

We thus have $\mathring{\sigma}^+ = \pi h_2(x_1\nu_1^+ + x_2\nu_2^+) - \mathring{\mu}_2^+ = \pi h_2 x_1 \nu_1^+ - \pi D^- \nu_2^+ + \sigma^+$.

Our aim is now to show that, if σ^+ becomes zero while going backward in time, switching u to zero remains optimal down to the initial time. Let the superscript $-$ denote the region of the state-space where we conjecture that the optimal strategy is $u = 0$. For instance, let $D^- := D(0) = \alpha + h_1 x_1$.

In the latter region, the Hamilton–Jacobi–Bellman equation states that $\forall x$,

$$\pi D^- \mathsf{V}^-(x) = x_1(\delta_1 - \mu_1^-) - \gamma^*\alpha\,. \tag{5.3}$$

Via a similar calculation of the characteristics:

$$\begin{cases} \mathring{\mu}_1^- = \pi D^- \nu_1^- - (\delta_1 - \pi h_1 \mathsf{V}^-(x) - \mu_1)\,, & \mu_1^-(\hat{s}) = \hat{\mu}_1, \\ \mathring{\mu}_2^- = \pi D^- \nu_2^-\,, & \mu_2^-(\hat{s}) = \hat{\mu}_2, \end{cases}$$

where $\hat{s}$ is the time at which the switch-manifold is reached.

We thus have $\mathring{\sigma}^- := \pi h_2 x_1 \nu_1^- - \mathring{\mu}_2^- = \pi h_2 x_1 \nu_1^- - \pi D^- \nu_2^-$. Hence on the switch-manifold, $\mathring{\sigma}^-(\hat{s}) = \mathring{\sigma}^+(\hat{s}) \geq 0$.

Using (5.3), we have

$$\begin{aligned} \mathring{\mu}_1 &= \pi D^- \nu_1^- - (\alpha\pi \mathsf{V}^-(x) + \gamma^*\alpha)/x_1 \\ &= \pi D^- \mu_1^- - \left[(\alpha\pi \mathsf{V}^-(x) + \gamma^*\alpha)/x_1 + \pi D^- \lambda_1\right]\,. \end{aligned}$$

Let $\Theta(x) := [(\alpha\pi \mathsf{V}^-(x) + \gamma^*\alpha)/x_1 + \pi D^- \lambda_1] \geq 0$,

$$\chi(s,\hat{s}) := \exp\left(\pi \int_{\hat{s}}^{s} D^- \mathrm{d}\ell\right) = \mathrm{e}^{\pi[t(s) - t(\hat{s})]}\,,$$

and

$$\phi(s,\hat{s}) := \alpha\pi \int_s^{\hat{s}} \Theta(x)\chi(s,\ell)\mathrm{d}\ell \geq 0\,,$$

as it is clear from section 3 that $\forall x,\ \lambda_1(x) \geq 0$.

We have $\forall s \in [0,\hat{s}]$,

$$\begin{cases} \mu_1^-(s) = \hat{\mu}_1\chi(s,\hat{s}) + \phi(s,\hat{s}), \\ \mu_2^-(s) = \hat{\mu}_2\chi(s,\hat{s}). \end{cases}$$

Let $\psi(x_1) := \pi h_2 x_1 \hat{\mu}_1^- - \pi D^- \hat{\mu}_2^- = \pi(h_2\hat{\mu}_1 - h_1\hat{\mu}_2)x_1 - \pi\alpha\hat{\mu}_2$, which yields

$$\mathring{\sigma}^- = \psi(x_1)\chi(s,\hat{s}) + \phi(s,\hat{s})\pi h_2 x_1\,.$$

As $\psi(\hat{x}_1) = \mathring{\sigma}^-(\hat{s}) \geq 0$, $(h_2\hat{\mu}_1 - h_1\hat{\mu}_2)$ is clearly positive. Thus $\psi(x_1)$ is increasing in x_1. Therefore $\mathring{\sigma}^-$ remains positive in region $-$. As an expected consequence, the trajectory generated by taking $u = 0$ backward from the switch-manifold implies that σ^- remains negative down to the initial time. Hence the optimal strategy is indeed at most one-switch bang-bang. □

5.1. A digression on the random end time problem. In this subsection, our aim is to numerically characterize the switch-manifold for a random end time problem, i.e., taking $\forall x\,,\ V_2(x) = 0$.

Integrating forward the trajectory field where $u = 1$ yields

$$\mathcal{V}^+(x) = \Delta B_1(\beta) - \gamma^* \alpha B_0(\beta)\,,$$

with $\Delta := \delta_1 x_1 + \delta_2 x_2$, $\beta := \pi(h_1 x_1 + h_2 x_2)$, and the function B_n is defined as follows, with $n \in \mathbb{N}$:

$$B_n(\beta) := \int_0^{s^*} \mathrm{e}^{-(n+\pi\alpha)s - \beta(1-\mathrm{e}^{-s})} \mathrm{d}s\,.$$

Integrating by parts easily yields

$$\left| \begin{array}{l} B_1(\beta) = (1 - z_0(\beta) - \pi\alpha B_0(\beta))/b, \\ B_2(\beta) = (1 - z_1(\beta) - (1+\pi\alpha)B_1(\beta))/\beta \end{array} \right.$$

with $z_n(\beta) := \mathrm{e}^{-(n+\pi\alpha)s^* - \beta(1-\mathrm{e}^{-s^*})}$.

Using either the explicit form of s^*[9] or the remark that it maximizes $\mathcal{G}$ and the envelope lemma, we have that

$$\forall i \in \{1,2\}\,, \quad \mu_i^+ = \delta_i B_1(\beta) - \Delta\pi h_i (B_1(\beta) - B_2(\beta)) + \gamma^* \alpha \pi h_i (B_0(\beta) - B_1(\beta))\,.$$

As $\sigma^+ := \delta_2 - \pi h_2 \mathcal{V}^+ - \mu_2^+$, we have that

$$\begin{aligned} \sigma^+ &= \delta_2 - \delta_2 B_1(\beta) - \Delta\pi h_2 B_2(\beta) + \gamma^* \alpha \pi h_2 B_1(\beta) \\ &= (\delta_2 - \Delta\pi h_2 (1 - z_1(\beta))/\beta) - (\delta_2 - (1+\pi\alpha)\Delta\pi h_2/\beta - \gamma^* \alpha \pi h_2) B_1(\beta)\,. \end{aligned}$$

The switch-manifold is thus given by $\sigma^+(x) = 0$. Figure 5.1 shows the switch-manifolds associated with various values of π.

5.2. Implications for the original problem. It is clear that the switch-manifold corresponding to the original problem is bounded by that of the synchronous foraging game characterized in section 3, as it corresponds to being disturbed by a conspecific with a probability of one. Besides the latter point, it is likely that qualitatively, the optimal policy remains equivalent to the random end time problem, i.e., switching at a given x_1, depending on the intensity of the Poisson process.

6. Conclusion. Our aim was to determine the evolutionarily stable strategy [20] that foragers competing for a plural and depleting resource should adopt, both in terms of diet selection and patch-leaving decision [3, 33].

First, we reformulated the optimal diet selection policy [23] for a lone forager, in a similar fashion to [26], except that we allow for a free patch-leaving time. On our way, we solved the optimal diet selection problem for a single forager with an end time either fixed or possibly random.

Next, we investigated the foraging game involving several foragers arriving simultaneously at a patch containing two distinct types of resources. The resulting differential game involves discontinuous state feedback strategies constructed via a classical synthesis technique, and hence requires for its solution a careful analysis of

[9] As s^* is the time to leave the patch as a function of the current state, we have $s^* = -\ln(\gamma^*\alpha/\Delta)$; see (3.6) and the dynamics taken with $u = 1$.

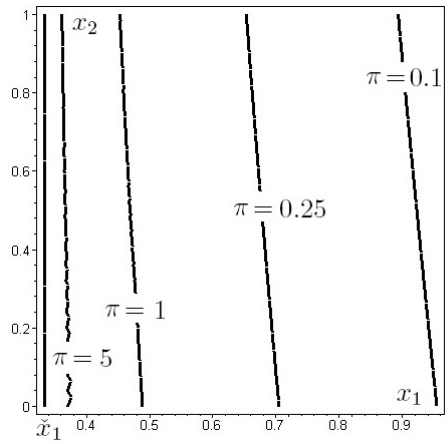


FIG. 5.1. *The switch-manifolds in the state-space* (x_1, x_2) *associated with several values of* π*; from right to left,* $\pi = \{0.1, 0.25, 1, 5\}$*, with* $\alpha = 1$*,* $h_1 = 1$*,* $h_2 = 1$*,* $e_1 = 1$*,* $e_2 = 0.25$*, and* $\gamma^* = 0.1$*. The left bound corresponds to the intake rate maximization switch-manifold, i.e.,* $\check{x}_1 = e_2\alpha/\zeta$*.*

the induced discontinuities of the adjoint variables. The end result is a one-switch bang-bang evolutionarily stable strategy. This is in agreement with [15, 26, 35] and is more precise in several respects.

As there is a qualitative gap between the optimal behavior of an isolated forager and that of competing foragers, the question which thus arose was that of the optimal strategy of a single forager subject to a potentially imminent competition.

As a preliminary approach, we solved an asynchronous two-forager game, where the interarrival time was assumed deterministic. Partial preferences arose in several fashions.

Finally, we no longer considered a deterministic interarrival time but let the probability that an opponent enters the game follow a Poisson distribution. We showed that the optimal policy belongs to a qualitative continuum which fills[10] the gap that separates the two extremal policies found previously.

Thus, although the classical diet selection policy states that a lone forager should take both resources indiscriminately during the whole time spent on the patch (see section 2), we showed that it suffices to add some stochasticity in the model to predict a qualitatively different behavior. Indeed, under the risk of viewing a predator (or a conspecific) shortening its time spent (alone) on the patch, a lone forager should be selective for a while, at least if the probability of being disturbed is nonnegligible.

7. Discussion and prospects. Our results are based on the assumption that foragers are identical in terms of their ability to find and consume resources and of the relative values they attribute to resources with respect to both other resources and their environment. The game is symmetric in this sense. In recent years, however, foraging theory has looked more at the effect of the foragers' state; see [16]. Thus arises the question of the robustness of our results when considering relevant differences in forager state such as

- competitive ability, which may be correlated to the size of the animal [25, 24];
- level of satiation or body reserves and their effect on the relative values of food

[10] By playing on the intensity of the Poisson variable.

resources compared to other resources or opportunities in the environment, such as finding a mate;
- time away from the nest and its effect on the "cost of the time";
- life expectancy of the animal [36], i.e., a time horizon.

In this sense the game would no longer be symmetric. Moreover, the question of the information on the opponent state would also arise. Yet, we conjecture that, of course, the foragers would probably not switch or leave simultaneously, but a qualitatively similar behavior would persist; i.e., switching from being selective to opportunistic and leaving according to their own Charnov's rule.

Also, in our model, foragers' ability to gather resources is not affected by the presence of conspecifics; the present paper ignores interference (i.e., contacts or fights) that could occur among them. As in the single resource case, as long as there is no interference, the evolutionarily stable strategy is pure, in particular in terms of the patch-leaving policy. We conjecture that including interference in the model would result in a war of attrition, or random patch-leaving times, but would not qualitatively affect the resource acceptance policy. More accurately, we conjecture that a war of attrition occurs after foragers have switched from a selective diet to an opportunistic one. Yet, this makes sense to us as long as interference intensity does not depend on the resource acceptance policy. If it does, i.e., if interference is greater when both foragers focus on the best quality resource, the question is open (see [25, 24] for experimental evidence of a competition avoidance behavior).

Last, but not least, the question arises of the relevance of this model with respect to real life. A field study in the Negev Desert, Israel [12] was conducted on Nubian ibex *Capra ibex nubiana*, wild social goats that actually compete for resources. Interestingly, an indirect observation based on "giving-up densities" [2] tends to show that Nubian ibex "forage selectively on plants of higher quality until a certain threshold density, switching later to a more opportunistic foraging." Also, such an interpretation may hold with respect to similar observations made on kangaroo rats *Dipodomys merriami* foraging on the same patch during a study done in Arizona [3]. As pointed out by [11], diet selection dynamics are rarely directly observed.[11] Nevertheless, the authors of [27] observed, through laboratory experiments with the cichlid fish *Haplochromis piceatus* (a predator accustomed to foraging in group), a switch in their resource acceptance policy, regardless of whether they were foraging alone or by pair. However, the switch-point occurred at a higher density of the preferred resource when foraging by pair than when foraging alone [35]. In light of the present model, an inverse ranking of the switch-points would have been expected; i.e., if uncertainty with respect to the time available to exploit resources—possibly before the expected arrival of a competitor—makes a single forager focus on the best resources first, such a selectivity is expected to be exacerbated, or at least unaffected, under competition. Our simple model is therefore falsifiable and seems to be so in this species. Moreover, white king pigeons, which also forage in a group under natural conditions, have been shown [29, 28] to be more "choosy" alone than in the presence of a competitor. More accurately, the authors actually observed that pigeons switch "earlier" under competition. It may be that interference occurring when focusing on the preferred resource qualitatively changes the resource acceptance policy. Further theoretical investigations are thus needed to better understand how competition affects the dynamics of diet selection.

[11] Yet, this article refers to other (of a physiological nature) dynamics, ignoring resource depletion and competition.

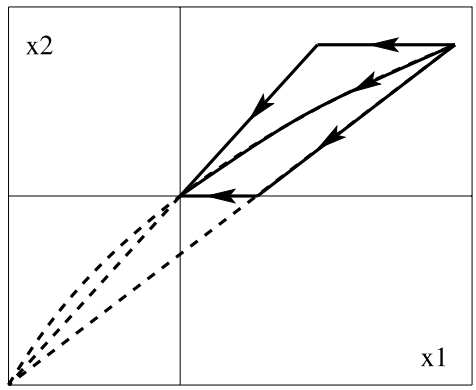


FIG. A.1. *Three of the optimal trajectories in the state-space* (x_1, x_2)*: "accepting the less profitable resource with a constant optimal acceptance rate"; "switching from selective to opportunistic"; and vice versa. The intersection of the vertical and horizontal lines represents the point* $(\hat{x}_1, \hat{x}_2)$*.*

Appendix A. Optimal diet selection with a fixed end time. In the body of the paper, we allow the forager a free patch-leaving time, given that the time itself has a cost of γ^* per unit. This appendix is a digression on the optimal diet selection problem under a fixed residence time constraint, as already addressed by some authors; see [26, 19], where the example of an intertidal forager is mentioned. Our patch dynamics are closer to that of [26],[12] whose author argues that there is a partial preference region in the state-space; our aim is to prove this statement.

Although the question addressed is not the same, our basic model remains that detailed in the body of the paper. As we consider a lone forager in an environment that offers two types of resources, the notations we shall use are those introduced in section 3. For instance, we assume that resource 1 is more profitable than resource 2, i.e., $e_1/h_1 \geq e_2/h_2$.

We claim the following result (see Figure A.1).

THEOREM A.1. *In the fixed end time problem, the optimal strategy is as follows: Let* $\hat{x}_1 = e_2\alpha/\zeta$*.*

- *Any admissible strategy that leads to* $x_1(T) = \hat{x}_1$*, provided that it is feasible. If so, there exists an infinite number of optimal trajectories, all of them reaching the same point* $(\hat{x}_1, \hat{x}_2)$*;* $\hat{x}_2$ *is given by* (A.1)*.*
- *Otherwise, take* $u = 0$*, respectively,* $u = 1$*, all along the trajectory if this leads to* $x_1(T) > e_2\alpha/\zeta$*, respectively,* $x_1(T) < e_2\alpha/\zeta$*.*

Proof. As we need to consider the variable t explicitly, we let it be a state variable. We thus have the following dynamics $f(x, u)$:

$$\begin{cases} \mathring{x}_1 = -x_1\,, & x_1(0) = x_1^0, \\ \mathring{x}_2 = -ux_2\,, & x_2(0) = x_2^0 = 1 - x_1^0, \\ \mathring{t} = D(u)\,, & t(0) = 0. \end{cases}$$

Our criterion is

$$\mathcal{J} = \mathcal{K}(x(T)), \quad \text{where} \quad \mathcal{K}(x) := e_1(x_1^0 - x_1) + e_2(x_2^0 - x_2)$$

[12]The dynamics of [19] are, then, really stochastic, as they allow for "a run of back luck" that leads the forager to "become more selective as the time left in the patch runs out."

and T is a fixed final time.

Let S be the final s, i.e., $t(S) = T$. Let λ be the adjoint vector. It yields the Hamiltonian

$$\mathcal{H} = \langle \lambda, f(x,u) \rangle = -\lambda_1 x_1 - \lambda_2 u x_2 + \lambda_3 D(u)\,.$$

According to Pontryagin's maximum principle, if a policy $u^*(s)$ generating a trajectory $x^*(s)$ is optimal, then there exists an adjoint trajectory $\lambda(s)$ such that

$$\left\{ \begin{array}{l} \mathring{\lambda} = -\nabla_x \mathcal{H}(\lambda, u^*, x^*), \\ \lambda(s^*) = \nabla_x \mathcal{K}(x^*) + \mathsf{v}, \\ \mathcal{H}(s^*) = 0 \\ \left| \begin{array}{l} \forall s \in [0, s^*], \text{ where } u^*(\cdot) \text{ is continuous}, \\ \mathcal{H}(\lambda(s), u^*(s), x^*(s)) = \max_{u \in [0,1]} \mathcal{H}(\lambda(s), u, x^*(s)), \end{array} \right. \end{array} \right.$$

where v is a vector normal to the target manifold. As the latter is the plane $t = T$, the only nonzero component of v is that in t, say, ν. We have

$$\left\{ \begin{array}{ll} \mathring{\lambda}_1 = -\partial\mathcal{H}/\partial x_1 = \lambda_1 - \lambda_3 h_1\,, & \lambda_1(S) = \partial\mathcal{K}/\partial x_1 = -e_1, \\ \mathring{\lambda}_2 = -\partial\mathcal{H}/\partial x_2 = u(\lambda_2 - \lambda_3 h_2)\,, & \lambda_2(S) = \partial\mathcal{K}/\partial x_2 = -e_2, \\ \mathring{\lambda}_3 = -\partial\mathcal{H}/\partial t = 0\,, & \lambda_3(S) = \partial\mathcal{K}/\partial t + \nu = \nu. \end{array} \right.$$

The last condition above translates into the switch-function

$$\sigma = \partial\mathcal{H}/\partial u = x_2(\lambda_3 h_2 - \lambda_2)\,.$$

As S is free, the final value of the Hamiltonian is zero. It yields

$$\forall s, \quad \lambda_3(s) = \nu = -\frac{e_1 x_1(S) + u e_2 x_2(S)}{\alpha + h_1 x_1(S) + h_2 x_2(S)}\,.$$

Thus

$$\sigma(S) = x_2(S)\left(\frac{e_2\alpha + x_1(S)(e_2 h_1 - e_1 h_2)}{\alpha + h_1 x_1(S) + h_2 x_2(S)}\right)\,.$$

It is easy to show that $\forall s,\ \mathring{\sigma} = 0$. Let $\hat{x}_1 = e_2\alpha/\zeta$. Hence

$$\left\{ \begin{array}{ll} \text{if } x_1(S) > \hat{x}_1, & \text{then } \forall s, \sigma < 0 \Rightarrow u^* = 0; \\ \text{if } x_1(S) = \hat{x}_1, & \text{then } \forall s, \sigma = 0 \Rightarrow u^* \in [0,1]; \\ \text{if } x_1(S) < \hat{x}_1, & \text{then } \forall s, \sigma > 0 \Rightarrow u^* = 1. \end{array} \right.$$

We now investigate the possible geometry of the trajectory fields, referring the reader to the four sketches of Figure A.2. Therefore, if being opportunistic yields a ratio of the best resource that remains lower than $\hat{x}_1$ (first quadrant), then this is optimal. In a similar fashion, if being selective yields a ratio of the best resource that remains greater than $\hat{x}_1$ (second quadrant), then this is optimal. Otherwise (third quadrant), the optimal policy is such that the ratio of best resources equals $\hat{x}_1$ at the end time. As it is clear that being opportunistic does not deplete the best resource as much as being selective during the same time, the fourth quadrant represents an impossible scenario.

However, Pontryagin's maximum principle provides only necessary conditions; it does not prove that *any* policy that leads to $x_1(S) = \hat{x}_1$ is optimal.

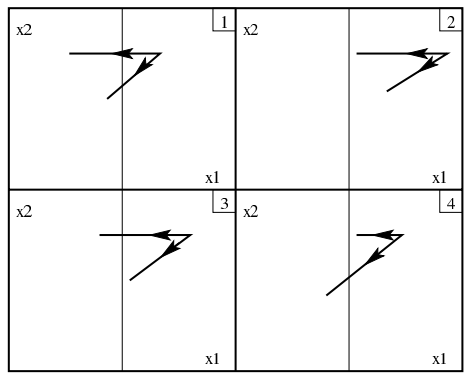


FIG. A.2. *Each quadrant represents a state-space* (x_1, x_2). *Vertical lines indicate the manifold* $x_1 = \hat{x}_1$. *The temporal horizon is the same for both trajectories plotted: "being selective" and "being opportunistic."*

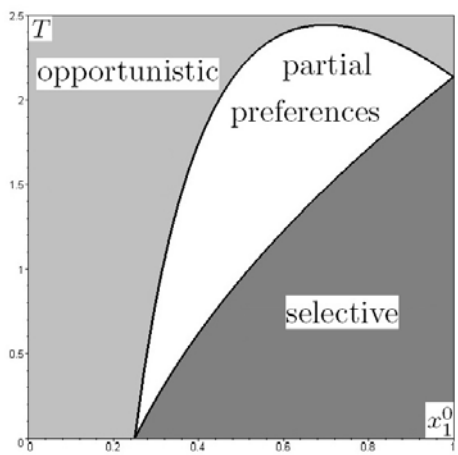


FIG. A.3. *The differing regions in the parameter space* (x_1^0, T). *We took* $\alpha = 1$, $e_1 = 1$, $e_2 = 1$, $h_1 = 1$, *and* $h_2 = 5$.

It is easy to see that

$$\forall u, s, \quad t(s) = \alpha s + [x_1^0 - x_1(s)]h_1 + [x_2^0 - x_2(s)]h_2 \,,$$

and that $S = \ln(x_1^0/\hat{x}_1)$. Hence there is a unique $x_2(S)$ such that $x_1(S) = \hat{x}_1$; i.e., $x_2(S) =: \hat{x}_2$ is such that

$$T = \alpha \ln(x_1^0/\hat{x}_1) + [x_1^0 - \hat{x}_1]h_1 + [x_2^0 - \hat{x}_2]h_2 \,. \tag{A.1}$$

However, it does not mean that the optimal trajectory is unique, as the time needed to move from a point (x_1, x_2) to another does not depend on the path followed. Therefore, any strategy that leads to $x_1(S) = \hat{x}_1$ is indeed optimal. For instance, Figure A.1 represents some possible optimal trajectories in the state-space (x_2, x_1). □

Figure A.3 shows the differing regions in the parameter space (x_1^0, T) that correspond to each policy: being selective, being opportunistic, and having "partial

preferences." The manifold separating the selective region and the partial preferences is given by the application

$$x_1^0 \mapsto \alpha \ln(x_1^0/\hat{x}_1) + [x_1^0 - \hat{x}_1]h_1 \,,$$

as $x_2^0 = \hat{x}_2$ on this boundary, and the other boundary (separating the partial preferences region from the opportunistic one) is given by

$$x_1^0 \mapsto \alpha \ln(x_1^0/\hat{x}_1) + (x_1^0 - \hat{x}_1)[h_1 + h_2 x_2^0/x_1^0],$$

as $x_1^0/x_2^0 = \hat{x}_1/\hat{x}_2$ on this boundary.

Acknowledgments. We thank Minus van Baalen, Esteban Freidin, Michel de Lara, Jean-Sébastien Pierre, and two anonymous referees for their comments.

REFERENCES

[1] P. Bernhard, *Singular surfaces in differential games. An introduction*, in Differential Games and Applications, Lecture Notes in Control and Information Sci. 3, P. Hagedorn, H. Knobloch, and G. Olsder, eds., Springer, Berlin, 1977.

[2] J. Brown, *Patch use as an indicator of habitat preference, predation risk, and competition*, Behavioral Ecology and Sociobiology, 22 (1988), pp. 37–47.

[3] J. Brown and W. Mitchell, *Diet selection on depletable resources*, Oikos, 54 (1989), pp. 33–43.

[4] C. Carathéodory, *Calculus of Variations and Partial Differential Equations of the First Order*, Holden–Day, San Francisco, 1965.

[5] E. Charnov, *Optimal foraging: Attack strategy of a mantid*, The American Naturalist, 110 (1976), pp. 141–151.

[6] E. Charnov, *Optimal foraging: The marginal value theorem*, Theoretical Population Biology, 9 (1976), pp. 129–136.

[7] L.-A. Giraldeau and G. Beauchamp, *Food exploitation: Searching for the optimal joining policy*, Trends in Ecology and Evolution, 14 (1999), pp. 102–106.

[8] F. Hamelin, P. Bernhard, P. Nain, and E. Wajnberg, *Foraging under competition: Evolutionarily stable patch-leaving strategies with random arrival times.* 1. *Scramble competition*, in Advances in Dynamic Game Theory, S. Jørgensen, M. Quincampoix, and T. Vincent, eds., Annals of the ISDG 9, Birkhaüser Boston, Boston, MA, 2007, pp. 327–348.

[9] F. Hamelin, P. Bernhard, A. Shaiju, and E. Wajnberg, *Foraging under competition: Evolutionarily stable patch-leaving strategies with random arrival times.* 2. *Interference competition*, in Advances in Dynamic Game Theory, S. Jørgensen, M. Quincampoix, and T. Vincent, eds., Annals of the ISDG 9, Birkhaüser Boston, Boston, MA, pp. 349–365.

[10] R. Heller, *On optimal diet in a patchy environment*, Theoretical Population Biology, 17 (1980), pp. 201–214.

[11] H. Hirvonen and E. Ranta, *Within-bout dynamics of diet choice*, Behavioral Ecology, 7 (1996), pp. 494–500.

[12] V. Hochman and B. Kotler, *Effects of food quality, diet preference and water on patch use by nubian ibex*, Oikos, 112 (2006), pp. 547–554.

[13] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.

[14] C. Holling, *Some characteristics of simple types of predation and parasitism*, The Canadian Entomologist, 91 (1959), pp. 385–398.

[15] R. Holt and B. Kotler, *Short-term apparent competition*, The American Naturalist, 130 (1987), pp. 412–430.

[16] A. Houston and J. McNamara, *Models of Adaptive Behavior: An Approach Based on State*, Cambridge University Press, Cambridge, UK, 1999.

[17] R. Hughes, ed., *Diet Selection: An Interdisciplinary Approach to Foraging Behavior*, Blackwell, Oxford, UK, 1993.

[18] J. Krebs and N. Davies, *Behavioral Ecology: An Evolutionary Approach*, Blackwell Science, Oxford, UK, 1997.

[19] J. Lucas and P. Schmid-Hempel, *Diet choice in patches: Time-constraint and state-space solutions*, Journal of Theoretical Biology, 131 (1988), pp. 307–332.

[20] J. Maynard Smith, *Evolution and the Theory of Games*, Cambridge University Press, Cambridge, UK, 1982.
[21] J. McNamara and A. Houston, *Partial preferences and foraging*, Animal Behavior, 35 (1987), pp. 1084–1099.
[22] J.M. McNamara, A.I. Houston, and E.J. Collins, *Optimality models in behavioral biology*, SIAM Rev., 43 (2001), pp. 413–466.
[23] J. McNamara, A. Houston, and J. Webb, *Combining prey choice and patch use—what does rate-maximising predict?*, Journal of Theoretical Biology, 164 (1993), pp. 219–238.
[24] V. Mikheev and J. Wanzenböck, *Satiation-dependent, intra-cohort variations in prey size selection of young roach (Rutilus rutilus),* Oecologia, 121 (1999), pp. 499–505.
[25] M. Milinski, *Optimal foraging: The influence of intraspecific competition on diet selection*, Behavioral Ecology and Sociobiology, 11 (1982), pp. 109–115.
[26] W. Mitchell, *An optimal control theory of diet selection: The effects of resource depletion and exploitative competition*, Oikos, 58 (1990), pp. 16–24.
[27] L. Peijnenburg and F. Galis, *Aspects of prey preference in Haplochromis piceatus*, Annales du Musée Royal d'Afrique Centrale Sciences Zoologiques, 257 (1989), pp. 85–88.
[28] C. Plowright and F. Landry, *A direct effect of competition on food choice by pigeons*, Behavioral Processes, 50 (2000), pp. 59–64.
[29] C. Plowright and D. Redmond, *The effect of competition on choice by pigeons: Foraging rate, resource availability and learning*, Behavioral Processes, 38 (1996), pp. 277–285.
[30] L. Pontryagin, V. Boltayanskii, R. Gamkrelidze, and E. Mishchenko, *The Mathematical Theory of Optimal Processes*, John Wiley, 1962 (translated from Russian).
[31] A. Sih and B. Christensen, *Optimal diet theory: When does it work, and when and why does it fail?*, Animal Behavior, 61 (2001), pp. 379–390.
[32] M. Sjerps and P. Haccou, *Effects of competition on optimal patch leaving: A war of attrition*, Theoretical Population Biology, 3 (1994), pp. 300–318.
[33] D. Stephens and J. Krebs, *Foraging theory*, Monographs in Behavior and Ecology, Princeton University Press, Princeton, NJ, 1986.
[34] M. van Baalen, V. Krivan, P. van Rijn, and M. Sabelis, *Alternative food, switching predators, and the persistence of predator-prey systems*, The American Naturalist, 157 (2001), pp. 512–524.
[35] M. Visser, *Prey selection by predators depleting a patch: An ess model*, Netherlands Journal of Zoology, 41 (1991), pp. 63–79.
[36] E. Wajnberg, P. Bernhard, F. Hamelin, and G. Boivin, *Optimal patch-time allocation for time-limited foragers*, Behavioral Ecology and Sociobiology, 60 (2006), pp. 1–10.
[37] J. Yearsley, *Optimal diet selection, frequency dependence and prey renewal*, Theoretical Population Biology, 64 (2003), pp. 129–139.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1562–1577

SINGULAR PERTURBATIONS IN ERGODIC CONTROL OF DIFFUSIONS*

VIVEK S. BORKAR† AND VLADIMIR GAITSGORY‡

Abstract. Ergodic control of a nondegenerate diffusion with two time-scales is studied in the limiting case as the time-scale separation increases to infinity. It is shown that the limit problem is another ergodic control problem for the slow time-scale component alone, with its dynamics averaged over the (controlled) invariant probability measures for the fast component. These measures in turn can be treated as the "effective control variable."

Key words. controlled diffusions, two time-scales, singular perturbations, ergodic control, invariant probability measures

AMS subject classification. 93E20

DOI. 10.1137/060657327

1. Introduction.

In this paper, we consider a long run average (ergodic) problem of optimal control of nonlinear singularly perturbed (SP) stochastic differential equations (SDEs), in which the singular perturbations parameter $\epsilon > 0$ is introduced in such a way that the state variables are decomposed into a group of slow variables that change their values with rates of the order $O(1)$, and a group of fast variables that change their values with rates of the order $O(\frac{1}{\epsilon})$.

Singularly perturbed problems of control and optimization have been studied intensively in both deterministic and stochastic settings (see the classic texts [7], [23], [25], [31] and the most recent publications [1], [2], [3], [4], [5], [6], [12], [14], [15], [16], [17], [18], [20], [21], [22], [26], [30], [32], [33]). Problems of optimal control of SP SDEs have been studied in [1], [7], [12], [21], [25], where earlier references can also be found.

In [12], in particular, it has been established in a very general setup that, for the problem of optimal control of SP SDEs considered on a finite time interval, the limiting problem (obtained when the singular perturbation parameter tends to zero) is an averaged problem, in which the slow dynamics is controlled by stationary marginal distributions of the fast dynamics, obtained with the slow state variables kept "frozen" (note that a deterministic counterpart of this result has been obtained in [17]).

In this article, we continue the line of research started in [12] by establishing the validity of a similar limit behavior for long run average problems of optimal control of SP SDEs (referred to in what follows as SP ergodic control problems). Note that in our study we restrict ourselves to the case of nondegenerate diffusions, and thus our results complement earlier results obtained in the purely deterministic setting in [18]. Our analysis is largely based on the stability and control theory for nondegenerate diffusions established in [8], [9], and [11].

*Received by the editors April 15, 2006; accepted for publication (in revised form) March 2, 2007; published electronically October 10, 2007.

http://www.siam.org/journals/sicon/46-5/65732.html

†School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India (borkar@tifr.res.in). The research of this author was supported in part by grant III.5(157)/99-ET from the Department of Science and Technology, Government of India.

‡Centre for Industrial and Applied Mathematics, University of South Australia, Mawson Lakes, SA 5095, Australia (v.gaitsgory@unisa.edu.au). The research of this author was supported by Australian Research Council Discovery grant DP0664330.

The paper is organized as follows. We introduce the singularly perturbed ergodic control problem in the next section. Our objective will be to relate this problem to the ergodic control problem for the "averaged" system obtained in the $\epsilon \to 0$ limit, i.e., to prove that the latter (lower dimensional) problem is a valid approximation to the above problem for small ϵ. The exact definition of the averaged problem is deferred until after the appropriate terminology has been introduced. A key assumption here is a condition on the running cost (see (4) below) which penalizes large excursions of the state process.

Section 3 recalls some known facts about ergodic control, notably the basic existence result (Theorem 3.1 below). This is stated in both the "almost sure" and the "expectation" form. It uses a characterization of limit points of the joint empirical process for state and control processes. This in turn shows that with probability one, none of these limit points can improve over the best attainable cost over the so-called Markov controls that depend only on the present state.

Section 4 is devoted to some preliminaries, in particular the definition of the averaged control problem. It then gives an important characterization of the set of the so-called "ergodic occupation measures," which have the property that the cost under stable Markov controls can be expressed as an average of the running cost function w.r.t. these. This result plays an important role throughout, as it helps us characterize limits of sequences of ergodic occupation measures and invariant measures. Section 4 also specializes the result concerning existence of optimal controls from section 3 to the present scenario.

Section 5 shows that the optimal cost for the averaged problem serves in general as an asymptotic lower bound for the optimal cost for the original problem in the $\epsilon \downarrow 0$ limit (Corollary 5.1). This argument is based on the tightness of the optimal ergodic occupation measures as ϵ is reduced to zero, implying their relative compactness in Prohorov topology. Section 6 shows that in the special case of the control entering the drift in an affine manner and the running cost strictly convex in the control, it is in fact the exact limit (Theorem 6.1). This crucially uses the fact that under these conditions, the expression being minimized over the control parameter in the associated Hamilton–Jacobi–Bellman equation is strictly convex, and therefore the minimizer is unique and continuously varying with ϵ. This result is extended to a more general case in section 7 under some technical assumptions (Theorem 7.1). These assumptions basically allow us, given an optimal control for the averaged problem, to approximate the optimal process at process level (i.e., in law) by the ϵ-indexed processes we start with, along with the associated control processes.

Section 8 discusses the "stable case," where a blanket stability condition is imposed on the controlled diffusion instead of the aforementioned condition on the running cost. This is sketched in the barest outline, as most of the details will be repetitive. Section 9 concludes with some discussion, which includes some directions for future research. The main tools used throughout are relative compactness in Prohorov topology and, in some cases, in the total variation norm, of the invariant distributions/ergodic occupation measures as the case may be, when appropriate parameters are varied.

2. The control problem. Let $\epsilon > 0$. We consider the coupled pair of SDEs in $\mathcal{R}^d \times \mathcal{R}^s$ given by

$$dz^\epsilon(t) = h(z^\epsilon(t), x^\epsilon(t), u(t))dt + \gamma(z^\epsilon(t))dB(t), \tag{1}$$

$$dx^\epsilon(t) = \frac{1}{\epsilon} m(z^\epsilon(t), x^\epsilon(t), u(t))dt + \frac{1}{\sqrt{\epsilon}} \sigma(z^\epsilon(t), x^\epsilon(t))dW(t). \tag{2}$$

Here

- for a prescribed compact metric action space A, $h : \mathcal{R}^d \times \mathcal{R}^s \times A \to \mathcal{R}^d$, $\gamma : \mathcal{R}^d \to \mathcal{R}^{d\times d}$, $m : \mathcal{R}^d \times \mathcal{R}^s \times A \to \mathcal{R}^s$, $\sigma : \mathcal{R}^d \times \mathcal{R}^s \to \mathcal{R}^{s\times s}$ are Lipschitz in the first and second (if any) arguments uniformly w.r.t. the third (if any);
- the least eigenvalues of $\gamma(z)\gamma(z)^T$, $\sigma(z,x)\sigma(z,x)^T$ are uniformly bounded away from zero (nondegeneracy assumption);
- the initial values are fixed: $(z^\epsilon(0), x^\epsilon(0)) = (z_0, x_0)$;
- $B(\cdot), W(\cdot)$ are, resp., d- and s-dimensional independent standard Brownian motions;
- $u(\cdot)$ is an A-valued control process with measurable paths satisfying the following *nonanticipativity* condition: for $t \geq s$, $(B(t) - B(s), W(t) - W(s))$ is independent of $\mathcal{F}_s \stackrel{def}{=}$ the completion of

$$\cap_{s'>s}\sigma(z^\epsilon(y), x^\epsilon(y), u(y), y \leq s').$$

We call such $u(\cdot)$ an *admissible control*.

We shall impose further restrictions on A, h, m later. The ergodic control problem is to minimize over all admissible $u(\cdot)$ the "ergodic cost"

$$\lim_{t\uparrow\infty} \frac{1}{t} \int_0^t E[k(z^\epsilon(s), x^\epsilon(s), u(s))]ds. \tag{3}$$

Here $k : \mathcal{R}^d \times \mathcal{R}^s \times A \to \mathcal{R}^+$ is a continuous map satisfying

$$\lim_{||(z,x)||\to\infty} \inf_u k(z,x,u) = \infty. \tag{4}$$

We shall discuss a possible relaxation of this condition later. We also assume the following:

(†) There exists an $\infty > M^* > 0$ such that for each $\epsilon \in (0,1)$, the cost for at least one admissible $u(\cdot)$ is $\leq M^*$.

We shall work with the *weak formulation* of the above control problem and assume that $u(\cdot)$ is a *relaxed* control. That is, for some compact metric space A', $A = \mathcal{P}(A') \stackrel{def}{=}$ the space of probability measures on A' with the Prohorov topology. Moreover, all functions above of the form $f(\ldots, u(t))$ (specifically, k and the components of h, m) are of the form $\int f'(\ldots, y)u(t, dy)$ for an f' satisfying the same conditions as f, except that the factor A of its domain is replaced by A'. This relaxation, originally introduced by L. C. Young in deterministic control, is a true relaxation in this context, in the sense that the infima of the cost over relaxed and original setup coincide. See [9, Chapter I] for more on this. As above, $\mathcal{P}(Z)$ for a Polish space Z will denote the Polish space of probability measures on Z with the Prohorov topology [10, Chapter 2].

Furthermore, we assume that the following "stochastic Liapunov" condition holds: Define $\mathcal{L} : C^2(\mathcal{R}^s)(\stackrel{def}{=}$ the space of twice continuously differentiable functions $\mathcal{R}^s \to \mathcal{R}) \to C_b(\mathcal{R}^d \times \mathcal{R}^s \times A')$ by

$$\mathcal{L}f(z,x,u) \stackrel{def}{=} \frac{1}{2} tr\left(\sigma(z,x)\sigma^T(z,x)\nabla^2 f(x)\right) + \langle \nabla f(x), m'(z,x,u)\rangle \tag{5}$$

$\forall\ f \in C^2(\mathcal{R}^s)$. Then there exists a $V \in C^2(\mathcal{R}^s)$, $g \in C(\mathcal{R}^d \times \mathcal{R}^s)$ such that $\lim_{||x||\to\infty} V(x) = \infty$, $\lim_{||x||\to\infty} g(z,x) = \infty$ uniformly in z belonging to any compact subset of $\mathcal{R}^d$, and

$$\mathcal{L}V(z,x,u) \leq -g(z,x). \tag{6}$$

3. Ergodic control. We now recall from [9, Chapter VI] some facts about ergodic control applicable to the above framework. For this purpose, we introduce the notion of a Markov control as a $u(\cdot)$ of the form $u(t) = v(z^\epsilon(t), x^\epsilon(t))\ \forall t$ for a measurable $v : \mathcal{R}^d \times \mathcal{R}^s \to A$. By a standard abuse of terminology, we identify this $u(\cdot)$ with the map v. Note that under a Markov control, $(z^\epsilon(\cdot), x^\epsilon(\cdot))$ will be a time-homogeneous Markov process. In turn, v will be said to be a stable Markov control if the resulting Markov diffusion is positive recurrent and thus has a unique invariant probability measure $\zeta_v(dzdx)$. Furthermore, (3) will then equal $\int k'(z,x,u)v(du|z,x)\zeta_v(dzdx)$. We call $\Phi_v(dzdxdu) \stackrel{def}{=} \zeta_v(dzdx)v(du|z,x)$ the *ergodic occupation measure* associated with v and denote by $\mathcal{G}$ the set of all ergodic occupation measures Φ_v as v varies over all stable Markov controls. This has another characterization as follows: Let

$$\begin{aligned}&\hat{\mathcal{L}}f(z,x,u)\\ &\stackrel{def}{=} \frac{1}{2}tr\left(\gamma(z)\gamma^T(z)\nabla_z^2 f(z,x)\right) + \langle \nabla_z f(z,x), h'(z,x,u)\rangle\\ &\qquad + \frac{1}{2\epsilon}tr\left(\sigma(z,x)\sigma^T(z,x)\nabla_x^2 f(z,x)\right) + \frac{1}{\epsilon}\langle \nabla_x f(z,x), m'(z,x,u)\rangle,\end{aligned}$$

where ∇_y, ∇_y^2 denote, resp., the gradient and the Hessian in the variable y. Also let $C_0^2(\mathcal{R}^{d+s}) \stackrel{def}{=}$ the space of twice continuously differentiable functions $\mathcal{R}^{d+s} \to \mathcal{R}$ that vanish at infinity, along with their first and second order partial derivatives.

We recall now Lemma 1.1 from [9, p. 144] (see [8] for a more general result).

LEMMA 3.1. $\mathcal{G} = \{\Phi \in \mathcal{P}(\mathcal{R}^{d+s} \times A') : \int \hat{\mathcal{L}} f d\Phi = 0\ \forall f \in C_0^2(\mathcal{R}^{d+s})\}$.

Proof. Let $\Phi(dzdxdu) = \zeta(dzdx)v(du|z,x) \in \mathcal{G}$ and denote by $\hat{\mathcal{L}}_v$ the extended generator of the diffusion controlled by the Markov control v, i.e., $\hat{\mathcal{L}}_v f \stackrel{def}{=} \int \hat{\mathcal{L}} f(z,x,u)v(du|z,x)$. The corresponding diffusion is a time-homogeneous Markov process with a transition semigroup $T_t, t \geq 0$, of positive operators on the Banach space $\mathcal{B}$ of bounded measurable functions $\mathcal{R}^{d+s} \to \mathcal{R}$ (with essential supremum norm) and with (operator) norm 1. Its infinitesimal generator $\mathcal{L}'_v$ is an extension to $\mathcal{D}(\mathcal{L}'_v)$ of $\hat{\mathcal{L}}_v : C_0^2(\mathcal{R}^{d+s}) \subset \mathcal{B} \to \mathcal{B}$. Thus, in particular,

$$\frac{d}{dt}T_t f = \hat{\mathcal{L}}_v T_t f = T_t \hat{\mathcal{L}}_v f, \quad f \in C_0^2(\mathcal{R}^{d+s}). \tag{7}$$

Also, ζ is invariant under $\{T_t\}$:

$$\int f d\zeta = \int T_t f d\zeta \quad \forall t \geq 0, \quad f \in C_b(\mathcal{R}^{d+s}). \tag{8}$$

Then for $t > s$, $f \in C_0^2(\mathcal{R}^{d+s})$,

$$T_t f = T_s f + \int_s^t T_y \hat{\mathcal{L}}_v f dy.$$

Integrating w.r.t. ζ and using (8), we get

$$\int f d\zeta = \int f d\zeta + \int_s^t \int \hat{\mathcal{L}}_v f d\zeta dy, \qquad t > s.$$

It follows that $\zeta(dzdx)v(du|z,x) = \Phi(dzdxdu)$ satisfies

$$\int \hat{\mathcal{L}} f d\Phi \quad \left(= \int \hat{\mathcal{L}}_v f d\zeta\right) = 0, \quad f \in C_0^2(\mathcal{R}^{d+s}). \tag{9}$$

Conversely, let $\Phi(dzdxdu) = \zeta(dzdx)v(du|z,x)$ satisfy (9). Then $\zeta(t) \equiv \zeta \ \forall t \geq 0$ satisfies the forward equation

$$\int f d\zeta(t) = \int f d\zeta(0) + \int_0^t \int \hat{\mathcal{L}}_v f d\zeta(s) ds. \tag{10}$$

From (7) and (10), direct calculation shows that $\frac{d}{ds}\int T_{t-s} f d\zeta(s) = 0$. Integrating this from 0 to t, we have $\int f d\zeta(t) = \int T_t f d\zeta(0)$. Since $\zeta(t) = \zeta(0) = \zeta$, this implies that ζ is stationary under v. In particular, in view of our nondegeneracy assumption, the resulting Markov diffusion process is ergodic. That is, $\Phi \in \mathcal{G}$. □

Define the empirical measures $\nu_t, t > 0$, and the average empirical measures $\bar{\nu}_t, t > 0$, by

$$\int f d\nu_t \stackrel{def}{=} \frac{1}{t}\int_0^t f(z^\epsilon(s), x^\epsilon(s), u(s)) ds,$$

$$\int f d\bar{\nu}_t \stackrel{def}{=} \frac{1}{t}\int_0^t E[f(z^\epsilon(s), x^\epsilon(s), u(s))] ds$$

for $f \in C_b(\mathcal{R}^{d+s} \times A')$. Let $\mathcal{R}^*$ denote the one point compactification of $\mathcal{R}^d \times \mathcal{R}^s$ with ∞ the point at infinity. Finally, let

$\mathcal{G}^* \stackrel{def}{=} \{\Phi \in \mathcal{P}(R^* \times A') :$ there exist some $0 \leq a \leq 1, \phi \in \mathcal{G},$ and $\phi' \in \mathcal{P}(\{\infty\} \times A')$ such that$\Phi(B \times B') = a\phi((B \times B') \cap (\mathcal{R}^{d+s} \times A')) + (1-a)\phi'((B \times B') \cap (\{\infty\} \times A')) \ \forall$ B Borel in $\mathcal{R}^*$, B' Borel in $A'\}$.

The following is adapted from [9, Chapter VI].

LEMMA 3.2. *As $t \uparrow \infty$, $\bar{\nu}_t \to \mathcal{G}^*$ and $\nu_t \to \mathcal{G}^*$ a.s. in $\mathcal{P}(R^* \times A')$.*

Proof. Consider $f \in C_0^2(\mathcal{R}^{d+s})$. Then

$$\frac{E[f(z^\epsilon(t), x^\epsilon(t))]}{t} - \frac{E[f(z^\epsilon(0), x^\epsilon(0))]}{t} = \frac{1}{t}\int_0^t E[\hat{\mathcal{L}} f(z^\epsilon(s), x^\epsilon(s), u(s))] ds$$

$$= \int \hat{\mathcal{L}} f d\bar{\nu}_t.$$

Let ν be a limit point of $\bar{\nu}_t$ in $\mathcal{P}(\mathcal{R}^* \times A')$ as $t \to \infty$. It can then be written as $a\nu_1 + (1-a)\nu_2$, where $\nu_1 \in \mathcal{P}(\mathcal{R}^{d+s} \times A')$, $\nu_2 \in \mathcal{P}(\infty \times A')$, and $a \in [0,1]$. Letting $t \to \infty$ in the above equation along an appropriate subsequence, it follows that $\int \hat{\mathcal{L}} f d\nu_1 = 0$ when $a > 0$. When $a = 0$, this may be imposed without any loss of generality. By Lemma 3.1, it then follows that $\nu_1 \in \mathcal{G}$. For the second claim, one similarly has

$$\frac{f(z^\epsilon(t), x^\epsilon(t))}{t} - \frac{f(z^\epsilon(0), x^\epsilon(0))}{t}$$

$$= \frac{1}{t}\int_0^t \hat{\mathcal{L}} f(z^\epsilon(s), x^\epsilon(s), u(s)) ds$$

$$= \frac{1}{t}\int_0^t \langle \nabla f(z^\epsilon(s), x^\epsilon(s)), (\gamma(z^\epsilon(s)) dB(s), \frac{1}{\sqrt{\epsilon}}\sigma(z^\epsilon(s), x^\epsilon(s)) dW(s))\rangle$$

$$= \int \hat{\mathcal{L}} f d\nu_t + \frac{M(t)}{t},$$

where $M(t)$ is a continuous square-integrable martingale with quadratic variation $\langle M\rangle(t)$ which is $O(t)$. Since $M(t)/\langle M\rangle(t) \to 0$ a.s. on $\{\langle M\rangle(t) \uparrow \infty\}$ (see, e.g., [28]), it follows that

$$\frac{M(t)}{t} = \frac{M(t)}{\langle M\rangle(t)} \frac{\langle M\rangle(t)}{t} \to 0$$

a.s., whence the second claim follows by arguments similar to those used for proving the first. □

The following consequence thereof follows easily.

THEOREM 3.1. *There exists a stable optimal Markov control v_ϵ^* such that if Φ_ϵ^* is the corresponding ergodic occupation measure, then under any admissible $u(\cdot)$,*

$$\liminf_{t\uparrow\infty} \frac{1}{t} \int_0^t k(z^\epsilon(s), x^\epsilon(s), u(s))ds \geq \int k' d\Phi_\epsilon^* \quad a.s.,$$

$$\liminf_{t\uparrow\infty} \frac{1}{t} \int_0^t E[k(z^\epsilon(s), x^\epsilon(s), u(s))]ds \geq \int k' d\Phi_\epsilon^*.$$

Proof. From Lemma 3.2 and (4), it follows that the left-hand side in both cases is ("a.s." in the former case) at least as much as $\inf_{\Phi\in\mathcal{G}} \int k' d\Phi$. By (4), it also follows that the set $\{\Phi \in \mathcal{G} : \int k' d\Phi \leq c\}$ is compact for any $c > 0$. Since $\Phi \to \int k' d\Phi$ is lower semicontinuous on $\mathcal{G}$, it follows that the infimum is attained at some $\Phi_\epsilon^* \in \mathcal{G}$. The claim follows. □

Remark. One can in fact show that the v_ϵ^* can be taken to be *precise*, i.e., $v_\epsilon^*(z, x)$ is a Dirac measure $\forall\ z, x$. This is because the extreme points of $\mathcal{G}$ correspond to precise controls, as proved in [13].

4. The averaged system. Setting $\tau = \frac{t}{\epsilon}$, $\bar{x}(\tau) = x^\epsilon(\epsilon\tau)$, $\bar{z}(\tau) = z^\epsilon(\epsilon\tau)$, $\bar{u}(\tau) = u(\epsilon\tau)$, $\bar{W}(\tau) = \frac{1}{\sqrt{\epsilon}} W(\epsilon\tau)$, (2) becomes

$$d\bar{x}(\tau) = m(\bar{z}(\tau), \bar{x}(\tau), \bar{u}(\tau))d\tau + \sigma(\bar{z}(\tau), \bar{x}(\tau))d\bar{W}(\tau),$$

which does not depend on ϵ explicitly. To this we associate the *associated system*

$$d\bar{x}(\tau) = m(\bar{z}, \bar{x}(\tau), \bar{u}(\tau))d\tau + \sigma(\bar{z}, \bar{x}(\tau))d\bar{W}(\tau), \tag{11}$$

where $\bar{z}$ is fixed, $\bar{W}(\cdot)$ is a standard Brownian motion independent of $\bar{x}(0)$, and admissibility of $\bar{u}(\cdot)$ is defined by the following: For $t > s$, $\bar{W}(t) - \bar{W}(s)$ is independent of $\Xi_s \stackrel{def}{=}$ the completion of $\cap_{s'>s}\sigma(\bar{x}(\tau), \bar{u}(\tau), \bar{W}(\tau), \tau \leq s')$.

Remark. Equation (11) above is the relaxed control form of the associated system. One also has the prerelaxation form

$$d\bar{x}(\tau) = m'(\bar{z}, \bar{x}(\tau), \bar{u}'(\tau))d\tau + \sigma(\bar{z}, \bar{x}(\tau))d\bar{W}(\tau), \tag{12}$$

which we shall have occasion to use later.

Let $D_z \stackrel{def}{=} \{\mu \in \mathcal{P}(\mathcal{R}^s \times A') : \int \mathcal{L}f(z, x, u)\mu(dxdu) = 0\ \forall\ f \in C_0^2(\mathcal{R}^s)\}$, where $\mathcal{L}$ is as in (5). The next lemma in particular characterizes this as the set of ergodic occupation measures for the associated system.

LEMMA 4.1. *D_z = the set of $\mu(dxdu)$ of the form $\mu(dxdu) = \eta(dx)v(du|x)$, where η is the unique stationary distribution for the time-homogeneous Markov diffusion $X(\cdot)$ given by (11) when $u(\cdot) = v(X(\cdot)) \stackrel{def}{=} v(du|X(\cdot))$. The set valued map $z \to D_z$*

is convex compact valued and continuous. Furthermore, for compact $B \subset \mathcal{R}^d$, $\cup_{z\in B} D_z$ *is compact.*

Proof. The first claim follows exactly as in Lemma 3.1. That D_z is convex closed for each z is easily verified from the definition. Thus we need to verify its relative compactness in $\mathcal{P}(\mathcal{R}^s \times A')$. Since A' is compact, it suffices to verify the relative compactness of the corresponding marginals $\eta(dx)$ in $\mathcal{P}(\mathcal{R}^s)$. Under our assumption (6), this is proved in [11]. (The key step is to prove that $\int g d\eta$ is uniformly bounded over all such η, from which the claim is immediate in view of the condition $\lim_{\|(z,x)\|\to\infty} g(z,x) = \infty$, and the Chebyshev inequality.) Next, let $z_n \to z_\infty$ and $\mu_n \in D_{z_n}\ \forall\ n, 1 \le n < \infty$. Then

(i) $\{\mu_n\}$ are tight by arguments similar to those used in [11] (see above), and

(ii) any limit point μ thereof is in D_{z_∞}—this is easily verified from the definition of D_z. Thus $z \to D_z$ is upper semicontinuous.

Now fix a $\mu(dxdu) = \eta_{z_\infty}(dx)v(du|x) \in D_{z_\infty}$. Under $z = z_n$, the stationary Markov control $v(du|x)$ leads to a unique stationary distribution $\eta_{z_n}(dx), 1 \le n \le \infty$. By our nondegeneracy assumption, the transition probabilities for $t > 0$ of the corresponding time-homogeneous Markov processes have densities w.r.t. the Lebesgue measure. Therefore so do the corresponding invariant probability measures $\{\eta_{z_n}\}$. Let $\{\chi_{z_n}(\cdot)\}$ denote these densities. We claim that they are pointwise bounded and equicontinuous. If pointwise boundedness does not hold, $\chi_{z_n}(x^*) \uparrow \infty$ for some x^*. But $\{\chi_{z_n}\}$ satisfy $(\mathcal{L}_v^{z_n})^*\chi_{z_n} \equiv 0$, where

$$\mathcal{L}_v^z \stackrel{def}{=} \frac{1}{2} tr\left(\sigma(z,x)\sigma^T(z,x)\nabla_x^2\right) + \langle m(z,x,v(x)), \nabla_x\rangle,$$

and $(\mathcal{L}_v^z)^*$ denotes its formal adjoint given by

$$(\mathcal{L}_v^z)^* f \stackrel{def}{=} \frac{1}{2}\sum_{i,j,k} \frac{\partial^2}{\partial x_i \partial x_j}\left(\sigma_{ik}(z,x)\sigma_{jk}(z,x)f(x)\right) - \sum_i \frac{\partial}{\partial x_i}\left(m_i(z,x,v(x))f\right).$$

By Harnack's inequality (see Theorem 8.20 on page 199 of [19]), the ratio of the maximum to the minimum of $\chi_{z_n}(\cdot)$ on any compact set must remain bounded uniformly in n. Thus $\chi_{z_n}(\cdot) \uparrow \infty$ uniformly on compacts, which contradicts the fact that they are probability densities. Hence they are pointwise bounded. By [19, Theorem 8.24, p. 202], they satisfy a uniform Hölder continuity condition on compacts, which gives equicontinuity. In particular, $\chi_{z_n}(\cdot)$ are uniformly continuous on compacts. The equation

$$\int \mathcal{L}_v^{z_n} f(x)\eta_{z_n}(dx) = 0\ \forall\ f \in C_0^2(\mathcal{R}^s) \tag{13}$$

characterizes $\eta_{z_n}(dx)$ and therefore $\chi_{z_n}(\cdot)$. Let η^* denote a limit point of $\eta_{z_n}(dx)$ in $\mathcal{P}(\mathcal{R}^s)$ as $n \uparrow \infty$. By the Arzela–Ascoli theorem, we may drop to a subsequence if necessary and suppose that $\chi_{z_n}(\cdot) \to \chi^*(\cdot)$ in $C(\mathcal{R}^s)$. Then for compactly supported $f \in C(\mathcal{R}^s)$,

$$\int f(x)\chi_{z_n}(x)dx \to \int f(x)\chi^*(x)dx,$$

implying that $\eta^*(dx) = \chi^*(x)dx$. By Scheffe's theorem (see Borkar [10, p. 26]), we have $\eta_{z_n}(dx) \to \eta^*$ in total variation. Hence we can let $n \to \infty$ in (13) to obtain

$$\int \mathcal{L}_v^{z_\infty} f d\eta^* = 0\ \forall\ f \in C_0^2(\mathcal{R}^s),$$

implying $\eta^*(dx) = \eta_{z_\infty}(dx)$. Thus the lower semicontinuity of $z \to D_z$ follows. Together, upper and lower semicontinuity imply continuity of this set-valued map. Compactness of $\cup_{z\in B} D_z$ is proved by an argument similar to the one used for proving upper semicontinuity. □

In particular, it follows that $\{(z,\mu) : z \in \mathcal{R}^s, \mu \in D_z\}$ is closed and $\{(z,\mu) : z \in B, \mu \in D_z\}$ is compact for compact $B \subset \mathcal{R}^s$. Define

$$\bar{h}(z,\mu) \stackrel{def}{=} \int h'(z,x,u)\mu(dxdu),$$

$$\bar{k}(z,\mu) \stackrel{def}{=} \int k'(z,x,u)\mu(dxdu).$$

The *averaged system* is defined by

$$dz(t) = \bar{h}(z(t),\mu(t))dt + \gamma(z(t))dB'(t), \tag{14}$$

$$\mu(t) \in D_{z(t)} \quad \forall\, t. \tag{15}$$

Here $z(0) = z_0$ (the same as in (1)), $B'(\cdot)$ is a standard Brownian motion in $\mathcal{R}^d$, and $\mu(\cdot)$ satisfies (15) and the following nonanticipativity condition: For $t \geq s \geq 0$, $B'(t) - B'(s)$ is independent of the completion of $\cap_{s'>s}\sigma(z(y), B'(y), \mu(y), y \leq s')$. We may view $\mu(\cdot)$ as the "effective control process" for the averaged system. The objective for the averaged control problem is to minimize

$$\limsup_{t\uparrow\infty} \frac{1}{t} E\left[\int_0^t \bar{k}(z(s),\mu(s))\right] ds$$

over all admissible $\mu(\cdot)$. By analogy with section 2, we call $\mu(\cdot)$ a Markov control if $\mu(t) = q(z(t)) \stackrel{def}{=} q(dxdu|z(t))\ \forall t$, identified with the measurable map q. Call it a stable Markov control if, in addition, the resulting time-homogeneous Markov process $z(\cdot)$ is positive recurrent. In the latter case, $z(\cdot)$ will have a unique invariant probability distribution $\varphi_q(dz)$ and the corresponding ergodic occupation measure $\Gamma(dzdxdu) \stackrel{def}{=} \varphi_q(dz)q(dxdu|z)$. Let $\mathcal{Q}$ denote the set of such Γ. Then as before, one has the following characterization: Define $\tilde{\mathcal{L}} : C_0^2(\mathcal{R}^d) \to C_b(\mathcal{R}^d \times \mathcal{P}(\mathcal{R}^s \times A'))$ by

$$\tilde{\mathcal{L}}f(z,\mu) = \frac{1}{2} tr\left(\gamma(z)\gamma^T(z)\nabla^2 f(z)\right) + \langle \nabla f(z), \bar{h}(z,\mu)\rangle.$$

Lemma 4.2.

$$\mathcal{Q} = \Bigg\{ \xi = q(dxdu|z)\phi(dz) \in \mathcal{P}(\mathcal{R}^d \times \mathcal{R}^s \times A') : q(\cdot|z) \in D_z\ \forall z,$$

$$\int \tilde{\mathcal{L}}f(z, q(dxdu|z))\phi(dz) = 0\ \forall\ f \in C_0^2(\mathcal{R}^d) \Bigg\}.$$

This again follows exactly as in Lemma 3.1. We have then the following counterpart of Theorem 3.1, proved analogously.

Theorem 4.1. *There exists a stable optimal Markov control q^* for the averaged system such that if $\Gamma^*(dzdxdu) = q^*(dxdu|z)\varphi^*(dz)$ is the corresponding ergodic*

occupation measure, then for any admissible $\mu(\cdot)$ as above,

$$\liminf_{t\uparrow\infty} \frac{1}{t}\int_0^t \bar{k}(z(s),\mu(s))ds \geq \int k'd\Gamma^* \ \ a.s.,$$

$$\liminf_{t\uparrow\infty} \frac{1}{t}\int_0^t E[\bar{k}(z(s),\mu(s))]ds \geq \int k'd\Gamma^*.$$

Let $\mathcal{Q}_{opt}$ denote the set of optimal ergodic occupation measures, i.e.,

$$\text{Argmin}\left\{\int k'd\xi : \xi \in \mathcal{Q}\right\}. \tag{16}$$

Also, write

$$q^*(dxdu|z) = v^*(du|z,x)\eta^*(dx|z). \tag{17}$$

5. A lower bound. We now consider the $\epsilon \downarrow 0$ limit. Let Φ^*_ϵ be as in Theorem 3.1 above. Then by (†), it follows that $\sup \int k d\Phi^*_\epsilon < \infty$. In turn, by (4) and the Chebyshev inequality, it then follows that $\{\Phi^*_\epsilon, \epsilon \in (0,1)\}$ is tight. Let Φ^*_0 be a limit point thereof in $\mathcal{P}(\mathcal{R}^{d+s}\times A')$.

THEOREM 5.1. $\Phi^*_0 \in \mathcal{Q}$.

Proof. Disintegrate Φ^*_0 as

$$\begin{aligned}\Phi^*_0(dzdxdu) &= \varphi(dz)\mu(dxdu|z)\\ &= \varphi(dz)\eta(dx|z)v(du|z,x).\end{aligned}$$

(In particular, $\mu(dxdu|z) = \eta(dx|z)v(du|z,x)$.) Let $f_1 \in C^2_0(\mathcal{R}^d), f_2 \in C^2_0(\mathcal{R}^s)$. Let $\epsilon \downarrow 0$ in the equation $\epsilon\int \hat{\mathcal{L}}(f_1 f_2)d\Phi^*_\epsilon = 0$ to obtain

$$\int f_1(z)\int \mathcal{L}f_2(z,x,u)\mu(dxdu|z)\varphi(dz) = 0. \tag{18}$$

Then as (18) holds $\forall\ f_1 \in C^2_0(\mathcal{R}^d)$, we conclude that for φ-a.s. z,

$$\int \mathcal{L}f_2(z,x,u)d\mu(dxdu|z) = 0,$$

implying that $\mu(dxdu|z) \in D_z$. The qualification "φ-a.s." may be dropped by choosing a suitable version. Now for $h \in C^2_0(\mathcal{R}^d)$ (i.e., h is a function of $z \in \mathcal{R}^d$ alone), let $\epsilon \downarrow 0$ in $\int \hat{\mathcal{L}}h d\Phi^*_\epsilon = 0$ to obtain

$$\int \hat{\mathcal{L}}h d\Phi^*_0 = \int \tilde{\mathcal{L}}h(z,\mu(\cdot|z))\varphi(dz) = 0. \tag{19}$$

By Lemma 4.2, (19) implies that φ is the unique stationary distribution under μ for the averaged system. It follows that $\Phi^*_0 \in \mathcal{Q}$. ∎

COROLLARY 5.1. $\liminf_{\epsilon\downarrow 0}\int k'd\Phi^*_\epsilon \geq \int \bar{k}d\Gamma^*$.

This shows that the optimal ergodic cost for the averaged problem provides an asymptotic lower bound (as $\epsilon \downarrow 0$) for the optimal ergodic cost of the original problem. To show that it is in fact a valid approximation, we must replace the "lim inf" by "lim" in the above and replace the inequality by an equality. We shall do so under additional assumptions in the following sections.

6. Main results—the affine case. Assume the following:

(*) A' is a compact subset of $\mathcal{R}^m$ for some $m \geq 1$, and for each z, x, $h'(z,x,\cdot)$, $m'(z,x,\cdot)$ are componentwise affine, and $k'(z,x,\cdot)$ is strictly convex.

(**) $||h'(z,x,u)|| = o(k'(z,x,u))$ as $||(z,x)|| \uparrow \infty$ and

$$\sup_u |k'(z,x,u)|^{1+a} \leq Kg(z,x)$$

for some $K, a > 0$, and g as in (6).

The next lemma, which uses only (*) and (**), shows in particular that v^* in (17) is unique. Thus we can state our third assumption as follows:

(***) $v^*(z,x) \stackrel{def}{=} v^*(du|z,x)$ in (17) is a stable Markov control for (1), (2) for sufficiently small $\epsilon > 0$ (say, $\epsilon < \epsilon_0$) and the corresponding stationary distributions, denoted $\zeta^\epsilon(dzdx), 0 < \epsilon < \epsilon_0$, are tight.

A stochastic Liapunov condition along the lines of (6) can be given to ensure this.

LEMMA 6.1. *$v^*(du|z,x)$ in (17) is unique (that is, $\mathcal{Q}_{opt}$ is a singleton: $\mathcal{Q}_{opt} = \{\Gamma^*\}$) and continuous in z, x.*

Proof. By [9, Theorem 3.3, p. 163], a necessary and sufficient condition for the optimality of q^* is that $q^*(z)$ minimize the function

$$\mu \to \bar{k}(z,\mu) + \langle \nabla\Psi(z), \bar{h}(z,\mu)\rangle \tag{20}$$

over D_z for a.e. z, where $\Psi \in C^2(\mathcal{R}^d)$ is the value function for the ergodic control problem for the averaged system.[1] We may drop the qualification "for a.e. z" by taking an appropriate version. Now for fixed z, consider the ergodic control problem for the associated system (12) with cost

$$\limsup_{t\uparrow\infty} \frac{1}{t} \int_0^t E[\ell_z(\bar{x}(s), a(t))]ds,$$

where $\ell_z \in C(\mathcal{R}^s \times A')$ is defined by

$$\ell_z(x,a) \stackrel{def}{=} k'(z,x,a) + \langle \nabla\Psi(z), h'(z,x,a)\rangle.$$

Since D_z is precisely the set of ergodic occupation measures for the associated system, q^* is the optimal ergodic occupation measure for the above problem. By (**), [9, Theorem 3.3, p. 163] can be applied again to this new control problem in order to conclude as above that $v^*(du|z,x)$ minimizes

$$\kappa \to \int (\ell_z(x,\cdot) + \langle \nabla\tilde{\Psi}_z(x), m'(z,x,\cdot)\rangle) d\kappa,$$

where $\tilde{\Psi}_z \in C^2(\mathcal{R}^s)$ is the value function for this new ergodic control problem. (Note that for each z, the cost function ℓ_z satisfies a condition akin to (4) because of the first half of (**), and thus the above remarks apply.) By [9, Theorem 2.1, p. 183], it follows

[1] [9] proves the existence of a C^2 value function and the associated "verification theorem" for non-degenerate diffusions with bounded coefficients and the so-called "near-monotone" cost for the case when the control space is state-independent. The latter would correspond to D_z being independent of z in the present setup. Condition (4) is a special case of near-monotonicity. The modifications required to handle the more general Lipschitz coefficients and state-dependent control space needed here are minor in view of the continuity of the set-valued map $z \to D_z$ already established. The details, though routine, would be a significant digression and are therefore omitted.

that the map $(z,x) \to \nabla\tilde{\Psi}_z(x)$ is continuous. By (*), the above minimum is attained at a unique point. It is easy to see then that this point will depend continuously on z,x. That is, $(z,x) \to v^*(z,x)$ is continuous. □

Remark. Note that $v^*(z,x)$ will in fact be Dirac $\forall\ z,x$. Also note that the assumption of affine dependence of k', m' on u is crucial in proving the strict convexity claims above: The sum of a strictly convex function and an affine function is strictly convex. Weakening it to convexity would not do because we do not have control on the signs of the components of $\nabla\Psi, \nabla\tilde{\Psi}$.

Recall the measures q^*, Γ^* from Theorem 4.1.

COROLLARY 6.1. *q^*, Γ^* are unique.*

Proof. Recall that $q^*(dxdu|z) = \eta^*(dx|z)v^*(du|z,x)$, where $\eta^*(dx|z)$ is the unique stationary distribution for the associated system under $v^*(du|z,x)$. Uniqueness of q^* follows. In turn, $\Gamma^*(dzdxdu) = q^*(dxdu|z)\varphi^*(dz)$, where φ^* is the unique stationary distribution of the averaged system under $q^*(dxdu|z)$. Thus Γ^* is unique. □

Let

$$\tilde{\Phi}_\epsilon(dzdxdu) \stackrel{def}{=} \zeta^\epsilon(dzdx)v^*(du|z,x), \qquad \epsilon \in (0,\epsilon_0),$$

and let v^* be as in (17).

THEOREM 6.1. $\lim_{\epsilon\downarrow 0}\tilde{\Phi}_\epsilon = \Gamma^*$, *the convergence being in* $\mathcal{P}(\mathcal{R}^d \times \mathcal{R}^s \times A')$. *Also*

$$\lim_{\epsilon\downarrow 0}\int k'd\tilde{\Phi}_\epsilon = \int k'd\Gamma^* \tag{21}$$

and

$$\lim_{\epsilon\downarrow 0}\int k'd\Phi^*_\epsilon = \int k'd\Gamma^*. \tag{22}$$

Proof. In view of Theorem 5.1, (21) implies (22). Below, we will prove the convergence of $\tilde{\Phi}_\epsilon$ to Γ^* and then show the validity of (21).

Let $\zeta^\epsilon(dzdx) \to \hat{\zeta}(dzdx) = \hat{\varphi}(dz)\hat{\eta}(dx|z)$ along a subsequence as $\epsilon \downarrow 0$. In view of the continuity of $v^*(du|\cdot,\cdot)$, we may pass to the limit along this subsequence in

$$\epsilon\int \hat{\mathcal{L}}f(z,x,u)v^*(du|z,x)\zeta^\epsilon(dzdx) = 0,\ f \in C_0^2(\mathcal{R}^{d+s})$$

to obtain

$$\int \hat{\mathcal{L}}f(z,x,u)v^*(du|z,x)\hat{\zeta}(dzdx) = 0,\ f \in C_0^2(\mathcal{R}^{d+s}).$$

Argue as in Theorem 5.1 to conclude that $\hat{\eta}(dx|z)$ is in fact the unique stationary distribution for the associated system controlled by $v^*(du|z,x)$ (i.e., $\hat{\eta}(dx|z) = \eta^*(dx|z)$) for $\hat{\varphi}$-a.s. z. The latter qualification may be dropped by choosing an appropriate version. Recall that $q^*(dxdu|z) = \eta^*(dx|z)v^*(du|z,x)\ \forall\ z$. Let $\epsilon \downarrow 0$ in

$$\int \hat{\mathcal{L}}f(z,x,u)v^*(du|z,x)\zeta^\epsilon(dzdx) = 0$$

for $f \in C_0^2(\mathcal{R}^d)$ (i.e., f is a C^2 function of the z variable alone). An argument similar to the above then yields

$$\int \tilde{\mathcal{L}}f(z,q^*(\cdot|z))\hat{\varphi}(dz) = 0,\ f \in C_0^2(\mathcal{R}^d).$$

Thus $\hat{\varphi}(dz)$ is the unique stationary distribution for the averaged system controlled by the stable Markov control q^*, i.e., $\hat{\varphi} = \varphi^*$. Then

$$v^*(du|z,x)\hat{\zeta}(dzdx) = \Gamma^*(dzdxdu).$$

That is, $\tilde{\Phi}_\epsilon \to \Gamma^*$. By (6) and Theorem 8.3 of [11], $\int g d\Phi$ is uniformly bounded as Φ varies over $\mathcal{Q}$. By the second half of (**) and [10, Theorem 1.3.10, p. 10], it then follows that k' is uniformly integrable over $\mathcal{Q}$. Hence (21) holds. □

7. Main results—the general case. Now we drop (*). Define $v^*_\delta(du|z,x), \delta > 0$ small (say, $\delta \in (0, \delta_0]$), by

$$\int f v^*_\delta(du|z,x) \stackrel{def}{=} \int\int f v^*(du|z',x')\pi_\delta(z-z', x-x')dz'dx', \ f \in C(A'),$$

where $\{\pi_\delta : \mathcal{R}^{d+s} \to \mathcal{R}, \delta \in (0,\delta_0]\}$ are smooth approximations to the Dirac measure, i.e., compactly supported C^∞ probability density functions such that $\pi_\delta(z,x)dzdx \to \delta_{(0,0)}$ in $\mathcal{P}(\mathcal{R}^{d+s})$ as $\delta \downarrow 0$. In the following, $v^*_0(du|z,x) \stackrel{def}{=} v^*(du|z,x)$ and all quantities with subscript $\delta = 0$ correspond to it. Replace (***) by (A1), (A2) below.

(A1) $v^*_\delta(z,x) \stackrel{def}{=} v^*_\delta(du|z,x)$ is a stable Markov control for (1), (2) for $\delta \in [0,\delta_0], \epsilon \in (0,\epsilon_0)$. Furthermore, there exists a $\hat{g} \in C(\mathcal{R}^{d+s})$ satisfying

$$\sup_u |k'(z,x,u)|^{1+a} \le K\hat{g}(z,x) \tag{23}$$

such that the stationary distributions of (1), (2) corresponding to $\{v^*_\delta\}$, denoted by $\zeta^\epsilon_\delta(dzdx), 0 < \epsilon < \epsilon_0$, satisfy

$$\sup_{0<\epsilon<\epsilon_0} \int \hat{g}(z,x)\zeta^\epsilon_\delta(dzdx) < \infty \tag{24}$$

for each $\delta \in [0,\delta_0]$.

Once again in view of our nondegeneracy assumption, the transition probabilities for $t > 0$ of the time-homogeneous Markov process described by (11) under Markov control $v^*_\delta, \delta \in [0,\delta_0]$, have densities w.r.t. the Lebesgue measure. Therefore so do the corresponding invariant probability measures $\hat{\eta}_\delta(dx|z)$. Let $\chi_\delta(x|z)$ denote this density. Let $\bar{\mu}_\delta(dxdu|z) \stackrel{def}{=} \hat{\eta}_\delta(dx|z)v^*_\delta(du|z,x)$ and $\bar{\varphi}_\delta$ be the unique stationary distribution for (14) under the Markov control $\bar{\mu}_\delta$. Let $\zeta^0_\delta(dzdx) \stackrel{def}{=} \hat{\eta}_\delta(dx|z)\bar{\varphi}_\delta(dz)$ and $\Phi^0_\delta(dzdxdu) \stackrel{def}{=} \zeta^0_\delta(dzdx)v^*_\delta(du|z,x)$ for δ as above. Note that $\Phi^0_0 \in \mathcal{Q}_{opt}$. We also assume the following.

(A2) $\bar{\mu}_\delta(dxdu|z)$ is a stable Markov control for (14) for $\delta \in [0,\delta_0]$, and for $\hat{g}$ as above,

$$\sup_{\delta\in[0,\delta_0]} \int \hat{g}(z,x)\zeta^0_\delta(dzdx) < \infty. \tag{25}$$

LEMMA 7.1. *As $(\delta_n, z_n) \to (\delta, z)$ in $[0,\delta^*] \times \mathcal{R}^d$, $\hat{\eta}_{\delta_n}(dx|z_n) \to \hat{\eta}_\delta(dx|z)$ in total variation.*

Proof. This follows by an argument based on the Harnack inequality, as in the proof of Lemma 4.1, using the fact that $\chi_\delta(\cdot|z)$ will be equicontinuous and pointwise bounded. □

LEMMA 7.2. $\int k' d\Phi_\delta^0 \to \int k' d\Phi_0^0$ *as* $\delta \downarrow 0$.

Proof. By (4), (23), (25), and the Chebyshev inequality, $\zeta_\delta^0, \delta \in [0, \delta_0]$, and therefore $\bar{\varphi}_\delta, \delta \in [0, \delta_0]$, are tight. Let $\bar{\varphi}$ be any limit point of $\bar{\varphi}_\delta$ as $\delta \downarrow 0$. Since $\bar{\varphi}_\delta$ is characterized by

$$\int \tilde{\mathcal{L}} f(z, \bar{\mu}_\delta(\cdot|z)) \bar{\varphi}_\delta(dz) = 0, \ f \in C^2(\mathcal{R}^s), \tag{26}$$

an argument based on the Harnack inequality analogous to that of Lemma 4.1 implies that this convergence is in fact in total variation. Now for $f \in C_b(\mathcal{R}^d \times \mathcal{R}^s \times A')$,

$$\int f(z, x, u) v_\delta^*(du|z, x) \to \int f(z, x, u) v_0^*(du|z, x)$$

a.e. along a subsequence $\delta = \delta_m \downarrow 0$. (This convergence is in $L^1_{loc}(\mathcal{R}^{d+s})$ by [27, Theorem 2.16, p. 64] and therefore a.e. along a subsequence.) Hence by Lemma 7.1, along this subsequence,

$$\int\int f(z, x, u) v_\delta^*(du|z, x) \hat{\eta}_\delta(dx|z) \to \int\int f(z, x, u) v_0^*(du|z, x) \hat{\eta}_0(dx|z)$$

a.e., which in turn leads to

$$\begin{aligned} &\int\int\int f(z, x, u) v_\delta^*(du|z, x) \hat{\eta}_\delta(dx|z) \bar{\varphi}_\delta(dz) \\ &\quad \to \int\int\int f(z, x, u) v_0^*(du|z, x) \hat{\eta}_0(dx|z) \bar{\varphi}(dz). \end{aligned}$$

In particular, letting $\delta \downarrow 0$ along an appropriate subsequence in (26), we have

$$\int \tilde{\mathcal{L}} f(z, \bar{\mu}_0(\cdot|z)) \bar{\varphi}(dz) = 0, \ f \in C^2(\mathcal{R}^s), \tag{27}$$

i.e., $\bar{\varphi} = \bar{\varphi}_0$. Thus $\Phi_\delta^0 = \mu_\delta(dxdu|z)\bar{\varphi}_\delta(dz) \to \Phi_0^0 = \mu_0(dxdu|z)\bar{\varphi}_0(dz)$ as $\delta \downarrow 0$. Equations (23) and (25) and [10, Theorem 1.3.4, p. 10] ensure uniform integrability of k' under these, which in turn implies the claim. □

Going back to (1), (2), let $u(\cdot) = v_\delta^*(z^\epsilon(\cdot), x^\epsilon(\cdot))$ and let $\Phi_\delta^\epsilon \in \mathcal{P}(\mathcal{R}^d \times \mathcal{R}^s \times A')$ be the corresponding ergodic occupation measure for $\delta > 0$.

LEMMA 7.3. $\int k' d\Phi_\delta^\epsilon \to \int k' d\Phi_\delta^0$ *as* $\epsilon \downarrow 0$.

The proof follows along similar lines, using (24) in place of (25), and is omitted.

THEOREM 7.1. $\lim_{\epsilon \downarrow 0} \int k' d\Phi_\epsilon^* = \int k' d\Phi_0^*$.

Proof. Fix $\alpha > 0$ and take $\delta > 0$ small enough such that

$$\left| \int k' d\Phi_\delta^0 - \int k' d\Phi_0^0 \right| < \frac{\alpha}{2}.$$

Then pick $\epsilon > 0$ small enough so that

$$\left| \int k' d\Phi_\delta^\epsilon - \int k' d\Phi_\delta^0 \right| < \frac{\alpha}{2}.$$

Thus

$$\begin{aligned} \limsup_{\epsilon \downarrow 0} \int k' d\Phi_\epsilon^* &\le \limsup_{\epsilon \downarrow 0} \int k' d\Phi_\delta^\epsilon \\ &\le \int k' d\Phi_0^0 + \alpha. \end{aligned}$$

Since $\alpha > 0$ is arbitrary, the claim follows in view of Corollary 5.1. □

We conclude this section by pointing out a routine extension of the condition (4): It can be replaced by the weaker requirement

$$\lim_{||z||\to\infty} \inf_{x,u} k(z,x,u) > \beta^* \stackrel{def}{=} \sup_{0\le\epsilon<\epsilon_0} \beta^\epsilon \tag{28}$$

for some $\epsilon_0 > 0$, where $\beta^\epsilon, \epsilon > 0$, is the optimal cost for the ergodic control problem ($\epsilon = 0$ corresponds to the same for the averaged problem). This follows exactly along the lines of [9, Chapter VI]. The important thing to note is that the above condition suffices to ensure tightness of Q. Since in particular this presupposes that β^ϵ are uniformly bounded for $\epsilon \in (0, \epsilon_0)$, we may replace the "$\sup_{0\le\epsilon<\epsilon_0} \beta^\epsilon$" above by "$\sup_{0<\epsilon<\epsilon_0} \beta^\epsilon$" in view of Theorem 5.1.

8. The stable case. We briefly indicate the corresponding developments when a blanket stability condition is available. We do not assume (4) or its generalization (28), but require that k' be bounded from below. Suppose for $\epsilon \in (0, \epsilon_0)$ there exist $\Delta_\epsilon, a_\epsilon > 0, B \subset \mathcal{R}^{d+s}$ bounded and $V_\epsilon^{(i)} \in C^2(\mathcal{R}^{d+s}), i = 1, 2$, such that $V_\epsilon^{(i)} \ge 0, \lim_{||(z,x)||\to\infty} V_\epsilon^{(i)}(z,x) = \infty$ for $i = 1, 2$, and,

$$\hat{\mathcal{L}} V_\epsilon^{(1)}(z,x,u) \le -\Delta_\epsilon, \tag{29}$$

$$\hat{\mathcal{L}} V_\epsilon^{(2)}(z,x,u) \le -a_\epsilon V_\epsilon^{(1)}(z,x) \tag{30}$$

for $(z,x) \notin B$. Let $\tau \stackrel{def}{=} \inf\{t \ge 0 : (z^\epsilon(t), x^\epsilon(t)) \in B\}$ and $\bar{\tau}_N \stackrel{def}{=} \inf\{t \ge 0 : ||(z^\epsilon(t), x^\epsilon(t))|| > N\}, N \ge 1$. Then $\bar{\tau}_N \uparrow \infty$ a.s., and for $(z_0, x_0) \notin B$, the Ito–Dynkin formula and (29) lead to

$$E[V_\epsilon^{(1)}(z^\epsilon(\tau \wedge \bar{\tau}_N), x^\epsilon(\tau \wedge \bar{\tau}_N))] - V_\epsilon^{(1)}(z_0, x_0) \le -\Delta_\epsilon E[\tau \wedge \bar{\tau}_N].$$

Letting $N \uparrow \infty$ and rearranging terms, we have

$$E[\tau] \le \frac{V_\epsilon^{(1)}(z_0, x_0)}{\Delta_\epsilon}.$$

Similarly from (30) we get

$$E\left[\int_0^{\tau\wedge\bar{\tau}_N} V_\epsilon^{(1)}(z^\epsilon(t), x^\epsilon(t))dt\right] \le \frac{V_\epsilon^{(2)}(z_0, x_0)}{a_\epsilon},$$

and therefore

$$\begin{aligned}
&\frac{1}{2}E[(\tau \wedge \bar{\tau}_N)^2] \\
&= E\left[\int_0^{\tau\wedge\bar{\tau}_N} (\tau \wedge \bar{\tau}_N - t)dt\right] \\
&= E\left[\int_0^{\tau\wedge\bar{\tau}_N} E[\tau \wedge \bar{\tau}_N - t|\mathcal{F}_t]dt\right] \\
&\le \frac{1}{\Delta_\epsilon} E\left[\int_0^{\tau\wedge\bar{\tau}_N} V_\epsilon^{(1)}(z^\epsilon(t), x^\epsilon(t))dt\right] \\
&\le \frac{1}{a_\epsilon \Delta_\epsilon} V^{(2)}(z_0, x_0).
\end{aligned}$$

Letting $N \uparrow \infty$,

$$E[\tau^2] \leq \frac{2}{a_\epsilon \Delta_\epsilon} V^{(2)}(z_0, x_0).$$

In view of this, one can argue as in [9, Chapter VI] to conclude Theorem 3.1. The key step is that the above inequality ensures that $\{\nu_t\}$ remain a.s. tight (so do $\{\bar{\nu}_t\}$), and thus Lemma 3.2 can be strengthened to the claim

$$\nu_t \to \mathcal{G} \text{ a.s.}, \qquad \bar{\nu}_t \to \mathcal{G}.$$

Furthermore, $\mathcal{G}$ can be shown to be compact. This allows one to conclude Theorem 3.1 without the need to use (4). Conditions similar to (29), (30) imposed on (14) ensure Theorem 4.1. Next, for obtaining the counterparts of the results of section 6 above for the affine case, assume the additional conditions stipulated in [9, section VI.4] to ensure the existence of C^2 value functions for the two ergodic control problems featured in the proof of Lemma 6.1. The rest remains as before. We omit the details, as they are straightforward adaptations of the foregoing.

9. Some future directions. We conclude by pointing out some further possibilities. We have not allowed σ to depend on the control variable u, or γ to depend on either u or x. This is because such a dependence would lead to a diffusion matrix which is measurable but not necessarily continuous under a Markov control (for the averaged system in the latter case). Even in the nondegenerate case, only the existence of weak solutions is known for this level of generality, not their uniqueness [24], [29]. It may be possible to work with a "set of all weak solutions" in place of the unique weak solution and extend the foregoing. In the degenerate case, even with the existing form of (1), (2), there are problems. The results of [8] extend the characterization of ergodic occupation measures from Lemma 3.1, which allows us to prove Theorem 5.1 under suitable hypotheses. But Theorem 6.1 is a more difficult proposition due to a lack of ergodicity and other problems.

REFERENCES

[1] O. Alvarez and M. Bardi, *Viscosity solutions methods for singular perturbations in deterministic and stochastic control*, SIAM J. Control Optim., 40 (2001), pp. 1159–1188.

[2] O. Alvarez and M. Bardi, *Singular perturbations of nonlinear degenerate parabolic PDEs: A general convergence result*, Arch. Ration. Mech. Anal., 170 (2003), pp. 17–61.

[3] Z. Artstein, *Invariant measures and their projections in nonautonomous dynamical systems*, Stoch. Dyn., 4 (2004), pp. 439–459.

[4] Z. Artstein and V. Gaitsgory, *Convergence to convex sets in infinite dimensions*, J. Math. Anal. Appl., 284 (2003), pp. 471–480.

[5] Z. Artstein and A. Leizarowitz, *Singularly perturbed control systems with one-dimensional fast dynamics*, SIAM J. Control Optim., 41 (2002), pp. 641–658.

[6] K. E. Avrachenkov, J. Filar, and M. Haviv, *Singular perturbations of Markov chains and decision processes*, in Handbook of Markov Decision Processes. Methods and Applications, E. Feinberg and A. Shwartz, eds., Kluwer Acad. Publ., Boston, 2002, pp. 113–150.

[7] A. Bensoussan, *Perturbation Methods in Optimal Control*, John Wiley, Chichester, 1988.

[8] A. G. Bhatt and V. S. Borkar, *Occupation measures for controlled Markov processes: Characterization and optimality*, Ann. Probab., 24 (1996), pp. 1531–1562.

[9] V. S. Borkar, *Optimal Control of Diffusion Processes*, Pitman Res. Notes Math. Ser. 203, Longman Scientific and Technical, Harlow, UK, 1989.

[10] V. S. Borkar, *Probability Theory: An Advanced Course*, Springer-Verlag, New York, 1995.

[11] V. S. Borkar, *Uniform stability of controlled Markov processes*, in System Theory: Modeling, Analysis and Control, T. E. Djaferis and I. C. Schick, eds., Kluwer Acad. Publ., Boston, 2000, pp. 107–120.

[12] V. Borkar and V. Gaitsgory, *On existence of limit occupational measures set of a controlled stochastic differential equation*, SIAM J. Control Optim., 44 (2005), pp. 1436–1473.
[13] V. S. Borkar and M. K. Ghosh, *Controlled diffusions with constraints*, J. Math. Anal. Appl., 152 (1990), pp. 88–108.
[14] F. Colonius and R. Fabri, *Controllability for systems with slowly varying parameters*, ESAIM: Control Optim. Calc. Var., 9 (2003), pp. 207–216.
[15] T. D. Donchev and A. L. Dontchev, *Singular perturbations in infinite-dimensional control systems*, SIAM J. Control Optim., 42 (2003), pp. 1795–1812.
[16] J. A. Filar, V. Gaitsgory, and A. Haurie, *Control of singularly perturbed hybrid stochastic systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 179–190.
[17] V. Gaitsgory, *On a representation of the limit occupational measures set of a control system with applications to singularly perturbed control systems*, SIAM J. Control Optim., 43 (2004), pp. 325–340.
[18] V. Gaitsgory and M.-T. Nguyen, *Multiscale singularly perturbed control systems: Limit occupational measures sets and averaging*, SIAM J. Control Optim., 41 (2002), pp. 954–974.
[19] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, Heidelberg, 1998.
[20] G. Grammel, *On nonlinear control systems with multiple time scales*, J. Dynam. Control Systems, 10 (2004), pp. 11–28.
[21] Y. Kabanov and S. Pergamenschikov, *Two-scale Stochastic Systems*, Springer-Verlag, New York, Berlin, Heidelberg, 2002.
[22] R. Z. Khasminskii and G. Yin, *On averaging principles: An asymptotic expansion approach*, SIAM J. Math. Anal., 35 (2004), pp. 1534–1560.
[23] P. V. Kokotovic, H. K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, New York, 1986.
[24] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, Berlin, Heidelberg, 1980.
[25] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser Boston, Boston, MA, 1990.
[26] A. Leizarowitz, *Order reduction is invalid for singularly perturbed control problems with vector fast variables*, Math. Control Signals Systems, 15 (2002), pp. 101–119.
[27] E. H. Lieb and M. Loss, *Analysis*, 2nd ed., AMS, Providence, RI, 2001.
[28] R. S. Liptser and A. N. Shiryayev, *Theory of Martingales*, Kluwer Acad. Publ., Dordrecht, 1989.
[29] N. Nadirashvili, *Nonuniqueness in the martingale problem and the Dirichlet problem for uniformly elliptic operators*, Ann. Scoula Norm. Sup. Pisa Cl. Sc. (4), 24 (1997), pp. 537–549.
[30] S. D. Naidu, *Singular perturbations and time scales in control theory and applications: An overview*, Dyn. Contin. Discrete Impuls. Syst. Ser. B Appl. Algorithms, 9 (2002), pp. 233–278.
[31] R. E. O'Malley, Jr., *Singular perturbations and optimal control*, in Mathematical Control Theory, W. A. Copel, ed., Lecture Notes in Math. 680, Springer-Verlag, Berlin, 1978, pp. 170–218.
[32] M. Quincampoix and F. Watbled, *Averaging method for discontinuous Mayer's problem of singularly perturbed control systems*, Nonlinear Anal.: Theory Methods Appl., 54 (2003), pp. 819–837.
[33] G. G. Yin and Q. Zhang, *Discrete-Time Markov Chains: Two-Time-Scale Methods and Applications*, Springer-Verlag, New York, 2004.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1578–1614

EXACT CONTROLLABILITY FOR MULTIDIMENSIONAL SEMILINEAR HYPERBOLIC EQUATIONS*

XIAOYU FU†, JIONGMIN YONG‡, AND XU ZHANG§

Abstract. In this paper, we obtain a global exact controllability result for a class of multidimensional semilinear hyperbolic equations with a superlinear nonlinearity and variable coefficients. For this purpose, we establish an observability estimate for the linear hyperbolic equation with an unbounded potential, in which the crucial observability constant is estimated explicitly by a function of the norm of the potential. Such an estimate is obtained by a combination of a pointwise estimate and a global Carleman estimate for the hyperbolic differential operators and analysis on the regularity of the optimal solution to an auxiliary optimal control problem.

Key words. exact controllability, semilinear hyperbolic equation, superlinear growth, observability inequality, global Carleman estimate

AMS subject classifications. Primary, 93B05; Secondary, 93B07, 35B37

DOI. 10.1137/040610222

1. Introduction.

Given $T > 0$ and a bounded domain Ω of $\mathbb{R}^n$ ($n \in \mathbb{N}$) with C^2 boundary Γ, put $Q = (0,T)\times\Omega$ and $\Sigma = (0,T)\times\Gamma$. Let ω be a proper open nonempty subset of Ω and denote by χ_ω the characteristic function of ω. For any set $M \subset \mathbb{R}^n$ and $\delta > 0$, we define $\mathcal{O}_\delta(M) = \big\{x \in \mathbb{R}^n \,\big|\, |x - x'| < \delta \text{ for some } x' \in M\big\}$. Also, we denote $\sum_{i,j=1}^n$ and $\sum_{i=1}^n$ simply by $\sum_{i,j}$ and $\sum_i$, respectively. For simplicity, we will use the notation $y_i = y_{x_i}$, where x_i is the ith coordinate of a generic point $x = (x_1, \dots, x_n)$ in $\mathbb{R}^n$. In a similar manner, we use the notation w_i, v_i, etc. for the partial derivatives of w and v with respect to x_i. On the other hand, for any domain M in $\mathbb{R}^n$ (even without any regularity condition on its boundary ∂M), we refer to [1, Chap. 3] for the definition and basic properties of the Sobolev spaces $H_0^1(M)$, $H^{-1}(M)$, etc. (Hence, $H_0^1(Q)$ and $H^{-1}(Q)$ are particularly well defined in [1, Chap. 3].)

Let $a^{ij}(\cdot) \in C^1(\overline{\Omega})$ be fixed, satisfying

$$a^{ij}(x) = a^{ji}(x) \quad \forall\, x \in \overline{\Omega}, \qquad i,j = 1,2,\dots,n, \tag{1.1}$$

and for some constant $\beta > 0$,

$$\sum_{i,j} a^{ij}(x)\xi^i\xi^j \geq \beta|\xi|^2 \quad \forall\, (x,\xi) \in \overline{\Omega}\times\mathbb{R}^n, \tag{1.2}$$

where $\xi = (\xi^1, \dots, \xi^n)$. In what follows, put $A \stackrel{\triangle}{=} (a^{ij})_{n\times n}$. We define a differential

*Received by the editors June 19, 2004; accepted for publication (in revised form) March 7, 2007; published electronically October 17, 2007. This work was partially supported by NSFC grants 10131030 and 10525105, Chinese Education Ministry Science Foundation grant 2000024605, the Cheung Kong Scholars Programme, NSF grant DMS-0604309, NCET of China grant NCET-04-0882, and Spanish MEC grant MTM2005-00714.

http://www.siam.org/journals/sicon/46-5/61022.html

†School of Mathematics, Sichuan University, Chengdu 610064, China (rj_xy@163.com).

‡Department of Mathematics, University of Central Florida, Orlando, FL 32816, and School of Mathematical Sciences, Fudan University, Shanghai 200433, China (jyong@mail.ucf.edu).

§Yangtze Center of Mathematics, Sichuan University, Chengdu 610064, China, and Key Laboratory of Systems and Control, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100080, China (xuzhang@amss.ac.cn).

operator $\mathcal{P}$ by

$$\mathcal{P}y \triangleq y_{tt} - \sum_{i,j} \left(a^{ij}(x)y_i\right)_j . \tag{1.3}$$

Next, we fix a function $f(\cdot) \in C^1(\mathbb{R})$, satisfying the following condition:

$$\overline{\lim_{s\to\infty}} \frac{f(s)}{s \ln^{1/2} |s|} = 0. \tag{1.4}$$

Note that $f(\cdot)$ in the above can have a superlinear growth. We consider the following controlled semilinear hyperbolic equation with an internal local controller acting on ω:

$$\begin{cases} \mathcal{P}y = f(y) + \chi_\omega(x)\gamma(t,x) & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in } \Omega. \end{cases} \tag{1.5}$$

In (1.5), $(y(t,\cdot), y_t(t,\cdot))$ is the *state*, and $\gamma(t,\cdot)$ is the *control* which acts on the system through the subset ω of Ω. In what follows, we choose the *state space* and the *control space* of system (1.5) to be $H_0^1(\Omega) \times L^2(\Omega)$ and $L^2((0,T)\times\omega)$, respectively. We point out that some other choices of spaces are possible. But our choice is natural in the context of the hyperbolic equations. The space $H_0^1(\Omega) \times L^2(\Omega)$ is often referred to as the *finite energy space*. For any $(y_0, y_1) \in H_0^1(\Omega) \times L^2(\Omega)$ and $\gamma \in L^2((0,T)\times\Omega)$, using the method in [4] one can prove the global existence of a unique weak solution $y \in C([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega))$ for (1.5) under assumption (1.1)–(1.2), and under (1.4) with $f(\cdot) \in C^1(\mathbb{R})$.

The main purpose of this paper is to study the global exact controllability of (1.5), by which we mean the following: For any given $(y_0, y_1), (z_0, z_1) \in H_0^1(\Omega) \times L^2(\Omega)$, find a control $\gamma \in L^2((0,T)\times\omega)$ such that the corresponding weak solution y of (1.5) satisfies

$$y(T) = z_0, \quad y_t(T) = z_1 \qquad \text{in } \Omega. \tag{1.6}$$

Due to the finite propagation speed of solutions to hyperbolic equations, the "waiting time" T has to be large enough. The estimate of T is also a part of the problem.

The problem of exact controllability for linear hyperbolic equations (for example, $f(\cdot)$ is a linear function, or simply, $f(\cdot) \equiv 0$ in (1.5)) has been studied by many authors. We mention here some standard references, for example, [2, 29, 33].

The study of exact controllability problems for nonlinear hyperbolic equations began in the 1960s. Early works, including [5, 6, 10] and so on, were mainly devoted to the local controllability problem, by which we mean that the controllability property was proved under some smallness assumptions on the initial data and/or the final target. In [43], further local results were proved for the exact controllability of some semilinear wave equations in the form of (1.5) with $A = I$, the identity matrix, and under a very general assumption on the nonlinearity $f(\cdot)$ (which allows $f(\cdot)$ to be local Lipschitz continuous). We refer to [27] and the references cited therein for some recent local controllability results of certain quasi-linear hyperbolic systems.

A global boundary exact controllability result for semilinear wave equations, corresponding to (1.5), in the state spaces $H_0^r(\Omega) \times H^{r-1}(\Omega)$ $(r \in (0,1/2) \cup (1/2,1))$ or $H_{00}^{1/2}(\Omega) \times (H_{00}^{1/2}(\Omega))'$, with Dirichlet boundary control, was given in [44] under the

assumption that $A = I$ and that the nonlinearity $f(\cdot)$ is globally Lipschitz continuous, i.e., $f'(\cdot) \in L^\infty(\mathbb{R})$. In [23], this controllability result was improved to include the critical points $r = 0$ and 1, and also extended to the abstract setting. Recent progress in this respect can be found in [36] and [37]. In the case that $f(\cdot)$ is sublinear, we refer to [34] for the global exact controllability of (1.5).

As for the case that $f(\cdot)$ grows superlinearly at infinity, very little is known for the global exact controllability of the semilinear hyperbolic equation (1.5) except for the one space dimension, i.e., $n = 1$. We refer to [3, 9, 30, 45] for related one-dimensional results. To our best knowledge, in the superlinear setting, [26] is the only paper that discussed the global exact controllability for multidimensional system (1.5) (we refer to [42] for an updated survey on this problem). By assuming that $A = I$ and $\omega = \mathcal{O}_\delta(\Gamma) \cap \Omega$ for some $\delta > 0$, [26] shows that system (1.5) with $f(\cdot)$ satisfying (1.4) is exactly controllable. In this paper, based on a method which is different from [26], we shall consider a more general case by using a smaller controller $\omega = \mathcal{O}_\delta(\Gamma_+) \cap \Omega$ (see (2.5) for Γ_+) and allowing the coefficients matrix A to be nonconstant one. We refer the reader to Condition 2.1 and the subsequent remarks, and especially Proposition 2.1, for assumptions on matrix A.

In order to obtain the exact controllability of (1.5), one needs to consider, by the well-known duality argument (see [29], [28, Lemma 2.4, p. 282], and [39, Theorem 3.2, p. 19], for example), the following dual system of the linearized system of (1.5):

$$\left\{\begin{array}{ll} \mathcal{P}w = qw & \text{in } Q, \\ w = 0 & \text{on } \Sigma, \\ (w(0), w_t(0)) = & (w_0, w_1) \in L^2(\Omega) \times H^{-1}(\Omega), \end{array}\right. \tag{1.7}$$

with a potential q in some space (larger than $L^\infty(Q)$, in general). It follows from the standard perturbation theorem in the semigroup theory [31] that for a suitable q, say $q \in L^\infty(0, T; L^n(\Omega))$, (1.7) is well-posed in $L^2(\Omega) \times H^{-1}(\Omega)$.

Similar to [45] and [26], the above controllability problem may be reduced to an explicit observability estimate for system (1.7). Namely, we expect to find a constant $\mathcal{C}(q) > 0$ such that all weak solutions w of (1.7) satisfy

$$|(w_0, w_1)|_{L^2(\Omega)\times H^{-1}(\Omega)} \leq \mathcal{C}(q)|w|_{L^2((0,T)\times\omega)} \ \forall\ (w_0, w_1) \in L^2(\Omega) \times H^{-1}(\Omega). \tag{1.8}$$

The explicit estimate of $\mathcal{C}(q)$ in terms of a suitable norm of the potential q is an indispensable part of the problem, which is actually the key novelty in this paper. Similar problems for $A = I$ and bounded potentials q were considered in [36, 37]. However, in the present case we cannot assume that q in (1.7) is bounded since we do not assume that the nonlinearity $f(\cdot)$ in (1.5) is globally Lipschitz continuous. To overcome this difficulty, we need, among other things, to combine some ideas found in [18] and [37].

It is well known that the Carleman estimate is one of the major tools used in the study of unique continuation, observability, and controllability problems for various kinds of partial differential equations (PDEs). However, the "concrete" Carleman estimate for these problems is actually quite different! Indeed, in principle, among these problems unique continuation is the "easiest," and one may develop an abstract theory for the unique continuation property (usually, of local nature) for very general partial differential operators, based on a pseudoconvexity condition, the Carleman estimate, and by means of the microlocal analysis technique [16, 17, 35]. Observability is, however, a quantitative version of the global unique continuation, which is much

more difficult to establish than the classical (qualitative) unique continuation. For example, the unique continuation for the parabolic equations was known for a very long time, but the observability for the same equation was not established until the 1990s by means of a new Carleman estimate [11, 14]. Also, for the hyperbolic equations, the work of [20, 21] applied Carleman estimates for the proofs of the observability results. On the other hand, there are many equations (say, the hyperbolic-parabolic coupled systems in [32]), for which one can easily establish its unique continuation, but its observability is completely unknown for multidimensions (the analysis for the one-dimensional problem [41] is highly nontrivial, and some atypical phenomenon occurs). Finally, as for controllability problems, as mentioned before the classical duality argument reduces the problem to obtaining a suitable observability estimate. However, for the global controllability problems for semilinear PDEs with superlinear growing nonlinearity, the key point is the explicit estimate of the observability constant by a suitable function of the norm of the potential. For this purpose, one has to proceed more carefully than one would for the usual observability when using the Carleman estimate. Note also that the approach developed in this article seems to be virtually complete. Our key estimate on the observability constant $\mathcal{C}(q)$ is presented in (2.12) of Theorem 2.3. As suggested by [8, Theorem 1.2], it may well be that (2.12) is sharp (see also our Remark 2.1). In this respect, it is worth mentioning that one can also adopt the method developed in [20, 21, 22] to establish an explicit observability estimate for some special case of system (1.7) (i.e., $A = aI$ with a suitable positive function a), as done in [36]. However, it seems that the estimate obtained in this way is far from sharp. Indeed, the estimate on the observability constant $\mathcal{C}(q)$ obtained in [36] (for bounded potential q) reads as $C\exp(\exp(\exp(Cr_0)))$ with $r_0 = |q|_{L^\infty(Q)}$, which is much weaker than that in (2.12). It would be quite interesting to check whether the method in [20, 21, 22] can be adopted to derive the same estimate as that of (2.12) in Theorem 2.3. But this remains to be done.

The rest of this paper is organized as follows. In section 2, we shall state the main results. Some preliminary results are collected in section 3. In section 4, we derive an estimate for second order differential operators with symmetric coefficients that is of independent interest. This estimate will play a key role when we establish in section 5 a global Carleman estimate for the hyperbolic differential operators in $H_0^1(Q)$. The latter estimate, in turn, is one of the crucial preliminary results we derive in section 7, i.e., a similar global Carleman estimate for the hyperbolic differential operators in a larger space $L^2(Q)$. Another crucial preliminary we study, in section 6, is an auxiliary optimal control problem, where the key point is to obtain some regularity of the optimal solution. In sections 8–9, we will prove our main results. Finally, Appendices A, B, and C are devoted to proving some technical results that are used throughout the paper.

2. Statement of the main results. To begin, we introduce the following condition.

CONDITION 2.1. *There exists a function $d(\cdot) \in C^2(\overline{\Omega})$ satisfying the following:* (i) *For some constant $\mu_0 > 0$, it holds that*

$$\sum_{i,j}\left\{\sum_{i',j'}\left[2a^{ij'}(a^{i'j}d_{i'})_{j'} - a^{ij}_{j'}a^{i'j'}d_{i'}\right]\right\}\xi^i\xi^j \geq \mu_0\sum_{i,j}a^{ij}\xi^i\xi^j \quad \forall\,(x,\xi^1,\dots,\xi^n) \in \overline{\Omega}\times\mathbb{R}^n. \tag{2.1}$$

(ii) *There is no critical point of function* $d(\cdot)$ *in* $\overline{\Omega}$, *i.e.,*

$$\min_{x\in\overline{\Omega}} |\nabla d(x)| > 0. \tag{2.2}$$

Let us make some remarks on the above condition.

First, Condition 2.1 is really a restriction on the coefficient matrix A and the domain Ω. Indeed, as we shall see later, Condition 2.1 at least leads to the exact controllability of system (1.5) with $f(\cdot) \equiv 0$ and $\omega = \mathcal{O}_\delta(\Gamma)\bigcap\Omega$ for any given $\delta > 0$ and sufficiently large "waiting time" $T > 0$, while it is shown in [2] that, in order for the latter to hold, (T, Ω, ω) has to satisfy a geometric optics condition which is characterized by the null bicharacteristic of operator $\mathcal{P}$. But, for any $T > 0$, this condition may fail to be true for some $\mathcal{P}$ (with special coefficients) and some (Ω, ω) (see [2]). This condition is crucial in what follows, where we derive a Carleman estimate for the hyperbolic operators (see (11.4)). Nevertheless, to the best of our knowledge, there is no universal tractable Carleman estimates in the literature for general hyperbolic operators. We shall give below some tractable examples. However, a detailed analysis of Condition 2.1 is beyond the scope of this paper and will be presented elsewhere.

Second, by (1.1)–(1.2), one can check that (2.1) is equivalent to the uniform positivity of the following (symmetric) matrix:

$$\begin{aligned}\mathcal{A} &\triangleq \left(\sum_{i',j'}\left(a^{ij'}a^{i'j}d_{i'j'} + \frac{(a^{ij'}a_{j'}^{i'j} + a^{jj'}a_{j'}^{i'i} - a_{j'}^{ij}a^{i'j'})d_{i'}}{2}\right)\right)_{1\le i,j\le n} \\ &\equiv A\mathcal{H}_d A + \frac{1}{2}\left(\sum_{i',j'}(a^{ij'}a_{j'}^{i'j} + a^{jj'}a_{j'}^{i'i} - a_{j'}^{ij}a^{i'j'})d_{i'}\right)_{1\le i,j\le n},\end{aligned} \tag{2.3}$$

where $\mathcal{H}_d$ is the Hessian matrix of $d(\cdot)$. Hence, if A is a constant matrix, then $\mathcal{A} = A\mathcal{H}_d A$, and (2.1) is reduced to the (uniformly) strict convexity of $d(x)$. A little further, for any uniformly strict convex function $d(\cdot) \in C^2(\overline{\Omega})$, one can show that the matrix $A\mathcal{H}_d A$ is uniformly positive definite. Therefore, if

$$\max_{1\le i,j,k\le n} \sup_{x\in\overline{\Omega}} |a_k^{ij}(x)| \text{ is small enough,} \tag{2.4}$$

one concludes that $\mathcal{A}$ is uniformly positive definite. Consequently, if in addition, $d(\cdot)$ satisfies (2.2), then Condition 2.1 holds for $d(\cdot)$.

Third, the above remark, especially (2.4), does not mean that Condition 2.1 can hold only for coefficient matrices A which are close to constant matrices. To illustrate this, let us state the following proposition, whose proof is presented in Appendix A.

PROPOSITION 2.1. *Let* $n = 2$, *and let* $A = \operatorname{diag}[a^1, a^2]$ *with* $a^1 \in C^2(\overline{\Omega})$ *and* $a^2 \in C^1(\overline{\Omega})$ *being uniformly positive functions. Assume further that*

(i) $a^1(x_1, x_2) \equiv a^1(x_1)$, *i.e., it is independent of* x_2;

(ii) $a_1^1 a_1^2 \ge 0$ *in* Ω; *and*

(iii) *there is at most one point* $x_1^0 \in G \triangleq \overline{\{x_1 \in \mathbb{R} \big| (x_1, x_2) \in \Omega \text{ for some } x_2 \in \mathbb{R}\}}$ *so that* $a_1^1(x_1^0) = 0$. *Moreover, if such an* x_1^0 *exists, it satisfies* $a_{11}^1(x_1^0) < 0$. *Then Condition* 2.1 *holds.*

We emphasize that in the above, the derivatives $a_1^1(\cdot)$, $a_1^2(\cdot)$, and $a_2^2(\cdot)$ are not necessarily small. Therefore, the matrix A is not necessarily close to a constant matrix.

As a more concrete case, let us look at the following situation: Let $a(x_1) \in C^2(G)$ be a uniformly positive and strictly concave function. One may check that if $a^1(x_1, x_2) \equiv a^2(x_1, x_2) \equiv a(x_1)$, then a^1 and a^2 satisfy the conditions in Proposition 2.1. What is more interesting is that for this nonidentity matrix $A = aI$, if a_1, the derivative of a with respect to x_1, changes sign, then one may further check that it does not satisfy the geometric condition introduced in [22, Theorem 2.2.4] (which, in our notation, reads $\frac{1}{2} - (x - x_0) \cdot \nabla a \geq 0$ in $\overline{\Omega}$, for some $x_0 \in \mathbb{R}^n$) unless the length of G, or the positive part of a_1 in G (i.e., $\max_{x \in G} a_1^+(x)$), or the negative part of a_1 in G (i.e., $\max_{x \in G} a_1^-(x)$), is assumed to be sufficiently small. Hence, we have found a class of explicit and nontrivial examples satisfying our Condition 2.1. Also, we indicate that it is possible to construct nontrivial examples of nondiagonal coefficient matrices that satisfy Condition 2.1.

For the function $d(\cdot)$ satisfying Condition 2.1, we introduce the following set:

$$\Gamma_+ \triangleq \left\{ x \in \Gamma \;\middle|\; \sum_{i,j} a^{ij} \nu_i d_j > 0 \right\}, \tag{2.5}$$

where $\nu = \nu(x) = (\nu_1, \nu_2, \ldots, \nu_n)$ is the unit outward normal vector of Ω at $x \in \Gamma$.

Note that for the case $A = I$, by choosing $d(x) = |x - x_0|^2$ with any given $x_0 \in \mathbb{R}^n \setminus \overline{\Omega}$, we have Condition 2.1 with $\mu_0 = 4$, and (2.1) holds with an equality. In this case,

$$\Gamma_+ = \Big\{ x \in \Gamma \;\Big|\; (x - x_0) \cdot \nu(x) > 0 \Big\},$$

which coincides with the usual star-shaped part of the whole boundary of Ω [29].

On the other hand, it is easy to check that, if $d(\cdot) \in C^2(\overline{\Omega})$ satisfies (2.1), then for any given constants $a \geq 1$ and $b \in \mathbb{R}$, the function

$$\hat{d} = \hat{d}(x) \triangleq a d(x) + b \tag{2.6}$$

(scaling and translating $d(x)$) still satisfies Condition 2.1 with μ_0 replaced by $a\mu_0$; meanwhile, the scaling and translating $d(x)$ do not change the set Γ_+. Hence, by scaling and translating $d(x)$, if necessary, we may assume without loss of generality that

$$\begin{cases} \text{(2.1) holds with } \mu_0 \geq 4, \\ \dfrac{1}{4} \displaystyle\sum_{i,j} a^{ij}(x) d_i(x) d_j(x) \geq \max_{x \in \overline{\Omega}} d(x) \geq \min_{x \in \overline{\Omega}} d(x) > 0 \qquad \forall\, x \in \overline{\Omega}. \end{cases} \tag{2.7}$$

In what follows, we let

$$R_1 \triangleq \max_{x \in \overline{\Omega}} \sqrt{d(x)}, \qquad T_* \triangleq 2 \inf \Big\{ R_1 \;\Big|\; d(\cdot) \text{ satisfies (2.7)} \Big\}. \tag{2.8}$$

Concerning the controller ω in (1.5), we need the following assumption.

CONDITION 2.2. *There is a constant $\delta > 0$ such that*

$$\omega = \mathcal{O}_\delta(\Gamma_+) \bigcap \Omega. \tag{2.9}$$

Note that condition (2.9) can be replaced by

$$\omega \supseteq \overline{\Gamma}_+, \tag{2.10}$$

which looks much weaker. In fact, when (2.10) holds, one can find a $\delta > 0$ such that

$$\omega \supseteq \mathcal{O}_\delta(\Gamma_+) \bigcap \Omega. \tag{2.11}$$

It is not hard to see that if we can prove the controllability for (1.5) with a smaller controller ω satisfying (2.9), then we can do so for a larger controller ω satisfying (2.11) (in particular, we can choose ω to be $\mathcal{O}_\delta(\Gamma) \bigcap \Omega$, a neighborhood of the whole boundary Γ). We assume an equality in (2.9) only for simplicity of presentation.

The main controllability result in this paper is stated as follows.

THEOREM 2.2. *Let $a^{ij}(\cdot) \in C^1(\overline{\Omega})$ satisfy* (1.1)–(1.2)*, and let $f(\cdot) \in C^1(\mathbb{R})$ satisfy* (1.4)*. Let Conditions* 2.1–2.2 *hold. Then for any $T > T_*$, system* (1.5) *is exactly controllable in $H_0^1(\Omega) \times L^2(\Omega)$ at time T by using some control $\gamma \in L^2((0,T) \times \omega)$.*

In what follows, we will use C to denote a generic positive constant which may vary from line to line (unless otherwise stated). As we mentioned before, the proof of Theorem 2.2 can be reduced to the following observability estimate result for system (1.7).

THEOREM 2.3. *Let $a^{ij}(\cdot) \in C^1(\overline{\Omega})$ satisfy* (1.1)–(1.2)*, $q \in L^\infty(0,T; L^n(\Omega))$, and Conditions* 2.1–2.2 *hold. Then for any $T > T_*$, all weak solutions w of system* (1.7) *satisfy estimate* (1.8) *with an observability constant $\mathcal{C}(q) > 0$ of the form*

$$\mathcal{C}(q) = C \exp(Cr^2), \tag{2.12}$$

where

$$r = |q|_{L^\infty(0,T;L^n(\Omega))}. \tag{2.13}$$

Several remarks are in order.

Remark 2.1. By adopting the approach developed in this paper, Theorem 2.3 is strengthened in [8] as follows (see [8, Theorem 2.2]): Replace the assumption on q by $q \in L^\infty(0,T; L^s(\Omega))$ for any fixed $s \in [n, \infty]$ and let the other assumptions in Theorem 2.3 remain unchanged. Then for any $T > T_*$, all weak solutions w of system (1.7) satisfy estimate (1.8) with an observability constant $\mathcal{C}(q) > 0$ of the form

$$\mathcal{C}(q) = C \exp\left(C|q|^{\frac{1}{3/2-n/s}}_{L^\infty(0,T;L^s(\Omega))} \right). \tag{2.14}$$

On the other hand, it is shown in [8, Theorem 1.2] that the exponent $2/3$ in the estimate $|q|^{2/3}_{L^\infty(0,T;L^\infty(\Omega))}$ (in (2.14) for the special case $s = \infty$) is sharp. Although the problem of the optimality of the exponent $\frac{1}{3/2-n/s}$ in $|q|^{\frac{1}{3/2-n/s}}_{L^\infty(0,T;L^s(\Omega))}$ is unsolved when $s \in [n, \infty)$, [8, Theorem 1.2] does support the idea that the exponent 2 of the estimate r^2 in (2.12) might be sharp.

Remark 2.2. The "minimal" waiting time T_* in Theorems 2.2–2.3 is explicitly constructed (by (2.8)) but not sharp. The sharp T_*, as suggested by the special case $A = I$ considered in [36, 37], should be given as follows:

$$T_* \triangleq 2 \inf \Bigg\{ R_1 \;\Big|\; d(x) \text{ satisfies (2.1) with } \mu_0 \geq 4 \text{ and } \frac{1}{4} \sum_{i,j} a^{ij}(x) d_i(x) d_j(x) \geq d(x) \geq \min_{x \in \overline{\Omega}} d(x) > 0 \quad \forall\, x \in \overline{\Omega} \Bigg\},$$

i.e., one replaces the term $\max_{x\in\overline{\Omega}} d(x)$ in (2.7) by $d(x)$. Unfortunately, we are unable to obtain such a sharp waiting time at this moment. One will see that the inequality involving $\sum_{i,j} a^{ij}(x)d_i(x)d_j(x)$ and $\max_{x\in\overline{\Omega}} d(x)$ in (2.7) plays a key role in (11.7).

Remark 2.3. Condition (1.4) on the nonlinearity $f(\cdot)$ in Theorem 2.2 is not sharp. As suggested in [45] for the one-dimensional problem, it is reasonable to expect that (1.4) may be relaxed to the following:

$$\overline{\lim_{s\to\infty}} \frac{f(s)}{s\ln^2|s|} = 0.$$

But this remains unsolved for the time being.

Remark 2.4. Theorems 2.2–2.3 cover the main results in [26] except the minimal waiting time T_*.

Remark 2.5. Theorems 2.2 can be extended to the case when the nonlinearity $f(y)$ in (1.5) is replaced by $f(t,x,y)$, under suitable growth conditions on (t,x,y). However, it seems to us that in the case when nonlinearity is $f(y,y_t,\nabla y)$, the technique developed in this paper is not enough, and one might have to employ the Nash–Moser–Hörmander iteration method [15] to overcome the difficulty due to the "loss of derivatives." The detailed study of this problem will be presented elsewhere. Note, however, that for purely PDE problems (existence and uniqueness of solutions, etc.) of the hyperbolic equations, the treatment on the nonlinearity $f(y,y_t,\nabla y)$ is almost the same as the simpler one, $f(y)$. This means that for the controllability problem of nonlinear systems, there exist some extra difficulties.

3. Some preliminaries. Let us consider the following linear inhomogeneous hyperbolic equation:

$$\begin{cases} \mathcal{P}z = f & \text{in } Q,\\ z = 0 & \text{on } \Sigma. \end{cases} \tag{3.1}$$

In what follows, we call $z\in L^2(Q)$ a weak solution to (3.1) if

$$(z,\mathcal{P}\eta)_{L^2(Q)} = \int_0^T \langle\, f(t,\cdot),\eta(t,\cdot)\,\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}dt \quad \forall\ \eta\in C_0^2((0,T);H^2(\Omega)\cap H_0^1(\Omega)).$$

Note that in (3.1), no initial conditions are specified. Similarly to [40, Lemma 5.1], one can prove the following regularity result for system (3.1).

LEMMA 3.1. *Let $0<t_1<t_2<T$, $f\in L^1(0,T;H^{-1}(\Omega))$, and $g\in L^2((t_1,t_2)\times\Omega)$ be given. Assume that $z\in L^2(Q)$ is a weak solution to* (3.1)*, and $z=g$ in $(t_1,t_2)\times\Omega$. Then $z\in C([0,T];L^2(\Omega))\cap C^1([0,T];H^{-1}(\Omega))$, and there exists a constant $C>0$, depending only on T, t_1, t_2, Ω, and a^{ij}, such that*

$$|z|_{C([0,T];L^2(\Omega))\cap C^1([0,T];H^{-1}(\Omega))} \le C\Big[|f|_{L^1(0,T;H^{-1}(\Omega))} + |g|_{L^2((t_1,t_2)\times\Omega)}\Big]. \tag{3.2}$$

From the above, we see that g plays the role of initial value for the weak solution z. Next, similarly to [36, Lemma 3.3] we have the following result.

LEMMA 3.2. *Let $a^{ij}\in C^1(\overline{\Omega})$ satisfy* (1.1)*, and let $g\triangleq(g^1,\dots,g^n):\mathbb{R}_t\times\mathbb{R}_x^n\to$*

$\mathbb{R}^n$ *be a vector field of class* C^1. *Then for any* $z \in C^2(\mathbb{R}_t \times \mathbb{R}^n_x)$, *we have*

$$\begin{aligned}
&-\sum_j \left[2(g\cdot\nabla z)\sum_i a^{ij} z_i + g^j \left(z_t^2 - \sum_{i,k} a^{ik} z_i z_k \right) \right]_j \\
&= 2\left[(\mathcal{P}z) g\cdot\nabla z - (z_t g\cdot\nabla z)_t + z_t g_t\cdot\nabla z - \sum_{i,j,k} a^{ij} z_i z_k \frac{\partial g^k}{\partial x_j} \right] \\
&\quad - (\nabla\cdot g) z_t^2 + \sum_{i,j} z_i z_j \nabla\cdot(a^{ij} g).
\end{aligned} \tag{3.3}$$

Next, we denote the energy of system (1.7) by

$$E(t) \triangleq \frac{1}{2}\Big[|w_t(t,\cdot)|^2_{H^{-1}(\Omega)} + |w(t,\cdot)|^2_{L^2(\Omega)} \Big]. \tag{3.4}$$

Using the usual energy method, one obtains the following result.

LEMMA 3.3. *Let* $T > 0$, $q \in L^\infty(0,T;L^n(\Omega))$, $w_0 \in L^2(\Omega)$, *and* $w_1 \in H^{-1}(\Omega)$. *Then the weak solution* $w(\cdot) \in C([0,T];L^2(\Omega)) \cap C^1([0,T];H^{-1}(\Omega))$ *of* (1.7) *satisfies (recall* (2.13) *for* r*)*

$$E(t) \le CE(s)e^{Cr} \qquad \forall\, t,s \in [0,T]. \tag{3.5}$$

Further, proceeding as in [36, Lemma 3.4], we conclude the following.

LEMMA 3.4. *Let* $0 \le S_1 < S_2 < T_2 < T_1 \le T$ *and* $q \in L^\infty(0,T;L^n(\Omega))$. *Then the weak solution* $w(\cdot) \in C([0,T];L^2(\Omega)) \cap C^1([0,T];H^{-1}(\Omega))$ *of* (1.7) *satisfies*

$$\int_{S_2}^{T_2} E(t)dt \le C(1+r)\int_{S_1}^{T_1} |w(t,\cdot)|^2_{L^2(\Omega)} dt. \tag{3.6}$$

Finally, the following proposition will be useful.

PROPOSITION 3.5. *For any* $h > 0$, $m = 2,3,\ldots$, *and* $q_m^i, w_m^i \in \mathbb{C}$ $(i = 0,1,\ldots,m)$ *with* $q_m^0 = q_m^m = 0$, *one has*

$$\begin{aligned}
-\sum_{i=1}^{m-1} q_m^i \frac{(w_m^{i+1} - 2w_m^i + w_m^{i-1})}{h^2} &= \sum_{i=0}^{m-1} \frac{(q_m^{i+1} - q_m^i)}{h} \frac{(w_m^{i+1} - w_m^i)}{h} \\
&= \sum_{i=1}^{m} \frac{(q_m^i - q_m^{i-1})}{h} \frac{(w_m^i - w_m^{i-1})}{h}.
\end{aligned} \tag{3.7}$$

Proof.

$$\begin{aligned}
-\sum_{i=1}^{m-1} q_m^i \frac{(w_m^{i+1} - 2w_m^i + w_m^{i-1})}{h^2} &= -\sum_{i=1}^{m-1} q_m^i \frac{(w_m^{i+1} - w_m^i)}{h^2} + \sum_{i=1}^{m-1} q_m^i \frac{(w_m^i - w_m^{i-1})}{h^2} \\
&= \sum_{i=1}^{m-1} \frac{(q_m^{i+1} - q_m^i)}{h} \frac{(w_m^{i+1} - w_m^i)}{h} - \sum_{i=1}^{m-1} \frac{q_m^{i+1}}{h} \frac{(w_m^{i+1} - w_m^i)}{h} + \sum_{i=0}^{m-2} \frac{q_m^{i+1}}{h} \frac{(w_m^{i+1} - w_m^i)}{h} \\
&= \sum_{i=1}^{m-1} \frac{(q_m^{i+1} - q_m^i)}{h} \frac{(w_m^{i+1} - w_m^i)}{h} + \frac{q_m^1}{h} \frac{(w_m^1 - w_m^0)}{h} \\
&= \sum_{i=0}^{m-1} \frac{(q_m^{i+1} - q_m^i)}{h} \frac{(w_m^{i+1} - w_m^i)}{h},
\end{aligned}$$

which gives the desired equality. ∎

4. Second order differential operators with symmetric coefficients. In this section, we consider second order differential operators with symmetric coefficients. Our hyperbolic differential operator $\mathcal{P}$ is of such a type. We will establish a pointwise equality and a couple of inequalities for such differential operators, which will play important roles. First, we have the following identity.

THEOREM 4.1. *Let* $m \in \mathbb{N}$,

$$b^{ij} = b^{ji} \in C^1(\mathbb{R}^m), \qquad i,j = 1,2,\dots,m, \tag{4.1}$$

and $u,\ \ell,\ \Psi \in C^2(\mathbb{R}^m)$. *Set* $\theta = e^{\ell}$ *and* $v = \theta u$. *Then*

$$\begin{aligned}
&\theta^2 \left| \sum_{i,j} (b^{ij} u_i)_j \right|^2 + 2 \sum_j \left\{ 2 \sum_{i,i',j'} b^{ij} b^{i'j'} \ell_{i'} v_i v_{j'} - \sum_{i,i',j'} b^{ij} b^{i'j'} \ell_i v_{i'} v_{j'} \right. \\
&\qquad \left. + \Psi \sum_i b^{ij} v_i v - \sum_i b^{ij} \left[(\Lambda + \Psi)\ell_i + \frac{\Psi_i}{2} \right] v^2 \right\}_j \\
&= 2 \sum_{i,j} \left\{ \sum_{i',j'} \left[2 b^{ij'} \left(b^{i'j} \ell_{i'} \right)_{j'} - \left(b^{ij} b^{i'j'} \ell_{i'} \right)_{j'} \right] + \Psi b^{ij} \right\} v_i v_j + B v^2 \\
&\qquad + \left| \sum_{i,j} (b^{ij} v_i)_j - \Lambda v \right|^2 + 4 \left| \sum_{i,j} b^{ij} \ell_i v_j \right|^2,
\end{aligned} \tag{4.2}$$

where

$$\begin{cases}
\Lambda \triangleq - \displaystyle\sum_{i,j} (b^{ij} \ell_i \ell_j - b^{ij}_j \ell_i - b^{ij} \ell_{ij}) - \Psi, \\
B \triangleq 2 \left[\Lambda \Psi - \displaystyle\sum_{i,j} \left((\Lambda + \Psi) b^{ij} \ell_i \right)_j \right] + \Psi^2 - \displaystyle\sum_{i,j} (b^{ij} \Psi_j)_i.
\end{cases} \tag{4.3}$$

We see that only the symmetry condition (4.1) is assumed in the above. Hence, Theorem 4.1 is applicable to hyperbolic and ultrahyperbolic operators.

Theorem 4.1 looks similar to [25, Lemma 1, p. 124](which is devoted to a similar problem for a class of ultrahyperbolic operators). The main difference is that we leave the function v on the right-hand side of (4.2) without returning to u, unlike the result of [25] mentioned above, which has only the variable u on both sides. Our result greatly simplifies the computation. Also, a similar idea played a key role in establishing the observability estimate for the wave equations with Neumann boundary conditions in [24] (which should be compared with [19]). We refer the reader to [12, 13] for further application of Theorem 4.1 and its generalization, and to [7] for related work.

Proof of Theorem 4.1. The proof is divided into several steps.

Step 1. Recalling $\theta = e^{\ell}$ and $v = \theta u$, one has $u_i = \theta^{-1}(v_i - \ell_i v)$ $(i = 1,2,\dots,m)$. By the symmetry condition (4.1), it is easy to see that

$$\sum_{i,j} b^{ij} (\ell_i v_j + \ell_j v_i) = 2 \sum_{i,j} b^{ij} \ell_i v_j.$$

Thus, we obtain

$$
\begin{aligned}
(4.4)\qquad \sum_{i,j}(b^{ij}u_i)_j &= \sum_{i,j}[\theta^{-1}b^{ij}(v_i-\ell_i v)]_j \\
&= \theta^{-1}\sum_{i,j}[b^{ij}(v_i-\ell_i v)]_j - \theta^{-1}\sum_{i,j} b^{ij}(v_i-\ell_i v)\ell_j \\
&= \theta^{-1}\sum_{i,j}\Big[(b^{ij}v_i)_j - b^{ij}(\ell_i v_j+\ell_j v_i) + (b^{ij}\ell_i\ell_j - b^{ij}_j\ell_i - b^{ij}\ell_{ij})v\Big] \\
&= \theta^{-1}\sum_{i,j}\Big[(b^{ij}v_i)_j - 2b^{ij}\ell_i v_j + (b^{ij}\ell_i\ell_j - b^{ij}_j\ell_i - b^{ij}\ell_{ij})v\Big] \\
&\equiv -\theta^{-1}(I_1+I_2+I_3),
\end{aligned}
$$

where

$$
(4.5)\qquad \begin{cases}
I_1 \triangleq -\displaystyle\sum_{i,j}\Big[(b^{ij}v_i)_j + (b^{ij}\ell_i\ell_j - b^{ij}_j\ell_i - b^{ij}\ell_{ij})v\Big] - \Psi v \\
\qquad\ = -\displaystyle\sum_{i,j}(b^{ij}v_i)_j + \Lambda v, \\
I_2 \triangleq 2\displaystyle\sum_{i,j} b^{ij}\ell_i v_j, \qquad I_3 \triangleq \Psi v.
\end{cases}
$$

Then, by (4.4) and (4.5), we get

$$
(4.6)\qquad \theta^2\Big|\sum_{i,j}(b^{ij}u_i)_j\Big|^2 = |I_1|^2+|I_2|^2+|I_3|^2+2(I_1I_2+I_2I_3+I_1I_3).
$$

Step 2. Let us compute I_1I_2. Using (4.1) again, and noting

$$
\sum_{i,j,i',j'}\Big(b^{ij}b^{i'j'}\ell_{i'}v_iv_j\Big)_{j'} = \sum_{i,j,i',j'}\Big(b^{ij}b^{i'j'}\ell_i v_{i'}v_{j'}\Big)_j,
$$

we get

$$
\begin{aligned}
&2\sum_{i,j,i',j'} b^{ij}b^{i'j'}\ell_{i'}v_iv_{jj'} \\
&= \sum_{i,j,i',j'} b^{ij}b^{i'j'}\ell_{i'}(v_iv_{jj'}+v_jv_{ij'}) = \sum_{i,j,i',j'} b^{ij}b^{i'j'}\ell_{i'}(v_iv_j)_{j'} \\
&= \sum_{i,j,i',j'}\Big(b^{ij}b^{i'j'}\ell_{i'}v_iv_j\Big)_{j'} - \sum_{i,j,i',j'}\Big(b^{ij}b^{i'j'}\ell_{i'}\Big)_{j'}v_iv_j \\
&= \sum_{i,j,i',j'}\Big(b^{ij}b^{i'j'}\ell_i v_{i'}v_{j'}\Big)_j - \sum_{i,j,i',j'}\Big(b^{ij}b^{i'j'}\ell_{i'}\Big)_{j'}v_iv_j.
\end{aligned}
\qquad (4.7)
$$

Hence, by (4.5) and (4.7), and noting

$$
\sum_{i,j,i',j'} b^{ij}\Big(b^{i'j'}\ell_{i'}\Big)_j v_iv_{j'} = \sum_{i,j,i',j'} b^{ij'}\Big(b^{i'j}\ell_{i'}\Big)_{j'}v_iv_j,
$$

we get

$$
\begin{aligned}
I_1I_2 &= 2\sum_{i,j} b^{ij}\ell_i v_j\Big[-\sum_{i,j}(b^{ij}v_i)_j+\Lambda v\Big]\\
&= -2\sum_{i,j,i',j'}(b^{ij}b^{i'j'}\ell_{i'}v_iv_{j'})_j+2\sum_{i,j,i',j'}b^{ij}(b^{i'j'}\ell_{i'})_jv_iv_{j'}\\
&\quad +2\sum_{i,j,i',j'}b^{ij}b^{i'j'}\ell_{i'}v_iv_{j'j}+\Lambda\sum_{i,j}b^{ij}\ell_i(v^2)_j \qquad (4.8)\\
&= -\sum_j\Big(2\sum_{i,i',j'}b^{ij}b^{i'j'}\ell_{i'}v_iv_{j'}-\sum_{i,i',j'}b^{ij}b^{i'j'}\ell_iv_{i'}v_{j'}-\Lambda\sum_i b^{ij}\ell_iv^2\Big)_j\\
&\quad +\sum_{i,j,i',j'}\Big[2b^{ij'}(b^{i'j}\ell_{i'})_{j'}-(b^{ij}b^{i'j'}\ell_{i'})_{j'}\Big]v_iv_j-\sum_{i,j}(\Lambda b^{ij}\ell_i)_jv^2.
\end{aligned}
$$

Step 3. Let us compute I_2I_3 and I_1I_3. By (4.5), we see that

$$
\begin{aligned}
I_2I_3 &= 2\Psi v\sum_{i,j}b^{ij}\ell_iv_j=\Psi\sum_{i,j}b^{ij}\ell_i(v^2)_j\\
&= \sum_{i,j}\Big(\Psi b^{ij}\ell_iv^2\Big)_j-\sum_{i,j}\Big(\Psi b^{ij}\ell_i\Big)_jv^2. \qquad (4.9)
\end{aligned}
$$

Similarly, by (4.5), we get

$$
\begin{aligned}
2I_1I_3 &= 2\Psi v\Big[-\sum_{i,j}(b^{ij}v_i)_j+\Lambda v\Big]\\
&= -2\sum_{i,j}\Big(\Psi b^{ij}vv_i\Big)_j+2\Psi\sum_{i,j}b^{ij}v_iv_j+\sum_{i,j}b^{ij}\Psi_j(v^2)_i+2\Lambda\Psi v^2\\
&= -\sum_{i,j}\Big(2\Psi b^{ij}vv_i-b^{ij}\Psi_iv^2\Big)_j+2\Psi\sum_{i,j}b^{ij}v_iv_j+\Big[-\sum_{i,j}(b^{ij}\Psi_j)_i+2\Lambda\Psi\Big]v^2. \qquad (4.10)
\end{aligned}
$$

Step 4. Finally, combining (4.6), (4.8), (4.9), and (4.10), we immediately conclude with the desired equality (4.2). This completes the proof of Theorem 4.1. ∎

As a consequence of Theorem 4.1, we have the following.

COROLLARY 4.2. *Let $a^{ij}\in C^1(\overline{\Omega})$ satisfy* (1.1)*, and let $u,\ell,\Psi\in C^2(\mathbb{R}^{1+n})$. Let $\theta=e^\ell$ and $v=\theta u$. Then*

$$
\begin{aligned}
&\theta^2|\mathcal{P}u|^2 \qquad (4.11)\\
&+2\Big[\ell_t\Big(v_t^2+\sum_{i,j}a^{ij}v_iv_j\Big)-2\sum_{i,j}a^{ij}\ell_iv_jv_t-\Psi vv_t+\Big((\Lambda+\Psi)\ell_t+\frac{\Psi_t}{2}\Big)v^2\Big]_t\\
&+2\sum_j\Big\{2\sum_{i,i',j'}a^{ij}a^{i'j'}\ell_{i'}v_iv_{j'}-\sum_{i,i',j'}a^{ij}a^{i'j'}\ell_iv_{i'}v_{j'}+\Psi v\sum_i a^{ij}v_i\\
&-2\ell_tv_t\sum_i a^{ij}v_i+\sum_i a^{ij}\ell_iv_t^2-\sum_i a^{ij}\Big[(\Lambda+\Psi)\ell_i+\frac{\Psi_i}{2}\Big]v^2\Big\}_j\\
&\geq 2\Big(\ell_{tt}+\sum_{i,j}(a^{ij}\ell_i)_j-\Psi\Big)v_t^2-8\sum_{i,j}a^{ij}\ell_{jt}v_iv_t\\
&+2\sum_{i,j}\Big\{a^{ij}\ell_{tt}+\sum_{i',j'}\Big[2a^{ij'}(a^{i'j}\ell_{i'})_{j'}-(a^{ij}a^{i'j'}\ell_{i'})_{j'}\Big]+\Psi a^{ij}\Big\}v_iv_j+Bv^2,
\end{aligned}
$$

where

$$\begin{cases} \Lambda = (\ell_t^2 - \ell_{tt}) - \sum_{i,j}(a^{ij}\ell_i\ell_j - a^{ij}_j\ell_i - a^{ij}\ell_{ij}) - \Psi, \\ B = 2\left[\Lambda\Psi + \Big((\Lambda+\Psi)\ell_t\Big)_t - \sum_{i,j}\Big((\Lambda+\Psi)a^{ij}\ell_i\Big)_j\right] \\ \qquad + \Psi^2 + \Psi_{tt} - \sum_{i,j}(a^{ij}\Psi_j)_i. \end{cases} \tag{4.12}$$

In particular, if

$$\begin{cases} \phi = \phi(t,x) \triangleq d(x) - c(t-T/2)^2, \\ \Psi \triangleq \lambda\left[\sum_{i,j}(a^{ij}d_i)_j - 2c - 1 + k\right], \\ \ell \triangleq \lambda\phi, \quad v \triangleq \theta u, \quad \theta \triangleq e^{\ell}, \end{cases} \tag{4.13}$$

with $\lambda, T > 0$, $c \in (0,1)$, *and* $k \in \mathbb{R}$, *then*

$$\begin{aligned} &\text{(left-hand side of (4.11))} \geq 2\lambda(1-k)v_t^2 \\ &+ 2\lambda\sum_{i,j}\left\{(k-1-4c)a^{ij} + \sum_{i',j'}\Big[2a^{ij'}(a^{i'j}d_{i'})_{j'} - a^{ij}_{j'}a^{i'j'}d_{i'}\Big]\right\}v_iv_j + Bv^2, \end{aligned} \tag{4.14}$$

where

$$\begin{cases} \Lambda = \lambda^2\left[4c^2(t-T/2)^2 - \sum_{i,j}a^{ij}d_id_j\right] + \lambda(4c+1-k), \\ B = 2\lambda^3\Big[(4c+1-k)\sum_{i,j}a^{ij}d_id_j + \sum_{i,j}a^{ij}d_i\left(\sum_{i',j'}a^{i'j'}d_{i'}d_{j'}\right)_j \\ \qquad -4(8c+1-k)c^2(t-T/2)^2\Big] + O(\lambda^2). \end{cases} \tag{4.15}$$

Proof. Using Theorem 4.1 with $m = 1+n$, and

$$(b^{ij})_{m\times m} = \begin{pmatrix} -1 & 0 \\ 0 & A \end{pmatrix},$$

by a direct calculation, we obtain (4.11). The inequality occurs because we have dropped the last two nonnegative terms (see (4.2)). Next, by the choice of (4.13), we can obtain (4.14). ∎

5. Global Carleman estimate for the hyperbolic operators in $H_0^1(Q)$. Recall (2.8) for the definitions of R_1 and T_*. Let $T > T_*$ be given. We may assume that

$$T > 2R_1. \tag{5.1}$$

By (5.1), one may choose a constant $c \in (0,1)$ so that

$$\left(\frac{2R_1}{T}\right)^2 < c < \frac{2R_1}{T}. \tag{5.2}$$

Henceforth, we choose $\phi(t,x)$ as in (4.13) with T and c satisfying, respectively, (5.1) and (5.2).

Remark 5.1. The function ϕ constructed above, together with Condition 2.1, will play a similar role in establishing the Carleman estimate for the hyperbolic operators to that of the function ψ in [18, Condition 1.1], both of which are pseudoconvex in the sense of [17, Definition 28.3.1]. We refer to the classical monographs [16, 17] for more extensive treatment of the Carleman estimate for general partial differential operators, based on pseudoconvex assumptions. Note, however, that our more concrete and explicit choice of ϕ has the following advantages:

(1) It avoids the complicated verification of the pseudoconvex assumption, say, Condition 1.1 in [18]. Indeed, we need only check the "convexity" condition (2.1) and the nonvanishing condition (2.2) (see Proposition 2.1 for an example).
(2) Our ϕ is more natural. In this respect, we note that the time variable t and spatial variables x are separate, which matches the very fact that for the principal operator $\mathcal{P}$, the time derivative ∂_{tt} and the spatial derivatives $-\sum_{i,j}\partial_j(a^{ij}(x)\partial_i)$ have a similar separation property.
(3) The explicit form of $\phi(\cdot)$ or $d(\cdot)$ is useful in the definition of the "controlled/observed" subboundary Γ_+ in (2.5). Also, it plays a key role by scaling and translating $d(x)$ as in (2.6) to achieve (2.7).
(4) What is more important, as mentioned before, is that with our assumption of Condition 2.1, we can give an explicit formula for the waiting time T_*, but this seems to be impossible in the setting of [18] and in that of [16, 17]. Meanwhile, as we shall see later (in the proof of Theorem 2.3), the explicit form of $\phi(\cdot)$ will play a crucial role in deducing the key estimate (2.12) on the observability constant $\mathcal{C}(q)$.

The following Carleman estimate will play a crucial role in section 7.

THEOREM 5.1. *Let $a^{ij} \in C^1(\overline{\Omega})$ satisfy* (1.1)–(1.2)*, and let Conditions* 2.1–2.2 *hold. Then there exists a $\lambda_0 > 1$ such that for all $\lambda \geq \lambda_0$ and all $u \in H_0^1(Q)$ with $\mathcal{P}u \in L^2(Q)$, it holds that*

$$\begin{aligned} &\lambda \int_Q (\lambda^2 u^2 + u_t^2 + |\nabla u|^2) e^{2\lambda\phi} dxdt \\ &\leq C\left[|e^{\lambda\phi}\mathcal{P}u|^2_{L^2(Q)} + \lambda^2 \int_0^T \int_\omega (\lambda^2 u^2 + u_t^2) e^{2\lambda\phi} dxdt\right]. \end{aligned} \tag{5.3}$$

For the reader's convenience, in Appendix B we will give a proof of Theorem 5.1 which is close to the spirit of [24].

Remark 5.2. In the above theorem, the main element, which enables one to integrate over the entire cylinder Q instead of the "conventional" case of its subdomain

bounded by a level surface of the function ϕ, is that $u(0,x)=u(T,x)=0$ in Ω. From the proof of Theorem 5.1, one can see that this point is achieved via (11.10). In the cases $A=I$, and more generally $A=a(x)I$, with a quite restrictive positive function $a(x)$, inequality (11.10) actually follows from [22, equation (2.2.51)] if we introduce (in this paper) a new variable $\tau=t-T/2$ instead of the time variable t.

6. An auxiliary optimal control problem. In this section, we will present an auxiliary optimal control problem which will be useful later. Although some ideas are taken from [18, pp. 190–199], our presentation seems to be easier to understand.

Throughout this section, we fix ϕ as in (4.13), a parameter $\lambda>0$, and a function $u\in C([0,T];L^2(\Omega))$ satisfying $u(0,x)=u(T,x)=0$ for $x\in\Omega$. For any $K>1$, we choose a function $\varrho\equiv\varrho^K(x)\in C^2(\overline{\Omega})$ with $\min_{x\in\overline{\Omega}}\varrho(x)=1$ so that (recall Condition 2.2 for ω)

$$\varrho(x)=\begin{cases}1 & \text{for } x\in\omega,\\ K & \text{for } \operatorname{dist}(x,\omega)\geq\frac{1}{\ln K}.\end{cases}\tag{6.1}$$

Next, fix any integer $m\geq 3$. Let $h=\frac{T}{m}$. Define

$$u_m^i\equiv u_m^i(x)=u(ih,x),\quad \phi_m^i\equiv\phi_m^i(x)=\phi(ih,x),\qquad i=0,1,\dots,m.\tag{6.2}$$

Let $\{(z_m^i,r_{1m}^i,r_{2m}^i,r_m^i)\}_{i=0}^m\in(H_0^1(\Omega)\times(L^2(\Omega))^3)^{m+1}$ satisfy the following system:

$$\begin{cases}\dfrac{z_m^{i+1}-2z_m^i+z_m^{i-1}}{h^2}-\displaystyle\sum_{j_1,j_2=1}^n\partial_{x_{j_2}}(a^{j_1j_2}\partial_{x_{j_1}}z_m^i)\\ \qquad=\dfrac{r_{1m}^{i+1}-r_{1m}^i}{h}+r_{2m}^i+\lambda u_m^i e^{2\lambda\phi_m^i}+r_m^i,\quad(1\leq i\leq m-1) & \text{in }\Omega,\\ z_m^i=0,\qquad(0\leq i\leq m) & \text{on }\Gamma,\\ z_m^0=z_m^m=r_{2m}^0=r_{2m}^m=r_m^0=r_m^m=0,\quad r_{1m}^0=r_{1m}^1 & \text{in }\Omega.\end{cases}\tag{6.3}$$

Note that we do not assume r_{1m}^0 and r_{1m}^m vanish; instead we assume $r_{1m}^0=r_{1m}^1$. In system (6.3), $(r_{1m}^i,r_{2m}^i,r_m^i)\in(L^2(\Omega))^3$ $(i=0,1,\dots,m)$ can be regarded as controls. The set of *admissible sequences* for (6.3) is defined as

$$\mathcal{A}_{ad}\stackrel{\triangle}{=}\Big\{\{(z_m^i,r_{1m}^i,r_{2m}^i,r_m^i)\}_{i=0}^m\in(H_0^1(\Omega)\times(L^2(\Omega))^3)^{m+1}\;\Big|\;\{(z_m^i,r_{1m}^i,r_{2m}^i,r_m^i)\}_{i=0}^m\text{ satisfy (6.3)}\Big\}.$$

Since $\{(0,0,0,-\lambda u_m^i e^{2\lambda\phi_m^i})\}_{i=0}^m\in\mathcal{A}_{ad}$, one sees that $\mathcal{A}_{ad}\neq\emptyset$.

Next, let us introduce the cost functional

$$\begin{aligned}&J(\{(z_m^i,r_{1m}^i,r_{2m}^i,r_m^i)\}_{i=0}^m)\\ &=\frac{h}{2}\int_\Omega\varrho\frac{|r_{1m}^m|^2}{\lambda^2}e^{-2\lambda\phi_m^m}dx\\ &\quad+\frac{h}{2}\sum_{i=1}^{m-1}\left[\int_\Omega|z_m^i|^2e^{-2\lambda\phi_m^i}dx+\int_\Omega\varrho\left(\frac{|r_{1m}^i|^2}{\lambda^2}+\frac{|r_{2m}^i|^2}{\lambda^4}\right)e^{-2\lambda\phi_m^i}dx+K\int_\Omega|r_m^i|^2dx\right].\end{aligned}\tag{6.4}$$

We pose the following *optimal control problem*: Find a $\{(\hat{z}_m^i, \hat{r}_{1m}^i, \hat{r}_{2m}^i, \hat{r}_m^i)\}_{i=0}^m \in \mathcal{A}_{ad}$ such that

$$\begin{aligned} &J(\{(\hat{z}_m^i, \hat{r}_{1m}^i, \hat{r}_{2m}^i, \hat{r}_m^i)\}_{i=0}^m) \\ &= \min_{\{(z_m^i, r_{1m}^i, r_{2m}^i, r_m^i)\}_{i=0}^m \in \mathcal{A}_{ad}} J(\{(z_m^i, r_{1m}^i, r_{2m}^i, r_m^i)\}_{i=0}^m). \end{aligned} \tag{6.5}$$

Note that for any $\{(z_m^i, r_{1m}^i, r_{2m}^i, r_m^i)\}_{i=0}^m \in \mathcal{A}_{ad}$, by standard regularity results of elliptic equations, one has that $z_m^i \in H^2(\Omega) \cap H_0^1(\Omega)$. The following technical result will play a crucial role in section 7.

PROPOSITION 6.1. *For any $K > 1$ and $m \geq 3$, problem* (6.5) *admits a unique solution $\{(\hat{z}_m^i, \hat{r}_{1m}^i, \hat{r}_{2m}^i, \hat{r}_m^i)\}_{i=0}^m \in \mathcal{A}_{ad}$ (which depends on K). Furthermore, for*

$$p_m^i \equiv p_m^i(x) \triangleq K\hat{r}_m^i(x), \qquad 0 \leq i \leq m, \tag{6.6}$$

one has

$$\hat{z}_m^0 = \hat{z}_m^m = p_m^0 = p_m^m = 0 \text{ in } \Omega, \hat{z}_m^i, p_m^i \in H^2(\Omega) \cap H_0^1(\Omega) \text{ for } 1 \leq i \leq m-1, \tag{6.7}$$

and the following optimality conditions hold:

$$\begin{cases} \dfrac{p_m^i - p_m^{i-1}}{h} + \varrho \dfrac{\hat{r}_{1m}^i}{\lambda^2} e^{-2\lambda\phi_m^i} = 0 & \text{in } \Omega, \\ p_m^i - \varrho \dfrac{\hat{r}_{2m}^i}{\lambda^4} e^{-2\lambda\phi_m^i} = 0 & \text{in } \Omega, \end{cases} \qquad 1 \leq i \leq m, \tag{6.8}$$

$$\begin{cases} \dfrac{p_m^{i+1} - 2p_m^i + p_m^{i-1}}{h^2} - \displaystyle\sum_{j_1, j_2 = 1}^n \partial_{x_{j_2}} (a^{j_1 j_2} \partial_{x_{j_1}} p_m^i) + \hat{z}_m^i e^{-2\lambda\phi_m^i} = 0 & \text{in } \Omega, \\ p_m^i = 0 & \text{on } \Gamma, \\ & 1 \leq i \leq m-1. \end{cases} \tag{6.9}$$

Moreover, there is a constant $C = C(K, \lambda) > 0$, independent of m, such that

$$h \sum_{i=1}^{m-1} \int_\Omega \left[|\hat{z}_m^i|^2 + |\hat{r}_{1m}^i|^2 + |\hat{r}_{2m}^i|^2 + K|\hat{r}_m^i|^2 \right] dx + h \int_\Omega |\hat{r}_{1m}^m|^2 dx \leq C, \tag{6.10}$$

and

$$\begin{aligned} h \sum_{i=0}^{m-1} \int_\Omega \Bigg[&\frac{(\hat{z}_m^{i+1} - \hat{z}_m^i)^2}{h^2} + \frac{(\hat{r}_{1m}^{i+1} - \hat{r}_{1m}^i)^2}{h^2} + \frac{(\hat{r}_{2m}^{i+1} - \hat{r}_{2m}^i)^2}{h^2} \\ &+ K \frac{(\hat{r}_m^{i+1} - \hat{r}_m^i)^2}{h^2} \Bigg] dx \leq C. \end{aligned} \tag{6.11}$$

We refer to Appendix C for a proof of this proposition.

7. Global Carleman estimate for hyperbolic operators in $L^2(Q)$. In order to prove Theorem 2.3, we need the following result.

THEOREM 7.1. *Let $a^{ij} \in C^1(\overline{\Omega})$ satisfy (1.1)–(1.2). Let Conditions 2.1–2.2 hold. Then for any $\lambda \geq \lambda_0 \geq 1$, and any $u \in C([0,T];L^2(\Omega))$ satisfying $u(0,x) = u(T,x) = 0$ for $x \in \Omega$, $\mathcal{P}u \in H^{-1}(Q)$, and*

$$(7.1) \quad (u, \mathcal{P}\eta)_{L^2(Q)} = \langle \mathcal{P}u, \eta \rangle_{H^{-1}(Q),H_0^1(Q)} \quad \forall\, \eta \in H_0^1(Q) \text{ with } \mathcal{P}\eta \in L^2(Q),$$

it holds that

$$(7.2) \qquad \lambda \int_Q u^2 e^{2\lambda\phi} dxdt \leq C \left(|e^{\lambda\phi}\mathcal{P}u|^2_{H^{-1}(Q)} + \lambda^2 \int_0^T \int_\omega u^2 e^{2\lambda\phi} dxdt \right),$$

where ϕ is the same as in Theorem 5.1.

Proof. The proof is close to that of [18, Theorem 1.1]. However, for the reader's convenience, we give the details here.

The main idea is to apply (7.1) to some special η with $\mathcal{P}\eta = \cdots + \lambda u e^{2\lambda\phi}$, which yields the desired term $\lambda \int_Q u^2 e^{2\lambda\phi} dxdt$ and reduces the estimate to that for $|\eta|_{H_0^1(Q)}$. We shall employ Proposition 6.1 to provide the desired η. The proof is divided into several steps.

Step 1. First, recall the functions $\{(\hat{z}^i_m, \hat{r}^i_{1m}, \hat{r}^i_{2m}, \hat{r}^i_m)\}_{i=0}^m$ in Proposition 6.1. We define

$$\begin{aligned}
\tilde{z}^m(t,x) &= \frac{1}{h} \sum_{i=0}^{m-1} \Big[(t-ih)\hat{z}^{i+1}_m(x) - \Big(t-(i+1)h\Big)\hat{z}^i_m(x) \Big] \chi_{(ih,(i+1)h]}(t), \\
\tilde{r}^m_1(t,x) &= \hat{r}^0_{1m}(x)\chi_{\{0\}}(t) \\
&\quad + \frac{1}{h} \sum_{i=0}^{m-1} \Big[(t-ih)\hat{r}^{i+1}_{1m}(x) - \Big(t-(i+1)h\Big)\hat{r}^i_{1m}(x) \Big] \chi_{(ih,(i+1)h]}(t), \\
\tilde{r}^m_2(t,x) &= \frac{1}{h} \sum_{i=0}^{m-1} \Big[(t-ih)\hat{r}^{i+1}_{2m}(x) - \Big(t-(i+1)h\Big)\hat{r}^i_{2m}(x) \Big] \chi_{(ih,(i+1)h]}(t), \\
\tilde{r}^m(t,x) &= \frac{1}{h} \sum_{i=0}^{m-1} \Big[(t-ih)\hat{r}^{i+1}_{m}(x) - \Big(t-(i+1)h\Big)\hat{r}^i_{m}(x) \Big] \chi_{(ih,(i+1)h]}(t).
\end{aligned}$$

By (6.10)–(6.11), one can find a subsequence of $(\tilde{z}^m, \tilde{r}^m_1, \tilde{r}^m_2, \tilde{r}^m)$, which converges weakly to some $(\tilde{z}, \tilde{r}_1, \tilde{r}_2, \tilde{r}) \in (H^1(0,T;L^2(\Omega)))^4$, as $m \to \infty$.

For any constant $K > 1$, put

$$\tilde{p} \triangleq K\tilde{r}.$$

In what follows, we shall choose K to be sufficiently large (see (7.19)). By (6.3), (6.8)–(6.11), and noting Lemma 3.1, we see that

$$\tilde{z},\ \tilde{p} \in C([0,T];H_0^1(\Omega)) \cap C^1([0,T];L^2(\Omega))$$

and

$$
\begin{cases}
\mathcal{P}\tilde{z} = \tilde{r}_{1,t} + \tilde{r}_2 + \lambda u e^{2\lambda\phi} + \tilde{r} & \text{in } Q,\\
\mathcal{P}\tilde{p} + \tilde{z}e^{-2\lambda\phi} = 0 & \text{in } Q,\\
\tilde{p} = \tilde{z} = 0 & \text{on } \Sigma,\\
\tilde{p}(0) = \tilde{p}(T) = \tilde{z}(0) = \tilde{z}(T) = 0 & \text{in } \Omega,\\
\tilde{p}_t + \varrho\dfrac{\tilde{r}_1}{\lambda^2}e^{-2\lambda\phi} = 0 & \text{in } Q,\\
\tilde{p} - \varrho\dfrac{\tilde{r}_2}{\lambda^4}e^{-2\lambda\phi} = 0 & \text{in } Q.
\end{cases}
\tag{7.3}
$$

Step 2. Applying Theorem 5.1 to $\tilde{p}$ in (7.3), one gets

$$
\begin{aligned}
&\lambda\int_Q (\lambda^2\tilde{p}^2 + \tilde{p}_t^2 + |\nabla\tilde{p}|^2)e^{2\lambda\phi}dxdt\\
&\le C\left[\int_Q \tilde{z}^2 e^{-2\lambda\phi}dxdt + \lambda^2\int_0^T\int_\omega (\lambda^2\tilde{p}^2 + \tilde{p}_t^2)e^{2\lambda\phi}dxdt\right]\\
&\le C\left[\int_Q \tilde{z}^2 e^{-2\lambda\phi}dxdt + \int_0^T\int_\omega \left(\frac{\tilde{r}_1^2}{\lambda^2} + \frac{\tilde{r}_2^2}{\lambda^4}\right)e^{-2\lambda\phi}dxdt\right].
\end{aligned}
\tag{7.4}
$$

Here and henceforth, C is a constant, independent of K and λ.

By (7.3) again, one finds that $\tilde{p}_t$ satisfies

$$
\begin{cases}
\mathcal{P}\tilde{p}_t + (\tilde{z}e^{-2\lambda\phi})_t = 0 & \text{in } Q,\\
\tilde{p}_t = 0 & \text{on } \Sigma,\\
\tilde{p}_{tt} + \dfrac{\varrho}{\lambda}\Big(\dfrac{\tilde{r}_{1,t}}{\lambda} - 2\phi_t\tilde{r}_1\Big)e^{-2\lambda\phi} = 0 & \text{in } Q,\\
\tilde{p}_t - \dfrac{\varrho}{\lambda^2}\Big(\dfrac{\tilde{r}_{2,t}}{\lambda^2} - \dfrac{2}{\lambda}\phi_t\tilde{r}_2\Big)e^{-2\lambda\phi} = 0 & \text{in } Q.
\end{cases}
\tag{7.5}
$$

Applying Theorem 5.1 to $\tilde{p}_t$ and noting (7.5), we obtain

$$
\begin{aligned}
&\lambda\int_Q \Big(\lambda^2\tilde{p}_t^2 + \tilde{p}_{tt}^2 + |\nabla\tilde{p}_t|^2\Big)e^{2\lambda\phi}dxdt\\
&\le C\left[|e^{\lambda\phi}(e^{-2\lambda\phi}\tilde{z})_t|^2_{L^2(Q)} + \lambda^2\int_0^T\int_\omega \left(\lambda^2\tilde{p}_t^2 + \tilde{p}_{tt}^2\right)e^{2\lambda\phi}dxdt\right]\\
&\le C\left[\int_Q (\tilde{z}_t^2 + \lambda^2\tilde{z}^2)e^{-2\lambda\phi}dxdt + \int_0^T\int_\omega \left(\frac{\tilde{r}_{1,t}^2}{\lambda^2} + \frac{\tilde{r}_{2,t}^2}{\lambda^4} + \tilde{r}_1^2 + \frac{\tilde{r}_2^2}{\lambda^2}\right)e^{-2\lambda\phi}dxdt\right].
\end{aligned}
\tag{7.6}
$$

Step 3. From (7.3), and noting that

$$
-\int_Q (\tilde{r}_{1,t}+\tilde{r}_2)\tilde{p}dxdt = \int_Q \Big(\tilde{r}_1\tilde{p}_t - \tilde{r}_2\tilde{p}\Big)dxdt = -\int_Q \varrho\left(\frac{\tilde{r}_1^2}{\lambda^2} + \frac{\tilde{r}_2^2}{\lambda^4}\right)e^{-2\lambda\phi}dxdt,
\tag{7.7}
$$

and recalling $\tilde{p} = K\tilde{r}$, we get

$$\begin{aligned} 0 &= (\mathcal{P}\tilde{z} - \tilde{r}_{1,t} - \tilde{r}_2 - \lambda u e^{2\lambda\phi} - \tilde{r}, \tilde{p})_{L^2(Q)} \\ &= -\int_Q \tilde{z}^2 e^{-2\lambda\phi} dxdt - \int_Q \varrho \left(\frac{\tilde{r}_1^2}{\lambda^2} + \frac{\tilde{r}_2^2}{\lambda^4} \right) e^{-2\lambda\phi} dxdt \\ &\quad - \lambda \int_Q u\tilde{p} e^{2\lambda\phi} dxdt - K \int_Q \tilde{r}^2 dxdt. \end{aligned} \tag{7.8}$$

Hence

$$\begin{aligned} &\int_Q \tilde{z}^2 e^{-2\lambda\phi} dxdt + \int_Q \varrho \left(\frac{\tilde{r}_1^2}{\lambda^2} + \frac{\tilde{r}_2^2}{\lambda^4} \right) e^{-2\lambda\phi} dxdt + K \int_Q \tilde{r}^2 dxdt \\ &= -\lambda \int_Q u\tilde{p} e^{2\lambda\phi} dxdt. \end{aligned} \tag{7.9}$$

Combining (7.4) and (7.9), we arrive at

$$\begin{aligned} &\int_Q \tilde{z}^2 e^{-2\lambda\phi} dxdt + \int_Q \varrho \left(\frac{\tilde{r}_1^2}{\lambda^2} + \frac{\tilde{r}_2^2}{\lambda^4} \right) e^{-2\lambda\phi} dxdt + K \int_Q \tilde{r}^2 dxdt \\ &\leq \frac{C}{\lambda} \int_Q u^2 e^{2\lambda\phi} dxdt. \end{aligned} \tag{7.10}$$

Step 4. Using (7.3) and (7.5) again, and noting $\tilde{p}_{tt}(0) = \tilde{p}_{tt}(T) = 0$ in Ω, we get

$$\begin{aligned} 0 &= (\mathcal{P}\tilde{z} - \tilde{r}_{1,t} - \tilde{r}_2 - \lambda u e^{2\lambda\phi} - \tilde{r}, \tilde{p}_{tt})_{L^2(Q)} \\ &= -\int_Q \tilde{z}(e^{-2\lambda\phi}\tilde{z})_{tt} dxdt - \int_Q (\tilde{r}_{1,t} + \tilde{r}_2)\tilde{p}_{tt} dxdt \\ &\quad - \lambda \int_Q u\tilde{p}_{tt} e^{2\lambda\phi} dxdt - \int_Q \tilde{r}\tilde{p}_{tt} dxdt. \end{aligned} \tag{7.11}$$

Note

$$\begin{aligned} -\int_Q \tilde{z}(e^{-2\lambda\phi}\tilde{z})_{tt} dxdt &= \int_Q \left(\tilde{z}_t^2 e^{-2\lambda\phi} - \frac{\tilde{z}^2}{2}(e^{-2\lambda\phi})_{tt} \right) dxdt \\ &= \int_Q (\tilde{z}_t^2 + \lambda\phi_{tt}\tilde{z}^2 - 2\lambda^2\phi_t^2\tilde{z}^2) e^{-2\lambda\phi} dxdt. \end{aligned} \tag{7.12}$$

Further, in view of the third and fourth equalities in (7.5), one has

$$\begin{aligned} &-\int_Q (\tilde{r}_{1,t} + \tilde{r}_2)\tilde{p}_{tt} dxdt = -\int_Q (\tilde{r}_{1,t}\tilde{p}_{tt} - \tilde{r}_{2,t}\tilde{p}_t) dxdt \\ &= \int_Q \tilde{r}_{1,t} \frac{\varrho}{\lambda} \left(\frac{\tilde{r}_{1,t}}{\lambda} - 2\phi_t \tilde{r}_1 \right) e^{-2\lambda\phi} dxdt + \int_Q \tilde{r}_{2,t} \frac{\varrho}{\lambda^2} \left(\frac{\tilde{r}_{2,t}}{\lambda^2} - \frac{2}{\lambda}\phi_t \tilde{r}_2 \right) e^{-2\lambda\phi} dxdt \\ &= \int_Q \varrho \left(\frac{\tilde{r}_{1,t}^2}{\lambda^2} + \frac{\tilde{r}_{2,t}^2}{\lambda^4} - \frac{2}{\lambda}\phi_t \tilde{r}_1 \tilde{r}_{1,t} - \frac{2}{\lambda^3}\phi_t \tilde{r}_2 \tilde{r}_{2,t} \right) e^{-2\lambda\phi} dxdt. \end{aligned} \tag{7.13}$$

Moreover, by $\tilde{p} \triangleq K\tilde{r}$ and integration by parts, one gets

$$-\int_Q \tilde{r}\tilde{p}_{tt}dxdt = K\int_Q \tilde{r}_t^2 dxdt. \tag{7.14}$$

Combining (7.11)–(7.14), we end up with

$$\begin{aligned}&\int_Q \varrho\left(\frac{\tilde{r}_{1,t}^2}{\lambda^2}+\frac{\tilde{r}_{2,t}^2}{\lambda^4}-\frac{2}{\lambda}\phi_t\tilde{r}_1\tilde{r}_{1,t}-\frac{2}{\lambda^3}\phi_t\tilde{r}_2\tilde{r}_{2,t}\right)e^{-2\lambda\phi}dxdt+K\int_Q \tilde{r}_t^2dxdt\\ &+\int_Q(\tilde{z}_t^2+\lambda\phi_{tt}\tilde{z}^2-2\lambda^2\phi_t^2\tilde{z}^2)e^{-2\lambda\phi}dxdt=\lambda\int_Q u\tilde{p}_{tt}e^{2\lambda\phi}dxdt.\end{aligned} \tag{7.15}$$

Now, by (7.15)$+C\lambda^2\cdot$(7.10) (with a sufficiently large $C>0$), using the Cauchy–Schwarz inequality and noting (7.6), we obtain

$$\begin{aligned}&\int_Q\left(\tilde{z}_t^2+\lambda^2\tilde{z}^2\right)e^{-2\lambda\phi}dxdt+\int_Q\varrho\left(\frac{\tilde{r}_{1,t}^2}{\lambda^2}+\frac{\tilde{r}_{2,t}^2}{\lambda^4}+\tilde{r}_1^2+\frac{\tilde{r}_2^2}{\lambda^2}\right)e^{-2\lambda\phi}dxdt\\ &\le C\lambda\int_Q u^2e^{2\lambda\phi}dxdt.\end{aligned} \tag{7.16}$$

Step 5. By (7.3), we have

$$\begin{aligned}&(\tilde{r}_{1,t}+\tilde{r}_2+\lambda ue^{2\lambda\phi}+\tilde{r},\tilde{z}e^{-2\lambda\phi})_{L^2(Q)}=(\mathcal{P}\tilde{z},\tilde{z}e^{-2\lambda\phi})_{L^2(Q)}\\ &=-\int_Q\tilde{z}_t(\tilde{z}e^{-2\lambda\phi})_tdxdt+\sum_{i,j}\int_Q a^{ij}\tilde{z}_i(\tilde{z}e^{-2\lambda\phi})_jdxdt\\ &=-\int_Q(\tilde{z}_t^2+\lambda\phi_{tt}\tilde{z}^2-2\lambda^2\phi_t^2\tilde{z}^2)e^{-2\lambda\phi}dxdt+\sum_{i,j}\int_Q a^{ij}\tilde{z}_i\tilde{z}_je^{-2\lambda\phi}dxdt\\ &\quad-2\lambda\sum_{i,j}\int_Q a^{ij}\tilde{z}_i\tilde{z}\phi_je^{-2\lambda\phi}dxdt.\end{aligned} \tag{7.17}$$

This, combined with (1.2), yields (recall $\lambda\ge\lambda_0>1$)

$$\begin{aligned}&\int_Q|\nabla\tilde{z}|^2e^{-2\lambda\phi}dxdt\\ &\le C\int_Q\left[|\tilde{r}_{1,t}+\tilde{r}_2+\tilde{r}||\tilde{z}|e^{-2\lambda\phi}+\lambda|u\tilde{z}|+(\tilde{z}_t^2+\lambda^2\tilde{z}^2)e^{-2\lambda\phi}\right]dxdt\\ &\le C\int_Q\left[u^2e^{2\lambda\phi}+\left(\frac{\tilde{r}_{1,t}^2}{\lambda^2}+\frac{\tilde{r}_2^2}{\lambda^2}+\tilde{r}^2+\tilde{z}_t^2+\lambda^2\tilde{z}^2\right)e^{-2\lambda\phi}\right]dxdt.\end{aligned} \tag{7.18}$$

Combining (7.10), (7.16), and (7.18); choosing the constant K in (7.10) so that

$$K\ge Ce^{2\lambda\max_{(t,x)\in Q}|\phi|} \tag{7.19}$$

(to absorb the term $C\int_Q\tilde{r}^2e^{-2\lambda\phi}dxdt$ in the right-hand side of (7.18)); and noting that $\varrho(x)\ge1$ in Ω, we deduce that

$$\begin{aligned}&\int_Q(|\nabla\tilde{z}|^2+\tilde{z}_t^2+\lambda^2\tilde{z}^2)e^{-2\lambda\phi}dxdt+\int_Q\varrho\left(\frac{\tilde{r}_{1,t}^2}{\lambda^2}+\frac{\tilde{r}_{2,t}^2}{\lambda^4}+\tilde{r}_1^2+\frac{\tilde{r}_2^2}{\lambda^2}\right)e^{-2\lambda\phi}dxdt\\ &\le C\lambda\int_Q u^2e^{2\lambda\phi}dxdt.\end{aligned} \tag{7.20}$$

Step 6. Recall that $(\tilde{z}, \tilde{r}_1, \tilde{r}_2, \tilde{r})$ depend on K. We now fix λ and let $K \to \infty$. By (7.10) and (7.20), we conclude that there exists a subsequence of $(\tilde{z}, \tilde{r}_1, \tilde{r}_2, \tilde{r})$ which converges weakly to some $(\check{z}, \check{r}_1, \check{r}_2, 0)$ in $H_0^1(Q) \times (H^1(0,T;L^2(\Omega)))^2 \times L^2(Q)$, with $\operatorname{supp} \check{r}_i \subset \overline{(0,T) \times \omega}$ $(i = 1, 2)$ since $\varrho(x) \equiv \varrho^K(x) \to \infty$ for any $x \notin \omega$, as $K \to \infty$. By (7.3), we deduce that $(\check{z}, \check{r}_1, \check{r}_2)$ satisfies

$$\begin{cases} \mathcal{P}\check{z} = \check{r}_{1,t} + \check{r}_2 + \lambda u e^{2\lambda\phi} & \text{in } Q, \\ \check{z} = 0 & \text{on } \partial Q. \end{cases} \tag{7.21}$$

Using (7.20) again, we find

$$|\check{z}e^{-\lambda\phi}|^2_{H_0^1(Q)} + \frac{1}{\lambda^2}\int_0^T \int_\omega (\check{r}_{1,t}^2 + \check{r}_2^2) e^{-2\lambda\phi} dxdt \le C\lambda \int_Q u^2 e^{2\lambda\phi} dxdt. \tag{7.22}$$

Now, by (7.1) with η replaced by the above $\check{z}$, one gets

$$\Big(u, \check{r}_{1,t} + \check{r}_2 + \lambda u e^{2\lambda\phi}\Big)_{L^2(Q)} = \langle\, \mathcal{P}u, \check{z}\,\rangle_{H^{-1}(Q), H_0^1(Q)}.$$

Hence, noting $\operatorname{supp} \check{r}_i \subset \overline{(0,T) \times \omega}$ $(i = 1, 2)$, we conclude that for any $\varepsilon > 0$, it holds that

$$\begin{aligned} \lambda \int_Q u^2 e^{2\lambda\phi} dxdt &= \langle\, \mathcal{P}u, \check{z}\,\rangle_{H^{-1}(Q), H_0^1(Q)} - (u, \check{r}_{1,t} + \check{r}_2)_{L^2((0,T)\times\omega)} \\ &\le C\Bigg\{ \frac{1}{\varepsilon}\left[|e^{\lambda\phi}\mathcal{P}u|^2_{H^{-1}(Q)} + \lambda^2 \int_0^T \int_\omega u^2 e^{2\lambda\phi} dxdt \right] \\ &\quad + \varepsilon \left[|\check{z}e^{-\lambda\phi}|^2_{H_0^1(Q)} + \frac{1}{\lambda^2}\int_0^T \int_\omega (\check{r}_{1,t}^2 + \check{r}_2^2) e^{-2\lambda\phi} dxdt \right] \Bigg\}. \end{aligned} \tag{7.23}$$

Finally, choosing ε in (7.23) sufficiently small and noting (7.22), we arrive at the desired estimate (7.2). This completes the proof of Theorem 7.1. ∎

8. Proof of Theorem 2.3. The main idea is to use the Carleman estimate in Theorem 7.1. Note, however, that our w satisfying (1.7) does not necessarily vanish at $t = 0, T$. Therefore we need to introduce a suitable cutoff function. To this end, set

$$\begin{cases} T_i \triangleq T/2 - \varepsilon_i T, \quad T_i' \triangleq T/2 + \varepsilon_i T, \\ R_0 \triangleq \min\limits_{x \in \overline{\Omega}} \sqrt{d(x)}\ (> 0), \end{cases} \tag{8.1}$$

where $i = 0, 1$; $0 < \varepsilon_0 < \varepsilon_1 < 1/2$ will be given below.

From (5.2) and (4.13), it is easy to see that

$$\phi(0,x) = \phi(T,x) < R_1^2 - cT^2/4 < 0 \qquad \forall\, x \in \Omega. \tag{8.2}$$

Therefore there exists an $\varepsilon_1 \in (0, 1/2)$ close to $1/2$ such that

$$\phi(t,x) \le R_1^2/2 - cT^2/8 < 0 \qquad \forall\, (t,x) \in \Big((0,T_1) \bigcup (T_1', T)\Big) \times \Omega \tag{8.3}$$

with T_1 and T_1' given by (8.1). Further, by (4.13), we see that

$$\phi(T/2,x)=d(x)\geq R_0^2 \qquad \forall\, x\in\Omega.$$

Hence, one can find an $\varepsilon_0\in(0,1/2)$, close to 0, such that

$$\phi(t,x)\geq R_0^2/2 \qquad \forall\,(t,x)\in(T_0,T_0')\times\Omega, \tag{8.4}$$

with T_0 and T_0' given by (8.1). We now choose a nonnegative function $\xi\in C_0^\infty(0,T)$ so that

$$\xi(t)\equiv 1 \quad \text{in } (T_1,T_1'). \tag{8.5}$$

Clearly, ξw vanishes at $t=0,T$. Hence, by Theorem 7.1, for any $\lambda\geq\lambda_0$, we have

$$\lambda\int_Q(\xi w)^2e^{2\lambda\phi}dxdt\leq C\left(|e^{\lambda\phi}\mathcal{P}(\xi w)|^2_{H^{-1}(Q)}+\lambda^2\int_0^T\int_\omega w^2e^{2\lambda\phi}dxdt\right). \tag{8.6}$$

By (1.7), we have

$$\begin{aligned}
&|e^{\lambda\phi}\mathcal{P}(\xi w)|_{H^{-1}(Q)}=|e^{\lambda\phi}(\xi\mathcal{P}w+2\xi_tw_t+w\xi_{tt}|_{H^{-1}(Q)}\\
&=|e^{\lambda\phi}(\xi qw+2\xi_tw_t+w\xi_{tt})|_{H^{-1}(Q)}\\
&=\sup_{|f|_{H_0^1(Q)}=1}\langle\, e^{\lambda\phi}(\xi qw+2\xi_tw_t+w\xi_{tt}),f\,\rangle_{H^{-1}(Q),H_0^1(Q)}\\
&\leq\sup_{|f|_{H_0^1(Q)}=1}\int_Q e^{\lambda\phi}\xi qwf\,dxdt\\
&\quad+\sup_{|f|_{H_0^1(Q)}=1}\langle\, e^{\lambda\phi}(2\xi_tw_t+w\xi_{tt}),f\,\rangle_{H^{-1}(Q),H_0^1(Q)}.
\end{aligned} \tag{8.7}$$

Using the Sobolev embedding theorem and the Hölder inequality, and recalling $r\stackrel{\triangle}{=}|q|_{L^\infty(0,T;L^n(\Omega))}$, we get

$$\sup_{|f|_{H_0^1(Q)}=1}\int_Q e^{\lambda\phi}\xi qwf\,dxdt\leq Cr|e^{\lambda\phi}w|_{L^2(Q)}. \tag{8.8}$$

On the other hand, by (8.3) and (8.5), we have

$$\begin{aligned}
&\sup_{|f|_{H_0^1(Q)}=1}\langle\, e^{\lambda\phi}(2\xi_tw_t+w\xi_{tt}),f\,\rangle_{H^{-1}(Q),H_0^1(Q)}\\
&=\sup_{|f|_{H_0^1(Q)}=1}\int_Q e^{\lambda\phi}w(-\xi_{tt}f-2\xi_tf_t-2\lambda\phi_t\xi_tf)dxdt\\
&\leq Ce^{(R_1^2/2-cT^2/8)\lambda}(1+\lambda)(|w|_{L^2((0,T_1)\times\Omega)}+|w|_{L^2((T_1',T)\times\Omega)}).
\end{aligned} \tag{8.9}$$

Further, by (8.3) and (8.5), we have

$$\begin{aligned}
&\int_Q (\xi w)^2 e^{2\lambda\phi} dxdt = \int_Q w^2 e^{2\lambda\phi} dxdt - \int_Q (1-\xi^2) w^2 e^{2\lambda\phi} dxdt \\
&= \int_Q w^2 e^{2\lambda\phi} dxdt - \int_0^{T_1} \int_\Omega (1-\xi^2) w^2 e^{2\lambda\phi} dxdt \\
&\qquad\qquad - \int_{T_1'}^{T} \int_\Omega (1-\xi^2) w^2 e^{2\lambda\phi} dxdt \\
&\geq \int_Q w^2 e^{2\lambda\phi} dxdt - Ce^{(R_1^2 - cT^2/4)\lambda} (|w|^2_{L^2((0,T_1)\times\Omega)} + |w|^2_{L^2((T_1',T)\times\Omega)}).
\end{aligned} \tag{8.10}$$

Combining (8.6)–(8.10), we arrive at

$$\begin{aligned}
&\lambda \int_Q w^2 e^{2\lambda\phi} dxdt \\
&\leq C_1 \bigg[r^2 \int_Q w^2 e^{2\lambda\phi} dxdt + \lambda^2 \int_0^T \int_\omega w^2 e^{2\lambda\phi} dxdt \\
&\qquad + e^{(R_1^2 - cT^2/4)\lambda} (1+\lambda^2) (|w|^2_{L^2((0,T_1)\times\Omega)} + |w|^2_{L^2((T_1',T)\times\Omega)}) \bigg],
\end{aligned} \tag{8.11}$$

for a constant $C_1 > 0$, independent of λ and r. Since $R_1^2 - cT^2/4 < 0$, one may find a $\lambda_1 \geq \lambda_0$ such that $e^{(R_1^2 - cT^2/4)\lambda}(1+\lambda^2) < 1$ for all $\lambda \geq \lambda_1$. Now, taking

$$\lambda \geq 2C_1(\lambda_1 + r^2), \tag{8.12}$$

it follows from (8.11) that

$$\begin{aligned}
&\lambda \int_Q w^2 e^{2\lambda\phi} dxdt \\
&\leq C \left(\lambda^2 \int_0^T \int_\omega w^2 e^{2\lambda\phi} dxdt + |w|^2_{L^2((0,T_1)\times\Omega)} + |w|^2_{L^2((T_1',T)\times\Omega)} \right).
\end{aligned} \tag{8.13}$$

From (8.4), we see that

$$\int_Q w^2 e^{2\lambda\phi} dxdt \geq e^{R_0^2\lambda} \int_{T_0}^{T_0'} \int_\Omega w^2 dxdt. \tag{8.14}$$

For any $S_0 \in (T_0, T/2)$ and $S_0' \in (T/2, T_0')$, by Lemma 3.4, we obtain (recall (3.4) for $E(t)$)

$$\int_{S_0}^{S_0'} E(t) dt \leq C(1+r) \int_{T_0}^{T_0'} \int_\Omega w^2 dxdt. \tag{8.15}$$

On the other hand, by Lemma 3.3, we have

$$|w|^2_{L^2((0,T_1)\times\Omega)} + |w|^2_{L^2((T_1',T)\times\Omega)} \leq CE(0)e^{Cr} \tag{8.16}$$

and

$$\int_{S_0}^{S_0'} E(t)dt \geq CE(0)e^{Cr}. \tag{8.17}$$

Combining (8.13)–(8.17), we end up with

$$\Big(C_2\lambda e^{R_0^2\lambda+C_2r} - C_3(1+r)e^{C_3r}\Big)E(0) \leq C\lambda^2(1+r)e^{C\lambda}\int_0^T\int_\omega w^2dxdt, \tag{8.18}$$

for two constants $C_2 > 0$ and $C_3 > 0$, independent of λ and r. We now choose λ so that

$$C_2\lambda \geq C_3(1+r), \qquad R_0^2\lambda + C_2r \geq C_3r. \tag{8.19}$$

Then, from (8.18), we obtain

$$E(0) \leq \mathcal{C}(q)|w|^2_{L^2((0,T)\times\omega)}. \tag{8.20}$$

Finally, noting (8.12) and (8.19), we conclude (2.12). This completes the proof of Theorem 2.3. □

9. Proof of Theorem 2.2. The proof is very close to that of [26, Theorem 3.1] and [38, Theorem 2.1]. However, for the reader's convenience, we give some details here.

Define a function $h(\cdot) \in C(\mathbb{R})$ by

$$h(s) \stackrel{\triangle}{=} \begin{cases} [f(s)-f(0)]/s & \text{if } s \neq 0, \\ f'(0) & \text{if } s = 0. \end{cases} \tag{9.1}$$

Let the initial and final data $(y_0, y_1), (z_0, z_1) \in H_0^1(\Omega) \times L^2(\Omega)$ be given. For any given $z(\cdot) \in L^\infty(0,T; L^2(\Omega))$, we look for a control $\gamma = \gamma(z(\cdot)) \in L^2((0,T)\times\omega)$ such that the solution $y = y(\cdot; z(\cdot))$ of

$$\begin{cases} \mathcal{P}y = h(z(\cdot))y + f(0) + \chi_\omega(x)\gamma(t,x) & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in } \Omega \end{cases} \tag{9.2}$$

satisfies

$$y(T) = z_0, \quad y_t(T) = z_1 \qquad \text{in } \Omega. \tag{9.3}$$

For this purpose, we use the classical duality argument [29, 28, 39]. First, we solve

$$\begin{cases} \mathcal{P}v = h(z(\cdot))v + f(0) & \text{in } Q, \\ v = 0 & \text{on } \Sigma, \\ v(T) = z_0, \quad v_t(T) = z_1 & \text{in } \Omega, \end{cases} \tag{9.4}$$

which admits a unique weak solution $v = v(\cdot; z(\cdot)) \in C([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega))$. Next, put $X \stackrel{\triangle}{=} L^2(\Omega) \times H^{-1}(\Omega)$. For any $(w_0, w_1) \in X$, we solve

$$\begin{cases} \mathcal{P}w = h(z(\cdot))w & \text{in } Q, \\ w = 0 & \text{on } \Sigma, \\ w(0) = w_0, \quad w_t(0) = w_1 & \text{in } \Omega \end{cases} \tag{9.5}$$

and

$$\begin{cases} \mathcal{P}\eta = h(z(\cdot))\eta + \chi_\omega(x)w(t,x) & \text{in } Q, \\ \eta = 0 & \text{on } \Sigma, \\ \eta(T) = \eta_t(T) = 0 & \text{in } \Omega. \end{cases} \tag{9.6}$$

Now, we define a linear and continuous operator $\Lambda : X \to X'$, the dual space of X, by

$$\Lambda(w_0, w_1) \stackrel{\triangle}{=} (-\eta_t(0), \eta(0)), \tag{9.7}$$

where $\eta \in C([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega))$ is the weak solution of (9.6).

Let us show the existence of some $(w_0, w_1) \in X$ such that

$$\Lambda(w_0, w_1) = (-y_1 + v_t(0), y_0 - v(0)). \tag{9.8}$$

For this purpose, we observe that, by multiplying the first equation in (9.6) by w; integrating it in Q; using integration by parts; and noting (9.5), $\eta(T) = \eta_t(T) = 0$ in Ω, and (9.7), it follows that

$$\langle\, \Lambda(w_0, w_1), (w_0, w_1)\, \rangle_{X',X} = \int_0^T \int_\omega w^2 dxdt. \tag{9.9}$$

However, by Theorem 2.3 and (9.9), we have

$$\langle\, \Lambda(w_0, w_1), (w_0, w_1)\, \rangle_{X',X} \geq \frac{1}{\mathcal{C}(h(z(\cdot)))} |(w_0, w_1)|_X^2 \qquad \forall\, (w_0, w_1) \in X, \tag{9.10}$$

where $\mathcal{C}(\cdot)$ is the constant given in (2.12). By the Lax–Milgram theorem, (9.8) admits a unique solution $(w_0, w_1) \in X$. It is easy to check that

$$\gamma = w \tag{9.11}$$

is the desired control such that the weak solution $y \equiv v + \eta$ of (9.2) satisfies (9.3).

Further, proceeding as in the proof of [38, Theorem 2.1], by (9.10) we end up with

$$\begin{aligned} &|w|_{C([0,T];L^2(\Omega))} \\ &\leq \mathcal{C}(h(z(\cdot)))(|f(0)| + |(y_0, y_1)|_{H_0^1(\Omega)\times L^2(\Omega)} + |(z_0, z_1)|_{H_0^1(\Omega)\times L^2(\Omega)}). \end{aligned} \tag{9.12}$$

Next, similarly to the proof of [26, Theorem 3.1] by applying the classical energy method to (9.2), noting (9.11)–(9.12), and recalling assumption (1.4), one concludes that there is a constant $C > 0$ such that, for any $\varepsilon \in (0, 4]$, it holds that

$$\begin{aligned} &|y|_{C([0,T];H_0^1(\Omega))\cap C^1([0,T];L^2(\Omega))} \\ &\leq C[|f(0)| + |(y_0, y_1)|_{H_0^1(\Omega)\times L^2(\Omega)} \\ &\qquad + |(z_0, z_1)|_{H_0^1(\Omega)\times L^2(\Omega)}]\Big(1 + |z|^{4/(1+\varepsilon)}_{L^\infty(0,T;L^2(\Omega))}\Big). \end{aligned} \tag{9.13}$$

Consequently if we take $\varepsilon = 4$ in (9.13), the desired exact controllability result follows from the fixed point technique. This completes the proof of Theorem 2.2. ∎

10. Appendix A. Proof of Proposition 2.1. Consider first the case when $A = \operatorname{diag}[a^1, \dots, a^n]$ with $a^i \in C^1(\overline{\Omega})$ $(i = 1, \dots, n)$. In this case, the matrix $\mathcal{A}$ (defined in (2.3)) reads

$$\mathcal{A} = \left(a^i a^j d_{ij} + \frac{a^i a^j_i d_j + a^j a^i_j d_i}{2} \right)_{1 \le i,j \le n} - \frac{1}{2} \operatorname{diag} \left[\sum_k a^k a^1_k d_k, \dots, \sum_k a^k a^n_k d_k \right].$$

In particular, when $n = 2$ and a^1 is independent of x_2 (hence $a^1_2 \equiv 0$), the above $\mathcal{A}$ is specialized as

$$\begin{aligned} \mathcal{A} &= \begin{pmatrix} (a^1)^2 d_{11} + \frac{a^1 a^1_1 d_1 - a^2 a^1_2 d_2}{2} & a^1 a^2 d_{12} + \frac{a^1 a^2_1 d_2 + a^2 a^1_2 d_1}{2} \\ a^1 a^2 d_{12} + \frac{a^1 a^2_1 d_2 + a^2 a^1_2 d_1}{2} & (a^2)^2 d_{22} + \frac{a^2 a^2_2 d_2 - a^1 a^2_1 d_1}{2} \end{pmatrix} \\ &= \begin{pmatrix} (a^1)^2 d_{11} + \frac{a^1 a^1_1 d_1}{2} & a^1 a^2 d_{12} + \frac{a^1 a^2_1 d_2}{2} \\ a^1 a^2 d_{12} + \frac{a^1 a^2_1 d_2}{2} & (a^2)^2 d_{22} + \frac{a^2 a^2_2 d_2 - a^1 a^2_1 d_1}{2} \end{pmatrix} \\ &\equiv \begin{pmatrix} \hat{a}^{11} & \hat{a}^{12} \\ \hat{a}^{12} & \hat{a}^{22} \end{pmatrix}. \end{aligned} \tag{10.1}$$

Put $L = 2\operatorname{diam}\Omega$. For any parameters $\tau > 0$ and $\mu > 0$, we now choose d to be of the form

$$d(x_1, x_2) = e^{-\tau a^1(x_1)} + e^{-\mu(L + x_2)}.$$

Then,

$$\begin{aligned} &d_1 = -\tau a^1_1 e^{-\tau a^1}, \quad d_{11} = \tau(\tau |a^1_1|^2 - a^1_{11}) e^{-\tau a^1}, \\ &d_{12} = 0, \quad d_2 = -\mu e^{-\mu(L + x_2)}, \quad d_{22} = \mu^2 e^{-\mu(L + x_2)}. \end{aligned} \tag{10.2}$$

We consider only the case when there is an $x_0 \in G$ such that $a^1_1(x^0_1) = 0$, $a^1_{11}(x^0_1) < 0$, $|a^1_1| \neq 0$ in $G \setminus \{x^0_1\}$ (the case when $a^1_1(x_1) \neq 0$ for any $x_1 \in G$ is easier to analyze). By (10.1), (10.2), and noting that a^1 is uniformly positive in $\overline{\Omega}$, one may choose a sufficiently large τ such that

$$\begin{aligned} \hat{a}^{11} &= (a^1)^2 d_{11} + \frac{a^1 a^1_1 d_1}{2} \\ &= \tau \left[\left(\tau (a^1)^2 - \frac{a^1}{2} \right) |a^1_1|^2 - a^1_{11} (a^1)^2 \right] e^{-\tau a^1} > 0 \end{aligned} \tag{10.3}$$

uniformly in $\overline{\Omega}$.

Further, by (10.1) and (10.2), by noting that $a^1_1 a^2_1 \ge 0$, and by noting that a^2 is uniformly positive in $\overline{\Omega}$, one may choose a sufficiently large μ such that

$$\begin{aligned} \hat{a}^{22} &= (a^2)^2 d_{22} + \frac{a^2 a^2_2 d_2 - a^1 a^2_1 d_1}{2} \\ &= \left[(a^2)^2 \mu^2 - \frac{a^2 a^2_2 \mu}{2} \right] e^{-\mu(L + x_2)} + \frac{a^1 a^1_1 a^2_1 \tau}{2} e^{-\tau a^1} > 0 \end{aligned} \tag{10.4}$$

uniformly in $\overline{\Omega}$.

Further, we have

$$\hat{a}^{12} = a^1 a^2 d_{12} + \frac{a^1 a_1^2 d_2}{2} = -\frac{a^1 a_1^2 \mu}{2} e^{-\mu(L+x_2)}.$$

Now, fixing the parameter τ, it is easy to see that

$$\hat{a}^{11}\hat{a}^{22} - (\hat{a}^{12})^2 > 0 \qquad \text{uniformly in } \overline{\Omega}, \tag{10.5}$$

provided that μ is large enough (because $(\hat{a}^{12})^2$ is an infinitesimal of higher order, compared to $\hat{a}^{11}\hat{a}^{12}$, with respect to large μ). By (10.3)–(10.5), we deduce that the matrix $\mathcal{A}$ in (10.1) is uniformly positive definite in $\overline{\Omega}$.

It is clear that $\min_{x\in\overline{\Omega}} |\nabla d(x)| > 0$. Therefore, Condition 2.1 holds for the above constructed function d. □

11. Appendix B. Proof of Theorem 5.1. The proof is long and we divide it into several steps.

Step 1. Applying Corollary 4.2 to our present u and d, we conclude that for any constants $\lambda > 0$ and $k \in (0,1)$, it holds that

$$\begin{aligned}
&\theta^2|\mathcal{P}u|^2 + M_t \\
&\quad + 2\sum_j \left\{ 2\sum_{i,i',j'} a^{ij}a^{i'j'}\ell_{i'}v_i v_{j'} - \sum_{i,i',j'} a^{ij}a^{i'j'}\ell_i v_{i'}v_{j'} + \Psi v \sum_i a^{ij}v_i \right. \\
&\qquad \left. -2\ell_t v_t \sum_i a^{ij}v_i + \sum_i a^{ij}\ell_i v_t^2 - \sum_i a^{ij}\left[(\Lambda+\Psi)\ell_i + \frac{\Psi_i}{2}\right]v^2 \right\}_j \\
&\quad \geq 2\lambda(1-k)v_t^2 + Bv^2 \\
&\quad + 2\lambda\sum_{i,j}\left\{ (k-1-4c)a^{ij} + \sum_{i',j'}\left[2a^{ij'}(a^{i'j}d_{i'})_{j'} - a^{ij}_{j'}a^{i'j'}d_{i'}\right]\right\} v_i v_j,
\end{aligned} \tag{11.1}$$

where

$$\left\{\begin{array}{l}
M \triangleq 2\left[\ell_t\left(v_t^2 + \sum_{i,j} a^{ij}v_i v_j\right) - 2\sum_{i,j} a^{ij}\ell_i v_j v_t \right. \\
\qquad\qquad \left. -\Psi v v_t + \left((\Lambda+\Psi)\ell_t + \frac{\Psi_t}{2}\right)v^2\right], \\
\Psi \triangleq \lambda\left[\sum_{i,j}(a^{ij}d_i)_j - 2c - 1 + k\right], \quad \ell \triangleq \lambda\phi, \quad v \triangleq \theta u, \quad \theta \triangleq e^{\ell}, \\
\Lambda = \lambda^2\left[4c^2(t-T/2)^2 - \sum_{i,j} a^{ij}d_i d_j\right] + \lambda(4c+1-k), \\
B = 2\lambda^3\left[(4c+1-k)\sum_{i',j'} a^{i'j'}d_{i'}d_{j'} + \sum_{i,j} a^{ij}d_i\left(\sum_{i',j'} a^{i'j'}d_{i'}d_{j'}\right)_j \right. \\
\qquad\qquad \left. -4(8c+1-k)c^2(t-T/2)^2\right] + O(\lambda^2).
\end{array}\right. \tag{11.2}$$

Next, fix a k with $4c-3<k<1$. Hence

$$1-k>0. \tag{11.3}$$

On the other hand, by Condition 2.1 and noting (2.1) with $\mu_0 \geq 4$, we get

$$\begin{aligned}
&\sum_{i,j}\left\{(k-1-4c)a^{ij}+\sum_{i',j'}\left[2a^{ij'}(a^{i'j}d_{i'})_{j'}-a^{ij}_{j'}a^{i'j'}d_{i'}\right]\right\}v_iv_j\\
&\geq (k-4c-1+\mu_0)\sum_{i,j}a^{ij}v_iv_j\\
&=\mu\sum_{i,j}a^{ij}v_iv_j \quad \forall\, x\in\Omega,
\end{aligned} \tag{11.4}$$

where

$$\mu=\mu_0-1+k-4c\geq 3+k-4c>0. \tag{11.5}$$

Recalling that d satisfies (2.1), and noting $a^{i'j'}=a^{j'i'}$, we find

$$\begin{aligned}
\mu_0\sum_{i,j}a^{ij}d_id_j &\leq \sum_{i,j,i',j'}\left[2a^{ij'}(a^{i'j}d_{i'})_{j'}-a^{ij}_{j'}a^{i'j'}d_{i'}\right]d_id_j\\
&=\sum_{i,j,i',j'}\left[2a^{ij'}a^{i'j}_{j'}d_{i'}d_id_j+2a^{ij'}a^{i'j}d_{i'j'}d_id_j-a^{ij}_{j'}a^{i'j'}d_{i'}d_id_j\right]\\
&=\sum_{i,j,i',j'}\left[a^{ij'}a^{i'j}_{j'}d_{i'}d_id_j+2a^{ij'}a^{i'j}d_{i'j'}d_id_j\right]\\
&=\sum_{i,j,i',j'}\left[a^{ij}a^{i'j'}_{j}d_{i'}d_id_{j'}+2a^{ij}a^{i'j'}d_{i'j}d_id_{j'}\right]\\
&=\sum_{i,j,i',j'}\left[a^{ij}a^{i'j'}_{j}d_{i'}d_id_{j'}+a^{ij}a^{i'j'}d_{i'j}d_id_{j'}+a^{ij}a^{j'i'}d_{j'j}d_id_{i'}\right]\\
&=\sum_{i,j}a^{ij}d_i\left(\sum_{i',j'}a^{i'j'}d_{i'}d_{j'}\right)_j.
\end{aligned} \tag{11.6}$$

Hence, recalling, respectively, (2.8) and (11.2) for R_1 and B, by (11.6) and using the third inequality in (2.7), and noting that A is positive definite and $4c+1-k+\mu_0>8c+1-k$, we arrive at

$$\begin{aligned}
B&\geq 2\lambda^3\left\{(4c+1-k+\mu_0)\sum_{i,j}a^{ij}d_id_j-4(8c+1-k)c^2(t-T/2)^2\right\}+O(\lambda^2)\\
&\geq 2\lambda^3(8c+1-k)\left[\sum_{i,j}a^{ij}d_id_j-4c^2(t-T/2)^2\right]+O(\lambda^2)\\
&\geq 16c(4R_1^2-c^2T^2)\lambda^3+O(\lambda^2).
\end{aligned} \tag{11.7}$$

Note that, by (5.2), the constant $16c(4R_1^2 - c^2T^2)$ in (11.7) is positive. Hence, by choosing a suitable $\lambda_0 > 1$, for any $\lambda \geq \lambda_0$, we have

$$B \geq 8c(4R_1^2 - c^2T^2)\lambda^3. \tag{11.8}$$

Step 2. Integrating (11.1) on Q, using integration by parts, recalling (11.3)–(11.5) and (11.8), and noting that $v_i = \frac{\partial v}{\partial \nu}\nu_i$ on Σ (which follows from $v|_\Sigma = 0$), we arrive at (recall (11.2) for $M = M(t,x)$)

$$\begin{aligned}
&\lambda \int_Q \left(\lambda^2 v^2 + v_t^2 + \sum_{i,j} a^{ij} v_i v_j \right) dxdt \\
&\leq C \Bigg[\int_Q \theta^2 |\mathcal{P}u|^2 dxdt + \int_\Omega M(T,x)dx - \int_\Omega M(0,x)dx \\
&\qquad + \lambda \int_\Sigma \left(\sum_{i,j} a^{ij}\nu_i\nu_j \right) \left(\sum_{i',j'} a^{i'j'} d_{i'}\nu_{j'} \right) \Big| \frac{\partial v}{\partial \nu} \Big|^2 dxdt \Bigg] \qquad \forall\, \lambda \geq \lambda_0.
\end{aligned} \tag{11.9}$$

By (4.13) and (11.2), and noting that $u(0,x) = u(T,x) \equiv 0$, we get

$$\begin{aligned}
&M(0,x) = 2\ell_t(0,x)[\theta(0,x)u_t(0,x)]^2 = 2cT\lambda[\theta(0,x)u_t(0,x)]^2 > 0, \\
&M(T,x) = 2\ell_t(T,x)[\theta(T,x)u_t(T,x)]^2 = -2cT\lambda[\theta(T,x)u_t(T,x)]^2 < 0.
\end{aligned} \tag{11.10}$$

Combining (11.9) and (11.10), and noting the definition of Γ_+ in (2.5), we obtain

$$\begin{aligned}
&\lambda \int_Q \left(\lambda^2 v^2 + v_t^2 + \sum_{i,j} a^{ij} v_i v_j \right) dxdt \\
&\leq C \Bigg[\int_Q \theta^2 |\mathcal{P}u|^2 dxdt \\
&\qquad + \lambda \int_0^T \int_{\Gamma_+} \left(\sum_{i,j} a^{ij}\nu_i\nu_j \right) \left(\sum_{i',j'} a^{i'j'} d_{i'}\nu_{j'} \right) \Big| \frac{\partial v}{\partial \nu} \Big|^2 dxdt \Bigg].
\end{aligned} \tag{11.11}$$

Recalling $u = \theta^{-1}v$ and $\theta = e^\ell$, noting (4.13) and (11.11), and noting (1.2) and $u|_\Sigma = 0$, we get

$$\begin{aligned}
&\lambda \int_Q \theta^2 (\lambda^2 u^2 + u_t^2 + |\nabla u|^2) dxdt \\
&\leq C \left(\int_Q \theta^2 |\mathcal{P}u|^2 dxdt + \lambda \int_0^T \int_{\Gamma_+} \theta^2 \Big| \frac{\partial u}{\partial \nu} \Big|^2 dxdt \right).
\end{aligned} \tag{11.12}$$

Step 3. Let us estimate

$$\int_0^T \int_{\Gamma_+} \theta^2 \Big| \frac{\partial u}{\partial \nu} \Big|^2 dxdt.$$

We choose a $g_0 \in C^1(\overline{\Omega}; \mathbb{R}^n)$ such that $g_0 = \nu$ on Γ, and a $\rho \in C^2(\overline{\Omega}; [0,1])$ such that (recall Condition 2.2 for δ)

$$\begin{cases} \rho(x) \equiv 1, & x \in \mathcal{O}_{\delta/3}(\Gamma_+) \cap \Omega, \\ \rho(x) \equiv 0, & x \in \Omega \setminus \mathcal{O}_{\delta/2}(\Gamma_+). \end{cases} \tag{11.13}$$

Put

$$g = g_0 \rho \theta^2. \tag{11.14}$$

Integrating (3.3) (in Lemma 3.2) in Q, with g defined by (11.14) and z replaced by u; using integration by parts; and noting (11.13), $u_i = \frac{\partial u}{\partial \nu}\nu_i$ on Σ (which follows from $u|_\Sigma = 0$), and $u(0,x) = u(T,x) \equiv 0$, we get

$$\begin{aligned}
&\int_\Sigma \left(\sum_{i,j} a^{ij}\nu_i\nu_j \right) \rho\theta^2 \Big| \frac{\partial u}{\partial \nu} \Big|^2 dxdt \\
&= \int_Q \sum_j \left[2(g\cdot\nabla u) \sum_i a^{ij} u_i + g^j \left(u_t^2 - \sum_{i,k} a^{ik} u_i u_k \right) \right]_j dxdt \\
&= -\int_Q \left\{ 2 \left[(\mathcal{P}u) g\cdot\nabla u - (u_t g\cdot\nabla u)_t + u_t g_t \cdot \nabla u - \sum_{i,j,k} a^{ij} u_i u_k \frac{\partial g^k}{\partial x_j} \right] \right. \\
&\qquad \left. -(\nabla\cdot g)\left(u_t^2 - \sum_{i,j} a^{ij} u_i u_j \right) \right\} dxdt \qquad (11.15) \\
&= -\int_Q \left\{ 2 \left[(\mathcal{P}u) g\cdot\nabla u + u_t g\cdot\nabla u + u_t g_t \cdot \nabla u - \sum_{i,j,k} a^{ij} u_i u_k \frac{\partial g^k}{\partial x_j} \right] \right. \\
&\qquad \left. -(\nabla\cdot g)\left(u_t^2 - \sum_{i,j} a^{ij} u_i u_j \right) \right\} dxdt \\
&\le C \left[\frac{1}{\lambda} |\theta \mathcal{P}u|^2_{L^2(Q)} + \lambda \int_0^T \int_{\mathcal{O}_{\delta/2}(\Gamma_+)\cap\Omega} \theta^2 (u_t^2 + |\nabla u|^2) dxdt \right].
\end{aligned}$$

Step 4. Let us estimate

$$\int_0^T \int_{\mathcal{O}_{\delta/2}(\Gamma_+)\cap\Omega} \theta^2 |\nabla u|^2 dxdt.$$

Put

$$\eta = \eta(t,x) \triangleq \rho_1^2 \theta^2, \tag{11.16}$$

where $\rho_1 \in C^2(\overline{\Omega}; [0,1])$ satisfies

$$\begin{cases} \rho_1(x) \equiv 1, & x \in \mathcal{O}_{\delta/2}(\Gamma_+) \cap \Omega, \\ \rho_1(x) \equiv 0, & x \in \Omega \setminus \omega. \end{cases} \tag{11.17}$$

By (1.3), we obtain

$$\begin{aligned}
\int_Q \eta u \mathcal{P} u dxdt &= \int_Q \eta u \left(u_{tt} - \sum_{i,j} (a^{ij} u_i)_j \right) dxdt \\
&= -\int_Q \Big[u_t (\eta_t u + \eta u_t) \Big] dxdt + \int_Q \eta \sum_{i,j} a^{ij} u_i u_j dxdt + \int_Q u \sum_{i,j} a^{ij} u_i \eta_j dxdt.
\end{aligned} \tag{11.18}$$

Hence, by (1.2) and (11.16)–(11.18), we find

$$\begin{aligned}&\int_0^T\int_{\mathcal{O}_{\delta/2}(\Gamma_+)\cap\Omega}\theta^2|\nabla u|^2dxdt\\&\le C\left[\frac{1}{\lambda^2}|\theta\mathcal{P}u|^2_{L^2(Q)}+\int_0^T\int_\omega\theta^2(\lambda^2u^2+u_t^2)dxdt\right].\end{aligned}\tag{11.19}$$

Finally, combining (11.12), (11.15), and (11.19), and noting (11.13), we get the desired estimate (5.3). □

12. Appendix C. Proof of Proposition 6.1. We borrow some ideas from [18]. The proof is split into several steps.

Step 1. Let $\{\{(z_m^{i,j},r_{1m}^{i,j},r_{2m}^{i,j},r_m^{i,j})\}_{i=0}^m\}_{j=1}^\infty\subset\mathcal{A}_{ad}$ be a minimizing sequence of $J(\cdot)$. Because of the coercivity of the cost functional and noting that $z_m^{i,j}$ solves an elliptic equation, it can be shown that $\{\{(z_m^{i,j},r_{1m}^{i,j},r_{2m}^{i,j},r_m^{i,j})\}_{i=0}^m\}_{j=1}^\infty$ is bounded in $\mathcal{A}_{ad}$. Therefore, there exists a subsequence of $\{\{(z_m^{i,j},r_{1m}^{i,j},r_{2m}^{i,j},r_m^{i,j})\}_{i=0}^m\}_{j=1}^\infty$ converging weakly in $(H_0^1(\Omega)\times(L^2(\Omega))^3)^{m+1}$ to some $\{(\hat z_m^i,\hat r_{1m}^i,\hat r_{2m}^i,\hat r_m^i)\}_{i=0}^m\in\mathcal{A}_{ad}$. Since the function J is strictly convex, this element is the unique solution of (6.5). By (6.6) and the definition of $\mathcal{A}_{ad}$, it is obvious that $\hat z_m^0=\hat z_m^m=p_m^0=p_m^m=0$ in Ω.

Step 2. Fix any $\delta_{0m}^i\in H^2(\Omega)\cap H_0^1(\Omega)$, $\delta_{1m}^i\in L^2(\Omega)$, and $\delta_{2m}^i\in L^2(\Omega)$ $(i=0,1,2,\dots,m)$ with $\delta_{0m}^0=\delta_{0m}^m=\delta_{2m}^0=\delta_{2m}^m\equiv0$ and $\delta_{1m}^0=\delta_{1m}^1$ in Ω. For $(\lambda_0,\lambda_1,\lambda_2)\in\mathbb{R}^3$, put

$$\begin{cases}r_m^i\triangleq\dfrac{\hat z_m^{i+1}-2\hat z_m^i+\hat z_m^{i-1}}{h^2}+\dfrac{\delta_{0m}^{i+1}-2\delta_{0m}^i+\delta_{0m}^{i-1}}{h^2}\lambda_0\\\qquad\quad-\displaystyle\sum_{j_1,j_2=1}^n\partial_{x_{j_2}}\Big(a^{j_1j_2}\partial_{x_{j_1}}(\hat z_m^i+\lambda_0\delta_{0m}^i)\Big)\\\qquad\quad-\dfrac{\hat r_{1m}^{i+1}-\hat r_{1m}^i}{h}-\dfrac{\delta_{1m}^{i+1}-\delta_{1m}^i}{h}\lambda_1-\hat r_{2m}^i-\lambda_2\delta_{2m}^i-\lambda u_m^ie^{2\lambda\phi_m^i},\quad1\le i\le m-1;\\r_m^0=r_m^m=0.\end{cases}$$

Then $\{(\hat z_m^i+\lambda_0\delta_{0m}^i,\hat r_{1m}^i+\lambda_1\delta_{1m}^i,\hat r_{2m}^i+\lambda_2\delta_{2m}^i,r_m^i)\}_{i=0}^m\in\mathcal{A}_{ad}$. Define a function in $\mathbb{R}^3$ by

$$g(\lambda_0,\lambda_1,\lambda_2)=J\big(\{(\hat z_m^i+\lambda_0\delta_{0m}^i,\hat r_{1m}^i+\lambda_1\delta_{1m}^i,\hat r_{2m}^i+\lambda_2\delta_{2m}^i,r_m^i)\}_{i=0}^m\big).$$

Obviously g has a minimum at $(0,0,0)$. Hence, $\nabla g(0,0,0)=0$. By $\frac{\partial g(0,0,0)}{\partial\lambda_1}=\frac{\partial g(0,0,0)}{\partial\lambda_2}=0$, and noting that $\{(\hat z_m^i,\hat r_{1m}^i,\hat r_{2m}^i,\hat r_m^i)\}_{i=0}^m$ satisfy the first equation in (6.3), one gets

$$\begin{aligned}&-K\sum_{i=1}^{m-1}\int_\Omega\hat r_m^i\frac{\delta_{1m}^{i+1}-\delta_{1m}^i}{h}dx+\sum_{i=1}^m\int_\Omega\varrho\frac{\hat r_{1m}^i\delta_{1m}^i}{\lambda^2}e^{-2\lambda\phi_m^i}dx=0,\\&-K\sum_{i=1}^{m-1}\int_\Omega\hat r_m^i\delta_{2m}^idx+\sum_{i=1}^{m-1}\int_\Omega\varrho\frac{\hat r_{2m}^i\delta_{2m}^i}{\lambda^4}e^{-2\lambda\phi_m^i}dx=0,\end{aligned}$$

which, combined with (6.6) and $p_m^0 = p_m^m = \hat{r}_{2m}^m = 0$ in Ω, gives (6.8). From $\frac{\partial g(0,0,0)}{\partial \lambda_0} = 0$, we obtain

$$\sum_{i=1}^{m-1} \int_\Omega \left\{ K\hat{r}_m^i \left[\frac{\delta_{0m}^{i+1} - 2\delta_{0m}^i + \delta_{0m}^{i-1}}{h^2} - \sum_{j_1,j_2=1}^{n} \partial_{x_{j_2}} (a^{j_1 j_2} \partial_{x_{j_1}} \delta_{0m}^i) \right] + \hat{z}_m^i \delta_{0m}^i e^{-2\lambda\phi_m^i} \right\} dx = 0, \tag{12.1}$$

which, combined with $p_m^0 = p_m^m = \delta_{0m}^0 = \delta_{0m}^m = 0$ in Ω, implies that $p_m^i = K\hat{r}_m^i$ is a weak solution of (6.9). By means of the regularity theory for elliptic equations of second order, one sees that $\hat{z}_m^i,\ p_m^i \in H^2(\Omega) \cap H_0^1(\Omega)$ for $1 \le i \le m-1$.

Step 3. Recalling that $\{(\hat{z}_m^i, \hat{r}_{1m}^i, \hat{r}_{2m}^i, \hat{r}_m^i)\}_{i=0}^m$ satisfy (6.3), and noting (6.7)–(6.9) and $p_m^i = K\hat{r}_m^i$, one gets

$$\begin{aligned}
0 &= \sum_{i=1}^{m-1} \int_\Omega \left(\frac{\hat{z}_m^{i+1} - 2\hat{z}_m^i + \hat{z}_m^{i-1}}{h^2} - \sum_{j_1,j_2=1}^{n} \partial_{x_{j_2}} (a^{j_1 j_2} \partial_{x_{j_1}} \hat{z}_m^i) \right. \\
&\qquad \left. - \frac{\hat{r}_{1m}^{i+1} - \hat{r}_{1m}^i}{h} - \hat{r}_{2m}^i - \lambda u_m^i e^{2\lambda\phi_m^i} - \hat{r}_m^i \right) p_m^i dx \\
&= \sum_{i=1}^{m-1} \int_\Omega \left(\frac{p_m^{i+1} - 2p_m^i + p_m^{i-1}}{h^2} - \sum_{j_1,j_2=1}^{n} \partial_{x_{j_2}} (a^{j_1 j_2} \partial_{x_{j_1}} p_m^i) \right) \hat{z}_m^i dx \\
&\quad + \sum_{i=1}^{m} \int_\Omega \frac{p_m^i - p_m^{i-1}}{h} \hat{r}_{1m}^i dx - \sum_{i=1}^{m-1} \int_\Omega \left(\hat{r}_{2m}^i + \lambda u_m^i e^{2\lambda\phi_m^i} + \hat{r}_m^i \right) p_m^i dx \\
&= -\sum_{i=1}^{m-1} \left[\int_\Omega |\hat{z}_m^i|^2 e^{-2\lambda\phi_m^i} dx + \int_\Omega \varrho \left(\frac{|\hat{r}_{1m}^i|^2}{\lambda^2} + \frac{|\hat{r}_{2m}^i|^2}{\lambda^4} \right) e^{-2\lambda\phi_m^i} dx \right. \\
&\qquad \left. + K \int_\Omega |\hat{r}_m^i|^2 dx \right] - \int_\Omega \varrho \frac{|\hat{r}_{1m}^m|^2}{\lambda^2} e^{-2\lambda\phi_m^m} dx - \lambda \sum_{i=1}^{m-1} \int_\Omega u_m^i e^{2\lambda\phi_m^i} p_m^i dx.
\end{aligned} \tag{12.2}$$

Using the Hölder inequality, by (12.2) and (6.8) we conclude that there is a constant $C = C(K, \lambda) > 0$, independent of m, such that

$$\begin{aligned}
&\sum_{i=1}^{m-1} \left[\int_\Omega |\hat{z}_m^i|^2 e^{-2\lambda\phi_m^i} dx + \int_\Omega \varrho \left(\frac{|\hat{r}_{1m}^i|^2}{\lambda^2} + \frac{|\hat{r}_{2m}^i|^2}{\lambda^4} \right) e^{-2\lambda\phi_m^i} dx + K \int_\Omega |\hat{r}_m^i|^2 dx \right] \\
&\qquad + \int_\Omega \varrho \frac{|\hat{r}_{1m}^m|^2}{\lambda^2} e^{-2\lambda\phi_m^m} dx \\
&\le C \sum_{i=1}^{m-1} \int_\Omega |u_m^i|^2 e^{2\lambda\phi_m^i} dx.
\end{aligned}$$

This yields (6.10).

Step 4. Noting that (6.9) holds for $i = 1, 2, \ldots, m-1$, and that $p_m^0 = \hat{z}_m^0 = p_m^m = \hat{z}_m^m = 0$, one gets

$$
\begin{aligned}
&\frac{p_m^3-4p_m^2+5p_m^1}{h^4}-\sum_{j_1,j_2=1}^{n}\partial_{x_{j_2}}\left(a^{j_1j_2}\partial_{x_{j_1}}\frac{(p_m^2-2p_m^1-p_m^0)}{h^2}\right)\\
&\quad+\frac{\hat z_m^2e^{-2\lambda\phi_m^2}-2\hat z_m^1e^{-2\lambda\phi_m^1}+\hat z_m^0e^{-2\lambda\phi_m^0}}{h^2}=0 \qquad \text{in } \Omega,\\
&\frac{5p_m^{m-1}-4p_m^{m-2}+p_m^{m-3}}{h^4}-\sum_{j_1,j_2=1}^{n}\partial_{x_{j_2}}\left(a^{j_1j_2}\partial_{x_{j_1}}\frac{(p_m^m-2p_m^{m-1}+p_m^{m-2})}{h^2}\right)\\
&\quad+\frac{\hat z_m^me^{-2\lambda\phi_m^m}-2\hat z_m^{m-1}e^{-2\lambda\phi_m^{m-1}}+\hat z_m^{m-2}e^{-2\lambda\phi_m^{m-2}}}{h^2}=0 \qquad \text{in } \Omega,
\end{aligned}
\tag{12.3}
$$

and for $i=2,\dots,m-2$,

$$
\begin{aligned}
&\frac{p_m^{i+2}-4p_m^{i+1}+6p_m^i-4p_m^{i-1}+p_m^{i-2}}{h^4}\\
&\quad-\sum_{j_1,j_2=1}^{n}\partial_{x_{j_2}}\left(a^{j_1j_2}\partial_{x_{j_1}}\frac{(p_m^{i+1}-2p_m^i+p_m^{i-1})}{h^2}\right)\\
&\quad+\frac{\hat z_m^{i+1}e^{-2\lambda\phi_m^{i+1}}-2\hat z_m^ie^{-2\lambda\phi_m^i}+\hat z_m^{i-1}e^{-2\lambda\phi_m^{i-1}}}{h^2}=0 \qquad \text{in } \Omega.
\end{aligned}
\tag{12.4}
$$

By (6.3), we find

$$
\begin{aligned}
0=\sum_{i=1}^{m-1}\int_\Omega\Bigg(&\frac{\hat z_m^{i+1}-2\hat z_m^i+\hat z_m^{i-1}}{h^2}-\sum_{j_1,j_2=1}^{n}\partial_{x_{j_2}}(a^{j_1j_2}\partial_{x_{j_1}}\hat z_m^i)\\
&-\frac{\hat r_{1m}^{i+1}-\hat r_{1m}^i}{h}-\hat r_{2m}^i-\lambda u_m^ie^{2\lambda\phi_m^i}-\hat r_m^i\Bigg)\frac{(p_m^{i+1}-2p_m^i+p_m^{i-1})}{h^2}dx.
\end{aligned}
\tag{12.5}
$$

Noting $\hat z_m^0=\hat z_m^m=p_m^0=p_m^m=0$ again, and using (12.3)–(12.4), we arrive at

$$
\begin{aligned}
&\sum_{i=1}^{m-1}\int_\Omega\frac{(\hat z_m^{i+1}-2\hat z_m^i+\hat z_m^{i-1})}{h^2}\frac{(p_m^{i+1}-2p_m^i+p_m^{i-1})}{h^2}dx\\
&=\sum_{i=2}^{m-1}\int_\Omega\hat z_m^i\frac{(p_m^i-2p_m^{i-1}+p_m^{i-2})}{h^4}dx-2\sum_{i=1}^{m-1}\int_\Omega\hat z_m^i\frac{(p_m^{i+1}-2p_m^i+p_m^{i-1})}{h^4}dx\\
&\quad+\sum_{i=1}^{m-2}\int_\Omega\hat z_m^i\frac{(p_m^{i+2}-2p_m^{i+1}+p_m^i)}{h^4}dx\\
&=\int_\Omega\hat z_m^1\frac{(p_m^3-4p_m^2+5p_m^1)}{h^4}dx+\int_\Omega\hat z_m^{m-1}\frac{(5p_m^{m-1}-4p_m^{m-2}+p_m^{m-3})}{h^4}dx\\
&\quad+\sum_{i=2}^{m-2}\int_\Omega\hat z_m^i\frac{(p_m^{i+2}-4p_m^{i+1}+6p_m^i-4p_m^{i-1}+p_m^{i-2})}{h^4}dx\\
&=\sum_{i=1}^{m-1}\int_\Omega\hat z_m^i\Bigg\{\sum_{j_1,j_2=1}^{n}\partial_{x_{j_2}}\left[a^{j_1j_2}\partial_{x_{j_1}}\frac{(p_m^{i+1}-2p_m^i+p_m^{i-1})}{h^2}\right]\\
&\qquad\qquad-\frac{\hat z_m^{i+1}e^{-2\lambda\phi_m^{i+1}}-2\hat z_m^ie^{-2\lambda\phi_m^i}+\hat z_m^{i-1}e^{-2\lambda\phi_m^{i-1}}}{h^2}\Bigg\}dx.
\end{aligned}
\tag{12.6}
$$

Next, noting $z_m^i|_\Gamma = p_m^i|_\Gamma = 0$, for $0 \le i \le m$, one has

$$\begin{aligned}(12.7)\quad &\sum_{i=1}^{m-1}\int_\Omega \left(\sum_{j_1,j_2=1}^{n} \partial_{x_{j_2}}(a^{j_1j_2}\partial_{x_{j_1}}\hat z_m^i)\right)\frac{(p_m^{i+1}-2p_m^i+p_m^{i-1})}{h^2}dx\\ &=\sum_{i=1}^{m-1}\int_\Omega \hat z_m^i \sum_{j_1,j_2=1}^{n}\partial_{x_{j_2}}\left(a^{j_1j_2}\partial_{x_{j_1}}\frac{(p_m^{i+1}-2p_m^i+p_m^{i-1})}{h^2}\right)dx.\end{aligned}$$

Combining (12.5)–(12.7), we obtain

$$\begin{aligned}(12.8)\quad 0=&-\sum_{i=1}^{m-1}\int_\Omega\left[\hat z_m^i\frac{(\hat z_m^{i+1}e^{-2\lambda\phi_m^{i+1}}-2\hat z_m^i e^{-2\lambda\phi_m^i}+\hat z_m^{i-1}e^{-2\lambda\phi_m^{i-1}})}{h^2}\right.\\ &\left.+\left(\frac{\hat r_{1m}^{i+1}-\hat r_{1m}^i}{h}+\hat r_{2m}^i+\lambda u_m^i e^{2\lambda\phi_m^i}+\hat r_m^i\right)\frac{(p_m^{i+1}-2p_m^i+p_m^{i-1})}{h^2}\right]dx.\end{aligned}$$

By Proposition 3.5 and noting $p_m^i = K\hat r_m^i$, one has

$$\begin{aligned}(12.9)\quad &-\sum_{i=1}^{m-1}\int_\Omega\left[\hat z_m^i\frac{(\hat z_m^{i+1}e^{-2\lambda\phi_m^{i+1}}-2\hat z_m^i e^{-2\lambda\phi_m^i}+\hat z_m^{i-1}e^{-2\lambda\phi_m^{i-1}})}{h^2}\right.\\ &\qquad\left.+\hat r_m^i\frac{(p_m^{i+1}-2p_m^i+p_m^{i-1})}{h^2}\right]dx\\ &=\sum_{i=0}^{m-1}\int_\Omega\left[\frac{(\hat z_m^{i+1}-\hat z_m^i)}{h}\frac{(\hat z_m^{i+1}e^{-2\lambda\phi_m^{i+1}}-\hat z_m^i e^{-2\lambda\phi_m^i})}{h}+K\frac{(\hat r_m^{i+1}-\hat r_m^i)^2}{h^2}\right]dx\\ &=\sum_{i=0}^{m-1}\int_\Omega\left[\frac{(\hat z_m^{i+1}-\hat z_m^i)^2}{h^2}e^{-2\lambda\phi_m^i}+\frac{(\hat z_m^{i+1}-\hat z_m^i)}{h}\frac{(e^{-2\lambda\phi_m^{i+1}}-e^{-2\lambda\phi_m^i})}{h}\hat z_m^{i+1}\right.\\ &\qquad\left.+K\frac{(\hat r_m^{i+1}-\hat r_m^i)^2}{h^2}\right]dx.\end{aligned}$$

Further, by (6.8), and using Proposition 3.5 again, we find

$$\begin{aligned}(12.10)\quad &-\sum_{i=1}^{m-1}\int_\Omega\left(\frac{\hat r_{1m}^{i+1}-\hat r_{1m}^i}{h}+\hat r_{2m}^i+\lambda u_m^i e^{2\lambda\phi_m^i}\right)\frac{(p_m^{i+1}-2p_m^i+p_m^{i-1})}{h^2}dx\\ &=-\sum_{i=1}^{m-1}\int_\Omega\left(\frac{\hat r_{1m}^{i+1}-\hat r_{1m}^i}{h}+\lambda u_m^i e^{2\lambda\phi_m^i}\right)\frac{1}{h}\left(\frac{p_m^{i+1}-p_m^i}{h}-\frac{p_m^i-p_m^{i-1}}{h}\right)dx\\ &\quad+\sum_{i=0}^{m-1}\int_\Omega\frac{(\hat r_{2m}^{i+1}-\hat r_{2m}^i)}{h}\frac{(p_m^{i+1}-p_m^i)}{h}dx\\ &=\sum_{i=1}^{m-1}\int_\Omega\frac{\varrho}{\lambda^2}\left(\frac{\hat r_{1m}^{i+1}-\hat r_{1m}^i}{h}+\lambda u_m^i e^{2\lambda\phi_m^i}\right)\frac{(\hat r_{1m}^{i+1}e^{-2\lambda\phi_m^{i+1}}-\hat r_{1m}^i e^{-2\lambda\phi_m^i})}{h}dx\\ &\quad+\sum_{i=0}^{m-1}\int_\Omega\frac{\varrho}{\lambda^4}\frac{(\hat r_{2m}^{i+1}-\hat r_{2m}^i)}{h}\frac{(\hat r_{2m}^{i+1}e^{-2\lambda\phi_m^{i+1}}-\hat r_{2m}^i e^{-2\lambda\phi_m^i})}{h}dx\\ &=\sum_{i=1}^{m-1}\int_\Omega\frac{\varrho}{\lambda^2}\left[\frac{(\hat r_{1m}^{i+1}-\hat r_{1m}^i)^2}{h^2}e^{-2\lambda\phi_m^i}\right.\\ &\qquad\left.+\frac{(\hat r_{1m}^{i+1}-\hat r_{1m}^i)}{h}\frac{(e^{-2\lambda\phi_m^{i+1}}-e^{-2\lambda\phi_m^i})}{h}\hat r_{1m}^{i+1}\right]dx\end{aligned}$$

$$+\lambda\sum_{i=1}^{m-1}\int_\Omega \frac{\varrho}{\lambda^2}u_m^i\left[\frac{(\hat r_{1m}^{i+1}-\hat r_{1m}^i)}{h}e^{-2\lambda\phi_m^i}+\frac{(e^{-2\lambda\phi_m^{i+1}}-e^{-2\lambda\phi_m^i})}{h}\hat r_{1m}^{i+1}\right]dx$$
$$+\sum_{i=0}^{m-1}\int_\Omega \frac{\varrho}{\lambda^4}\left[\frac{(\hat r_{2m}^{i+1}-\hat r_{2m}^i)^2}{h^2}e^{-2\lambda\phi_m^i}\right.$$
$$\left.+\frac{(\hat r_{2m}^{i+1}-\hat r_{2m}^i)}{h}\frac{(e^{-2\lambda\phi_m^{i+1}}-e^{-2\lambda\phi_m^i})}{h}\hat r_{2m}^{i+1}\right]dx.$$

Combining (12.8)–(12.10), and noting that $\hat r_{1m}^1=\hat r_{1m}^0$, $u_m^0=0$, we end up with

$$\begin{aligned}
&\sum_{i=0}^{m-1}\int_\Omega\left[\frac{(\hat z_m^{i+1}-\hat z_m^i)^2}{h^2}e^{-2\lambda\phi_m^i}+\frac{\varrho}{\lambda^2}\frac{(\hat r_{1m}^{i+1}-\hat r_{1m}^i)^2}{h^2}e^{-2\lambda\phi_m^i}\right.\\
&\qquad\left.+\frac{\varrho}{\lambda^4}\frac{(\hat r_{2m}^{i+1}-\hat r_{2m}^i)^2}{h^2}e^{-2\lambda\phi_m^i}+K\frac{(\hat r_m^{i+1}-\hat r_m^i)^2}{h^2}\right]dx\\
&=-\sum_{i=0}^{m-1}\int_\Omega\frac{(\hat z_m^{i+1}-\hat z_m^i)}{h}\frac{(e^{-2\lambda\phi_m^{i+1}}-e^{-2\lambda\phi_m^i})}{h}\hat z_m^{i+1}dx\\
&\quad-\sum_{i=1}^{m-1}\int_\Omega\frac{\varrho}{\lambda^2}\frac{(\hat r_{1m}^{i+1}-\hat r_{1m}^i)}{h}\frac{(e^{-2\lambda\phi_m^{i+1}}-e^{-2\lambda\phi_m^i})}{h}\hat r_{1m}^{i+1}dx\\
&\quad-\lambda\sum_{i=1}^{m-1}\int_\Omega\frac{\varrho}{\lambda^2}u_m^i\left[\frac{(\hat r_{1m}^{i+1}-\hat r_{1m}^i)}{h}e^{-2\lambda\phi_m^i}+\frac{(e^{-2\lambda\phi_m^{i+1}}-e^{-2\lambda\phi_m^i})}{h}\hat r_{1m}^{i+1}\right]dx\\
&\quad-\sum_{i=0}^{m-1}\int_\Omega\frac{\varrho}{\lambda^4}\frac{(\hat r_{2m}^{i+1}-\hat r_{2m}^i)}{h}\frac{(e^{-2\lambda\phi_m^{i+1}}-e^{-2\lambda\phi_m^i})}{h}\hat r_{2m}^{i+1}dx.
\end{aligned}\tag{12.11}$$

Using the Hölder inequality and noting that ϕ is a smooth function, from (12.11) we conclude that there is a positive constant $C=C(K,\lambda)$, independent of m, such that

$$\begin{aligned}
&\sum_{i=0}^{m-1}\int_\Omega\left[\frac{(\hat z_m^{i+1}-\hat z_m^i)^2}{h^2}e^{-2\lambda\phi_m^i}+\frac{\varrho}{\lambda^2}\frac{(\hat r_{1m}^{i+1}-\hat r_{1m}^i)^2}{h^2}e^{-2\lambda\phi_m^i}\right.\\
&\qquad\left.+\frac{\varrho}{\lambda^4}\frac{(\hat r_{2m}^{i+1}-\hat r_{2m}^i)^2}{h^2}e^{-2\lambda\phi_m^i}+K\frac{(\hat r_m^{i+1}-\hat r_m^i)^2}{h^2}\right]dx\\
&\le C\left[\sum_{i=1}^{m-1}\int_\Omega\Big(|\hat z_m^i|^2+|\hat r_{1m}^i|^2+|\hat r_{2m}^i|^2+K|\hat r_m^i|^2+|u_m^i|^2\Big)dx\right.\\
&\qquad\left.+\int_\Omega|\hat r_{1m}^m|^2dx\right].
\end{aligned}\tag{12.12}$$

Finally, combining (12.12) and (6.10), and recalling that $u\in C([0,T];L^2(\Omega))$, we establish the desired estimate (6.11). This completes the proof of Proposition 6.1. □

Acknowledgments. The authors acknowledge the anonymous referees for their comments which led to this improved version. The third author also thanks Professor M. Yamamoto for stimulating discussion.

REFERENCES

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

[2] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[3] P. Cannarsa, V. Komornik, and P. Loreti, *One-sided and internal controllability of semilinear wave equations with infinitely iterated logarithms*, Discrete Contin. Dyn. Syst. B, 8 (2002), pp. 745–756.

[4] T. Cazenave and A. Haraux, *Equations d'évolution avec non-linéarité logarithmique*, Ann. Fac. Sci. Toulouse, 2 (1980), pp. 21–51.

[5] W. C. Chewning, *Controllability of the nonlinear wave equation in several space variables*, SIAM J. Control Optim., 14 (1976), pp. 19–25.

[6] M. Cirinà, *Boundary controllability of nonlinear hyperbolic systems*, SIAM J. Control, 7 (1969), pp. 198–212.

[7] A. Doubova and A. Osses, *Rotated weights in global Carleman estimates applied to an inverse problem for the wave equation*, Inverse Problems, 33 (2006), pp. 265–296.

[8] T. Duyckaerts, X. Zhang, and E. Zuazua, *On the optimality of the observability inequalities for parabolic and hyperbolic systems with potentials*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[9] O. Yu. Émanuilov, *Boundary controllability of semilinear evolution equations*, Russian Math. Surveys, 44 (1989), pp. 183–184.

[10] H. O. Fattorini, *Local controllability of a nonlinear wave equation*, Math. Systems Theory, 9 (1975), pp. 30–45.

[11] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.

[12] X. Fu, *A weighted identity for partial differential operators of second order and its applications*, C. R. Math. Acad. Sci. Paris, 342 (2006), pp. 579–584.

[13] X. Fu, J. Yong, and X. Zhang, *Exact Controllability of the Heat Equation with Hyperbolic Memory Kernel in Anisotropic and Non-Homogeneous Media*, preprint, 2005.

[14] A. V. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Series 34, Research Institute of Mathematics, Seoul National University, Seoul, Korea, 1994.

[15] L. Hörmander, *On the Nash-Moser implicit function theorem*, Ann. Acad. Sci. Fenn. Ser. A I Math., 10 (1985), pp. 255–259.

[16] L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1969.

[17] L. Hörmander, *The Analysis of Linear Partial Differential Operators* IV, Springer, Berlin, 1985.

[18] O. Yu. Imanuvilov, *On Carleman estimates for hyperbolic equations*, Asymptot. Anal., 32 (2002), pp. 185–220.

[19] V. Isakov and M. Yamamoto, *Carleman estimate with the Neumann boundary condition and its applications to the observability inequality and inverse problems*, in Differential Geometric Methods in the Control of Partial Differential Equations (Boulder, CO, 1999), Contemp. Math. 268, AMS, Providence, RI, 2000, pp. 191–225.

[20] M. Kazemi and M. V. Klibanov, *Stability estimates for ill-posed Cauchy problem involving hyperbolic equations and inequalities*, Appl. Anal., 50 (1993), pp. 93–102.

[21] M. V. Klibanov and J. Malinsky, *Newton-Kantorovich method for* 3*-dimensional potential inverse scattering problem and stability of the hyperbolic Cauchy problem with time dependent data*, Inverse Problems, 7 (1991), pp. 577–595.

[22] M. V. Klibanov and A. Timonov, *Carleman Estimates for Coefficient Inverse Problems and Numerical Applications*, VSP, Utrecht, 2004.

[23] I. Lasiecka and R. Triggiani, *Exact controllability of semilinear abstract systems with application to waves and plates boundary control problems*, Appl. Math. Optim., 23 (1991), pp. 109–154.

[24] I. Lasiecka, R. Triggiani, and X. Zhang, *Nonconservative wave equations with unobserved Neumann B.C.: Global uniqueness and observability in one shot*, in Differential Geometric Methods in the Control of Partial Differential Equations (Boulder, CO, 1999), Contemp. Math. 268, AMS, Providence, RI, 2000, pp. 227–325.

[25] M. M. Lavrent'ev, V. G. Romanov, and S. P. Shishatskii, *Ill-posed problems of mathematical physics and analysis*, translated from the Russian by J. R. Schulenberger, Transl. Math. Monogr. 64, AMS, Providence, RI, 1986.

[26] L. Li and X. Zhang, *Exact controllability for semilinear wave equations*, J. Math. Anal. Appl., 250 (2000), pp. 589–597.
[27] T.-T. Li and B.-P. Rao, *Exact boundary controllability for quasi-linear hyperbolic systems*, SIAM J. Control Optim., 41 (2003), pp. 1748–1755.
[28] X. Li and J. Yong, *Optimal Control Theory for Infinite-Dimensional Systems*, Systems Control Found. Appl., Birkhäuser Boston, Inc., Boston, MA, 1995.
[29] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués, Tome* 1. *Contrôlabilité exacte*, Rech. Math. Appl. 8, Masson, Paris, 1988.
[30] P. Martinez and J. Vancostenoble, *Exact controllability in "arbitrarily short time" of the semilinear wave equation*, Discrete Contin. Dyn. Syst. B, 9 (2003), pp. 901–924.
[31] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.
[32] J. Rauch, X. Zhang, and E. Zuazua, *Polynomial decay of a hyperbolic-parabolic coupled system*, J. Math. Pures Appl., 84 (2005), pp. 407–470.
[33] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
[34] T. I. Seidman, *Invariance of the reachable set under nonlinear perturbations*, SIAM J. Control Optim., 25 (1987), pp. 1173–1191.
[35] D. Tataru, *Carleman estimates and unique continuation for solutions to boundary value problems*, J. Math. Pures Appl., 75 (1996), pp. 367–408.
[36] X. Zhang, *Explicit observability estimate for the wave equation with potential and its application*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 456 (2000), pp. 1101–1115.
[37] X. Zhang, *Explicit observability inequalities for the wave equation with lower order terms by means of Carleman inequalities*, SIAM J. Control Optim., 39 (2000), pp. 812–834.
[38] X. Zhang, *Exact controllability of the semilinear plate equations*, Asymptot. Anal., 27 (2001), pp. 95–125.
[39] X. Zhang, *Exact Controllability of Semi-Linear Distributed Parameter Systems*, Higher Education Press, Beijing, 2004 (in Chinese).
[40] X. Zhang and E. Zuazua, *Decay of solutions of the system of thermoelasticity of type* III, Commun. Contemp. Math., 5 (2003), pp. 25–83.
[41] X. Zhang and E. Zuazua, *Polynomial decay and control of a* 1*-d hyperbolic-parabolic coupled system*, J. Differential Equations, 204 (2004), pp. 380–438.
[42] X. Zhang and E. Zuazua, *Exact controllability of the semi-linear wave equation*, in Sixty Open Problems in the Mathematics of Systems and Control, V. D. Blondel and A. Megretski, eds., Princeton University Press, Princeton, NJ, 2004, pp. 173–178.
[43] E. Zuazua, *Exact controllability for the semilinear wave equation*, J. Math. Pures Appl., 69 (1990), pp. 1–31.
[44] E. Zuazua, *Exact boundary controllability for the semilinear wave equation*, in Nonlinear Partial Differential Equations and their Applications, Collège de France Seminar, Vol. X (Paris, 1987–1988), Pitman Res. Notes Math. Ser. 220, Longman Sci. Tech., Harlow, 1991, pp. 357–391.
[45] E. Zuazua, *Exact controllability for semilinear wave equations in one space dimension*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 109–129.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1615–1636

CONVERGENCE IN NONLINEAR FILTERING FOR STOCHASTIC DELAY SYSTEMS*

ANTONELLA CALZOLARI†, PATRICK FLORCHINGER‡, AND GIOVANNA NAPPO§

Abstract. We study an approximation scheme for a nonlinear filtering problem when the state process $\boldsymbol{X}$ is the solution of a stochastic delay diffusion equation and the observation process is a noisy function of $X(s)$ for $s \in [t-\tau, t]$, where τ is a constant. The approximating state is the piecewise linear Euler–Maruyama scheme, and the observation process is a noisy function of the approximating state. The rate of convergence of this scheme is computed.

Key words. conditional laws, strong approximation, stochastic delay differential equations, rate of convergence

AMS subject classifications. 60G35, 62M20, 93E10, 60H35

DOI. 10.1137/050646135

1. Introduction.

Stochastic diffusion processes with delay have been used as models in many applications: in population dynamics (see Goel, Maitra, and Montroll [16]), in respiratory systems (see Longtin et al. [30]), in eye movement control (see Vasilakos and Beuter [42]), in postural control (see Peterka [37]), and in transmission delays for neural networks and/or ensemble of coupled neural oscillators (see Niebur, Schuster, and Kammen [35]).

In most of the literature the stochastic process is assumed to be completely observable. However, this cannot always be the case, since measurement errors may occur. This difficulty can be overcome by modelling this situation as a nonlinear filtering problem.

The aim of stochastic nonlinear filtering is to compute the conditional law at time t of a state process, which cannot be directly observed, given an observation process up to time t. This task can be achieved only in a few specific cases, and therefore the problem of the approximation of the conditional law naturally arises.

A classical model of partially observed system extensively studied in the last few years arises when both the state and the observation processes are diffusion processes.

For this model, under suitable hypotheses on the coefficients, the filtering equations have been established by Kushner [23], Duncan [10], Mortensen [34], and Zakai [43] and have been studied since then by many authors (see, for example, Pardoux [36] or Kallianpur [18] and the references therein). Different approximation schemes for the filter have been studied in various frameworks by many authors (see, for example, Kushner [24], Le Gland [28], or Del Moral [9] and the references therein).

In this paper we are interested in nonlinear filtering of partially observed delay systems of the following form.

*Received by the editors November 26, 2005; accepted for publication (in revised form) April 16, 2007; published electronically November 2, 2007. This work was partially supported by PRIN 2006 project "Stochastic Methods in Finance" of the Ministero dell'Università e della Ricerca (MIUR).
http://www.siam.org/journals/sicon/46-5/64613.html

†Dipartimento di Matematica, Università di Roma "Tor Vergata," via della Ricerca Scientifica 1, I 00133 Roma, Italy (calzolar@axp.mat.uniroma2.it).

‡Département de Mathématiques, Université de Metz, 23 Allée des Oeillets, F 57160 Moulins les Metz, France (patrick.florchinger@wanadoo.fr).

§Dipartimento di Matematica, Università di Roma "La Sapienza," piazzale A. Moro 2, I 00185 Roma, Italy (nappo@mat.uniroma1.it).

The state process $\boldsymbol{X} = \big(X(t)\big)_{t\in[-\tau,T]}$ satisfies the stochastic delay differential equation on the probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, P)$

$$\begin{cases} X(t) = \eta(t), & -\tau \le t \le 0, \\ X(t) = \eta(0) + \displaystyle\int_0^t a(u, \Pi_u X)du + \int_0^t b(u, \Pi_u X)d\tilde{W}_u, & 0 \le t \le T, \end{cases} \tag{1.1}$$

where τ is a positive constant, $\big(\Pi_t X\big)_{t\in[0,T]}$ is a $C([-\tau,0],\mathbb{R})$-valued random process defined by

$$\Pi_t X(s) = X(t+s), \qquad -\tau \le s \le 0,$$

$\tilde{\boldsymbol{W}} = \big(\tilde{W}(t)\big)_{t\in[0,T]}$ is a standard Brownian motion, and $\boldsymbol{\eta} = \big(\eta(s)\big)_{s\in[-\tau,0]}$ is a $C([-\tau,0],\mathbb{R})$-valued random variable.

The observation process $\boldsymbol{Y} = \big(Y(t)\big)_{t\in[0,T]}$ is given by

$$Y(t) = \int_0^t h(u, \Pi_u X)du + W(t), \qquad 0 \le t \le T, \tag{1.2}$$

where $\boldsymbol{W} = \big(W(t)\big)_{t\in[0,T]}$ is a standard Brownian motion, independent of $\tilde{\boldsymbol{W}}$, and $h : [0,T] \times C([-\tau,0],\mathbb{R}) \to \mathbb{R}$ is a Borel measurable function.

As an example the functions $a(t,\theta)$, $b(t,\theta)$, and $h(t,\theta)$ for $\theta \in C([-\tau,0],\mathbb{R})$ can be taken of the form

$$g\left(t, \max_{u\in[\tau_{i-1},\tau_i]} \theta(u);\, i = 1,\dots,r\right), \tag{1.3}$$

where $-\tau = \tau_0 < \tau_1 < \cdots < \tau_r = 0$, or

$$g\left(t, \int_{-\tau}^0 \psi_i(u,\theta(u))\gamma_i(du);\, i = 1,\dots,r\right), \tag{1.4}$$

where γ_i are finite measures on $[-\tau,0]$, and g and ψ_i are continuous functions.

By taking in (1.4) $\psi_i(u,x) = x$ for all i, and $\gamma_i(ds)$ in the set $\{e^{\lambda s}ds, \delta_{-\tau}(ds), \delta_0(ds)\}$, we recover the stochastic delay differential equation considered in a control framework by Larssen and Risebro [27] and by Elsanousi and Larssen [11]

$$\begin{aligned} dX(t) = {} & g_a\left(t, X(t),\, X(t-\tau),\, \int_{t-\tau}^t e^{\lambda(u-t)}X(u)\,du,\, w(t)\right)dt \\ & + g_b\left(t, X(t),\, X(t-\tau),\, \int_{t-\tau}^t e^{\lambda(u-t)}X(u)\,du,\, w(t)\right)d\widetilde{W}_t, \end{aligned}$$

where $w(t)$ is the control.

Moreover, when the function $g_a(t,x_1,x_2,x_3,w)$, as well as $g_b(t,x_1,x_2,x_3,w)$, depends only on (x_1,x_2), we recover the fixed time delay model

$$dX(t) = g_a\big(X(t),\, X(t-\tau)\big)dt + g_b\big(X(t),\, X(t-\tau)\big)d\widetilde{W}_t, \tag{1.5}$$

which is studied in Frank [14], and in particular, for $g_a(x_1,x_2) = \big(H - K\log(x_2)\big)\,x_1$ and $g_b(x_1,x_2) = x_1$, we recover the stochastic Gompertz model with delay; see Goel,

Maitra, and Montroll [16] and Frank and Beek [15]. The stochastic Gompertz model without delay has been used in population growth (see Ricciardi [40]) or in biomedical sciences (see Ferrante et al. [12]).

At first the aim was to obtain a computable approximation for $E\big[\varphi\big(X(t)\big)/\mathcal{F}_t^Y\big]$ for all functions φ belonging to a determining class, i.e., the best estimate of $\varphi\big(X(t)\big)$ given the σ-algebra of the observations up to time t, $\mathcal{F}_t^Y = \sigma\{Y(s),\ s \le t\}$. In fact, since $\Pi_t X(0) = X(t)$, we shall give a computable approximation for the filter π_t associated with the delay system $\big(\Pi_t X, Y(t)\big)_{t\in[0,T]}$, defined for any measurable and bounded functions ϕ mapping $C([-\tau,0],\mathbb{R})$ into $\mathbb{R}$ by

$$\pi_t(\phi) = E[\phi(\Pi_t X)|\mathcal{F}_t^Y]. \tag{1.6}$$

In our paper the state is approximated by the piecewise linear Euler–Maruyama scheme (see (2.7) and (2.10)), while the observation is approximated by a diffusion (see (2.8)), which can be considered as a continuous Euler–Maruyama scheme (see Remark 2.1 for the peculiarities due to the delay). The filter process of this approximating system is the first approximation scheme of $\boldsymbol{\pi} = (\pi_t)_{t\in[0,T]}$ we consider (see (2.14)). This approximation scheme has a drawback: it depends on the approximating observation process, which is not the actual observation process $\boldsymbol{Y}$. Instead, the other approximation scheme we consider depends on the actual observation process (see (2.15)).

As the time step converges to zero, the two approximating filters converge in probability to $\boldsymbol{\pi}$ as measure-valued processes (see Theorem 2.2).

To our knowledge there are only three papers dealing with nonlinear filtering for delay systems: Kwong and Willsky [26], Chang [8], and Kallianpur and Mandal [20].

In [26] Kwong and Willsky give a characterization of the optimal filter when dealing with nonlinear delay systems with Gaussian noises, i.e., with b depending only on time. A Fujisaki–Kallianpur–Kunita equation for the filter is deduced from a representation result which characterizes conditional moment functionals of nonlinear delay systems. However, the uniqueness of the solution of this equation is not guaranteed.

In [8] Chang gives a computable approximation for the optimal filter when dealing with one-dimensional nonlinear delay filtering systems with $b = 1$. The original model is approximated by a discrete-time model obtained by applying an Euler discretization scheme. An optimal filter for the approximate system is obtained by an explicit procedure, and the weak convergence of the approximating process and the approximating filter to the original ones is verified.

In [20] Kallianpur and Mandal study a nonlinear filtering problem where the state process is solution of the stochastic delay differential equation (1.1), in the homogeneous case, and the observation process is given by (1.2). By using some extensions of results obtained by Mohammed [32] for stochastic delay differential equations, they prove that the signal process is the unique solution to an appropriate martingale problem. By taking this fact into account the authors prove that the optimal filter corresponding to the nonlinear filtering problem solves a Zakai-type equation. The uniqueness for the solution of the Zakai equation is deduced from the results of Bhatt, Kallianpur, and Karandikar [2], and a Fujisaki–Kallianpur–Kunita equation for the filter is deduced from the Zakai equation by usual arguments in nonlinear filtering theory.

In addition to the previous references we also quote the paper [3] by Bhatt, Kallianpur, and Karandikar, which is the starting point in some of our analysis, and Bhatt and Karandikar [4]. Though none of these papers is explicitly connected with

filtering models involving delays, the results achieved by these authors can be used in the delay context.

This paper is divided into six sections and is organized as follows. In section 2, we introduce the approximation scheme for the system we are dealing with in this paper and state the main results. The first result concerns the convergence of the approximation schemes for the filter, while the second result deals with the rate of convergence w.r.t. the bounded Lipschitz metric. In section 3, we prove the convergence for the filter by making use of a convergence result deduced from the papers by Bhatt, Kallianpur, and Karandikar [3] and Bhatt and Karandikar [4]. In section 4, we compute an upper bound for the rate of convergence w.r.t. the bounded Lipschitz metric for our approximation scheme by combining filter approximation techniques similar to those in Calzolari, Florchinger, and Nappo [7] with a convergence result for the approximation of stochastic delay differential equations. In section 5, we give the proofs of some technical results that we use in sections 3 and 4. Moreover we recall a result on the expectation of the modulus of continuity for diffusion due to Słomiński [41], which we use in the proof of the approximation result for stochastic delay differential equations. In section 6, we conclude by giving a comparison between our result and the one by Chang [8], and we discuss briefly some works on approximation for stochastic delay differential equations, in particular the one by Mao and Sabanis [31].

2. Approximation scheme and main results. For the partially observed delay system (1.1) and (1.2) stated above we assume the following standing hypotheses:

(A1) η is a $\mathcal{F}_0$-measurable $C([-\tau,0],\mathbb{R})$-valued random variable, with

$$E\left(\|\Pi_0 X\|^{2k}\right) = E\left(\sup_{s\in[-\tau,0]}|\eta(s)|^{2k}\right) < \infty, \qquad k=1,2.$$

(A2) The functionals $a(t,\theta)$ and $b(t,\theta)$ on $[0,T]\times C([-\tau,0],\mathbb{R})$ are jointly globally continuous, Hölder in time, and Lipschitz in space, i.e.,

$$|a(t,\theta)-a(t',\bar\theta)|^2 + |b(t,\theta)-b(t',\bar\theta)|^2 \le K\left(|t-t'|^{2\alpha} + \|\theta-\bar\theta\|^2\right), \qquad \alpha>0, \tag{2.1}$$

and satisfy the growth condition

$$|a(t,\theta)|^2 + |b(t,\theta)|^2 \le K\left(1+\|\theta\|^2\right) \tag{2.2}$$

for some constant $K>0$.

(A3) $h:[0,T]\times C([-\tau,0],\mathbb{R})\to\mathbb{R}$ is jointly continuous and sublinear, i.e.,

$$|h(t,\theta)|^2 \le K(1+\|\theta\|^2).$$

Conditions (A1) and (A2), with $t'=t$ in (2.1), assure the existence and the uniqueness of the solution of (1.1) together with

$$E\left[\sup_{u\in[0,T]}\|\Pi_u X\|^{2k}\right] < \infty, \qquad k=1,2 \tag{2.3}$$

(see [32, Theorem II.2.1 and Lemma III.1.2] and [33, Theorem I.2]). Note that, under condition (A2), with $t'=t$ in (2.1), the existence and the uniqueness of the solution of (1.1) follow without condition (A1) (see Kallianpur and Mandal [20]). The latter

condition is used to obtain (2.3), which together with the sublinearity of h implies that

$$\int_0^T E[|h(u,\Pi_u X)|^{2k}]du < \infty, \qquad k=1,2.$$

The above condition, for $k=1$, together with the independence of the noises, is a classical assumption in nonlinear filtering theory which guarantees that the filter π_t can be represented via a Kallianpur–Striebel formula

$$\pi_t(\phi) = \frac{\sigma_t(\phi)}{\sigma_t(\mathbf{1})},$$

with

$$\sigma_t(\phi) = E^0\left[\phi(\Pi_t X)\exp\left\{\int_0^t h(s,\Pi_s X)dY_s - \frac{1}{2}\int_0^t |h(s,\Pi_s X)|^2\,ds\right\}\Big|\mathcal{F}_t^Y\right], \tag{2.4}$$

where E^0 denotes the expectation w.r.t. the reference probability measure P^0, defined by the Radon–Nikodym derivative

$$\frac{dP^0}{dP} = \exp\left\{-\int_0^T h(s,\Pi_s X)dY_s + \frac{1}{2}\int_0^T |h(s,\Pi_s X)|^2\,ds\right\}. \tag{2.5}$$

The independence of $\boldsymbol{X}$ and $\boldsymbol{W}$ under P implies the independence of $\boldsymbol{X}$ and $\boldsymbol{Y}$ under P^0; furthermore the law of $\boldsymbol{X}$ is the same under P and P^0 (see, e.g., Bhatt, Kallianpur, and Karandikar [3] and Kallianpur and Mandal [20]). This fact will play a fundamental role in the proof of our approximation results. In particular it implies that there exists a deterministic functional, with values in $\mathcal{P}(C([-\tau,0],\mathbb{R}))$, the metric space of probability measures on $C([-\tau,0],\mathbb{R})$, endowed with the Prohorov metric,

$$\begin{aligned} U:\ [0,T]\times C([0,T],\mathbb{R}) &\to \mathcal{P}(C([-\tau,0],\mathbb{R})),\\ (t,\boldsymbol{y}) &\mapsto U\big(t,\boldsymbol{y}\big)\end{aligned}$$

with the property $U\big(t,\boldsymbol{y}\big) = U\big(t,y(\cdot\wedge t)\big)$, and such that

$$\pi_t = U\big(t,\boldsymbol{Y}\big). \tag{2.6}$$

We recall that the paths $t\mapsto U(t,\boldsymbol{y})$ are right continuous with left limits, i.e., belong to the Skorohod space $D_{\mathcal{P}(C([-\tau,0],\mathbb{R}))}([0,T])$ of *càdlàg* functions with values in $\mathcal{P}(C([-\tau,0],\mathbb{R}))$, and therefore the same property holds for $t\mapsto\pi_t$.

In this paper we consider the following approximation scheme.

The approximation $\boldsymbol{X^n} = \big(X^n(t)\big)_{t\in[-\tau,T]}$ of the state process $\boldsymbol{X} = \big(X(t)\big)_{t\in[-\tau,T]}$ is the piecewise linear Euler–Maruyama scheme, that is, the linear interpolation of the Euler discretization scheme with step $\delta=\delta_n = T/n$, with $\tau = m\delta$ (as in Chang [8], for the sake of simplicity, we assume that T/τ is rational):[1]

$$\begin{cases} X^n(\ell\delta) = \eta(\ell\delta), & -m\le\ell\le 0,\\ X^n((\ell+1)\delta) = X^n(\ell\delta) + a(\ell\delta,\Pi_{\ell\delta}X^n)\delta & \\ \qquad\qquad + b(\ell\delta,\Pi_{\ell\delta}X^n)\big[\tilde W((\ell+1)\delta) - \tilde W(\ell\delta)\big], & 0\le\ell\le n-1. \end{cases} \tag{2.7}$$

[1] It is clear that, assuming $T=\frac{p}{q}\tau$, we first fix $m=kq$ a multiple of q and then set $\delta=\tau/m$, so that $T=kp\delta$ and $n=kp$. Or better, we first fix m, then set $\delta=\tau/m$, and finally take the interval $[-\tau,\lfloor T/\delta\rfloor\delta]$, instead of $[-\tau,T]$, so that $n=n(m)$.

With this approximation for the state process $\boldsymbol{X}$, we can consider the piecewise-constant $C([-\tau,0],\mathbb{R})$-valued process $\big(\Pi_{\lfloor t/\delta\rfloor\cdot\delta}X^n\big)_{t\in[0,T]}$ as an approximation of the $C([-\tau,0],\mathbb{R})$-valued process $\big(\Pi_tX\big)_{t\in[0,T]}$.

For the approximation of the observation process, we define $\boldsymbol{Y^n}=\big(Y^n(t)\big)_{t\in[0,T]}$ by

$$Y^n(t)=\int_0^t h(\lfloor s/\delta\rfloor\cdot\delta,\Pi_{\lfloor s/\delta\rfloor\cdot\delta}X^n)ds+W(t),\qquad 0\le t\le T, \tag{2.8}$$

where $\lfloor x\rfloor$ is the integer part of x.

Remark 2.1. *Note that, unlike in the finite-dimensional Euler scheme, the interpolation has to be performed at every step in order to evaluate $\Pi_{\ell\delta}X^n$. Nevertheless it is clear that*

$$\big\{\big(X^n(\ell\delta),X^n((\ell-1)\delta),\dots,X^n((\ell-m)\delta)\big)\big\}_{0\le\ell\le n} \tag{2.9}$$

is an $(m+1)$-dimensional Markov chain, and for $t\in[\ell\delta,(\ell+1)\delta]$, $0\le\ell\le n-1$,

$$\begin{aligned}X^n(t)&=X^n(\ell\delta)+a(\ell\delta,\Pi_{\ell\delta}X^n)(t-\ell\delta)\\&\quad+b(\ell\delta,\Pi_{\ell\delta}X^n)\big[\tilde W((\ell+1)\delta)-\tilde W(\ell\delta)\big](t-\ell\delta)/\delta,\end{aligned} \tag{2.10}$$

with $X^n(0)=\eta(0)$, and

$$Y^n(t)=Y^n(\ell\delta)+h(\ell\delta,\Pi_{\ell\delta}X^n)(t-\ell\delta)+\big[W(t)-W(\ell\delta)\big], \tag{2.11}$$

with $Y^n(0)=0$.

When the state is given by fixed time delay model (1.5)*, the linear interpolation, in the above discrete Euler–Maruyama scheme, is not needed in order to compute the sequence $\{X^n(\ell\delta)\}_{0\le\ell\le n}$. Indeed in this case*

$$\begin{aligned}X^n((\ell+1)\delta)&=X^n(\ell\delta)+g_a\big(\ell\delta,X^n(\ell\delta),X^n((\ell-m)\delta)\big)\delta\\&\quad+g_b\big(\ell\delta,X^n(\ell\delta),X^n((\ell-m)\delta)\big)\big[\tilde W((\ell+1)\delta)-\tilde W(\ell\delta)\big],\end{aligned}$$

with $X^n(0)=\eta(0)$, and therefore the computation of the discrete Markov chain (2.9) *is much simpler.*

The process $\boldsymbol{X^n}$ is neither adapted nor Markov, and therefore one cannot use the results of [2] to characterize the filter as the unique solution of the Zakai equation. Nevertheless the signal noise $\boldsymbol{W}$ is independent of $\boldsymbol{X^n}$, and the same holds for the approximated state process $\big(\Pi_{\lfloor t/\delta\rfloor\cdot\delta}X^n\big)_{t\in[0,T]}$, which is an adapted process; therefore one can compute the filter π_t^n associated with the approximating delay system $\big(\Pi_{\lfloor t/\delta\rfloor\cdot\delta}X^n,Y^n(t)\big)_{t\in[0,T]}$ by means of the classical Kallianpur–Striebel formula (see, e.g., [19] and [3]). The filter π_t^n is then given by

$$\pi_t^n(\phi)=\frac{\sigma_t^n(\phi)}{\sigma_t^n(\mathbf{1})}=\frac{E^{0,n}\big[\phi(\Pi_{\lfloor t/\delta\rfloor\cdot\delta}X^n)\mathcal{L}_t^n|\mathcal{F}_t^{Y^n}\big]}{E^{0,n}\big[\mathcal{L}_t^n|\mathcal{F}_t^{Y^n}\big]}, \tag{2.12}$$

where $E^{0,n}$ denotes the expectation w.r.t. the reference probability measure $P^{0,n}$, defined by the Radon–Nikodym derivative

$$\frac{dP^{0,n}}{dP}=\big(\mathcal{L}_T^n\big)^{-1},$$

with

(2.13)
$$\mathcal{L}_t^n = \exp\left\{\int_0^t h(\lfloor s/\delta\rfloor\cdot\delta, \Pi_{\lfloor s/\delta\rfloor\cdot\delta}X^n)dY_s^n - \frac{1}{2}\int_0^t |h(\lfloor s/\delta\rfloor\cdot\delta, \Pi_{\lfloor s/\delta\rfloor\cdot\delta}X^n)|^2\,ds\right\},$$

which is well defined thanks to the sublinearity of h and (A1).

Taking into account that $s \mapsto h(\lfloor s/\delta\rfloor\cdot\delta, \Pi_{\lfloor s/\delta\rfloor\cdot\delta}X^n)$ is piecewise constant, we have that

$$\mathcal{L}_t^n = L_t^n(X^n(\cdot), Y_0^n, Y_\delta^n, \dots, Y_{\lfloor t/\delta\rfloor}^n, Y_t^n),$$

where, for $0 \le \ell \le n$,

$$\log L_{\ell\delta}^n(x(\cdot), y_0, y_1, \dots, y_\ell) = \sum_{k=0}^{\ell-1} h(k\delta, \Pi_{k\delta}x(\cdot))(y_{k+1} - y_k) - \frac{1}{2}\sum_{k=0}^{\ell-1} |h(k\delta, \Pi_{k\delta}x(\cdot))|^2\delta,$$

and, for $t \in (\ell\delta, (\ell+1)\delta)$, $0 \le \ell \le n-1$,

$$\begin{aligned}\log L_t^n(x(\cdot), y_0, y_1, \dots, y_\ell, y) &= \log L_{\ell\delta}^n(x(\cdot), y_0, y_1, \dots, y_\ell)\\ &\quad + h(\lfloor t/\delta\rfloor\cdot\delta, \Pi_{\lfloor t/\delta\rfloor\cdot\delta}x(\cdot))(y - y_\ell) - \frac{1}{2}|h(\lfloor t/\delta\rfloor\cdot\delta, \Pi_{\lfloor t/\delta\rfloor\cdot\delta}x(\cdot))|^2(t - \ell\delta).\end{aligned}$$

Moreover, under $P^{0,n}$, the processes $\boldsymbol{X^n}$ and $\boldsymbol{Y^n}$ are independent and the law of the approximated state process is invariant under P and $P^{0,n}$, and hence, for $t \in [\ell\delta, (\ell+1)\delta)$, $0 \le \ell \le n-1$,

$$\sigma_t^n(\phi) = E\left[\phi(\Pi_{\lfloor t/\delta\rfloor\cdot\delta}X^n)L_t^n(X^n(\cdot), y_0, y_1, \dots, y_\ell, y)\right]\Big|_{y_0=Y_0^n, y_1=Y_\delta^n, \dots, y_\ell=Y_{\ell\delta}^n, y=Y_t^n}.$$

Therefore, by taking the above equality into account, one can explicitly obtain a deterministic functional

$$\begin{aligned}U^n :\ [0,T]\times C([0,T],\mathbb{R}) &\to \mathcal{P}(C([-\tau,0],\mathbb{R}))\\ (t, \boldsymbol{y}) &\mapsto U^n\big(t, \boldsymbol{y}\big),\end{aligned}$$

with the property that $U^n\big(t, \boldsymbol{y}\big)$ depends only on $\big(y(k\delta)\big)_{0\le k\le\lfloor\frac{t}{\delta}\rfloor}$ and $y(t)$, such that

$$\pi_t^n = U^n\big(t, \boldsymbol{Y^n}\big). \tag{2.14}$$

So the filter π_t^n defined above depends explicitly on the approximated observation process $\boldsymbol{Y^n}$, which, however, is not directly observable. To overcome this difficulty we also consider the following approximation $\tilde{\pi}_t^n$ for the filter:

$$\tilde{\pi}_t^n = U^n\big(t, \boldsymbol{Y}\big) = \frac{\tilde{\sigma}_t^n}{\tilde{\sigma}_t^n(\mathbf{1})}, \tag{2.15}$$

where

(2.16)
$$\tilde{\sigma}_t^n(\phi) = E\left[\phi(\Pi_{\lfloor t/\delta\rfloor\cdot\delta}X^n)L_t^n(X^n(\cdot), y_0, y_1, \dots, y_\ell, y)\right]\Big|_{y_0=Y_0, y_1=Y_\delta, \dots, y_\ell=Y_{\ell\delta}, y=Y_t}.$$

Then the following convergence result, which will be proved in the following section, holds.

THEOREM 2.2. *Let* $\boldsymbol{\pi} = (\pi_t; t \geq 0)$, $\boldsymbol{\pi^n} = (\pi_t^n; t \geq 0)$, *and* $\tilde{\boldsymbol{\pi}}^{\boldsymbol{n}} = (\tilde{\pi}_t^n; t \geq 0)$ *be the càdlàg probability measure-valued processes defined by* (1.6), (2.12), *and* (2.15), *respectively. Then the following hold:*

1. *The sequence of filters* $\boldsymbol{\pi^n}$ *converges in probability (and therefore weakly) to the original filter* $\boldsymbol{\pi}$ *in* $D_{\mathcal{P}(C([-\tau,0],\mathbb{R}))}([0,T])$.

2. *The sequence of measure-valued processes* $\tilde{\boldsymbol{\pi}}^{\boldsymbol{n}}$ *converges in probability to the original filter* $\boldsymbol{\pi}$.

3. *The sequence* $\max_{k=1,\dots,n} d(\tilde{\pi}_{k\delta}^n, \pi_{k\delta})$, *where* d *denotes the Prohorov metric, converges in probability to zero.*

From the practical point of view, the interest of this method relies on the fact that $\sigma_t(\phi)$ in (2.4) cannot usually be computed explicitly, or with a Monte Carlo method, while $\tilde{\sigma}_t^n(\phi)$ in (2.16) can always be computed by means of a Monte Carlo method. In addition, under further hypotheses on a, b, and h, the following result concerning the rate of convergence w.r.t. the bounded Lipschitz metric of our approximation scheme will be proved in section 4. For the ease of the reader we recall that, for any metric space S, and for given probability measures ν_1 and ν_2 on S,

$$d_{BL}(\nu_1, \nu_2) = \sup \left\{ \frac{|\nu_1(\varphi) - \nu_2(\varphi)|}{\|\varphi\| \vee L_\varphi}; \varphi \text{ bounded and Lipschitz} \right\},$$

where $\|\varphi\|$ denotes the sup-norm, and L_φ is the Lipschitz constant of φ.

THEOREM 2.3. *Assume further that the functions* a *and* b *are bounded,* $1/2 \leq \alpha \leq 1$ *in* (2.1), *the function* h *is jointly globally Lipschitz, and there exists a constant* C_η *such that the modulus of continuity of the initial condition* η *satisfies*

$$E\big[\omega_\eta^2\big(\delta; [-\tau, 0]\big)\big] \leq C_\eta\, \delta\, \log\big(\tfrac{1}{\delta}\big). \tag{2.17}$$

Then there exists a constant C *such that*

$$E\big[d_{BL}(\pi_t, \tilde{\pi}_t^n)\big] \leq C\big(\tfrac{\log n}{n}\big)^{\frac{1}{2}}, \tag{2.18}$$

where d_{BL} *is the bounded Lipschitz metric on the space* $\mathcal{P}(C([-\tau, 0], \mathbb{R}))$.

REMARK 2.4. *As will be shown in the example at the end of section* 4, *by considering the case where* $\eta = 0$, $a = 0$, $b = 1$, *and* $h = 0$, *the upper bound for the rate of convergence given by* (2.18) *appears to be the best we can obtain in our context.*

Furthermore, as is clear from the proof (see (5.13)*), if* $E\big[\omega_\eta^2\big(\delta; [-\tau, 0]\big)\big]$ *converges to zero with an order of convergence lower than* $O\big(\delta \log\big(\frac{1}{\delta}\big)\big)$, *then*

$$E\big[d_{BL}(\pi_t, \tilde{\pi}_t^n)\big] \leq C\,\big(E\big[\omega_\eta^2\big(\delta; [-\tau, 0]\big)\big]\big)^{\frac{1}{2}}$$

for a suitable constant C.

Finally, as will be clear from the proof, condition (A1) *for* $k = 2$ *is not necessary (see Proposition* 4.2*).*

To conclude this section, note that in order to evaluate π_t^n and $\tilde{\pi}_t^n$ we need to compute

(a) the transition probability of the $(m+1)$-dimensional Markov chain

$$\big(X^n(\ell\delta), X^n((\ell-1)\delta), \dots, X^n((\ell-m)\delta)\big),$$

(b) the explicit expression of L_t^n.

Consequently we need the explicit expression of $a(\ell\delta, \Pi_{\ell\delta} X^n)$, $b(\ell\delta, \Pi_{\ell\delta} X^n)$, and $h(\ell\delta, \Pi_{\ell\delta} X^n)$.

When the functionals a, b, and h are taken to be of the form (1.4), with $r = 1$, we need to evaluate expressions such as

$$g\left(\ell\delta, \int_{-\tau}^{0} \psi(u, \Pi_{\ell\delta} X^n(u))\, \gamma(du)\right),$$

where

$$\int_{-\tau}^{0} \psi(u, \Pi_{\ell\delta} X^n(u))\gamma(du) = \int_{-\tau}^{0} \psi(u, X^n(\ell\delta + u))\, \gamma(du)$$
$$= \sum_{k=-m}^{-1} \int_{k\delta}^{(k+1)\delta} \psi\big(u, X^n((\ell+k)\delta) + \tfrac{u - k\delta}{\delta}\big[X^n((\ell+k+1)\delta) - X^n((\ell+k)\delta)\big]\big)\, \gamma(du).$$

3. The convergence result. This section is dedicated to the proof of Theorem 2.2. With this aim, we will make use of a result deduced from the papers by Bhatt, Kallianpur, and Karandikar [3] and Bhatt and Karandikar [4] in the following context.

Consider a signal process $\boldsymbol{\mathcal{X}} = (\mathcal{X}_t)_{t\in[0,T]}$, with values in a complete separable metric space (S, d_S), defined on $(\Omega, \mathcal{F}, P)$, with *càdlàg* paths and being continuous in probability, and the observation process $\boldsymbol{Y} = (Y(t))_{t\in[0,T]}$ given by

$$Y(t) = \int_0^t \boldsymbol{h}(\mathcal{X}_s) ds + W(t), \tag{3.1}$$

where $\boldsymbol{W} = (W(t))_{t\in[0,T]}$ is a standard Brownian motion, defined on $(\Omega, \mathcal{F}, P)$, independent of $\boldsymbol{\mathcal{X}}$, and $\boldsymbol{h}$ is a measurable function on S with values in $\mathbb{R}^k$, such that

$$P\left(\int_0^T |\boldsymbol{h}(\mathcal{X}_s)|^2\, ds < \infty\right) = 1.$$

The approximation signal processes $\boldsymbol{\mathcal{X}^n} = (\mathcal{X}_t^n)_{t\in[0,T]}$ are defined on $(\Omega, \mathcal{F}, P)$ and take values in S as well. The approximation observation processes $\boldsymbol{Y^n} = (Y^n(t))_{t\in[0,T]}$ are defined by

$$Y^n(t) = \int_0^t \boldsymbol{h}^n(\mathcal{X}_s^n) ds + W(t), \tag{3.2}$$

where $\boldsymbol{W}$ is independent of $\boldsymbol{\mathcal{X}^n}$, and $\boldsymbol{h}^n$ are measurable functions on S with values in $\mathbb{R}^k$, such that

$$P\left(\int_0^T |\boldsymbol{h}^n(\mathcal{X}_s^n)|^2\, ds < \infty\right) = 1.$$

Then the following result is an easy consequence of [3] and Remark 7.4 in [4].

THEOREM 3.1. *Assume that*

(B1) $\boldsymbol{h}^n$ *converges to* $\boldsymbol{h}$ *uniformly on compact sets;*

(B2) $\boldsymbol{h}$ *is continuous;*

(B3) $\boldsymbol{\mathcal{X}^n}$ *converges in* P*-probability (and therefore weakly) to* $\boldsymbol{\mathcal{X}}$ *in* $D_S([0,T])$*;*

(B4) $\lim_{n\to\infty} E(\int_0^T |\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)|^2\, ds) = 0.$

Then the filters of the system $(\boldsymbol{\mathcal{X}^n}, \boldsymbol{Y^n})$ *converge in probability (and therefore weakly) to the filter of the system* $(\boldsymbol{\mathcal{X}}, \boldsymbol{Y})$ *as processes with values in* $D_{\mathcal{P}(S)}([0,T])$*, where* $\mathcal{P}(S)$ *is the metric space of probability measures on* S*, endowed with the Prohorov metric.*

Note that the above conditions (B1)–(B4) are only sufficient conditions and that in [3] weaker conditions and different frameworks can be found.

In order to be in the framework described above, we take $S = [0,T] \times C([-\tau, 0], \mathbb{R})$, endowed with the distance

$$\|(t,\theta) - (t',\theta')\|_S = |t - t'| + \|\theta - \theta'\|,$$

and we consider the limit model $\big(\mathcal{X}_t, Y(t)\big)_{t\in[0,T]}$, where

$$\mathcal{X}_t = (t, \Pi_t X),$$

and $Y(t)$ is given by (1.2), and the approximating model $\big(\mathcal{X}^n_t, Y^n(t)\big)_{t\in[0,T]}$, where

$$\mathcal{X}^n_t = (\delta \cdot \lfloor t/\delta \rfloor, \Pi_{\delta\cdot\lfloor t/\delta\rfloor} X^n), \tag{3.3}$$

and $Y^n(t)$ is given by (2.11).

Then with this choice, the processes $\big(\mathcal{X}_t\big)_{t\in[0,T]}$ and $\big(\mathcal{X}^n_t\big)_{t\in[0,T]}$ have paths in $D_S([0,T])$, and conditions (B1) and (B2) are obviously satisfied with $\boldsymbol{h}^n = \boldsymbol{h} = h$.

Condition (B3), which asserts that $\boldsymbol{\mathcal{X}^n} = \big((\delta\cdot\lfloor t/\delta\rfloor, \Pi_{\delta\cdot\lfloor t/\delta\rfloor} X^n)\big)_{t\in[0,T]}$ converges in probability to $\boldsymbol{\mathcal{X}} = \big((t, \Pi_t X)\big)_{t\in[0,T]}$ in $D_S([0,T]) = D_{[0,T]\times C([-\tau,0],\mathbb{R})}([0,T])$, is also satisfied thanks to the following proposition, which will be proved in section 5.

PROPOSITION 3.2. *Assume that conditions* (A1)*, for* $k = 1$*, and* (A2) *are satisfied. Then*

$$\lim_{n\to\infty} E\left[\sup_{t\in[0,T]} \big\|(\delta\cdot\lfloor t/\delta\rfloor, \Pi_{\delta\cdot\lfloor t/\delta\rfloor} X^n) - (t, \Pi_t X)\big\|^2_S\right] = 0. \tag{3.4}$$

In our approximation scheme, since the function h appears in the definitions of both $Y(t)$ and $Y^n(t)$, condition (B4) reads

$$\lim_{n\to\infty} \int_0^T E\Big(|h(\delta\cdot\lfloor s/\delta\rfloor, \Pi_{\delta\cdot\lfloor s/\delta\rfloor} X^n) - h(s, \Pi_s X)|^2\Big) ds = 0. \tag{3.5}$$

When the observation function h is jointly globally Lipschitz, then condition (3.5) immediately follows by Proposition 3.2. Furthermore, condition (A1) is used for $k = 1$, in the proof of Proposition 3.2, and, for $k = 2$, to obtain (3.5) in the general case considered in (A3). In this general case, since h is jointly continuous and the convergence condition (3.4) holds, the sequence $|h(\delta\cdot\lfloor s/\delta\rfloor, \Pi_{\delta\cdot\lfloor s/\delta\rfloor} X^n) - h(s, \Pi_s X)|^2$ in (3.5) converges to zero in probability. Moreover a uniform integrability condition holds by property (2.3), for $k = 2$, and by

$$\sup_{s\in[0,T]} \sup_n E\left[|h(\delta\cdot\lfloor s/\delta\rfloor, \Pi_{\delta\cdot\lfloor s/\delta\rfloor} X^n)|^4\right] \le \sup_n E\left[C'\left(1 + \sup_{t\in[-\tau,T]} |X^n(t)|^4\right)\right] < \infty.$$

The above inequalities are immediate, thanks to the sublinear growth conditions in (A3) and Lemma 3.3, which will be proved in section 5. Then the dominated convergence theorem implies (3.5).

LEMMA 3.3. *Assume that conditions* (A1) *and* (A2) *are satisfied; then, for* $k = 1, 2$,

$$\sup_n E\left[\sup_{t\in[0,T]} \|\Pi_t X^n\|^{2k}\right] = \sup_n E\left[\sup_{u\in[-\tau,T]} |X^n(u)|^{2k}\right] < \infty.$$

Therefore, since all the conditions in Theorem 3.1 hold, the filters of the systems $(\boldsymbol{\mathcal{X}^n}, \boldsymbol{Y^n})$ converge in probability to the filter of the system $(\boldsymbol{\mathcal{X}}, \boldsymbol{Y})$ as random processes with values in $D_{\mathcal{P}(S)}([0,T])$, and this implies the convergence of $\boldsymbol{\pi^n}$ to $\boldsymbol{\pi}$ as random processes with values in $D_{\mathcal{P}(C([-\tau,0],\mathbb{R}))}([0,T])$.

To prove the second assertion of Theorem 2.2, we make use, as in section 2, of a representation result for the filters $(\boldsymbol{\mathcal{X}^n}, \boldsymbol{Y^n})$ and $(\boldsymbol{\mathcal{X}}, \boldsymbol{Y})$ in the state space S.

In this setting there exist functionals

$$V^n,\, V:\ [0,T] \times D_{\mathbb{R}^k}([0,T]) \mapsto \mathcal{P}(S), \tag{3.6}$$

with the properties $V^n\big(t, \boldsymbol{y}\big) = V^n\big(t, y(\cdot \wedge t)\big)$ and $V\big(t, \boldsymbol{y}\big) = V\big(t, y(\cdot \wedge t)\big)$, such that the filters of $(\boldsymbol{\mathcal{X}^n}, \boldsymbol{Y^n})$ and $(\boldsymbol{\mathcal{X}}, \boldsymbol{Y})$ are given by $V^n(t, \boldsymbol{Y^n})$ and $V(t, \boldsymbol{Y})$, respectively. This fact is true under very general conditions (see, for instance, Kurtz and Ocone [22]). Note that if one uses the general representation result, then one could only say that the functionals V^n and V are defined almost surely w.r.t. $P_{\boldsymbol{Y^n}}$, the law of $\boldsymbol{Y^n}$, and w.r.t. $P_{\boldsymbol{Y}}$, the law of $\boldsymbol{Y}$, respectively. Therefore, the approximation $V^n(t, \boldsymbol{Y})$ of the $(\boldsymbol{\mathcal{X}}, \boldsymbol{Y})$-filter (as the one provided in (2.15)) could be not well defined. However, in this case $P_{\boldsymbol{Y^n}}$ and $P_{\boldsymbol{Y}}$ are equivalent, and this problem does not occur.

In our delay case, $S = [0,T] \times C([-\tau, 0], \mathbb{R})$ and the functional V^n can be computed starting from the Kallianpur–Striebel formula with the Radon–Nikodym derivative $\mathcal{L}_t^n$ defined by (2.13). This fact implies also that, for any $(t, \boldsymbol{y})$ in $[0,T] \times C([0,T], \mathbb{R})$, the projection on the space $C([-\tau, 0], \mathbb{R})$ of the probability measure $V^n(t, \boldsymbol{y})$ coincides with $U^n(t, \boldsymbol{y})$, defined in section 2.

Moreover in [3] (see Theorem 3.3(a)), as a step in the proof of a weak convergence result, the authors prove that for any Wiener process $\boldsymbol{B} = (B(t))_{t\in[0,T]}$, the $\mathcal{P}(S)$-valued processes $(V^n(t, \boldsymbol{B}))_{t\in[0,T]}$ converge in probability to the $\mathcal{P}(S)$-valued process $(V(t, \boldsymbol{B}))_{t\in[0,T]}$. This amounts to saying that if P^0 is the reference probability measure defined by the Radon–Nikodym derivative

$$\frac{dP^0}{dP} = \exp\left\{-\int_0^T \boldsymbol{h}(\mathcal{X}_s) dY_s + \frac{1}{2}\int_0^T |\boldsymbol{h}(\mathcal{X}_s)|^2\, ds\right\}, \tag{3.7}$$

i.e., the measure under which the process $\boldsymbol{Y}$ is a Wiener process, independent of the state process $\boldsymbol{\mathcal{X}}$, then the $\mathcal{P}(S)$-valued processes $(V^n(t, \boldsymbol{Y}))_{t\in[0,T]}$ converge in P^0-probability to the $\mathcal{P}(S)$-valued process $(V(t, \boldsymbol{Y}))_{t\in[0,T]}$. In addition, since the measure P is also absolutely continuous w.r.t. P^0, the convergence also holds in P-probability. This implies that

$$\tilde{\boldsymbol{\pi}}^{\boldsymbol{n}} \text{ converges in } P\text{-probability to } \boldsymbol{\pi},$$

which is the second assertion in Theorem 2.2.

Since the filter $\boldsymbol{\pi}$ is continuous in time, the last statement of the theorem is an immediate consequence of the convergence in probability of $\tilde{\boldsymbol{\pi}}^{\boldsymbol{n}}$ to $\boldsymbol{\pi}$.

4. Rate of convergence. The aim of this section is to compute an upper bound for the rate of convergence of our scheme under the further hypotheses that h is jointly globally Lipschitz, that a and b are bounded, and that $1/2 \leq \alpha \leq 1$ in (2.1), i.e., to prove Theorem 2.3.

Let $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}^n}, \boldsymbol{Y}, \boldsymbol{Y^n})$ be the stochastic processes introduced at the beginning of section 3, with values in a complete separable metric space (S, d_S), and let P^n be the probability measure defined by

$$\frac{dP^n}{dP^0} = \exp\left\{ \int_0^T \boldsymbol{h}^n(\mathcal{X}_s^n) dY_s - \frac{1}{2} \int_0^T |\boldsymbol{h}^n(\mathcal{X}_s^n)|^2 \, ds \right\}, \tag{4.1}$$

where P^0 is the reference probability measure on $(\Omega, \mathcal{F})$ defined in (3.7).

Then the law of $(\boldsymbol{\mathcal{X}^n}, \boldsymbol{Y})$ under P^n is the same as the law of $(\boldsymbol{\mathcal{X}^n}, \boldsymbol{Y^n})$ under P, so that the processes $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}^n}, \boldsymbol{Y})$ and the probabilities P^0, P, and P^n satisfy conditions (a), ($\mathrm{a_n}$), (b1), and (b2) of Calzolari, Florchinger, and Nappo [7], apart from the fact that we are in a complete separable metric space (S, d_S). Therefore, with slight modifications in the proof of (32) in Theorem 2.3 of [7], we get

$$\begin{aligned} &E\big[d_{BL}^S\big(V(t,\boldsymbol{Y}), V^n(t,\boldsymbol{Y})\big)\big] \\ &\leq 2\, E^0\Big[\big|(dP^n/dP^0)|_{\widetilde{\mathcal{F}}_t} - (dP/dP^0)|_{\widetilde{\mathcal{F}}_t}\big|\Big] + E\big[d_S\big(\mathcal{X}_t, \mathcal{X}_t^n\big)\big], \end{aligned} \tag{4.2}$$

where V and V^n are the functionals defined as in (3.6), d_{BL}^S is the bounded Lipschitz metric on $\mathcal{P}(S)$, and $\widetilde{\mathcal{F}}_t = \mathcal{F}_t^{\mathcal{X}, \mathcal{X}^n, Y}$.

The above inequality is the starting point of the proof of Theorem 2.3, and, as a consequence, we need the estimates for the quantities in the right-hand side of (4.2) stated in the following proposition.

PROPOSITION 4.1. *For all $t \leq T$, we have*

$$\begin{aligned} &E^0\Big[\big|(dP^n/dP^0)|_{\widetilde{\mathcal{F}}_t} - (dP/dP^0)|_{\widetilde{\mathcal{F}}_t}\big|\Big] \\ &\leq 2\left(E\left[\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2 \, ds \right] \right)^{\frac{1}{2}} + E\left[\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2 \, ds \right]. \end{aligned} \tag{4.3}$$

In the particular case when $\boldsymbol{h}^n = \boldsymbol{h}$ and $\boldsymbol{h}$ is a globally Lipschitz function, then, for all $t \leq T$, we have, for a suitable constant $K(T)$,

$$\begin{aligned} &E^0\Big[\big|(dP^n/dP^0)|_{\widetilde{\mathcal{F}}_t} - (dP/dP^0)|_{\widetilde{\mathcal{F}}_t}\big|\Big] \\ &\leq 2\, K(T) \left\{ \sup_{s\in[0,T]} \big(E\big[d_S^2(\mathcal{X}_s^n, \mathcal{X}_s)\big]\big)^{\frac{1}{2}} + \sup_{s\in[0,T]} E\big[d_S^2(\mathcal{X}_s^n, \mathcal{X}_s)\big] \right\}. \end{aligned} \tag{4.4}$$

Proof. Define Λ_t and Λ_t^n by

$$\Lambda_t = \int_0^t \boldsymbol{h}(\mathcal{X}_s) dY_s - \frac{1}{2} \int_0^t |\boldsymbol{h}(\mathcal{X}_s)|^2 \, ds \tag{4.5}$$

and

$$\Lambda_t^n = \int_0^t \boldsymbol{h}^n(\mathcal{X}_s^n) dY_s - \frac{1}{2} \int_0^t |\boldsymbol{h}^n(\mathcal{X}_s^n)|^2 \, ds. \tag{4.6}$$

Then, using the fact that $|e^a - e^b| \le e^a\,|a-b| + e^b\,|a-b|$, we have

$$\begin{aligned} E^0\Big[\big|(dP^n/dP^0)|_{\widetilde{\mathcal{F}}_t} - (dP/dP^0)|_{\widetilde{\mathcal{F}}_t}\big|\Big] &= E^0\Big[\big|e^{\Lambda_t^n} - e^{\Lambda_t}\big|\Big] \\ &\le E^0\Big[e^{\Lambda_t^n}\big|\Lambda_t^n - \Lambda_t\big|\Big] + E^0\Big[e^{\Lambda_t}\big|\Lambda_t^n - \Lambda_t\big|\Big] \\ &= E^n\Big[\big|\Lambda_t^n - \Lambda_t\big|\Big] + E\Big[\big|\Lambda_t^n - \Lambda_t\big|\Big], \end{aligned}$$

where E^n is the expectation w.r.t. P^n.

An easy calculation gives

$$\begin{aligned} \Lambda_t^n - \Lambda_t &= \int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]\,dY_s - \frac{1}{2}\int_0^t \big[|\boldsymbol{h}^n(\mathcal{X}_s^n)|^2 - |\boldsymbol{h}(\mathcal{X}_s)|^2\big]\,ds \\ &= \int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]\big(dY_s - \boldsymbol{h}^n(\mathcal{X}_s^n)ds\big) + \frac{1}{2}\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2\,ds \\ &= \int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]\big(dY_s - \boldsymbol{h}(\mathcal{X}_s)ds\big) + \frac{1}{2}\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2\,ds. \end{aligned}$$

Therefore we have

$$\begin{aligned} &E^0\Big[\big|(dP^n/dP^0)|_{\widetilde{\mathcal{F}}_t} - (dP/dP^0)|_{\widetilde{\mathcal{F}}_t}\big|\Big] \le E^n\Big[\big|\Lambda_t^n - \Lambda_t\big|\Big] + E\Big[\big|\Lambda_t^n - \Lambda_t\big|\Big] \\ &\le E^n\left[\left|\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]\big(dY_s - \boldsymbol{h}^n(\mathcal{X}_s^n)ds\big)\right|\right] \\ &+ \frac{1}{2}E^n\left[\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2\,ds\right] \\ &+ E\left[\left|\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]\big(dY_s - \boldsymbol{h}(\mathcal{X}_s)ds\big)\right|\right] + \frac{1}{2}E\left[\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2\,ds\right]. \end{aligned}$$

By recalling that

$$Y_t - \int_0^t \boldsymbol{h}^n(\mathcal{X}_s^n)ds \quad \text{and} \quad Y_t - \int_0^t \boldsymbol{h}(\mathcal{X}_s)ds$$

are Wiener processes under P^n and P, respectively, we get, by Cauchy–Schwarz inequality and the isometry of stochastic integrals,

$$\begin{aligned} &E^0\Big[\big|(dP^n/dP^0)|_{\widetilde{\mathcal{F}}_t} - (dP/dP^0)|_{\widetilde{\mathcal{F}}_t}\big|\Big] \\ &\le \left(E^n\left[\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2\,ds\right]\right)^{\frac{1}{2}} + \frac{1}{2}E^n\left[\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2\,ds\right] \\ &+ \left(E\left[\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2\,ds\right]\right)^{\frac{1}{2}} + \frac{1}{2}E\left[\int_0^t \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2\,ds\right]. \end{aligned}$$

As the joint laws of $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{X}^n}$ under P^n and P coincide (with the joint law under P^0), the final upper bound is

$$\begin{aligned} &E^0\Big[\big|(dP^n/dP^0)|_{\widetilde{\mathcal{F}}_t} - (dP/dP^0)|_{\widetilde{\mathcal{F}}_t}\big|\Big] \\ &\le 2\left(E\left[\int_0^T \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2\,ds\right]\right)^{\frac{1}{2}} + E\left[\int_0^T \big[\boldsymbol{h}^n(\mathcal{X}_s^n) - \boldsymbol{h}(\mathcal{X}_s)\big]^2\,ds\right], \end{aligned}$$

which is inequality (4.3), and it immediately implies inequality (4.4). □

As in the previous section, we take $S = [0,T] \times C([-\tau,0],\mathbb{R})$, $\mathcal{X}_t = (t, \Pi_t X)$, $\mathcal{X}^n_t = (\delta \cdot \lfloor t/\delta \rfloor, \Pi_{\delta\cdot\lfloor t/\delta\rfloor} X^n)$, and $\boldsymbol{h}^n = \boldsymbol{h} = h$. Therefore (4.2) implies that

$$E\big[d_{BL}(\pi_t, \tilde{\pi}^n_t)\big] \le 2E^0\Big[\big|(dP^n/dP^0)|_{\tilde{\mathcal{F}}_t} - (dP/dP^0)|_{\tilde{\mathcal{F}}_t}\big|\Big] + E\big[\|\mathcal{X}_t - \mathcal{X}^n_t\|_S\big], \tag{4.7}$$

and, when h is a jointly globally Lipschitz function, inequality (4.4) holds. Finally the result of Theorem 2.3 is a direct consequence of the following improvement of the result of Proposition 3.2.

PROPOSITION 4.2. *Assume that conditions* (A1), *for* $k = 1$, *and* (A2), *with* $1/2 \le \alpha \le 1$, *are satisfied, and furthermore assume that the initial condition* η *satisfies* (2.17) *and that the functions* a *and* b *are bounded. Then there exists a constant* C_X *such that*

$$E\left[\sup_{t\in[0,T]} \big\|(\delta\cdot\lfloor t/\delta\rfloor, \Pi_{\delta\cdot\lfloor t/\delta\rfloor} X^n) - (t, \Pi_t X)\big\|^2_S\right] \le C_X \,\frac{\log n}{n}. \tag{4.8}$$

The proof of this result will be given in section 5.

REMARK 4.3. *Obviously, from* (4.8) *there exists a constant* $C'_X \le C_X$ *such that*

$$\sup_{t\in[0,T]} E\left[\big\|(\delta\cdot\lfloor t/\delta\rfloor, \Pi_{\delta\cdot\lfloor t/\delta\rfloor} X^n) - (t, \Pi_t X)\big\|^2_S\right] \le C'_X \,\frac{\log n}{n}, \tag{4.9}$$

and, by (4.4), *the above inequality is a sufficient condition to get the upper bound* (2.18) *in Theorem* 2.3. *From the following example it appears that the rate of convergence cannot be improved either in the right-hand side of the previous inequality or in* (2.18).

Example. Take $\eta = 0$, $a = 0$, $b = 1$, and $h = 0$. In this case $X = \tilde{W}$ and $X^n = \tilde{W}^n$, where $\tilde{W}^n$ is the piecewise linear interpolation of $\tilde{W}$. Moreover π_t and $\tilde{\pi}^n_t$ coincide with the laws of $\Pi_t\tilde{W}$ and $\Pi_{\delta\cdot\lfloor t/\delta\rfloor}\tilde{W}^n$, respectively, and therefore

$$\begin{aligned} E\big[d_{BL}(\pi_t, \tilde{\pi}^n_t)\big] &= d_{BL}(\pi_t, \tilde{\pi}^n_t) \\ &\le E\left[\big\|\Pi_{\delta\cdot\lfloor t/\delta\rfloor}\tilde{W}^n - \Pi_t\tilde{W})\big\|\right] \le \left(E\left[\big\|\Pi_{\delta\cdot\lfloor t/\delta\rfloor}\tilde{W}^n - \Pi_t\tilde{W})\big\|^2\right]\right)^{\frac{1}{2}}, \end{aligned}$$

where the first inequality follows by standard coupling techniques.

Furthermore,

$$E\left[\big\|\Pi_{\delta\cdot\lfloor t/\delta\rfloor}\tilde{W}^n - \Pi_t\tilde{W})\big\|^2\right] = O(\tfrac{\log n}{n})$$

for any $t \in [0,T]$ and uniformly in $[0,T]$. This fact can be shown by using the results established by Pickands in [38] (see also Fischer and Nappo [13]). This result could be expected thanks to Lévy's modulus of continuity, which implies that there exists a finite random variable M such that

$$\sup_{\substack{s,t\in[0,1]\\ |s-t|\le\delta,}} |\tilde{W}_s - \tilde{W}_t| \le M\sqrt{\delta\,\log(1/\delta)}$$

holds almost surely (see the paper by Pinsky [39] for a simple proof).

5. Technical results. This section is devoted to the proofs of Lemma 3.3, Proposition 3.2, and Proposition 4.2. In order to prove these results we introduce, as a technical tool,

• the operator P^δ, which gives the linear interpolation of a function $\big(x(s)\big)_{s\in[-\tau,T]}$, with step δ, so that $\big(P^\delta x(s)\big)_{s\in[-\tau,T]}$ is the linear interpolation of $(\ell\delta, x(\ell\delta))$ for $\ell = -m,\dots,n$,

and, as another approximation for the state,

• the continuous Euler–Maruyama scheme, i.e., the diffusion processes $\boldsymbol{Z^n} = \big(Z^n(t)\big)_{t\in[0,T]}$, where

$$\text{(5.1)}\qquad \begin{cases} Z^n(t) := \eta(t), & -\tau \le t \le 0,\\ Z^n(t) := \eta(0) + \displaystyle\int_0^t a(\delta\cdot\lfloor s/\delta\rfloor, \Pi_{\delta\cdot\lfloor s/\delta\rfloor}X^n)ds & \\ \qquad\qquad + \displaystyle\int_0^t b(\delta\cdot\lfloor s/\delta\rfloor, \Pi_{\delta\cdot\lfloor s/\delta\rfloor}X^n)d\tilde W_s, & 0\le t\le T, \end{cases}$$

which can be considered as intermediate approximation processes for the state $\boldsymbol{X}$.

The processes $\boldsymbol{Z^n}$ have the property that

$$\text{(5.2)}\qquad P^\delta Z^n(s) = X^n(s) \qquad \text{for } s\in[-\tau,0]\cup[0,T],$$

since

$$Z^n(\ell\delta) = X^n(\ell\delta) \qquad \text{for } \ell\ge -m.$$

Indeed $Z^n(\ell\delta) = X^n(\ell\delta) = \eta(\ell\delta)$ for $-m\le\ell\le 0$. The case $\ell\ge 0$ follows by observing that for $t\in[\ell\delta,(\ell+1)\delta]$

$$\begin{aligned} Z^n(t) &= Z^n(\ell\delta) + \int_{\ell\delta}^t a(\ell\delta,\Pi_{\ell\delta}X^n)ds + \int_{\ell\delta}^t b(\ell\delta,\Pi_{\ell\delta}X^n)d\tilde W_s\\ &= Z^n(\ell\delta) + a(\ell\delta,\Pi_{\ell\delta}X^n)(t-\ell\delta) + b(\ell\delta,\Pi_{\ell\delta}X^n)[\tilde W(t)-\tilde W(\ell\delta)], \end{aligned}$$

so that

$$Z^n((\ell+1)\delta) = Z^n(\ell\delta) + a(\ell\delta,\Pi_{\ell\delta}X^n)\delta + b(\ell\delta,\Pi_{\ell\delta}X^n)[\tilde W((\ell+1)\delta)-\tilde W(\ell\delta)],$$

and finally comparing the above recursive formula with the definition of $X^n((\ell+1)\delta)$ in (2.7).

We are now able to prove Lemma 3.3.

Proof of Lemma 3.3. First, observe that by (5.2)

$$\sup_{t\in[-\tau,T]}|X^n(t)| \le \max\left(\|\eta\|, \sup_{t\in[0,T]}|Z^n(t)|\right),$$

and set

$$\text{(5.3)}\qquad M_t^n := \int_0^t b_n(u)d\tilde W_u, \quad \text{where } b_n(u) := b(\delta\cdot\lfloor u/\delta\rfloor, \Pi_{\delta\cdot\lfloor u/\delta\rfloor}X^n).$$

For any $\ell \in \{1,2\}$ take $\alpha = 2\ell$ and $\beta = 2\ell/(2\ell-1)$, so that $(1/\alpha)+(1/\beta)=1$. Then for any stopping time σ, there exists a suitable constant C_ℓ such that

$$\begin{aligned}
&\sup_{u\in[-\tau,t\wedge\sigma]} |X^n(u)|^{2\ell} \\
&\le C_\ell \left\{ \|\eta\|^{2\ell} + \left[\int_0^t \sup_{u\in[0,s\wedge\sigma]} |a(\delta\cdot\lfloor u/\delta\rfloor, \Pi_{\delta\cdot\lfloor u/\delta\rfloor} X^n)| ds \right]^{2\ell} + \sup_{s\in[0,t\wedge\sigma]} |M_s^n|^{2\ell} \right\} \\
&\le C_\ell \left\{ \|\eta\|^{2\ell} + t^{\frac{2\ell}{\beta}} \left[\int_0^t \sup_{u\in[0,s\wedge\sigma]} |a(\delta\cdot\lfloor u/\delta\rfloor, \Pi_{\delta\cdot\lfloor u/\delta\rfloor} X^n)|^\alpha ds \right]^{\frac{2\ell}{\alpha}} + \sup_{s\in[0,t\wedge\sigma]} |M_s^n|^{2\ell} \right\} \\
&\le C_\ell \left\{ \|\eta\|^{2\ell} + t^{2\ell-1} \int_0^t \ell K^\ell \left[1 + \sup_{u\in[-\tau,s\wedge\sigma]} |X^n(u)|^{2\ell} \right] ds + \sup_{s\in[0,t]} |M_{s\wedge\sigma}^n|^{2\ell} \right\}.
\end{aligned}$$

If $M_{t\wedge\sigma}^n$ is a martingale, setting

$$\phi_{\sigma,\ell}^n(t) := E\left[\sup_{u\in[-\tau,t\wedge\sigma]} |X^n(u)|^{2\ell} \right]$$

and applying Doob's inequality for $p = 2\ell$ to $M_{t\wedge\sigma}^n$ yields

$$\phi_{\sigma,\ell}^n(t) \le C'_\ell \left\{ 1 + \|\eta\|^{2\ell} + \int_0^t \phi_{\sigma,\ell}^n(s) ds + E\left[|M_{t\wedge\sigma}^n|^{2\ell} \right] \right\}, \tag{5.4}$$

for all $t\in[0,T]$, for a suitable constant $C'_\ell = C'_\ell(T)$.

Taking $\sigma = \sigma_N^n := \inf\{s>0;\, \sup_{u\in[-\tau,s]} |X^n(u)| \ge N\}$, we have

$$\int_0^T E[\mathbf{1}_{s\le\sigma_N^n} b_n^{2\ell}(s)] ds < \infty, \tag{5.5}$$

and $M_{t\wedge\sigma}^n = M_{t\wedge\sigma_N^n}^n$ is a martingale. Indeed by the sublinearity condition (2.2) on b we have

$$\int_0^t E[\mathbf{1}_{s\le\sigma_N^n} b_n^{2\ell}(s)] ds \le \int_0^t E\left[\ell K^\ell \left(1 + \sup_{u\in[-\tau,s\wedge\sigma_N^n]} |X^n(u)|^{2\ell} \right) \right] ds, \tag{5.6}$$

which is finite since $\sup_{u\in[-\tau,s\wedge\sigma_N^n]} |X^n(u)| \le N$. Then

$$E[|M_{t\wedge\sigma_N^n}^n|^2] = E\left[\left(\int_0^t \mathbf{1}_{s\le\sigma_N^n} b_n(s) d\tilde{W}_s \right)^2 \right] = \int_0^t E[\mathbf{1}_{s\le\sigma_N^n} b_n^2(s)] ds$$

and (see, e.g., Lemma 4.12, page 125, in Liptser and Shiryayev [29])

$$E[|M_{t\wedge\sigma}^n|^4] = E\left[\left(\int_0^t \mathbf{1}_{s\le\sigma_N^n} b_n(s) d\tilde{W}_s \right)^4 \right] \le 6^2 t \int_0^t E[\mathbf{1}_{s\le\sigma_N^n} b_n^4(s)] ds.$$

Then, taking into account (5.4) and (5.6) and invoking Gronwall's inequality, we get a bound for $\phi_{\sigma,\ell}^n(T) = \phi_{\sigma_N^n,\ell}^n(T)$, uniform in n and N. Therefore, applying Fatou's lemma and making use of the fact that $\sigma_N^n \to \infty$ as $N\to\infty$, we get the results. □

REMARK 5.1. *Note that with the same technique one could prove that under the same assumptions of Lemma* 3.3

$$\sup_n E\left[\sup_{t\in[0,T]}\|\Pi_t Z^n\|^{2\ell}\right]=\sup_n E\left[\sup_{u\in[-\tau,T]}|Z^n(u)|^{2\ell}\right]<\infty \qquad \textit{for } \ell=1,2.$$

In order to prove Proposition 3.2 we need some intermediate results, stated in the following lemmas, which will be proved at the end of this section.

The first lemma concerns the behavior of the modulus of continuity.

LEMMA 5.2. *Denoting by*

$$\omega_x(\delta;[-\tau,T]):=\sup_{\substack{s,t\in[-\tau,T]\\ |s-t|\le\delta,}}|x(s)-x(t)|$$

the modulus of continuity of the function $\big(x(s)\big)_{s\in[-\tau,T]}$, *we have*

$$\sup_{t\in[0,T]}\|\Pi_{\delta\cdot\lfloor t/\delta\rfloor}P^\delta x-\Pi_t x\|\le 2\,\omega_x(\delta;[-\tau,T]), \tag{5.7}$$

and, for $\delta=\delta_n$,

$$\lim_{n\to\infty}E\left[\omega_X^2(\delta,[-\tau,T])\right]=0 \tag{5.8}$$

and

$$\lim_{n\to\infty}E\left[\sup_{t\in[0,T]}\|\Pi_{\delta\cdot\lfloor t/\delta\rfloor}P^\delta X-\Pi_t X\|^2\right]=0. \tag{5.9}$$

The second lemma concerns the convergence of the approximation $\boldsymbol{Z^n}$.

LEMMA 5.3. *Under the hypotheses of Proposition* 3.2

$$\lim_{n\to\infty}E\left[\sup_{t\in[0,T]}\|\Pi_t Z^n-\Pi_t X\|^2\right]=0.$$

With the above results the proof of Proposition 3.2 is straightforward.

Proof of Proposition 3.2. First, we note that

$$\left\|\left(\delta\cdot\lfloor t/\delta\rfloor,\Pi_{\delta\cdot\lfloor t/\delta\rfloor}X^n\right)-\left(t,\Pi_t X\right)\right\|_S^2\le 2\,\delta^2+2\,\left\|\Pi_{\delta\cdot\lfloor t/\delta\rfloor}X^n-\Pi_t X\right\|^2. \tag{5.10}$$

Then, by adding and subtracting $\Pi_{\delta\cdot\lfloor t/\delta\rfloor}P^\delta X$ in the second term on the right-hand side of the above expression, it yields

$$\left\|\Pi_{\delta\cdot\lfloor t/\delta\rfloor}X^n-\Pi_t X\right\|^2\le 2\left\|\Pi_{\delta\cdot\lfloor t/\delta\rfloor}X^n-\Pi_{\delta\cdot\lfloor t/\delta\rfloor}P^\delta X\right\|^2+2\left\|\Pi_{\delta\cdot\lfloor t/\delta\rfloor}P^\delta X-\Pi_t X\right\|^2. \tag{5.11}$$

Then, taking into account (5.9) and that

$$\begin{aligned}&\sup_{t\in[0,T]}\|\Pi_t X^n-\Pi_t P^\delta X\|=\sup_{k:k\delta\in[-\tau,T]}|X^n(k\delta)-X(k\delta)|\\ &=\sup_{k:k\delta\in[-\tau,T]}|Z^n(k\delta)-X(k\delta)|\le\sup_{t\in[-\tau,T]}|Z^n(t)-X(t)|=\sup_{t\in[0,T]}\|\Pi_t Z^n-\Pi_t X\|,\end{aligned}$$

the result follows by Lemma 5.3. □

Proof of Lemma 5.2. Noticing that $\Pi_{\delta\cdot\lfloor t/\delta\rfloor}P^\delta x = P^\delta \Pi_{\delta\cdot\lfloor t/\delta\rfloor}x$, we have

$$\begin{aligned}\|\Pi_{\delta\cdot\lfloor u/\delta\rfloor}P^\delta x - \Pi_u x\| &\le \|\Pi_{\delta\cdot\lfloor u/\delta\rfloor}P^\delta x - \Pi_{\delta\cdot\lfloor u/\delta\rfloor}x\| + \|\Pi_{\delta\cdot\lfloor u/\delta\rfloor}x - \Pi_u x\| \\ &= \|P^\delta \Pi_{\delta\cdot\lfloor u/\delta\rfloor}x - \Pi_{\delta\cdot\lfloor u/\delta\rfloor}x\| + \|\Pi_{\delta\cdot\lfloor u/\delta\rfloor}x - \Pi_u x\|.\end{aligned}$$

Furthermore, since

$$P^\delta\theta(v) = \lambda(v)\,\theta(\delta\cdot\lfloor v/\delta\rfloor + \delta) + (1-\lambda(v))\,\theta(\delta\cdot\lfloor v/\delta\rfloor)$$

with $\lambda(v) = v/\delta - \lfloor v/\delta\rfloor$, and since $\theta(v) = \lambda(v)\,\theta(v) + (1-\lambda(v))\,\theta(v)$, we deduce that

$$\|\Pi_{\delta\cdot\lfloor u/\delta\rfloor}P^\delta x - \Pi_u x\| \le 2\sup_{\substack{s,t\in[0,T]\\ |s-t|\le\delta,}}\|\Pi_s x - \Pi_t x\| = 2\sup_{\substack{s,t\in[-\tau,T]\\ |s-t|\le\delta,}}|x(s)-x(t)|,$$

which is the first assertion (5.7) of the lemma.

Equality (5.8) follows by the dominated convergence theorem: indeed

$$\omega_x(\delta;[-\tau,T]) \le 2\sup_{t\in[-\tau,T]}|x(t)|,$$

the integrability condition (2.3) for $k=1$ holds, and, finally, the modulus of continuity $\omega_X(\delta,[-\tau,T])$ converges to zero as $\delta=\delta_n$ converges to zero, as the paths of X are continuous.

The last assertion (5.9) is an interesting observation which is a straightforward consequence of (5.7) and (5.8). □

Proof of Lemma 5.3. Noticing that $P^\delta Z^n(s) = X^n(s)$, for $s\in[-\tau,0]\cup[0,T]$, then we can rewrite (5.1) as

$$Z^n(t) = \eta(0) + \int_0^t a(\delta\cdot\lfloor s/\delta\rfloor, \Pi_{\delta\cdot\lfloor s/\delta\rfloor}P^\delta Z^n)ds + \int_0^t b(\delta\cdot\lfloor s/\delta\rfloor, \Pi_{\delta\cdot\lfloor s/\delta\rfloor}P^\delta Z^n)d\tilde{W}_s.$$

Therefore, taking into account that $Z^n(t) = X(t) = \eta(t)$ for $t\in[-\tau,0]$,

$$\begin{aligned}&\sup_{s\in[-\tau,t]}|Z^n(s)-X(s)|^2 \\ &\le 2\left(\int_0^t |a(\delta\cdot\lfloor u/\delta\rfloor, \Pi_{\delta\cdot\lfloor u/\delta\rfloor}P^\delta Z^n) - a(u,\Pi_u X)|du\right)^2 \\ &+ 2\sup_{s\in[0,t]}\left(\int_0^s \left[b(\delta\cdot\lfloor u/\delta\rfloor, \Pi_{\delta\cdot\lfloor u/\delta\rfloor}P^\delta Z^n) - b(u,\Pi_u X)\right]d\tilde{W}_u\right)^2 \\ &\le 2t\int_0^t |a(\delta\cdot\lfloor u/\delta\rfloor, \Pi_{\delta\cdot\lfloor u/\delta\rfloor}P^\delta Z^n) - a(u,\Pi_u X)|^2du \\ &+ 2\sup_{s\in[0,t]}\left(\int_0^s \left[b(\delta\cdot\lfloor u/\delta\rfloor, \Pi_{\delta\cdot\lfloor u/\delta\rfloor}P^\delta Z^n) - b(u,\Pi_u X)\right]d\tilde{W}_u\right)^2.\end{aligned}$$

By Lemma 3.3, Remark 5.1, and the sublinearity of b,

$$\int_0^s \left[b(\delta\cdot\lfloor u/\delta\rfloor, \Pi_{\delta\cdot\lfloor u/\delta\rfloor}P^\delta Z^n) - b(u,\Pi_u X)\right]d\tilde{W}_u$$

is a martingale.

Then taking the expectations we can apply Doob's inequality, and we get for $t \in [0,T]$

$$
\begin{aligned}
&E\left[\sup_{u\in[0,t]} \|\Pi_u Z^n - \Pi_u X\|^2\right] = E\left[\sup_{s\in[-\tau,t]} |Z^n(s) - X(s)|^2\right] \\
&\le 2t \int_0^t E\left[\left|a(\delta\cdot\lfloor u/\delta\rfloor, \Pi_{\delta\cdot\lfloor u/\delta\rfloor} P^\delta Z^n) - a(u,\Pi_u X)\right|^2\right] du \\
&+ 8\int_0^t E\left[\left|b(\delta\cdot\lfloor u/\delta\rfloor, \Pi_{\delta\cdot\lfloor u/\delta\rfloor} P^\delta Z^n) - b(u,\Pi_u X)\right|^2\right] du \\
&\le \max(2T,8)\int_0^t KE\left[\sup_{u\in[0,s]} \|\Pi_{\delta\cdot\lfloor u/\delta\rfloor} P^\delta Z^n - \Pi_u X\|^2 + \delta^{2\alpha}\right] ds \\
&\le C(T)\int_0^t E\left[\sup_{u\in[0,s]} \Big(\|\Pi_{\delta\cdot\lfloor u/\delta\rfloor} P^\delta Z^n - \Pi_{\delta\cdot\lfloor u/\delta\rfloor} P^\delta X\|^2 \right. \\
&\qquad\qquad \left. + \|\Pi_{\delta\cdot\lfloor u/\delta\rfloor} P^\delta X - \Pi_u X\|^2\Big) + \delta^{2\alpha}\right] ds \\
&\le C(T)\int_0^t E\left[\sup_{u\in[0,s]} \|\Pi_u Z^n - \Pi_u X\|^2 + 4\omega_X^2(\delta;[-\tau,T]) + \delta^{2\alpha}\right] ds,
\end{aligned}
$$

where we have used (5.7).

Then Gronwall's inequality gives the upper bound

$$
E\left[\sup_{u\in[0,T]} \|\Pi_u Z^n - \Pi_u X\|^2\right] \le C_1(T)\left(E\left[\omega_X^2(\delta;[-\tau,T])\right] + \delta^{2\alpha}\right) \tag{5.12}
$$

and the proof is accomplished, since $\delta = \delta_n$ and (according to (5.8)) $E\left[\omega_X^2(\delta;[-\tau,T])\right]$ go to zero as n goes to infinity. □

We conclude this section with the proof of Proposition 4.2.

Proof of Proposition 4.2. First, we note that the inequalities in the proof of Proposition 3.2 together with (5.7) imply

$$
\begin{aligned}
&\left\|(\delta\cdot\lfloor t/\delta\rfloor, \Pi_{\delta\cdot\lfloor t/\delta\rfloor} X^n) - (t,\Pi_t X)\right\|_S^2 \\
&\le 4\left(\delta^2 + \sup_{t\in[0,T]} \|\Pi_t Z^n - \Pi_t X\|^2 + \omega_X^2(\delta;[-\tau,T])\right).
\end{aligned}
$$

Then by (5.12) we get

$$
E\left[\left\|(\delta\cdot\lfloor t/\delta\rfloor, \Pi_{\delta\cdot\lfloor t/\delta\rfloor} X^n) - (t,\Pi_t X)\right\|_S^2\right] \le C_2(T)\left(\delta^{2\alpha} + E\left[\omega_X^2(\delta;[-\tau,T])\right]\right).
$$

The thesis follows by observing that

$$
\omega_X(\delta;[-\tau,T]) \le \omega_\eta(\delta;[-\tau,0]) + \omega_X(\delta;[0,T]) \tag{5.13}
$$

and taking into account the result of Lemma A.4 of Słomiński [41], which is recalled in what follows for the ease of the reader. □

LEMMA 5.4 (Lemma A.4 of Słomiński [41]). *Let H, G be two adapted processes with values in $\mathbb{R}^d \otimes \mathbb{R}^d$ and $\mathbb{R}^d$, respectively, such that $\|H_t\|_{\mathbb{R}^d\otimes\mathbb{R}^d}$, $|G_t| \le L < \infty$,*

for some constant $L > 0$, and let $Y_t = \int_0^t H_s\, dW_s + \int_0^t G_s\, ds$, $t \in \mathbb{R}^+$. Then for every $p \in \mathbb{N}$

$$E\left[\omega_Y^{2p}(\tfrac{1}{n};[0,T])\right] = O\left(\left(\tfrac{\log n}{n}\right)^p\right). \tag{5.14}$$

REMARK 5.5. *We observe that the thesis holds for any real $p > 0$, as can be seen following the lines in the proof by Słomiński.*

6. Conclusion. As already recalled in the introduction, Chang in [8] gives a computable approximation for the optimal filter associated with the partially observable delay system (1.1) and (1.2), with $b(t,\theta) = 1$. The state process is approximated by the linear interpolation of $X^n(\ell\delta)$ as in (2.7) and the approximation for the observation process is the linear interpolation of $Y^n(\ell\delta)$ defined by (2.8), while our approximation of the observation process is a continuous time diffusion. However, the two approximation processes coincide at times $\ell\delta$.

The author proves weak convergence of the filters under the assumption that there exists a strictly positive constant k such that

$$E[\exp\{k\|\eta\|^2\}] < \infty,$$

and for any partition $-\tau \le \tau_0 < \cdots < \tau_n = 0$ the $(n+1)$-dimensional random vector $(\eta(\tau_i);\ 0 \le i \le n)$ has a density w.r.t. the Lebesgue measure in $\mathbb{R}^{n+1}$. There are other minor differences between our assumptions and the one by Chang, about the diffusion coefficients, which allow us to consider coefficients of the form (1.4) but not of the form (1.3).

Other weak approximation schemes, such as those recently used by Kushner in a stochastic control framework (see Kushner [25]), could also be used to get weak convergence of the filters. Nevertheless it seems difficult to compute the rate of convergence with these methods. For this reason we were interested in strong approximation schemes.

The problem of strong approximation for stochastic delay differential equations has been the subject of research for many authors in the last years. There is quite a substantial amount of work in this field, and in the following we mention only some of them. Küchler and Platen [21] have proposed a Taylor approximation scheme, besides the Euler scheme, and have proved the strong convergence of their scheme (see also Baker and Buckwar [1] and Buckwar [6]). Hu, Mohammed, and Yan [17] have studied strong convergence of Milstein schemes for stochastic delay differential equations with tame coefficient functions g_a and g_b as in (1.4), with $\gamma_i = \delta_{s_i}$, for $s_i \in [-\tau, 0]$. Though these schemes have better performances than Euler schemes, the authors do not deal with convergence of the expectation of the uniform norm on $[-\tau, T]$, which we need in order to get our first result. In [31] Mao and Sabanis have investigated the uniform norm for the continuous Euler–Maruyama scheme instead of the piecewise linear Euler–Maruyama scheme, in the framework of a variable delay, namely, when in (1.5) the term $X(t-\tau)$ is replaced by $X(\delta(t))$, where $\delta(t)$ is a Lipschitz function with $-\tau \le \delta(t) \le t$, and when g_a and g_b are locally Lipschitz functions. Moreover Mao and Sabanis get, under suitable assumptions, a rate of convergence of order less than or equal to $\sqrt{1/n}$. In our setting, and with our techniques, we cannot get a rate of convergence of order $\sqrt{1/n}$, as one can see from the example at the end of section 4, where the simple case of a Wiener process with redundant observation is considered. Note that in this example the Wiener process coincides with its continuous Euler–Maruyama approximation scheme, and hence the two filters also coincide, while this

does not happen with the piecewise linear Euler–Maruyama scheme. Nevertheless in the latter case the filters coincide with the expectations, and a rate of convergence of order $\sqrt{1/n}$ is achieved when restricting to functions $\phi(\theta)$ depending on a fixed number of times $s_i \in [-\tau, 0]$.

Finally, note that the boundedness condition on a in Proposition 4.2 (and as a consequence in Theorem 2.3) is not necessary, since

$$\left| \int_t^{t'} a(u, \Pi_u X) du \right| \leq |t' - t| \, K^{1/2} \left(1 + \sup_{u \in [0,T]} \|\Pi_u X\|^2 \right)^{1/2},$$

while the boundedness condition on b is fundamental to use Lemma 5.4. However, one expects that this condition can be dropped, since a rate of convergence of order $(\log n/n)^{1/2}$ holds for the piecewise linear approximation of solutions of ordinary stochastic differential equations (see Bouleau and Lépingle [5]). This fact is under investigation.

Note added in proof. As conjectured above, the results of Theorem 2.3 are still valid without the boundedness condition on the coefficients a and b: indeed one can use a generalization of Słomiński's lemma (see [13]).

REFERENCES

[1] C. T. H. Baker and E. Buckwar, *Numerical analysis of explicit one-step methods for stochastic delay differential equations*, LMS J. Comput. Math., 3 (2000), pp. 315–335.

[2] A. G. Bhatt, G. Kallianpur, and R. L. Karandikar, *Uniqueness and robustness of solution of measure-valued equations of nonlinear filtering*, Ann. Probab., 23 (1995), pp. 1895–1938.

[3] A. G. Bhatt, G. Kallianpur, and R. L. Karandikar, *Robustness of the nonlinear filter*, Stochastic Process. Appl., 81 (1999), pp. 247–254.

[4] A. G. Bhatt and R. L. Karandikar, *Robustness of the nonlinear filter: The correlated case*, Stochastic Process. Appl., 97 (2002), pp. 41–58.

[5] N. Bouleau and D. Lépingle, *Numerical Methods for Stochastic Processes*, Wiley Ser. Probab. Math. Statist. Appl. Probab. Statist., John Wiley and Sons, New York, 1994.

[6] E. Buckwar, *Introduction to the numerical analysis of stochastic delay differential equations*, J. Comput. Appl. Math., 125 (2000), pp. 297–307.

[7] A. Calzolari, P. Florchinger, and G. Nappo, *Approximation of nonlinear filters for Markov systems with delayed observations*, SIAM J. Control Optim., 45 (2006), pp. 599–633.

[8] M. H. Chang, *Discrete approximation of nonlinear filtering for stochastic delay equations*, Stochastic Anal. Appl., 5 (1987), pp. 267–298.

[9] P. Del Moral, *Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications*, Probab. Appl. (NY), Springer-Verlag, New York, 2004.

[10] T. Duncan, *Probability Density for Diffusion Processes with Applications to Nonlinear Filtering Theory*, Ph.D. thesis, Stanford University, Stanford, CA, 1967.

[11] S. A. Elsanousi and B. Larssen, *Optimal Consumption under Partial Observations for a Stochastic System with Delay*, Technical report 9, University of Oslo, Oslo, Norway, 2001.

[12] L. Ferrante, S. Bompadre, L. Possati, and L. Leone, *Parameter estimation in a Gompertzian stochastic model for tumor growth*, Biometrics, 56 (2000), pp. 1076–1081.

[13] M. Fischer and G. Nappo, *On the Moments of the Modulus of Continuity of Itô Diffusions*, manuscript; available online from http://www.mat.uniroma1.it/people/nappo/attivita-scientifica.html#preprint.

[14] T. D. Frank, *Multivariate Markov processes for stochastic systems with delays: Application to the stochastic Gompertz model with delay*, Phys. Rev. E (3), 66 (2002), 011914.

[15] T. D. Frank and P. J. Beek, *Stationary solutions of linear stochastic delay differential equations: Applications to biological system*, Phys. Rev. E (3), 64 (2001), 021917.

[16] N. S. Goel, S. C. Maitra, and E. W. Montroll, *On the Volterra and other nonlinear models of interacting populations*, Rev. Modern Phys., 43 (1971), pp. 231–276.

[17] Y. Hu, S.-E. A. Mohammed, and F. Yan, *Discrete-time approximations of stochastic delay equations: The Milstein scheme*, Ann. Probab., 32 (2004), pp. 265–314.

[18] G. Kallianpur, *Stochastic Filtering Theory*, Appl. Math. 13, Springer-Verlag, New York, 1980.
[19] G. Kallianpur and R. L. Karandikar, *White Noise Theory of Prediction, Filtering and Smoothing*, Stochastics Monographs 3, Gordon and Breach Science Publishers, New York, 1988.
[20] P. K. Kallianpur and G. Mandal, *Nonlinear filtering with stochastic delay equations*, in Advances on Theoretical and Methodological Aspects of Probability and Statistics (IISA 1998), Vol. 2, N. Balakrishnan, ed., Gordon and Breach Science Publishers, New York, 2002, pp. 3–36.
[21] U. Küchler and E. Platen, *Strong discrete time approximation of stochastic differential equations with time delay*, Math. Comput. Simulation, 54 (2000), pp. 189–205.
[22] T. G. Kurtz and D. L. Ocone, *Unique characterization of conditional distributions in nonlinear filtering*, Ann. Probab., 16 (1988), pp. 80–107.
[23] H. J. Kushner, *Dynamical equations for optimal nonlinear filtering*, J. Differential Equations, 3 (1967), pp. 179–190.
[24] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Systems Control Found. Appl. 3, Birkhäuser Boston, Boston, MA, 1990.
[25] H. J. Kushner, *Numerical approximations for nonlinear stochastic systems with delays*, Stochastics, 77 (2005), pp. 211–240.
[26] R. H. Kwong and A. S. Willsky, *Estimation and filter stability of stochastic delay systems*, SIAM J. Control Optim., 16 (1978), pp. 660–681.
[27] B. Larssen and N. H. Risebro, *When are HJB-equations in stochastic control of delay systems finite dimensional?*, Stochastic Anal. Appl., 21 (2003), pp. 643–671.
[28] F. Le Gland, *Time discretization of nonlinear filtering equations*, in Proceedings of the 28th IEEE Conference on Decision and Control (Tampa, FL, 1989), Vols. 1–3, IEEE, New York, 1989, pp. 2601–2606.
[29] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes.* I. *General Theory*, Appl. Math. 5, Springer-Verlag, Berlin, 1977.
[30] A. Longtin, J. G. Milton, J. Bos, and M. C. Mackey, *Noise and critical behavior of the pupil light reflex at oscillation onset*, Phys. Rev. A, 41 (1990), pp. 6992–7005.
[31] X. Mao and S. Sabanis, *Numerical solutions of stochastic differential delay equations under local Lipschitz condition*, J. Comput. Appl. Math., 151 (2003), pp. 215–227.
[32] S.-E. A. Mohammed, *Stochastic Functional Differential Equations*, Res. Notes in Math. 99, Pitman, Boston, MA, 1984.
[33] S.-E. A. Mohammed, *Stochastic differential systems with memory: Theory, examples and applications*, in Stochastic Analysis and Related Topics, VI (Geilo, 1996), Progr. Probab. 42, Birkhäuser Boston, Boston, MA, 1998, pp. 1–77.
[34] R. Mortensen, *Optional Control of Continuous Time Stochastic Systems*, Ph.D. thesis, University of California, Berkeley, CA, 1966.
[35] E. Niebur, H. G. Schuster, and D. Kammen, *Collective frequencies and metastability in networks of limit cycle oscillators with time delay*, Phys. Rev. Lett., 67 (1991), pp. 2753–2756.
[36] É. Pardoux, *Filtrage non linéaire et équations aux dérivées partielles stochastiques associées*, in École d'Été de Probabilités de Saint-Flour XIX—1989, Lecture Notes in Math. 1464, Springer-Verlag, Berlin, 1991, pp. 67–163.
[37] R. J. Peterka, *Postural control model interpretation of stabilogram diffusion analysis*, Biol. Cybernet., 82 (2000), pp. 335–343.
[38] J. Pickands, III, *Moment convergence of sample extremes*, Ann. Math. Statist., 39 (1968), pp. 881–889.
[39] M. A. Pinsky, *Brownian continuity modulus via series expansions*, J. Theoret. Probab., 14 (2001), pp. 261–266.
[40] L. M. Ricciardi, *Stochastic population theory: Diffusion processes*, in Mathematical Ecology (Miramare-Trieste, 1982), Biomathematics 17, Springer-Verlag, Berlin, 1986, pp. 191–238.
[41] L. Słomiński, *Euler's approximations of solutions of SDEs with reflecting boundary*, Stochastic Process. Appl., 94 (2001), pp. 317–337.
[42] K. Vasilakos and A. Beuter, *Effects of noise on a delayed visual feedback system*, J. Theoret. Biol., (1993), pp. 389–407.
[43] M. Zakai, *On the optimal filtering of diffusion processes*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 11 (1969), pp. 230–243.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1637–1663

DUAL NONLINEAR FILTERS AND ENTROPY PRODUCTION*

NIGEL J. NEWTON†

Abstract. This paper makes connections between nonlinear filtering and the entropic properties of Markov processes. It starts by developing information flow models for continuous-time, discrete-state filtering problems, identifying rates of information *supply* and *dissipation*. Time reversal yields a *dual* filtering problem in which these flows are interchanged. The dual problem comprises a diffusion signal with nonlinear dynamics, and observations of the point process variety, but yields a finite-dimensional nonlinear filter. The paper goes on to define an entropic time derivative for a general class of Markov processes and relates the entropic derivatives of the signal and filter to the rates of information supply and dissipation. This leads to the definition of a *rate of interactive entropy production*, which measures the time asymmetry of the interaction between the signal and filter. This asymmetry is of the same nature as that occurring in the theory of nonequilibrium statistical mechanics based on stochastic dynamics. In this context, the interaction between the signal and filter is *nondissipative*—a property intimately connected with the existence of a dual problem.

Key words. dual filters, entropic derivative, finite-dimensional nonlinear filters, information theory, nonequilibrium statistical mechanics, statistical filtering

AMS subject classifications. 93E11, 94A17, 62F15, 60J25, 60J60, 60G35, 82C31

DOI. 10.1137/050633809

1. Introduction.

This paper investigates the information-theoretic properties of nonlinear filters for discrete-state Markov processes. It builds on results appearing in [19], concerning linear Gaussian problems. Starting from conventional definitions of signal, X, and observation, Y, it defines information *supply*, *storage*, and *dissipation* processes that quantify the information value of the observation history, $(Y_s,\ s \le t)$, in the context of estimators of the entire signal process, X, its present and future, $(X_s,\ s \ge t)$, and its past, $(X_s,\ s < t)$, respectively. The filter state variable, Z_t, is thought of as *storing* that part of the information in the observation history useful for estimating the present and future values of the signal. As time progresses, information *flows* from the observation process into the filter state and is later dissipated. Rates of information supply and dissipation are identified.

The signal and filter processes have the following key properties: (i) they are *jointly* Markov; (ii) they are *marginally* Markov; (iii) the future of the signal is X_t-conditionally independent of the past of both processes; (iv) the past of the filter is Z_t-conditionally independent of the future of both processes. Property (iii) states that the evolution of the signal is not influenced by the filter process, and so the flow of information between the processes is strictly one-way, from signal to filter. Properties (i)–(iv) are retained if time is reversed and the signal and filter processes are interchanged. The processes thus obtained can be thought of as the signal and

*Received by the editors June 16, 2005; accepted for publication (in revised form) April 21, 2007; published electronically November 2, 2007. This work was partially supported by Leverhulme Trust Research Fellowship 2003/0426, by MURI grant F49620-02-1-0325 (Complex Adaptive Networks for Cooperative Control), by ARO-MURI grant DAAD19-00-1-0466 (Data Fusion in Large Arrays of Microsensors (sensor web)), and by NSF-ITR grant CCR-0325774 Collaborative Research: New Approaches to Experimental Design and Statistical Analysis of Genomic and Structural Biological Data from Multiple Sources.

http://www.siam.org/journals/sicon/46-5/63380.html

†Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK, and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 (njn@essex.ac.uk, njnewton@mit.edu).

filter of a *dual* problem. Like the original, this has information supply, storage, and dissipation processes, and these are simply related to those of the original. The dual problem has an $\mathbb{R}^n$-valued signal with nonlinear dynamics and a point process observation with signal dependent switching rates. Normally such a problem would require an infinite-dimensional filter [6], but because of the special relation between the signal and observation here, the filter is not only of finite dimension but even evolves on a finite space.

One of the motivations for studying such issues is their connection with statistical mechanics. In a modern theory of nonequilibrium statistical mechanics, based on stochastic dynamics (see, for example, [3], [11], [13], and [15]), invariant distributions of time-homogeneous Markov processes are used to model *stationary states* of statistical mechanical systems. The latter come in two varieties called *equilibrium* and *nonequilibrium* states. Nonequilibrium states are time-invariant states in which there is a nonzero flow of some physical quantity (often energy). In this paper we define an entropic time derivative for a general class of continuous-time Markov processes. This is defined either in or out of an invariant distribution and is also defined for time-inhomogeneous processes. For time-homogeneous processes admitting invariant distributions, it coincides with the *rate of entropy production*, as defined in [13] (although it is defined in a totally different way). Because our filtering problems have properties (i) and (ii) above, we can identify three entropic time derivatives: one for the signal, X, one for the filter, Z, and one for the joint process, (X, Z). Subtracting the first two from the third, we obtain a *rate of interactive entropy production* that characterizes the interaction between the signal and filter. (In fact, it is not quite that easy because of degeneracy issues.)

The statistical mechanical properties of Kalman–Bucy filters are developed in some detail in [18], and the interactive statistical mechanics of linear and nonlinear filters are investigated in [20]. In the analogies developed in those papers, the signal-filter pair describes a nondissipative system that makes statistical mechanical sense in reverse time. This property is intimately connected with the existence of a dual filtering problem. It is also connected with an information theoretic optimality of Bayesian estimation, investigated in its abstract form in [17].

The paper is structured as follows. Section 2 introduces the signal and filter models and evaluates the information theoretic quantities of interest. Section 3 derives the dual system and relates its information flows to those of the original. Section 4 defines entropic time derivatives for Markov processes and, in particular, a rate of interactive entropy production for the system comprising the signal and filter. Finally, a simple illustrative example is developed in section 5.

2. Information flow in Markov chain filters. This section investigates the information flows occurring between a discrete-state signal and its nonlinear filter based on observations of the "signal-plus-white-noise" variety. In the model we consider, both processes evolve over the finite time interval $[-T, T]$, and the observation comprises *initial* and *running* parts. This enables the study of transient effects since it permits the signal and filter to be initialized in any consistent state (including their joint invariant distribution, where one exists). By "consistent state" we mean that the average value of the filter variable (whose range is the space of probability measures on that of the signal) should equal the marginal distribution of the signal. The use of a *symmetric* time interval simplifies the discussion of the dual filter in section 3. All random quantities are defined on the complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The signal, X, is a measurable Markov process taking values in the finite set

$\mathbf{X} = \{1, 2, \ldots, n\}$, and having finite rate matrix A, so that

$$\mathbb{P}(X_{t+\epsilon} = i \mid X_t = j) = (I_n + A\epsilon)_{i,j} + o(\epsilon),$$

where I_n is the $n \times n$ matrix identity. Let $\mathcal{P}(\mathbf{X})$ be the set of probability measures on $\mathbf{X}$. We identify elements $Q \in \mathcal{P}(\mathbf{X})$ by the associated probability mass function in vector form:

$$c(Q) := \mathrm{vec}_i\{Q(\{i\})\}.$$

We shall be particularly interested in the convex subset of $\mathcal{P}(\mathbf{X})$ corresponding to probability mass functions that belong to the following set:

$$\mathbb{S}^{n-1} := \{z \in \mathbb{R}^n : z_i \in (0,1) \text{ for all } i;\ \Sigma_i z_i = 1\}.$$

We assume that the initial value of the signal, X_{-T}, has distribution $\theta \in \mathbb{S}^{n-1}$; the prior distribution of X_t, $p_X(t)$ $(= \exp(A(t+T))\theta)$, is then also in $\mathbb{S}^{n-1}$ for all t.

The initial observation is a random variable, ψ, taking values in a Borel space $(M, \mathcal{M}, \mu)$; it is X_{-T}-conditionally independent of X with X_{-T}-conditional density (with respect to μ) $q : \mathbf{X} \times M \to \mathbb{R}^+$. We assume that

$$\mathbf{E} \log(q(i,\psi)/q(j,\psi)) < \infty \quad \text{for all } i, j; \tag{2.1}$$

this ensures (among other properties) that the posterior distribution of X_{-T} is also in $\mathbb{S}^{n-1}$. The *running* observation takes the form

$$Y_t^r = \int_{-T}^t g(X_s)\, ds + W_t \quad \text{for } t \in [-T, T], \tag{2.2}$$

where $g : \mathbf{X} \to \mathbb{R}^d$, and W is a d-vector *shifted* Brownian motion, independent of (X, ψ); i.e., $(W_t = B_{T+t},\ t \in [-T, T])$, where B is a d-vector standard Brownian motion independent of (X, ψ). The *full* observation is then the $M \times \mathbb{R}^d$-valued process $Y = ((\psi, Y_t^r),\ t \in [-T, T])$.

Remark 2.1. *Multidimensional* running observations are included in this study since they play an essential role in section 4, where a relaxation argument is used to define interactive entropy production. Whatever the dimension of Y^r in the filtering problem of interest, a "relaxed" version, in which the dimension of the running observation is at least $n - 1$, is used in that context.

Throughout the paper, $\mathcal{F}_{s,t}^{\xi}$ will be used to signify the σ-field generated by a process ξ over the time interval $[s, t]$, where $-T \le s \le t \le T$:

$$\mathcal{F}_{s,t}^{\xi} := \sigma(\xi_r,\ r \in [s, t]). \tag{2.3}$$

Thus $(\mathcal{F}_{-T,t}^{Y},\ t \in [-T, T])$ is the filtration generated by the observations process.

Wonham's filter [21] is a recursive formula for calculating a continuous process $(Z_t \in \mathbb{R}^n,\ t \in [-T, T])$ with the property that Z_t is the $\mathcal{F}_{-T,t}^{Y}$-conditional distribution of X_t for all t. The initial value, Z_{-T}, is found from the initial prior, θ, the likelihood function, $q(\,\cdot\,, \psi)$, and Bayes's formula; subsequent values of Z are found from the following Itô equation, which has a strong solution (see [21] or [14]):

$$\begin{aligned} Z_t &= Z_{-T} + \int_{-T}^t (AZ_s - \sigma(Z_s)\bar{g}(Z_s))\, ds + \int_{-T}^t \sigma(Z_s)\, dY_s^r \\ &= Z_{-T} + \int_{-T}^t AZ_s\, ds + \int_{-T}^t \sigma(Z_s)\, d\nu_s. \end{aligned} \tag{2.4}$$

Here, $\bar{g} \in C_b^2(\mathbb{R}^n;\mathbb{R}^d)$, $\sigma \in C_b^2(\mathbb{R}^n;\mathbb{R}^{n\times d})$, and, for all $z \in \mathbb{S}^{n-1}$,

$$\begin{aligned} \bar{g}(z) &:= \Sigma_i g(i) z_i, \\ \sigma(z) &:= \mathrm{mat}_{i,l}\{(g(i) - \bar{g}(z))_l z_i\}, \end{aligned} \tag{2.5}$$

and ν is the innovations process,

$$\nu_t := Y_t^r - \int_{-T}^t \bar{g}(Z_s)\, ds. \tag{2.6}$$

Remark 2.2. The process Z remains in $\mathbb{S}^{n-1}$. (See Proposition 2.1.) The extension of the domains of definition of $\bar{g}$ and σ to the whole of $\mathbb{R}^n$ is purely a matter of convenience. It enables the use of standard results from the theory of stochastic differential equations in $\mathbb{R}^n$.

Remark 2.3. Unlike the problem addressed in this paper, the filtering problem in [21] (and most papers on nonlinear filtering) does not involve an initial observation. However, the extension of the results of that paper to the problem addressed here is straightforward. Since (2.4) has a strong solution, and Z_{-T} is X_{-T}-conditionally independent of X, the filtering problem can be thought of as a *family* of problems parametrized by the values of Z_{-T}, each problem having no initial observation. The "parameter" Z_{-T} is then the *prior* distribution for the initial signal value in the corresponding filtering problem. The family of distributions of the signal (respectively, observation, filter) processes of these problems is a regular Z_{-T}-conditional distribution for the signal (respectively, observation, filter) of this paper.

In order to derive some useful properties of X and Z, we introduce a measure transformation of the type first proposed by Duncan in [7]. Let $\mathbb{P}^M$ be the measure on $\mathcal{F}$, whose Radon–Nikodym derivative with respect to $\mathbb{P}$ is L_T, where

$$\begin{aligned} L_t = \frac{\chi(X_{-T})'\theta}{\chi(X_{-T})'Z_{-T}} \exp\Bigg(&- \int_{-T}^t (g(X_s) - \bar{g}(Z_s))'\, dW_s \\ &- \frac{1}{2}\int_{-T}^t |g(X_s) - \bar{g}(Z_s)|^2\, ds \Bigg) \quad \text{for } t \in [-T,T]. \end{aligned} \tag{2.7}$$

Here, and in what follows, $|\cdot|$ is the Euclidean norm in $\mathbb{R}^d$, and $\chi : \mathbf{X} \to \{0,1\}^n$ is the "occupancy" map

$$\chi(i) = \mathrm{vec}_j\{\mathbf{1}_{\{i\}}(j)\}, \tag{2.8}$$

where $\mathbf{1}_B$ is the indicator function of a set B.

PROPOSITION 2.1. (i) *The processes X, Z, and (X,Z) are Markov and have infinitesimal generators $\mathcal{L}^X$, $\mathcal{L}^Z$, and $\mathcal{L}^{X,Z}$, respectively, defined, for appropriate functions f, as follows:*

$$\begin{aligned} (\mathcal{L}^X f)(i) &= \sum_j A_{j,i}(f(j) - f(i)), \\ (\mathcal{L}^Z f)(z) &= \sum_j (Az)_j \frac{\partial f}{\partial z_j}(z) + \frac{1}{2}\sum_{j,k} (\sigma\sigma')(z)_{j,k} \frac{\partial^2 f}{\partial z_j \partial z_k}(z), \\ (\mathcal{L}^{X,Z} f)(i,z) &= \sum_j b(i,z)_j \frac{\partial f}{\partial z_j}(i,z) + \frac{1}{2}\sum_{j,k} (\sigma\sigma')(z)_{j,k} \frac{\partial^2 f}{\partial z_j \partial z_k}(i,z) \\ &\quad + \sum_j A_{j,i}(f(j,z) - f(i,z)), \end{aligned} \tag{2.9}$$

where, for each $i \in \mathbf{X}$, $b(i, \cdot) \in C_b^2(\mathbb{R}^n; \mathbb{R}^n)$ *and, for* $i \in \mathbf{X}$ *and* $z \in \mathbb{S}^{n-1}$,

$$b(i, z) := Az + \sigma(z)(g(i) - \bar{g}(z)). \tag{2.10}$$

(ii) *For each* $t \in [-T, T]$, $\mathcal{F}^X_{-T,T}$ *and* $\mathcal{F}^Y_{-T,t}$ *are* $\mathcal{F}^Z_{-T,t}$*-conditionally independent.*

(iii) *For each* $t \in [-T, T]$, $\mathcal{F}^{X,Z}_{-T,t}$ *and* $\mathcal{F}^X_{t,T}$ *are* X_t*-conditionally independent.*

(iv) *For each* $t \in [-T, T]$, $\mathcal{F}^Z_{-T,t}$ *and* $\mathcal{F}^{X,Z}_{t,T}$ *are* Z_t*-conditionally independent.*

(v) $\mathbb{P}(Z_t \in \mathbb{S}^{n-1} \text{ for all } t \in [-T, T]) = 1$, *and* $\sup_{i \in \mathbf{X},\, t \in [-T,T]} \mathbf{E} \log(Z^{-1}_{t,i}) < \infty$.

(vi) $\mathbb{P}^M$ *is a probability measure;* $\mathbb{P}$ *and* $\mathbb{P}^M$ *are mutually absolutely continuous.*

(vii) *Under* $\mathbb{P}^M$, $(L_t^{-1}, \mathcal{F}^{X,Z}_{-T,t})$ *is a martingale.*

(viii) *Under* $\mathbb{P}^M$, X *and* (Y, Z) *are independent processes but retain the marginal distributions they have under* $\mathbb{P}$.

Proof. See Appendix A. □

We define information *supply*, *storage*, and *dissipation* processes as follows: for each $t \in [-T, T]$,

$$\begin{aligned} S(t) &:= I(X\,;\,(Y_s,\, s \in [-T, t])), \\ C(t) &:= I((X_s,\, s \in [t, T])\,;\,(Y_s,\, s \in [-T, t])), \\ D(t) &:= S(t) - C(t), \end{aligned} \tag{2.11}$$

where, for random variables Θ and Φ taking values in measurable spaces and having joint and marginal distributions $\mathbb{P}_{\Theta,\Phi}$, $\mathbb{P}_\Theta$, and $\mathbb{P}_\Phi$, $I(\Theta\,;\,\Phi)$ is the *mutual information*:

$$I(\Theta\,;\,\Phi) := \begin{cases} \displaystyle\int \log\left(\frac{d\,\mathbb{P}_{\Theta,\Phi}}{d(\mathbb{P}_\Theta \otimes \mathbb{P}_\Phi)}\right) d\,\mathbb{P}_{\Theta,\Phi} & \text{if the integral exists,} \\ +\infty & \text{otherwise.} \end{cases} \tag{2.12}$$

Remark 2.4. Where one of the random quantities in $I(\,\cdot\,;\,\cdot\,)$ is a *stochastic process*, as is the case in (2.11), it is regarded as being a random variable taking values in a product space. Thus, if $(\eta_t,\ t \in [-T, T])$ is a stochastic process taking values in the space E, then η is regarded as being a random variable that takes values in the measurable space $(E^{[-T,T]}, \mathcal{B}_E^{[-T,T]})$, where $E^{[-T,T]}$ is the space of all maps $f : [-T, T] \to E$ and $\mathcal{B}_E^{[-T,T]}$ is the σ-field generated by the cylinder sets in $E^{[-T,T]}$. Thus, for example, the joint distribution $P_{X,Y}$ is a probability measure on $\mathcal{B}_{\mathbf{X}}^{[-T,T]} \times \mathcal{M} \times \mathcal{B}_{\mathbb{R}^d}^{[-T,T]}$. $S(t)$, $C(t)$, and $D(t)$ are insensitive to any choices we make about the nature of the sample paths of X, Y^r, and Z but depend only on their finite-dimensional distributions.

The reader should think of $S(t)$ (respectively, $C(t)$) as being the information about X (respectively, $(X_s,\, s \in [t, T])$) made available to the filter by the observations $(Y_s,\, s \in [-T, t])$. It follows from parts (ii), (iii), (vii), and (viii) of Proposition 2.1 and elementary manipulations of Radon–Nikodym derivatives that

$$\begin{aligned} S(t) &= I(X\,;\,(Z_s,\, s \in [-T, t])) \\ &= I((X_s,\, s \in [-T, t])\,;\,(Z_s,\, s \in [-T, t])) \\ &= \mathbf{E} \log(L_t^{-1}) \\ &= \mathbf{E} \sum_i \log\left(\frac{Z_{-T,i}}{\theta_i}\right) Z_{-T,i} + \frac{1}{2}\mathbf{E}\int_{-T}^{t} |g(X_s) - \bar{g}(Z_s)|^2\, ds. \end{aligned} \tag{2.13}$$

Similarly, it follows from parts (ii), (iii), and (iv) of Proposition 2.1 that

$$\begin{aligned} C(t) &= I((X_s,\, s\in[t,T])\,;\,(Z_s,\, s\in[-T,t])) \\ &= I(X_t\,;\,(Z_s,\, s\in[-T,t])) \\ &= I(X_t\,;\,Z_t) \\ &= \mathbf{E}\sum_i \log\left(\frac{Z_{t,i}}{p_X(t)_i}\right) Z_{t,i}. \end{aligned} \tag{2.14}$$

Itô's rule can be used to obtain a stochastic integral representation for the integrand here, and this shows that the storage $C(t)$ changes with rate

$$\dot{C}(t) = \mathbf{E}\sum_{i,j} \log\left(\frac{Z_{t,i}}{p_X(t)_i}\right) A_{i,j} Z_{t,j} + \frac{1}{2}\mathbf{E}\left|g(X_t) - \bar{g}(Z_t)\right|^2. \tag{2.15}$$

Thus the information dissipation, $D(t)$, takes the form

$$D(t) = -\mathbf{E}\int_{-T}^{t} \sum_{i,j} \log\left(\frac{Z_{s,i}}{p_X(s)_i}\right) A_{i,j} Z_{s,j}\, ds. \tag{2.16}$$

It follows from elementary properties of mutual information that both S and C take only nonnegative values. That the same is true of D follows from (2.18). Furthermore, $\dot{S}(t)$ is clearly nonnegative for all t, and since $\dot{D}(t) = \mathbf{E}f(Z_t) - f(\mathbf{E}Z_t)$, where f is the convex function $f(z) = -\sum_{i,j}\log(z_i)A_{i,j}z_j$, the same is true of $\dot{D}(t)$.

The connection between quantities equivalent to $S(t)$ and $D(t)$ for diffusion filters was investigated in [16]. It was shown there that the rate of change of $D(t)$ is essentially a Fisher information quantity. This is also the case for the discrete-state signal of this paper; in fact

$$\dot{D}(t) = -\mathbf{E}\left(\mathcal{L}^X \log\left(\frac{Z_{t,i}}{p_X(t)_i}\right)\right)(X_t), \tag{2.17}$$

and so $\dot{D}(t)$ is associated with the randomization of X as described by its infinitesimal generator, $\mathcal{L}^X$. The dissipation also has the following interpretation. Consider the problem of estimating $(X_s,\, s\in[t,T])$ given the "initial" observation $(Y_s,\, s\in[-T,t])$ and the running observation $(Y^r_s - Y^r_t,\, s\in[t,T])$. It follows from its definition that $C(t)$ is the initial information supply in this problem. Furthermore, arguments similar to those used to derive (2.13) show that the rate of information supply for this problem at time $s\in[t,T]$ is $\dot{S}(s)$. Thus, for any $t\in[-T,T]$,

$$\begin{aligned} I((X_s,\, s\in[t,T])\,;\,Y) &= C(t) + \frac{1}{2}\mathbf{E}\int_t^T \left|g(X_s) - \bar{g}(Z_s)\right|^2 ds \\ &= S(T) - D(t). \end{aligned} \tag{2.18}$$

So $D(t)$ is that part of the information on X, derived from Y, that is of no use in estimating the values of X at and beyond time t. It is shown in section 3 that $D(t)$ is itself an (average) mutual information.

3. The dual filter. This section explores the time-reversed dynamics of the joint process (X,Z) under both $\mathbb{P}$ and $\mathbb{P}^M$, and shows that they are the dynamics of a dual system, in which Z is a signal process and X its filter.

Let $(\tilde{Z}_t\in\mathbb{R}^{n-1},\, t\in[-T,T])$ be the following parametrization of the filter:

$$\tilde{Z}_t := \Gamma Z_t = [I_{n-1}\ 0]\, Z_t \quad \text{for all } t\in[-T,T]. \tag{3.1}$$

We assume the following:

(H1) $\tilde{Z}_{-T}$ has a square-integrable density with respect to Lebesgue measure.

The following technical lemma prepares the way for time reversal.

LEMMA 3.1. *Suppose that* (H1) *holds; then, for almost every* t*, the distribution of* $\tilde{Z}_t$ *has a density,* $p_{\tilde{Z}}(\,\cdot\,,t)$.

Proof. $\tilde{Z}$ is the unique solution on $(\Omega,\mathcal{F},\mathbb{P},Z_{-T},\nu)$ of the following equation:

$$\tilde{Z}_t = \Gamma Z_{-T} + \int_{-T}^{t} \Gamma A\tilde{\gamma}(\tilde{Z}_s)\,ds + \int_{-T}^{t} \Gamma\sigma\circ\tilde{\gamma}(\tilde{Z}_s)\,d\nu_s, \tag{3.2}$$

where $\tilde{\gamma}\in C_b^2(\mathbb{R}^{n-1};\mathbb{R}^n)$ and $\tilde{\gamma}(\Gamma z)=z$ for all $z\in\mathbb{S}^{n-1}$. The statement of the lemma follows from (H1), the Lipschitz continuity and boundedness of $A\tilde{\gamma}$ and $\sigma\circ\tilde{\gamma}$, the boundedness of the first two derivatives of $\sigma\circ\tilde{\gamma}$, and Theorem 3.1 in [12]. □

For each $t\in[-T,T]$ and each $z\in\mathbb{R}^n$, let

$$\begin{aligned} X_t^* &:= Z_{-t},\\ Z_t^* &:= X_{-t},\\ \bar{b}(z,t) &:= -Az + \operatorname{vec}_i\left\{\sum_j \frac{\partial(\sigma\sigma')_{i,j}}{\partial z_j}\right\}(z) + (\sigma\sigma')(z)\Gamma'\frac{(\nabla_{\tilde{z}}p_{\tilde{Z}})(\Gamma z,-t)}{p_{\tilde{Z}}(\Gamma z,-t)}, \end{aligned} \tag{3.3}$$

where $\bar{b}$ is taken to be zero at any (z,t) for which $p_{\tilde{Z}}(\Gamma z,-t)$ is zero, and $\nabla_{\tilde{z}}p_{\tilde{Z}}$ is the gradient of $p_{\tilde{Z}}$ with respect to its first argument; the latter is understood in the sense of distributions if $p_{\tilde{Z}}(\,\cdot\,,t)$ is not differentiable. For each $t\in[-T,T]$ and $z\in\mathbb{S}^{n-1}$, let $\bar{A}(t)$ and $\breve{A}(z)$ be $n\times n$ Markov rate matrices defined as follows:

$$\begin{aligned} \bar{A}(t)_{i,j} &= A_{j,i}\frac{p_X(-t)_i}{p_X(-t)_j} \quad\text{and}\quad \breve{A}(z)_{i,j} = A_{j,i}\frac{z_i}{z_j} \quad \text{if } j\neq i,\\ \bar{A}(t)_{i,i} &= -\sum_{j\neq i}\bar{A}(t)_{j,i} \quad\text{and}\quad \breve{A}(z)_{i,i} = -\sum_{j\neq i}\breve{A}(z)_{j,i}. \end{aligned} \tag{3.4}$$

THEOREM 3.2. *The processes* X^**,* Z^**, and* (X^*,Z^*) *are Markov and have infinitesimal generators* $\mathcal{L}_t^{X^*}$*,* $\mathcal{L}_t^{Z^*}$*, and* $\mathcal{L}_t^{X^*,Z^*}$*, respectively, defined, for appropriate functions* f*, as follows:*

$$\begin{aligned} (\mathcal{L}_t^{X^*}f)(z) &= \sum_j \bar{b}(z,t)_j\frac{\partial f}{\partial z_j}(z) + \frac{1}{2}\sum_{j,k}(\sigma\sigma')(z)_{j,k}\frac{\partial^2 f}{\partial z_j\partial z_k}(z),\\ (\mathcal{L}_t^{Z^*}f)(i) &= \sum_j \bar{A}(t)_{j,i}(f(j)-f(i)),\\ (\mathcal{L}_t^{X^*,Z^*}f)(z,i) &= \sum_j \bar{b}(z,t)_j\frac{\partial f}{\partial z_j}(z,i) + \frac{1}{2}\sum_{j,k}(\sigma\sigma')(z)_{j,k}\frac{\partial^2 f}{\partial z_j\partial z_k}(z,i)\\ &\quad + \sum_j \breve{A}(z)_{j,i}(f(z,j)-f(z,i)). \end{aligned} \tag{3.5}$$

Proof. See Appendix B. □

We can regard X^* and Z^* as being the signal and filter processes of a dual problem. The dual signal X^* is a Markov process in its own right, evolving on the state space $\mathbb{S}^{n-1}$ and having the infinitesimal generator $\mathcal{L}_t^{X^*}$. The $\mathbf{X}$-valued process Z^* is a filter for this dual signal in the sense that Z_t^* is a sufficient statistic for estimating

the future of X^* from the past of Z^*, as is shown by the following argument. For any bounded, measurable $f : \mathbb{S}^{n-1} \to \mathbb{R}$ and any $B \in \mathcal{F}^X_{t,T}$,

$$\begin{aligned}\int_B \mathbf{E}\left(f(Z_t) \,|\, \mathcal{F}^X_{t,T}\right) d\,\mathbb{P} &= \int_B f(Z_t)\, d\,\mathbb{P} \\ &= \int \mathbf{E}(\mathbf{1}_B f(Z_t) \,|\, X_t)\, d\,\mathbb{P} \\ &= \int \mathbf{E}(\mathbf{1}_B \,|\, X_t)\mathbf{E}(f(Z_t) \,|\, X_t)\, d\,\mathbb{P} \\ &= \int_B \mathbf{E}(f(Z_t) \,|\, X_t)\, d\,\mathbb{P},\end{aligned}$$

where the third step follows from part (iii) of Proposition 2.1. Thus

$$\mathbf{E}\left(f(X^*_t) \,|\, \mathcal{F}^{Z^*}_{-T,t}\right) = \mathbf{E}\left(f(X^*_t) \,|\, Z^*_t\right) \quad \text{a.s.}$$

A straightforward application of Bayes's formula shows that the Z^*_t-conditional distribution of X^*_t has the form

$$\begin{aligned}P_{X^*|Z^*}(B, i, t) &= \mathbb{P}(Z_{-t} \in B \mid X_{-t} = i) \\ &= \frac{\mathbf{E}Z_{-t,i}\mathbf{1}_B(Z_{-t})}{\mathbb{P}(X_{-t} = i)} \\ &= \int_{\Gamma B} \frac{\tilde{\gamma}(\tilde{z})_i p_{\tilde{Z}}(\tilde{z}, -t)}{p_X(-t)_i}\, d\tilde{z}.\end{aligned} \tag{3.6}$$

Clearly we could also regard Z^* as being the dual observation process. However, there are other possibilities. Any dual observation process $(Y^*_t,\ t \in [-T, T])$ must satisfy the following two conditions:

(O1) $\mathcal{F}^{Y^*}_{-T,t} \supseteq \mathcal{F}^{Z^*}_{-T,t}$ for all $t \in [-T, T]$;

(O2) X^* and $\mathcal{F}^{Y^*}_{-T,t}$ are $\mathcal{F}^{Z^*}_{-T,t}$-conditionally independent for all $t \in [-T, T]$.

The first of these requires that the dual filter should be derivable from the dual observations; the second requires that any randomness in Y^* that is not in Z^* should bear no additional information about X^*.

Remark 3.1. The observation process of the original filtering problem in section 2 obviously satisfies the equivalent of (O1); that it also satisfies the equivalent of (O2) is demonstrated in part (ii) of Proposition 2.1.

The following two examples of dual observation processes satisfy (O1) and (O2). For ease of construction, we (temporarily) assume that Z^* has right-continuous paths with left limits.

Example 3.1. Let $(Y^*_t,\ t \in [-T, T])$ be the n-vector process whose ith component is defined as follows:

$$Y^*_{t,i} = \mathbf{1}_{\{i\}}(Z^*_{-T}) + \sum_{s \in (-T,t]} \left(1 - \mathbf{1}_{\{i\}}(Z^*_{s-})\right)\left(\chi(Z^*_s) - \chi(Z^*_{s-})\right)_i. \tag{3.7}$$

Y^* is a vector of counting processes; its ith component increments by one whenever Z^* jumps into state i. Conditions (O1) and (O2) are trivially satisfied for this Y^* since $\mathcal{F}^{Y^*}_{-T,t} = \mathcal{F}^{Z^*}_{-T,t}$ for all t. Note, however, that $\mathcal{F}^{Y^*}_{t,t}$ is not in general the same as $\mathcal{F}^{Z^*}_{t,t}$, and so there is some work for the filter to do; in fact $Z^*_t = \chi^{-1}(Y^*_{-T})$ if no component of Y^* jumps during the interval $(-T, t]$; otherwise

$$Z^*_t = \arg\max_i \left\{\max\{s \in (-T, t] : (Y^*_{s,i} - Y^*_{s-,i}) = 1\}\right\}.$$

The observation process here generates the same filtration as does the filter process, a property that is not as unnatural as it may seem at first sight. In fact it is shared by the original filtering problem of section 2 in the special case that there is no initial observation, the dimension of the running observation d is one, and $g(i) \neq g(j)$ for at least one pair $i, j \in \mathbf{X}$.

Example 3.2. Let $(B_k \in \{0,1\},\ k = 0,1,\ldots)$ be a sequence of independent Bernoulli random variables, independent of (X^*, Z^*), with $\mathbb{P}(B_k = 0) = \mathbb{P}(B_k = 1) = 0.5$. Let

$$Y_t^* = 2Z_t^* + \sum_{k=0}^{N_t} B_k,$$

where $N_t = \#\{s \in (-T,t] : Z_s^* - Z_{s-}^* \neq 0\}$. The observation filtration here, $\mathcal{F}_{-T,t}^{Y^*}$, is strictly larger than that generated by the filter process. However, condition (O2) is satisfied since the source of the additional randomness, (B_k), is independent of (X^*, Z^*). Condition (O1) is satisfied since

$$\begin{aligned} Z_{-T}^* &= [0.5Y_{-T}^*], \\ Z_t^* - Z_{t-}^* &= \left[0.5(Y_t^* - Y_{t-}^*)\right] \quad \text{for } t \in (-T,T], \end{aligned}$$

where, for $u \in \mathbb{R}$, $[u]$ is the largest integer less than or equal to u. The additional randomness in the observations here may appear, at first sight, to be rather superficial. However, this is because the observation process is expressed in terms of the *filter* process, rather than the signal. Of course, having constructed the observation in this way we could express it purely in terms of X^* by integrating out Z^* according to its X^*-conditional distribution.

The dual filter is a very particular example of a *finite-dimensional* nonlinear filter. The dual signal is a multidimensional diffusion process with nonlinear dynamics, and the components of the dual observation process are Markov-modulated point processes. In general, problems of this nature lead to infinite-dimensional nonlinear filters. Here, because of the special connection between the prior distribution and dynamics of X^*, and the observation mechanism, the nonlinear filter is not only finite dimensional but even evolves on a finite space.

In analogy with the definitions of $S(t)$, $C(t)$, and $D(t)$ in section 2, we can identify the information supply, storage, and dissipation of the dual filter as follows:

$$\begin{aligned} S^*(t) &:= I(X^*\,;\,(Y_s^*,\ s \in [-T,t])), \\ C^*(t) &:= I((X_s^*,\ s \in [t,T])\,;\,(Y_s^*,\ s \in [-T,t])), \\ D^*(t) &:= S^*(t) - C^*(t). \end{aligned} \tag{3.8}$$

The dual supply, like $S(t)$, could be found by a measure transformation technique involving an integral formula for the dual martingale

$$L_t^* = \mathbf{E}\left(L_T \,|\, \mathcal{F}_{-T,t}^{X^*,Z^*}\right), \tag{3.9}$$

where L is as defined in (2.7). However, there is an easier way of finding S^*. It follows from (O2), part (ii) of Proposition 2.1, and (2.18) that

$$\begin{aligned} S^*(t) &= I((X_s,\ s \in [-t,T])\,;\,Z) \\ &= S(T) - D(-t). \end{aligned} \tag{3.10}$$

Furthermore, as in (2.14), it follows from (O2) that

$$\begin{aligned} C^*(t) &= I((X_s,\, s\in[-t,T])\,;\,(Z_s,\, s\in[-T,-t])) \\ &= C(-t), \end{aligned} \tag{3.11}$$

and so

$$D^*(t) = S(T) - S(-t). \tag{3.12}$$

This shows that the time-reversed information flows for the original problem can be interpreted as the forward-time information flows for the dual problem, with information supply and dissipation swapping roles.

Like $\dot{D}$ in (2.17), $\dot{D}^*$ is associated with the randomization of the dual signal, as described by its infinitesimal generator, $\mathcal{L}^{X^*}$. In fact

$$\dot{D}^*(t) = -\mathbf{E}\left(\mathcal{L}^{X^*}\log\left(\frac{\chi(Z^*_t)'z}{\chi(Z^*_t)'p_X(-t)}\right)\right)(X^*_t). \tag{3.13}$$

It may seem, at first sight, that $\dot{S}^*(-t)$ should equal $\dot{S}(t)$ since both concern the supply of information between X and Z over the short time interval $[t,t+\delta t]$. That this is not so is because S and S^* are *cumulative* information quantities with derivatives that represent the flow rate of *new* information. What constitutes new information between X and Z depends on the time direction over which it is being accumulated. Over the time interval $[t,t+\delta t]$ an amount $S(t+\delta t)-S(t)\approx\dot{S}(t)\delta t$ of new information is supplied in the original problem. Because this is new, it relates to a dependency between $X_{t+\delta t}$ and $Z_{t+\delta t}$ that is not present between $(X_s,\, s\in[-T,t])$ and $(Z_s,\, s\in[-T,t])$. It is present in $C(t+\delta t)$ $(=C^*(-t-\delta t))$, and therefore useful to the dual filter at time $-t-\delta t$, but not present in $C(t)$ $(=C^*(-t))$, and therefore no longer useful to the dual filter at time $-t$. For this reason the dual filter dissipates it and $\dot{D}^*(-t)=\dot{S}(t)$.

4. Entropic derivatives. This section further investigates the interaction between the components of the joint process (X,Z). Parts (iii) and (iv) of Proposition 2.1 show that, in forward time, the evolution of X at time t is influenced only by the first component of (X_t,Z_t) and that, in reverse time, the evolution of Z is influenced only by the second component. These properties result in unidirectional flows of entropy between the two components. Let $S_X(t)$, $S_Z(t)$, and $S_{X,Z}(t)$ be the entropies of X_t, Z_t, and (X_t,Z_t), respectively:

$$\begin{aligned} S_X(t) &:= -\mathbf{E}\log\left(\chi(X_t)'p_X(t)\right), \\ S_Z(t) &:= -\mathbf{E}\log\left(p_{\tilde{Z}}(\Gamma Z_t,t)\right), \\ S_{X,Z}(t) &:= -\mathbf{E}\log\left(\chi(X_t)'Z_t p_{\tilde{Z}}(\Gamma Z_t,t)\right). \end{aligned} \tag{4.1}$$

(S_Z is defined in terms of a *uniform* measure on $\mathbb{S}^{n-1}$, whose total mass is $\int_{\Gamma\mathbb{S}^{n-1}} d\tilde{z}$.)

As t increases, $S_X(t)$ can increase, decrease, or remain constant according to the distribution $p_X(t)$. Whichever of these happens, X_t acquires "new" entropy from the random switching and loses "old" entropy as it "forgets its past." If X has an invariant distribution, then these two effects are balanced in that distribution. The same is true of $S_Z(t)$, except that part of the "new" entropy acquired by Z_t comes from X_t, (the remainder having its origin in the observation noise). This intuition can be made precise from a consideration of the *side* entropies:

$$S_{X|Z}(t) := S_X(t) - I(X_t\,;\,Z_t) \quad\text{and}\quad S_{Z|X}(t) := S_Z(t) - I(X_t\,;\,Z_t).$$

The common term here $I(X_t\,;\,Z_t)$ $(= C(t))$ is a *shared* component of the joint entropy $S_{X,Z}(t)$. In fact

$$S_{X,Z}(t) = S_{X|Z}(t) + C(t) + S_{Z|X}(t). \tag{4.2}$$

The sum of the first two components on the right-hand side of (4.2) is the signal entropy $S_X(t)$. Its rate of change would not be altered if we were to "turn off" the observation mechanism at time t (which could be achieved by setting the observation function, g, to zero); however, such action would reduce the rate of change of C by the amount $\dot{S}(t)$. From this we can deduce that there is a *flow* of entropy from the first component to the second component on the right-hand side of (4.2) of rate $\dot{S}(t)$. (The term "flow" is justified because the entropy changes in question are conservative.) This flow is strictly one way because of the optimality of the filter, which never discards any information that is relevant to the present and future of the signal.

Similarly, the sum of the second and third components on the right-hand side of (4.2) is the entropy of the dual signal $S_{X^*}(-t)$ $(= S_Z(t))$. Its rate of change would not be affected if we were to "turn off" the dual observation mechanism at (reverse) time $-t$ (which could be achieved by replacing $\breve{A}(X^*_{-t})$ by $\bar{A}(-t)$). However, a simple rearrangement of (2.17) shows that

$$\dot{S}^*(-t) = \mathbf{E} \sum_{i,j:\, A_{i,j}>0} \log\left(\frac{\breve{A}(X^*_{-t})_{i,j}}{\bar{A}(-t)_{i,j}}\right) A_{j,i} X^*_{-t,i}, \tag{4.3}$$

which in turn shows that this action would reduce the rate of change of C^* by the amount $\dot{S}^*(-t)$. From this we can deduce that, in forward time, there is a flow of entropy from the second component to the third component on the right-hand side of (4.2) of rate $\dot{S}^*(-t)$ $(= \dot{D}(t))$. Once again, this is conservative and strictly one way. Of course, these entropy flows are the *information* flows of sections 2 and 3.

In the theory of nonequilibrium statistical mechanics based on stochastic dynamics (see, for example, [3], [11], [13], and [15]), invariant distributions of Markov processes are used to model *stationary states* of statistical mechanical systems. The latter come in two varieties called *equilibrium* and *nonequilibrium* states. Nonequilibrium states represent time-invariant states in which there is a nonzero flow of some physical quantity (often energy). For example, if the two ends of a cylinder of gas are held at different fixed temperatures, a stationary nonequilibrium state is eventually set up in the gas in which energy flows from the hotter end to the cooler end. The state is *stationary* in the sense that the temperature profile along the cylinder does not change with time; it is *nonequilibrium* in the sense that there is a nonzero flow of energy through the system in the state. One of the fundamental features of thermodynamics is that energy flow down a temperature gradient is accompanied by entropy increase.

The joint process (X, Z) of this paper is Markov, and so it can be associated with an abstract statistical mechanical system. This analogy is developed in [18] and [20], where it is shown that the signal-filter pair models a nondissipative system lying on the boundary between systems that do and do not satisfy the second law of thermodynamics. In the analogy, the unidirectional flow of entropy identified above is accompanied by a unidirectional flow of energy that is not driven by temperature gradients. This does not cause entropy increase, and so it makes statistical mechanical sense in reverse time—a property connected with the existence of a dual system.

In what follows, we define an *entropic time derivative* for Markov processes and explore its connections with the information flows of sections 2 and 3. In [13], a

rate of entropy flow is identified for stationary nonequilibrium states of certain time-homogeneous Markov processes. This involves large-time limits of the forward and backward dynamics of the process in its stationary state. A *rate of entropy production* for nonstationary states is then defined in terms of this quantity and the rate of change of entropy of the state (cf. $\dot{S}_X(t)$ above). Despite the name, a positive rate of entropy production does not always imply that the physical system modelled causes entropy increase, merely that its dynamics reveal the direction of time. The rate of entropy production of [13] subsumes entropy changes and (conservative) entropy flows. The entropic time derivative defined below coincides with the rate of entropy production of [13] when the process in question is time-homogeneous and admits an invariant distribution. However, since it involves small-time limits, it is inherently dynamical in nature and is equally applicable to time-inhomogeneous processes such as the dual signal process of section 3. We call the entropic time derivative a "rate of entropy production" because of this connection with statistical mechanics.

Let $(\eta_t,\ t \in [a,b])$ be a Markov process defined on a complete probability space and taking values in a Borel space $(E,\mathcal{E})$. For any $t \in (a,b)$, let $\bar{\epsilon} := \min\{t-a,\, b-t\}$, and let $G := (\mathcal{G}_\epsilon \subseteq \mathcal{B}_E^{[0,\epsilon]},\ 0 < \epsilon < \bar{\epsilon})$ be a family of σ-fields parametrized by ϵ. For each ϵ, $\mathcal{G}_\epsilon$ is a sub-σ-field of $\mathcal{B}_E^{[0,\epsilon]}$ (the σ-field on $E^{[0,\epsilon]}$, generated by the cylinder sets). Let $\Pi^\eta_{t,t+\epsilon}$ and $\Pi^\eta_{t,t-\epsilon}$ be the distributions of the processes $(\eta_{t+s},\ s \in [0,\epsilon])$ and $(\eta_{t-s},\ s \in [0,\epsilon])$, respectively, on the common space $(E^{[0,\epsilon]}, \mathcal{B}_E^{[0,\epsilon]})$. Let $\Pi^{\eta,G}_{t,t+\epsilon}$ and $\Pi^{\eta,G}_{t,t-\epsilon}$ be the restrictions of $\Pi^\eta_{t,t+\epsilon}$ and $\Pi^\eta_{t,t-\epsilon}$ to $\mathcal{G}_\epsilon$. For probability measures $\mathbb{P}_a$ and $\mathbb{P}_b$ on a measurable space $(N,\mathcal{N})$, let $h(\mathbb{P}_a|\mathbb{P}_b)$ be the *relative entropy*:

$$h(\mathbb{P}_a|\mathbb{P}_b) := \begin{cases} \displaystyle\int \log\left(\frac{d\,\mathbb{P}_a}{d\,\mathbb{P}_b}\right) d\,\mathbb{P}_a & \text{if } \mathbb{P}_a \ll \mathbb{P}_b \text{ and the integral exists,} \\ +\infty & \text{otherwise.} \end{cases} \tag{4.4}$$

Definition 4.1. (i) *The rate of entropy production (entropic time derivative) of the process η at time t* with respect to the family G *is*

$$R_{\eta,G}(t) := \limsup_{\epsilon \downarrow 0} \epsilon^{-1} \mathbf{E} h\left(\Pi^{\eta,G}_{t,t+\epsilon} \,|\, \Pi^{\eta,G}_{t,t-\epsilon}\right). \tag{4.5}$$

(ii) *The* path *rate of entropy production of the process η at time t is its rate of entropy production with respect to the family $\{\mathcal{G}_\epsilon = \mathcal{B}_E^{[0,\epsilon]},\ 0 < \epsilon < \bar{\epsilon}\}$.*

(iii) *The* point *rate of entropy production of the process η at time t is its rate of entropy production with respect to the family $\{\mathcal{G}_\epsilon = (\phi_0, \phi_\epsilon)^{-1}(\mathcal{E}^2),\ 0 < \epsilon < \bar{\epsilon}\}$, where ϕ is the coordinate function on $E^{[0,\epsilon]}$.*

Remark 4.1. $R_{\eta,G}$ depends only on the finite-dimensional distributions of η, and is not influenced by the choice of *versions* of η, for the same reason that this is true of S, C, and D. (See Remark 2.4.)

Remark 4.2. The initial-time marginals of $\Pi^\eta_{t,t+\epsilon}$ and $\Pi^\eta_{t,t-\epsilon}$ are the same (they are both the distribution of η_t). So, if the paths of η take values in a Polish space (for example the Skorohod space $D([0,\infty);E)$, where E is a Polish space), then the relative entropy in part (ii) (respectively, part (iii)) can also be thought of as that between regular conditional forward and backward path (respectively, point) *transition* measures.

$R_{\eta,G}(t)$ is a measure of the *time asymmetry* of the process η at time t. Consider a game in which one player secretly takes a segment of the sample path of η between

times $t-\epsilon$ and $t+\epsilon$, tosses a coin, reversing the time direction of the segment if "heads" occurs, and then shows the segment to the second player, asking whether or not it has been reversed. $R_{\eta,G}(t)$ is a measure of how easily the second player, knowing only the forward and backward generators of the process, could answer correctly. It follows from the convexity of the function $x \mapsto x \log x$ and Jensen's inequality that if families G_1 and G_2 are such that $\mathcal{G}_{1,\epsilon} \subseteq \mathcal{G}_{2,\epsilon}$ for all ϵ, then $R_{\eta,G_1}(t) \leq R_{\eta,G_2}(t)$, and so the measures of time asymmetry, $R_{\eta,G}$, become sharper as the families of σ-fields, G, become richer. The sharpest measure is the *path* rate of entropy production.

We do not dwell here on the properties of $R_{\eta,G}$ in the general case but proceed to investigate it in the context of the filtering problems of sections 2 and 3. Since the information quantities $S(t)$, $D(t)$, $S^*(t)$, and $D^*(t)$ are defined in terms of path measures, and since we aim to connect rates of entropy production with these information quantities, we shall concentrate on *path* rates of entropy production and omit the subscript G from R_η. We shall make use of the following conditions; these are *not* all required in the main result of this section (Proposition 4.4).

(H2) $A_{i,j} > 0$ for any $i, j \in \mathbf{X}$ for which $A_{j,i} > 0$.

(H3) $d = n-1$, and $\operatorname{rank}\{\sigma(z)\} = n-1$ for all $z \in \mathbb{S}^{n-1}$.

(H4) $\max_{i,j} \mathbf{E}(q(i,\psi)/q(j,\psi))^2 < \infty$, where ψ is the initial observation and q is the likelihood function of (2.1).

(H5) For a specified $t \in (-T,T)$, there is an $\epsilon > 0$ such that $p_{\tilde{Z}}(\tilde{z},s)$ is differentiable with respect to $\tilde{z}$ for all $s \in [t-\epsilon, t]$, and $\mathbf{E}\,|\sigma(Z_{t+r})'\Gamma'(\nabla_{\tilde{z}} \log p_{\tilde{Z}})(\Gamma Z_{t+r}, t-|r|)|^2$ is bounded and continuous as a function of r on the interval $r \in (-\epsilon, \epsilon)$.

PROPOSITION 4.2. (i) *Let R_X and R_{Z^*} be the path rates of entropy production of the signal process, X, and its time reversal, Z^*. Then, for each $t \in (-T,T)$,*

(4.6)
$$R_X(t) = R_{Z^*}(-t) = \begin{cases} \displaystyle\sum_{i,j:\, A_{i,j}>0} \log\left(\frac{A_{i,j}}{\bar{A}(-t)_{i,j}}\right) A_{i,j} p_X(t)_j & \text{if (H2) is satisfied,} \\ +\infty & \text{otherwise.} \end{cases}$$

(ii) *Let R_Z and R_{X^*} be the path rates of entropy production of the filter process, Z, and its time reversal, X^*. If* (H1), (H3), *and* (H4) *are satisfied, then, for each $t \in (-T,T)$ for which* (H5) *is satisfied,*

$$R_Z(t) = R_{X^*}(-t) = \frac{1}{2}\mathbf{E}\left|(\Gamma\sigma(Z_t))^{-1}\Gamma(AZ_t - \bar{b}(Z_t,-t))\right|^2 < \infty. \tag{4.7}$$

(iii) *Let $R_{X,Z}$ and R_{X^*,Z^*} be the path rates of entropy production of the joint process, (X,Z), and its time reversal, (Z^*,X^*). If* (H1)–(H4) *are satisfied, then, for each $t \in (-T,T)$ for which* (H5) *is satisfied,*

$$\begin{aligned} R_{X,Z}(t) = R_{X^*,Z^*}(-t) = {} & \frac{1}{2}\mathbf{E}\left|(\Gamma\sigma(Z_t))^{-1}\Gamma(b(X_t,Z_t) - \bar{b}(Z_t,-t))\right|^2 \\ & + \mathbf{E}\sum_{i,j:\, A_{i,j}>0} \log\left(\frac{A_{i,j}}{\breve{A}(Z_t)_{i,j}}\right) A_{i,j} Z_{t,j} < \infty. \end{aligned} \tag{4.8}$$

Proof. See Appendix C. □

Remark 4.3. Parts (ii) and (iii) of Proposition 4.2 are not in their ripest form, in that condition (H5) is not easy to verify. It would clearly be preferable to replace it by explicit conditions on the parameters n, A, d, g, etc. The main difficulty with

such a result is that of establishing suitable regularity properties for $p_{\tilde{Z}}$—the forward equation for $p_{\tilde{Z}}$ is not uniformly parabolic, and this is a condition in many results concerning explicit bounds on solutions and their derivatives. (See, for example, [10].) Where explicit bounds are available, for example in [1], they are not sufficiently tight for the purposes of establishing (H5). However, (H5) is only marginally stronger than the finiteness of the right-hand sides in (4.7) and (4.8). The term in (H5) also occurs in the Itô expansion of $\log(p_{\tilde{Z}}(\Gamma Z_t, t))$, and so (H5) is connected with the finiteness of the rate of change of entropy of the filter, $\dot{S}_Z(t)$. For example, if (X_t, Z_t) is in an invariant distribution with twice differentiable $\tilde{Z}_t$-marginal density $p_{\tilde{Z},SS}$ and $S_{Z,SS} > -\infty$, then a comparison of the terms in $\log(M_1(x))$ of Appendix C with those of the Itô expansion of $\log(p_{\tilde{Z},SS}(\Gamma Z_t))$ shows that (H5) is satisfied, and

$$\frac{1}{2}\mathbf{E}\left|\sigma(Z_t)'\Gamma'(\nabla_{\tilde{z}}\log p_{\tilde{Z},SS})(\Gamma Z_t)\right|^2 = -\operatorname{tr}(A) + \frac{1}{2}\sum_{i,j}\mathbf{E}\frac{\partial^2(\sigma\sigma')_{i,j}}{\partial z_i \partial z_j}(Z_t).$$

Remark 4.4. When regarded as an equation on $\Gamma B_m \times (-T, T)$, where $B_m = \{z \in \mathbb{S}^{n-1} : z_i > m^{-1} \text{ for all } i\}$, the forward equation for $p_{\tilde{Z}}$ *is* uniformly parabolic; this observation can be used as the basis for a definition of a *local* path rate of entropy production for Z and (X, Z) with respect to the sequence of sets (B_m). This is the lim sup, over m, of the path rate of entropy production of the processes Z and (X, Z), *stopped* at the exit time of the forward and backward processes $(Z_{t+s},\ s \in [0, \epsilon])$ and $(X^*_{-t+s},\ s \in [0, \epsilon])$ from B_m. Condition (H5) is easily verified for these stopped processes, and Z, X^*, (X, Z), and (X^*, Z^*) can be shown to have local path rates of entropy production equal to the derivatives of Proposition 4.2 under mild, explicit conditions. However, unlike R_η of Definition 4.1, this local rate depends on the choice of a specific version of Z. We do not pursue it further in this paper.

Remark 4.5. If (H3) is not satisfied, then, except in special cases, R_Z, R_{X^*}, $R_{X,Z}$, and R_{X^*,Z^*} are all infinite. This is because the components of the vector fields $\Gamma A Z_t$ and $\Gamma\bar{b}(Z_t, -t)$ in the kernel of $\Gamma\sigma(Z_t)\sigma(Z_t)'\Gamma'$ can differ, resulting in the mutual singularity of $\Pi^Z_{t,t+\epsilon}$ and $\Pi^Z_{t,t-\epsilon}$.

Remark 4.6. Proposition 4.2 shows that (at least for the processes of this paper under (H1)–(H5)) the path rate of entropy production is *time symmetric*—the rates for the forward and time-reversed processes are identical, even though $h(\Pi^\eta_{t,t+\epsilon} \,|\, \Pi^\eta_{t,t-\epsilon}) \neq h(\Pi^\eta_{t,t-\epsilon} \,|\, \Pi^\eta_{t,t+\epsilon})$ in general. This is because the negation caused by the inversion of the Radon–Nikodym derivative in $h(\Pi^\eta_{t,t+\epsilon} \,|\, \Pi^\eta_{t,t-\epsilon})$ is countered, in the limit of small ϵ, by the change of integrating measure.

The fact that X and Z are both marginally and jointly Markov can be used to isolate a component of their joint path rate of entropy production that is purely associated with their *interaction*. If conditions (H1)–(H4) are satisfied, and (H5) is satisfied for a given t, then an *interactive* path rate of entropy production can be defined by simple subtraction. A straightforward calculation then shows that

$$R_I(t) := R_{X,Z}(t) - R_X(t) - R_Z(t) = \dot{S}(t) + \dot{D}(t). \tag{4.9}$$

Clearly, the interactive path rate of the dual problem at time $-t$ also takes this value. Thus the time asymmetry of the *interaction* between the signal and filter processes, as measured by the path rate of entropy production, is entirely associated with the flow of information between the processes. The reason for this property becomes apparent when the Radon–Nikodym derivative $d\Pi^{X,Z}_{t,t+\epsilon}/d\Pi^{X,Z}_{t,t-\epsilon}$ is factorized in terms of the likelihood functions of the original and dual filters. (See (C.3), (C.6), (C.7), and

(C.8) in Appendix C.) It is closely related to the time symmetry property discussed in Remark 4.6. Since R_I is purely associated with the conservative flows of information, $\dot{S}$ and $\dot{D}$, it could also be called a rate of interactive entropy *flow*.

The sum of the supply and dissipation rates, $\dot{S}(t) + \dot{D}(t)$, is finite under much weaker conditions than (H1)–(H5), and so it is natural to ask whether one can define a rate of interactive entropy production for filtering problems not satisfying these conditions. In fact this can be done via a relaxation argument.

Let $(\Phi^k := (\Omega^k, \mathcal{F}^k, \mathbb{P}^k, X^k, \psi^k, Y^{k,r}, Z^k),\ k \in \mathbb{N})$ be a sequence of filtering problems similar to that of section 2. More precisely, for each k, let X^k be an $\mathbf{X}$-valued Markov process with initial distribution $\theta^k \in \mathbb{S}^{n-1}$ and rate matrix A^k; let ψ^k $(\in M^k)$ be an initial observation that is X^k_{-T}-conditionally independent of X^k; and let $Y^{k,r}$ be an $\mathbb{R}^m$-valued running observation process of the form

$$Y^{k,r}_t = \int_{-T}^{t} g^k(X_s)\, ds + W^k_t \quad \text{for } t \in [-T, T], \tag{4.10}$$

where $m \geq \max\{d, n-1\}$, $g^k : \mathbf{X} \to \mathbb{R}^m$, and W^k is an m-vector shifted Brownian motion, independent of (X^k, ψ^k). Suppose, further, that the following hold:

(A1) for each k, Φ^k satisfies (2.1) and (H1)–(H4);

(A2) $A^k \to A$;

(A3) $(X^k_{-T}, Z^k_{-T}) \Rightarrow (X_{-T}, Z_{-T})$ and $\mathbf{E}^k \log(Z^k_{-T,i}) \to \mathbf{E} \log(Z_{-T,i})$ for all i;

(A4) for each $i \in \mathbf{X}$, $g^k(i) \to g^0(i)$, where g^0 is the $\mathbb{R}^m$-valued function on $\mathbf{X}$ whose first d components are those of g and whose remaining components are zero.

Example 4.1. For each k, let $\zeta^k : \Omega^k \to \mathbb{S}^{n-1}$ and $\eta^k : \Omega^k \to \mathbb{R}^n$ be independent, ζ^k having the same distribution as Z_{-T} and η^k having the standard n-variate Gaussian distribution. Let

$$\begin{aligned}
A^k &:= A + k^{-1}(J_n J_n' - nI_n),\\
g^k_l &:= g_l \qquad\qquad \text{for } l = 1, 2, \ldots, d,\\
&\quad\ k^{-1}\mathbf{1}_{\{l-d\}} \qquad \text{for } l = d+1, d+2, \ldots, d+n-1,\\
\psi^{k,1} &:= (kn)^{-1} J_n + (1 - k^{-1})\zeta^k,\\
\psi^{k,2} &:= \chi(X^k_{-T}) + k\eta^k,
\end{aligned} \tag{4.11}$$

where J_n is the n-vector whose entries are all 1, $\mathbf{1}$ is the indicator function, X^k_{-T} has $\psi^{k,1}$-conditional distribution $\psi^{k,1}$, and X^k is X^k_{-T}-conditionally independent of (ζ^k, η^k) with rate matrix A^k. Let the running observation be as in (4.10), and let the initial observation be ψ^k $(:= (\psi^{k,1}, \psi^{k,2}))$. Then (Φ^k) satisfies (A1)–(A4). (It can also be shown by means of a number of theorems in [10] that $\tilde{Z}_t$ has a strictly positive differentiable density for all $t \in (-T, T]$, and so Φ^k satisfies at least part of (H5).)

Definition 4.3. *The* path rate of interactive entropy production *of the process* (X, Z) *at time* t *is*

$$\begin{aligned}
R_I(t) := \limsup_{k\uparrow\infty} \limsup_{\epsilon\downarrow 0} \epsilon^{-1} \mathbf{E}^k \Bigg(& h\left(\Pi^{X,Z,k}_{t,t+\epsilon} \,|\, \Pi^{X,Z,k}_{t,t-\epsilon}\right) - h\left(\Pi^{X,k}_{t,t+\epsilon} \,|\, \Pi^{X,k}_{t,t-\epsilon}\right)\\
& - h\left(\Pi^{Z,k}_{t,t+\epsilon} \,|\, \Pi^{Z,k}_{t,t-\epsilon}\right)\Bigg),
\end{aligned} \tag{4.12}$$

where $\infty - \infty$ *is taken to be* $+\infty$, (Φ^k) *is a sequence of filtering problems satisfying* (A1)–(A4), *and* $\Pi^{X,Z,k}_{t,t+\epsilon}$ *is the equivalent of* $\Pi^{X,Z}_{t,t+\epsilon}$ *for the filtering problem* Φ^k, *etc.*

The following proposition justifies Definition 4.3, and shows that it is consistent, in the sense that the limits in (4.12) do not depend on the specific sequence satisfying (A1)–(A4). It also evaluates R_I.

PROPOSITION 4.4. *Let* (Φ^k) *be a sequence of filtering problems satisfying* (A1)–(A4).

(i) *The processes* $((X^k, Z^k), k = 1, 2, \ldots)$, *considered as random variables in the product space* $\mathbf{X}^{[-T,T]} \times C([-T, T]; \mathbb{S}^{n-1})$, *converge weakly as* $k \uparrow \infty$ *to* (X, Z).

(ii) *For each* $t \in (-T, T)$, *the forward and backward generators of the processes* (X^k, Z^k) *at time* t *converge to those of* (X, Z) *in the following sense. For any* $f, h : \mathbf{X} \times \mathbb{R}^n \to \mathbb{R}$ *such that* $f(i, \cdot), h(i, \cdot) \in C^2(\mathbb{R}^n; \mathbb{R})$ *for all* i,

$$\begin{aligned} &\lim_{k\uparrow\infty} \lim_{\epsilon\downarrow 0} \epsilon^{-1} \mathbf{E}^k \left(f(X^k_{t\pm\epsilon}, Z^k_{t\pm\epsilon}) - f(X^k_t, Z^k_t) \right) h(X^k_t, Z^k_t) \\ &\qquad = \lim_{\epsilon\downarrow 0} \epsilon^{-1} \mathbf{E} \left(f(X_{t\pm\epsilon}, Z_{t\pm\epsilon}) - f(X_t, Z_t) \right) h(X_t, Z_t). \end{aligned} \tag{4.13}$$

(iii) *For each* $t \in (-T, T)$, *for which the filtering problems in the sequence* (Φ^k) *satisfy* (H5),

$$R_I(t) = \dot{S}(t) + \dot{D}(t). \tag{4.14}$$

Proof. See Appendix D. □

Remark 4.7. If (H1)–(H5) are satisfied by the original filtering problem, then the problems in the sequence (Φ^k) can all be chosen to be the same as the original, which shows that Definition 4.3 is consistent with that in (4.9).

5. A simple example. The ideas of sections 2 to 4 are illustrated here through a simple example in which $n = 2$ and the signal process X has the following symmetrical rate matrix for some positive constant λ:

$$A = \begin{bmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{bmatrix}.$$

The filtering problem will be investigated in its invariant distribution, and so the initial distribution of X is chosen to be $\theta = [0.5\ 0.5]'$.

The running observation has dimension $d = 1$, and $g(i) = \alpha u(i)$, where α is a constant and $u(i) := 2i - 3$. In order to exploit the symmetry of the example, we will represent the signal value by the $\{-1, +1\}$-valued statistic $u(X_t)$, and the filter value, Z_t, by the $(-1, 1)$-valued conditional mean of $u(X_t)$,

$$\mu_t := Z_{t,2} - Z_{t,1}.$$

The filter formulae of (2.4) now take the form

$$\begin{aligned} d\mu_t &= \left(-2\lambda\mu_t - \alpha^2(1 - \mu_t^2)\mu_t \right) dt + \alpha(1 - \mu_t^2)\, dY^r_t \\ &= -2\lambda\mu_t\, dt + \alpha(1 - \mu_t^2)\, d\nu_t. \end{aligned} \tag{5.1}$$

μ has the following invariant density (for an appropriate normalizing constant K):

$$p_{\mu,SS}(x) = \frac{K}{(1 - x^2)^2} \exp\left(\frac{-2\lambda x^2}{\alpha^2(1 - x^2)} \right). \tag{5.2}$$

In order to initialize the filter in this distribution, we choose an initial observation $\psi : \Omega \to (-1, 1)$ with the following X_{-T}-conditional density:

$$p_{\psi|X_{-T}}(x, i) = (1 + u(i)x) p_{\mu,SS}(x).$$

With this initial observation, $\mu_{-T} = \psi$.

The storage, $C(t)$, and the supply and dissipation rates, $\dot{S}(t)$ and $\dot{D}(t)$, can now be evaluated from (2.14), (2.13), and (2.16):

$$\begin{aligned} C(t) &= \frac{1}{2}\int_{-1}^{1}\left(\log(1-x^2) + x\log\left(\frac{1+x}{1-x}\right)\right)p_{\mu,SS}(x)\,dx, \\ \dot{S}(t) &= \frac{1}{2}\alpha^2\int_{-1}^{1}(1-x^2)p_{\mu,SS}(x)\,dx, \\ \dot{D}(t) &= \lambda\int_{-1}^{1}x\log\left(\frac{1+x}{1-x}\right)p_{\mu,SS}(x)\,dx. \end{aligned} \tag{5.3}$$

(Since the signal and filter are initialized in their joint invariant distribution, these quantities do not depend on t, and $\dot{S} = \dot{D}$.) One can easily define "disintegrated," μ_t-conditional information quantities by removing the averaging operation in (5.3). This reveals that the disintegrated supply rate is proportional to the power signal-to-noise ratio, α^2; it is largest when μ_t is close to zero and decreases to zero as μ_t approaches ± 1. Similarly, the disintegrated dissipation rate is proportional to the signal switching rate, λ; it is large when μ_t approaches ± 1 and zero when μ_t is zero. The disintegrated storage is zero when $\mu_t = 0$ and equal to the self information of X_t (1 bit) if $\mu_t = \pm 1$.

It is clear that (H1)–(H5) are all satisfied for this example. Time reversing the pair (X, μ), we obtain the dual system. The dual signal X^* takes values in the interval $(-1, 1)$ and evolves according to the following equation:

$$dX_t^* = -2\lambda X_t^*\,dt + \alpha(1 - X_t^*)^2\,dV_t^*, \tag{5.4}$$

where $((V_t^*, \mathcal{F}_{-T,t}^{X^*,Z^*}),\ t \in [-T,T])$ is a scalar shifted Brownian motion; it has initial distribution $p_{\mu,SS}$. The dual observation is the counting process Y^* of (3.7) in Example 3.1. The dual filter is a two-state Markov process with initial X^*-conditional distribution

$$P(Z_{-T}^* = i \mid X_{-T}^*) = (1 + u(i)X_{-T}^*)/2 \quad \text{for } i = 1, 2$$

and X^*-conditional rate matrix at time t

$$\breve{A}(X_t^*) = \lambda\begin{bmatrix} -(1+X_t^*)/(1-X_t^*) & (1-X_t^*)/(1+X_t^*) \\ (1+X_t^*)/(1-X_t^*) & -(1-X_t^*)/(1+X_t^*) \end{bmatrix}.$$

The observation-conditional distribution of X_t^* is as follows:

$$P(X_t^* \in B \mid Y_s^*,\ s \in [-T,t]) = \int_B (1 + u(Z_t^*)x)p_{\mu,SS}(x)\,dx.$$

The disintegrated dual information supply rate is the same as the disintegrated primal dissipation rate, and so it is large when the switching rate scaling factor λ is large. Thus a high switching rate for Y^* is associated with a high information supply rate. Similarly, the disintegrated dual dissipation rate is the same as the disintegrated primal supply rate, and so it is large when the dual signal noise factor, α^2, is large. We can also "disintegrate" the dual supply and dissipation rates with respect to X_t^* to obtain Z_t^*-conditional rates. However, because of the symmetry of the simple problem considered here, these are equal to $\dot{S}^*$ and $\dot{D}^*$ for both values of Z_t^*.

Both processes X and μ are *self-adjoint*, in that their dynamics are the same in both time directions. This results in their (marginal) path rates of entropy production, R_X and R_μ, being zero. Of course the dynamics of their *interaction* are not time symmetric, and

$$R_I = R_{X,\mu} = \dot{S} + \dot{D} = 2\dot{S}. \tag{5.5}$$

6. Conclusions. This paper has developed dual nonlinear filtering problems and defined and related information flows for them. It has defined an entropic time derivative for a general class of Markov processes and connected the entropic derivatives of the signal and filter processes with these information flows. In the special case of time-homogeneous processes admitting invariant distributions, the entropic time derivative coincides with the "rate of entropy production" of [13]. This connection enables the construction of statistical mechanical analogies for the signal and filter (reported in [18] and [20]) in which the signal and filter interact in a nondissipative way. The existence of dual filtering problems is thus connected with the second law of thermodynamics.

The derivation of dual filters by the techniques of this paper is not restricted to systems with discrete-state signals. The same approach can be taken with filters for diffusion processes and leads to systems in which the optimal filter for an *infinite-dimensional* signal is of *finite* dimension. A notable exception to this is the linear Gaussian case in which the signal and filter are both of finite dimension. The signal and Kalman–Bucy filter equations are *self-dual* in a certain sense. (See [19].) Other special cases are the finite-dimensional filters of Beneš [2].

Appendix A. Proof of Proposition 2.1. It follows from the boundedness of $g - \bar{g}$ and the Novikov criterion (e.g., Theorem 6.1 in [14]) that, under $\mathbb{P}$, $(L_t, \mathcal{F}^{X,Y}_{-T,t})$ is a martingale, and so $\mathbb{P}^M(\Omega) = \mathbf{E}L_T = \mathbf{E}L_{-T} = 1$, which shows that $\mathbb{P}^M$ is a probability measure. By the same argument $(L_t^{-1}, \mathcal{F}^{X,Y}_{-T,t})$ is a martingale under $\mathbb{P}^M$, and this proves part (vi).

For any $\epsilon > 0$, let

$$\begin{aligned}\tau_\epsilon &:= \inf\{t \in [-T,T] : Z_{t,i} \le \epsilon \text{ for any } i \in \mathbf{X}\} \wedge T,\\ \xi^\epsilon_t &:= Z_{t\wedge\tau_\epsilon};\end{aligned}$$

then, from Itô's rule applied to the second equation in (2.4),

$$\begin{aligned}\log(\xi^\epsilon_{T,i}) &= \log(\xi^\epsilon_{-T,i}) + \int_{-T}^{\tau_\epsilon}\left((\xi^\epsilon_{t,i})^{-1}(A\xi^\epsilon_t)_i - \frac{1}{2}(\xi^\epsilon_{t,i})^{-2}(\sigma\sigma'(\xi^\epsilon_t))_{i,i}\right)dt\\ &\quad + \int_{-T}^{\tau_\epsilon}(\xi^\epsilon_{t,i})^{-1}\left(\sigma(\xi^\epsilon_t)\,d\nu_t\right)_i\\ &\ge \log(\xi^\epsilon_{-T,i}) + \int_{-T}^{\tau_\epsilon}\left(A_{i,i} - \frac{1}{2}|g(i) - \bar{g}(Z_t)|^2\right)ds + \int_{-T}^{\tau_\epsilon}(g(i) - \bar{g}(Z_t))'\,d\nu_t\\ &\ge \zeta(\omega) > -\infty \quad \text{a.s.},\end{aligned}$$

where ζ does not depend on ϵ or i. Thus, if $\epsilon < \exp(\zeta(\omega))$, then $\tau_\epsilon(\omega) = T$. This, and the fact that $\mathbb{P}(\Sigma_i Z_{t,i} = 1 \text{ for all } t) = 1$ (which follows immediately from (2.4)), proves part (v).

For any Borel measurable $x : [-T,T] \to \mathbf{X}$ and any $t \in [-T,T]$, let

$$M(x)_t := Z_t - Z_{-T} - \int_{-T}^{t}(AZ_s + \sigma(Z_s)(g(x_s) - \bar{g}(Z_s)))\,ds;$$

then $(M(x)_t, \mathcal{F}^Z_{-T,t})$ and $(M(X)_t, \mathcal{F}^{X,Z}_{-T,t})$ are both continuous semimartingales with the same quadratic covariation:

$$\begin{aligned}[M(x)_i, M(x)_j]_t &= [M(X)_i, M(X)_j]_t \\ &= \int_{-T}^t (\sigma\sigma')(Z_s)_{i,j}\, ds \\ &= \int_{-T}^t Z_{s,i} Z_{s,j} (g(i) - \bar{g}(Z_s))'(g(j) - \bar{g}(Z_s))\, ds.\end{aligned}$$

Let

$$K(x)_t := \frac{\chi(x_{-T})'\theta}{\chi(x_{-T})' Z_{-T}} \exp\left(\int_{-T}^t (\chi(x_s)' Z_s)^{-1} \chi(x_s)'\, dM(x)_s \right. \\ \left. - \frac{1}{2} \int_{-T}^t |g(x_s) - \bar{g}(Z_s)|^2\, ds \right).$$

(The stochastic integral here is well defined in the L^2 sense because of part (v).) Then $L_t^{-1} = K(X)_t$, which shows that L is adapted to $(\mathcal{F}^{X,Z}_{-T,t})$. This, together with the fact that $(L_t^{-1}, \mathcal{F}^{X,Y}_{-T,t})$ is a martingale under $\mathbb{P}^M$, proves part (vii).

The abstract Bayes formula of nonlinear filtering (e.g., Theorem 7.23 in [14]) states that, for any $B \in \mathcal{B}^X_{-T,T}$,

$$\mathbb{P}\left(X \in B \mid \mathcal{F}^Y_{-T,t}\right) = \int_B K(x)_t\, P_X(dx), \tag{A.1}$$

where P_X is the distribution of X considered as a random variable taking values in $(\mathbf{X}^{[-T,T]}, \mathcal{B}^X_{-T,T})$. (In fact, Theorem 7.23 of [14] does not treat filtering problems having initial observations. However, its extension to the filtering problem of this paper is straightforward. See Remark 2.3.) Part (viii) follows from the fact that

$$\frac{d\,\mathbb{P}}{d\,\mathbb{P}^M} = K(X)_T = \frac{dP_{X,Y}}{d(P_X \otimes P_Y)}(X, Y),$$

where P_Y is the distribution of Y considered as a random variable taking values in $(M \times C([-T,T]; \mathbb{R}^d), \mathcal{M} \times \mathcal{B}^d_{-T,T})$, and $P_{X,Y}$ is the joint distribution of X and Y. (See Theorem 7.23 in [14].)

The signal and filter processes X and Z are Markov, X by definition, and Z since it is a solution of the second equation in (2.4), in which the process ν is a multivariate Brownian motion. (See, again, Theorem 7.23 in [14].) That their infinitesimal generators are as stated in (2.9) follows from Itô's rule. For any $B \in \mathcal{F}^{X,Z}_{-T,t}$ and any $C \in \mathcal{F}^{X,Z}_{t,T}$,

$$\begin{aligned}\int_B \mathbb{P}\left(C \mid \mathcal{F}^{X,Z}_{-T,t}\right) d\,\mathbb{P} &= \int_B \mathbf{E}^M\left(\mathbf{1}_C L_T^{-1} \mid \mathcal{F}^{X,Z}_{-T,t}\right) d\,\mathbb{P}^M \\ &= \int_B \mathbf{E}^M\left(\mathbf{1}_C L_T^{-1} L_t \mid \mathcal{F}^{X,Z}_{-T,t}\right) L_t^{-1}\, d\,\mathbb{P}^M \\ &= \int_B \mathbf{E}^M\left(\mathbf{1}_C L_T^{-1} L_t \mid X_t, Z_t\right) d\,\mathbb{P},\end{aligned} \tag{A.2}$$

where we have used part (vii), the fact that (X, Z) is Markov under $\mathbb{P}^M$, and the fact that $L_T^{-1} L_t = K(X)_T K(X)_t^{-1}$, which is $\mathcal{F}^{X,Z}_{-T,t}$-measurable, in the third step.

From this it follows that $\mathbb{P}(C \,|\, \mathcal{F}^{X,Z}_{-T,t})$ is (X_t, Z_t)-measurable, which shows that (X, Z) is Markov under $\mathbb{P}$. Once again, Itô's rule shows that $\mathcal{L}^J$ is its generator, and this proves part (i).

Part (ii) follows directly from (A.1), since $K(x)_t$ is $\mathcal{F}^Z_{-T,t}$-measurable. For any $B \in \mathcal{F}^{X,Z}_{-T,t}$ and any $C \in \mathcal{F}^X_{t,T}$,

$$\begin{aligned}\int_B \mathbb{P}\left(C \,|\, \mathcal{F}^{X,Z}_{-T,t}\right) d\,\mathbb{P} &= \int_B \mathbb{P}^M\left(C \,|\, \mathcal{F}^{X,Z}_{-T,t}\right) L_t^{-1}\, d\,\mathbb{P}^M \\ &= \int_B \mathbb{P}^M(C \,|\, X_t)\, d\,\mathbb{P}.\end{aligned}$$

Thus $\mathbb{P}(C \,|\, \mathcal{F}^{X,Z}_{-T,t})$ is X_t-measurable, and this proves part (iii). Finally, let $B \in \mathcal{F}^Z_{-T,t}$, $i \in \mathbf{X}$, and $C \in \mathcal{B}^n$; then

$$\begin{aligned}\int \mathbb{P}(B \,|\, X_t, Z_t)\mathbf{1}_{\{i\}\times C}(X_t, Z_t)\, d\,\mathbb{P} &= \int_B \mathbf{E}(\mathbf{1}_{\{i\}\times C}(X_t, Z_t) \,|\, \mathcal{F}^Z_{-T,t})\, d\,\mathbb{P} \\ &= \int_B Z_{t,i}\mathbf{1}_C(Z_t)\, d\,\mathbb{P} \\ &= \int \mathbb{P}(B \,|\, Z_t)\mathbf{1}_{\{i\}\times C}(X_t, Z_t)\, d\,\mathbb{P}.\end{aligned}$$

This shows that $\mathbb{P}(B \,|\, X_t, Z_t)$ is Z_t-measurable, and this, together with part (i), proves part (iv).

Appendix B. Proof of Theorem 3.2. Under the measure $\mathbb{P}^M$ (as introduced before Proposition 2.1), X^* and Z^* are independent Markov processes with the same marginal distributions they have under $\mathbb{P}$, and so, in order to prove the statement of the theorem concerning the marginal processes X^* and Z^*, it suffices to show that the following hold:

(a) $(\Omega, \mathcal{F}, \mathbb{P}^M, X^*)$ solves the martingale problem associated with $\mathcal{L}^{X^*}_t$;

(b) $(\Omega, \mathcal{F}, \mathbb{P}^M, Z^*)$ solves the martingale problem associated with $\mathcal{L}^{Z^*}_t$.

By (a), we mean that for all $f \in C^2_b(\mathbb{R}^n; \mathbb{R})$, all $h \in C_b(\mathbb{R}^n; \mathbb{R})$, and all $s \le t$,

$$\mathbf{E}\left(\left(f(X^*_t) - f(X^*_s) - \int_s^t (\mathcal{L}^{X^*}_r f)(X^*_r)\, dr\right) h(X^*_s)\right) = 0.$$

(Since ΓX^*_r has a density, this formulation of the martingale problem admits the case that $\nabla_{\tilde{z}} p_{\tilde{Z}}(\,\cdot\,, t)$ is defined only in the sense of distributions.) Statement (a) follows from an application of Theorems 2.1 and 3.1 in [12] to the process $\tilde{Z}$ of (3.1). To prove (b), we proceed as follows.

For any $f : \mathbf{X} \to \mathbb{R}$, any $i \in \mathbf{X}$, and any $s \le t$, let

$$\begin{aligned}\pi(i, t) &:= \sum_j f(j) \exp(A(t-s))_{i,j} p_X(s)_j, \\ \phi(i, t) &:= \pi(i, -t)/p_X(-t)_i.\end{aligned}$$

Then $\mathbf{E}(f(X_s) \mid X_t = i) = \phi(i, -t)$, and

$$\begin{aligned}\frac{\partial \phi}{\partial t}(i, t) &= -(p_X(-t)_i)^{-1} \frac{\partial \pi}{\partial t}(i, -t) + (p_X(-t)_i)^{-2} \pi(i, -t) \frac{\partial (p_X)_i}{\partial t}(-t), \\ &= -\sum_j \bar{A}(t)_{j,i}(\phi(j, t) - \phi(i, t)),\end{aligned}$$

and so Z^* is a Markov jump process on $\mathbf{X}$ with time-dependent rate matrix $\bar{A}$, and this proves (b).

In order to prove the statement of the theorem concerning the joint process (X^*, Z^*), it now suffices to establish that, under $\mathbb{P}$, Z^* is X^*-conditionally Markov with rate matrix $\breve{A}$ of (3.4). For some $t \in [-T, T]$, let $B \in \mathcal{F}^X_{t,T}$, $C \in \mathcal{F}^{X,Z}_{-T,t}$, and $D \in \mathcal{F}^Z_{t,T}$; then

$$\begin{aligned}
\int_{C\cap D} \mathbb{P}(B \mid \mathcal{F}^X_{-T,t} \vee \mathcal{F}^Z_{-T,T})\, d\mathbb{P} &= \int_C \mathbb{P}(B \cap D \mid \mathcal{F}^{X,Z}_{-T,t})\, d\mathbb{P} \\
&= \int_C \mathbb{P}(B \cap D \mid X_t, Z_t)\, d\mathbb{P} \\
&= \int_C \mathbf{E}\left(\mathbb{P}(B \mid X_t, \mathcal{F}^Z_{t,T})\mathbf{1}_D \mid X_t, Z_t\right) d\mathbb{P} \\
&= \int_C \mathbf{E}\left(\mathbb{P}(B \mid X_t, \mathcal{F}^Z_{t,T})\mathbf{1}_D \mid \mathcal{F}^{X,Z}_{-T,t}\right) d\mathbb{P} \\
&= \int_{C\cap D} \mathbb{P}(B \mid X_t, \mathcal{F}^Z_{t,T})\, d\mathbb{P},
\end{aligned}$$

where we have used the joint Markov property in the second and fourth steps. Since $(C \cap D)$ forms a Dynkin π-system that generates $\mathcal{F}^X_{-T,t} \vee \mathcal{F}^Z_{-T,T}$, the first and last integrands are equal a.s., and X is Z-conditionally Markov, so that Z^* is X^*-conditionally Markov.

For any $f : \mathbf{X} \to \mathbb{R}$, any $j \in \mathbf{X}$, and any $s \le t$, let

$$\begin{aligned}
\alpha^{s,j}_t &= \mathrm{vec}_i \left\{\mathbb{P}(X_t = i \mid X_s = j,\, \mathcal{F}^Y_{-T,t})\right\}, \\
\beta^s_t &= \mathrm{vec}_i \left\{\mathbb{P}(X_s = i \mid \mathcal{F}^Y_{-T,t})\right\}, \\
\Phi(i, -t) &= \mathbf{E}(f(X_s) \mid X_t = i,\, Z) \\
&= \sum_j f(j)\alpha^{s,j}_{t,i}\beta^s_{t,j}Z^{-1}_{t,i}.
\end{aligned}$$

Then it follows from Theorems 9.3 and 9.4 in [14], and Itô's rule, that

$$\frac{\partial}{\partial t}\alpha^{s,j}_{t,i}\beta^s_{t,j}Z^{-1}_{t,i} = \beta^s_{t,j}Z^{-2}_{t,i}(Z_{t,i}(A\alpha^{s,j}_t)_i - \alpha^{s,j}_{t,i}(AZ_t)_i),$$

and so

$$\begin{aligned}
\left(\frac{\partial \Phi}{\partial t}\right)(i, -t) &= \sum_{j,k} f(j)\beta^s_{t,j}Z^{-2}_{t,i}\left(Z_{t,i}A_{i,k}\alpha^{s,j}_{t,k} - \alpha^{s,j}_{t,i}A_{i,k}Z_{t,k}\right) \\
&= -\sum_k A_{i,k}\frac{Z_{t,k}}{Z_{t,i}}\sum_i f(j)\beta^s_{t,j}\left(\frac{\alpha^{s,j}_{t,i}}{Z_{t,i}} - \frac{\alpha^{s,j}_{t,k}}{Z_{t,k}}\right) \\
&= -\sum_k \breve{A}(Z_t)_{k,i}\left(\Phi(k, -t) - \Phi(i, -t)\right).
\end{aligned}$$

Appendix C. Proof of Proposition 4.2. We prove part (iii) only. Parts (i) and (ii) follow from similar (but simpler) arguments. Throughout the proof, X^t, Z^t, X^{*t}, and Z^{*t} will be used as shorthand for the processes $(X_{t+s},\ s \in [0, \epsilon])$, $(Z_{t+s},\ s \in [0, \epsilon])$, $(X^*_{-t+s},\ s \in [0, \epsilon])$, and $(Z^*_{-t+s},\ s \in [0, \epsilon])$. The proof involves the

construction of Radon–Nikodym derivatives between various regular X_t-, Z_t-, and (X_t, Z_t)-conditional distributions of these processes. We shall, therefore, use versions of X^t and Z^{*t} whose sample paths belong to the space, $D([0,\epsilon];\mathbf{X})$, of right-continuous functions having left limits. When equipped with the Skorohod metric corresponding to the Krönecker metric on $\mathbf{X}$ ($r(i,j) = 1$ if $i \neq j$), this is a Polish space. (See, for example, [4], [8], and [9].) Of course, the sample paths of Z^t and X^{*t} belong to the space $C([0,\epsilon];\overline{\mathbb{S}^{n-1}})$, where $\overline{\mathbb{S}^{n-1}}$ is the closure of $\mathbb{S}^{n-1}$ in $\mathbb{R}^n$; this becomes a Polish space when "metrized" by the supremum norm. For a process $(\eta_s, s \in [0,\epsilon])$ having sample paths in a Polish space, $\Pi^{\eta}_{\epsilon|0}(\,\cdot\,,\theta)$ will signify a regular $(\eta_0 = \theta)$-conditional distribution of η.

Let $G_{i,t} := g(i) - \bar{g}(Z_t)$; then it follows from Itô's rule that

$$dZ^{-1}_{t,i} = \left(|G_{i,t}|^2 - A_{i,i}\right) Z^{-1}_{t,i}\,dt - Z^{-1}_{t,i} G'_{i,t}\,d\nu_t - Z^{-2}_{t,i} \sum_{j\neq i} A_{i,j} Z_{t,j}\,dt.$$

Let $(\Theta_t,\ t \in [-T,T])$ be defined by the following equation:

$$\Theta_t = Z^{-1}_{-T,i} + \int_{-T}^{t} \left(|G_{i,s}|^2 - A_{i,i}\right)\Theta_s\,ds - \int_{-T}^{t} \Theta_s G'_{i,s}\,d\nu_s;$$

then

$$\begin{aligned}\Theta_t - Z^{-1}_{t,i} &= \int_{-T}^{t} \exp\left(\int_s^t \left(\left(\frac{1}{2}|G_{i,r}|^2 - A_{i,i}\right)dr - G'_{i,r}\,d\nu_r\right)\right) Z^{-2}_{s,i} \sum_{j\neq i} A_{i,j} Z_{s,j}\,ds \\ &\geq 0,\end{aligned}$$

and so $0 < Z^{-1}_{t,i} < \Theta_t$ for all t. It now follows from (H4) and a standard result on the moments of solutions of Itô equations (see, for example, Theorem 4.6 in [14]) that

$$\sup_{t\in[-T,T]} \mathbf{E} Z^{-2}_{t,i} \leq \sup_{t\in[-T,T]} \mathbf{E}\Theta^2_t < \infty. \tag{C.1}$$

Let $F : D([0,\epsilon];\mathbf{X}) \times \mathbb{S}^{n-1} \times C([0,\epsilon];\mathbb{R}^{n-1}) \to C([0,\epsilon];\mathbb{S}^{n-1})$ be the strong solution of the following equation (which is parametrized by $x \in D([0,\epsilon];\mathbf{X})$):

$$\rho_s = \rho_0 + \int_0^s b(x_r,\rho_r)\,dr + \int_0^s \sigma(\rho_r)\,dB_r \quad \text{for } s \in [0,\epsilon],$$

so that $Z^t = F(X^t, Z_t, W)$. For $x \in D([0,\epsilon];\mathbf{X})$ and $z_0 \in \mathbb{S}^{n-1}$, let $\Pi^{Z^t|X^t}_{\epsilon|0}(\,\cdot\,,x,z_0)$ be the distribution of $F(x,z_0,W)$. (This is a regular $(X^t = x,\ Z_t = z_0)$-conditional distribution for Z^t.)

For $x \in D([0,\epsilon];\mathbf{X})$, $z \in C([0,\epsilon];\mathbb{S}^{n-1})$, and $s \in [0,\epsilon]$, let

$$\begin{aligned}\alpha(x,z)_s &:= (\Gamma\sigma(z_s))^{-1}\Gamma(b(x_s,z_s) - \bar{b}(z_s,s-t)), \\ V^*_s &:= \int_0^s (\Gamma\sigma(X^{*t}_r))^{-1}\Gamma\left(dX^{*t}_r - \bar{b}(X^{*t}_r, r-t)dr\right), \\ \bar{V}(x)_s &:= -\int_0^s \alpha(x,X^{*t})_r\,dr + V^*_s, \\ M_1(x) &:= \exp\left(\int_0^\epsilon \alpha(x,X^{*t})'_s\,d\bar{V}(x)_s + \frac{1}{2}\int_0^\epsilon \left|\alpha(x,X^{*t})_s\right|^2 ds\right).\end{aligned}$$

It follows from (C.1) and (H5) that, for ϵ sufficiently small, the ordinary integral in the expression for M_1 has finite expected value. Furthermore, $X^{*t} = F(x, Z_t, \bar{V}(x))$, and so $(\alpha(x, X^{*t})_s,\ s \in [0,\epsilon])$ is adapted to $(\mathcal{F}^Z_{t,t} \vee \mathcal{F}^{\bar{V}(x)}_{0,s},\ s \in [0,\epsilon])$. Now $(V^*_s, \mathcal{F}^{X^{*t}}_{0,s},\ s \in [0,\epsilon])$ is a standard $(n-1)$-vector Brownian motion, and so it follows from Theorem 7.6 in [14] that $\bar{V}(x)$ has a regular $(Z_t = z_0)$-conditional distribution, $\mu_{\bar{V}(x)}(\,\cdot\,, z_0)$, that is absolutely continuous with respect to the Wiener measure, μ_W, and

$$\frac{d\mu_{\bar{V}(x)}(\,\cdot\,, Z_t)}{d\mu_W}(\bar{V}(x)) = M_1(x)^{-1} \quad \text{a.s.} \tag{C.2}$$

Now $F(x, z_0, V^*)$ has the distribution $\Pi^{Z^t|X^t}_{\epsilon|0}(\,\cdot\,, x, z_0)$, and so it follows from (C.1) and (H5) that, for almost all z_0 $(\Pi^Z_{t,t})$,

$$\mathbf{E}\int_0^\epsilon |\alpha(x, F(x, z_0, V^*))_s|^2\, ds < \infty$$

and from Theorem 7.7 in [14] that $\mu_W \ll \mu_{\bar{V}(x)}(\,\cdot\,, z_0)$ for almost all z_0 $(\Pi^Z_{t,t})$. Thus the measure on $\mathcal{F}$, defined by $d\,\mathbb{P}_1(x) = M_1(x) d\,\mathbb{P}$, is a probability measure under which $(\bar{V}(x)_s, \mathcal{F}^{X^{*t}}_{0,s},\ s \in [0,\epsilon])$ is an $(n-1)$-vector standard Brownian motion and X^{*t} has the Z_t-conditional distribution $\Pi^{Z^t|X^t}_{\epsilon|0}(\,\cdot\,, x, Z_t)$. So

$$\frac{d\Pi^{Z^t|X^t}_{\epsilon|0}(\,\cdot\,, x, Z_t)}{d\Pi^{X^{*t}}_{\epsilon|0}(\,\cdot\,, Z_t)}(X^{*t}) = M_1(x) \quad \text{a.s.} \tag{C.3}$$

We now apply a similar procedure to the jump processes X^t and Z^{*t}. For each $z \in C([0,\epsilon]; \mathbb{S}^{n-1})$ and $s \in (0,\epsilon]$, let

$$\begin{aligned}
\lambda_s &:= -\chi(X^t_{s-})' A \chi(X^t_{s-}),\\
\breve{\lambda}(z)_s &:= -\chi(X^t_{s-})' \breve{A}(z_s) \chi(X^t_{s-}),\\
\Phi_s &:= \begin{cases} \mathrm{vec}_i\left\{\lambda_s^{-1}(1 - \mathbf{1}_{\{i\}}(X^t_{s-}))\chi(i)' A \chi(X^t_{s-})\right\} & \text{if } \lambda_s > 0,\\ 0 & \text{otherwise,}\end{cases}\\
\breve{\Phi}(z)_s &:= \begin{cases} \mathrm{vec}_i\left\{\breve{\lambda}(z)_s^{-1}(1 - \mathbf{1}_{\{i\}}(X^t_{s-}))\chi(i)' \breve{A}(z_s) \chi(X^t_{s-})\right\} & \text{if } \breve{\lambda}(z)_s > 0,\\ 0 & \text{otherwise,}\end{cases}\\
\beta(z)_s &:= \mathrm{vec}_i\left\{\log\left(\frac{\lambda_s \Phi_{s,i}}{\breve{\lambda}(z)_s \breve{\Phi}(z)_{s,i}}\right)\right\} \quad \text{(where } 0/0 \text{ is taken to be 1)},\\
N_s &:= \mathrm{vec}_i\left\{\int_{(0,s]} (1 - \mathbf{1}_{\{i\}}(X^t_{r-}))\, d\mathbf{1}_{\{i\}}(X^t_r)\right\},\\
M_2(z) &:= \exp\left(-\int_{(0,\epsilon]} \beta(z)'_s\, dN_s + \int_0^\epsilon (\lambda_s - \breve{\lambda}(z)_s)\, ds\right).
\end{aligned}$$

In the terminology of Chapter VIII in [5], X^t is a $\mathbf{X}$-marked point process with local characteristics $(\lambda_s, \Phi_s, \mathcal{F}^{X^t}_{0,s-},\ s \in [0,\epsilon])$. Since z is continuous and $[0,\epsilon]$ is compact, there exists a $\delta > 0$ such that $\inf_{s,i}\{z_{s,i}\} > \delta$, and so

$$\sup_{s\in[0,\epsilon]} \breve{\lambda}(z)_s < \infty \quad \text{a.s.}, \tag{C.4}$$

$$\sup_{s\in[0,\epsilon]} |\beta(z)_s| \le n \sup_{s\in[0,\epsilon];\, i,j:\, A_{i,j}>0} \left| \log\left(\frac{A_{i,j}}{\breve{A}(z_s)_{i,j}} \right) \right| < \infty \quad \text{a.s.}, \tag{C.5}$$

and, for any $K < \infty$,

$$\mathbf{E} \exp\left(K \Sigma_i N_{\epsilon,i}\right) \le \sum_{k=0}^{\infty} \exp(Kk) \frac{(\Lambda\epsilon)^k}{k!} \exp(-\Lambda\epsilon) < \infty,$$

where $\Lambda := \max_i(-A_{i,i})$. It therefore follows from Theorem T11 in Chapter VIII of [5] that $\mathbf{E}M_2(z) = 1$ for all z, and from Theorem T10 of the same reference that the measure defined on $\mathcal{F}$ by $d\,\mathbb{P}_2(z) = M_2(z) d\,\mathbb{P}$ is a probability measure, under which X^t is an $\mathbf{X}$-marked point process with local characteristics $(\breve{\lambda}(z)_s, \breve{\Phi}(z)_s, \mathcal{F}^{X^t}_{0,s-},\, s \in [0,\epsilon])$. Under $\mathbb{P}_2(z)$, the regular $(X_t = i)$-conditional distribution of X^t is the same as the regular $(X^{*t} = z,\, X_t = i)$-conditional distribution of Z^{*t} under $\mathbb{P}$, $\Pi^{Z^{*t}|X^{*t}}_{\epsilon|0}(\,\cdot\,, z, i)$. The latter is therefore absolutely continuous with respect to $\Pi^{X^t}_{\epsilon|0}(\,\cdot\,, i)$, and

$$\frac{d\Pi^{Z^{*t}|X^{*t}}_{\epsilon|0}(\,\cdot\,, z, X_t)}{d\Pi^{X^t}_{\epsilon|0}(\,\cdot\,, X_t)}(X^t) = M_2(z) \quad \text{a.s.} \tag{C.6}$$

It now easily follows from (C.3) and (C.6) that

$$R_{X,Z}(t) = \limsup_{\epsilon\downarrow 0} \epsilon^{-1} \mathbf{E}\left(\mathbf{E}_1(x) \log(M_1(x))|_{x=X^t} - \log(M_2(Z^t)) \right) \tag{C.7}$$

and

$$R_{X^*,Z^*}(-t) = \limsup_{\epsilon\downarrow 0} \epsilon^{-1} \mathbf{E}\left(\mathbf{E}_2(z) \log(M_2(z))|_{z=X^{*t}} - \log(M_1(Z^{*t})) \right). \tag{C.8}$$

Now (C.1), (C.4), and (C.5) show that, for each $z \in C([0,\epsilon]; \mathbb{S}^{n-1})$ and each $i \in \mathbf{X}$,

$$\mathbf{E} \int_0^\epsilon \beta(Z^t)^2_{s,i} \lambda_s \Phi_{s,i}\, ds < \infty \quad \text{and} \quad \mathbf{E}_2(z) \int_0^\epsilon \beta(z)^2_{s,i} \breve{\lambda}(z)_s \breve{\Phi}(z)_{s,i}\, ds < \infty,$$

and so, from a standard result in the L_2 theory of stochastic integration,

$$\begin{aligned} \mathbf{E} \int_{(0,\epsilon]} \beta(Z^t)'_s \left(dN_s - \lambda_s \Phi_s\, ds \right) &= \mathbf{E}_2(z) \int_{(0,\epsilon]} \beta(z)'_s \left(dN_s - \breve{\lambda}(z)_s \breve{\Phi}(z)_s\, ds \right) \\ &= 0. \end{aligned} \tag{C.9}$$

A simple calculation shows that $\mathbf{E}_2(z)(\breve{\lambda}(z) - \lambda) \equiv \mathbf{E}(\breve{\lambda}(Z^t) - \lambda) \equiv 0$, and this, together with (C.9) and (C.1), shows that

$$\begin{aligned} \lim_{\epsilon\downarrow 0} \epsilon^{-1} \mathbf{E}\mathbf{E}_2(z) \log(M_2(z))|_{z=X^{*t}} &= -\lim_{\epsilon\downarrow 0} \epsilon^{-1} \mathbf{E} \log(M_2(Z^t)) \\ &= \mathbf{E} \sum_{i,j:\, A_{i,j}>0} \log\left(\frac{A_{i,j}}{\breve{A}(Z_t)_{i,j}} \right) A_{i,j} Z_{t,j}. \end{aligned} \tag{C.10}$$

Similarly, it follows from (H5) and (C.1) that

$$\begin{aligned}\lim_{\epsilon\downarrow 0}\epsilon^{-1}\mathbf{E}\mathbf{E}_1(x)\log(M_1(x))|_{x=X^t} &= -\lim_{\epsilon\downarrow 0}\epsilon^{-1}\mathbf{E}\log(M_1(Z^{*t}))\\ &= \frac{1}{2}\mathbf{E}\left|(\Gamma\sigma(Z_t))^{-1}\Gamma(b(X_t,Z_t)-\bar{b}(Z_t,-t))\right|^2.\end{aligned}\tag{C.11}$$

Equations (C.7), (C.8), (C.10), and (C.11) now prove part (iii).

Appendix D. Proof of Proposition 4.4. As in the proof of Proposition 4.2, we use versions of the jump processes X^k and X whose sample paths belong to the Skorohod space $D([-T,T];\mathbf{X})$. Throughout the proof the superscript 0 will be used to indicate a parameter or process of the original (unrelaxed) filtering problem.

For each $k=0,1,2,\ldots$, let $(T^k(t),\ t\in[-T,T])$ be the transition semigroup of the process $((X_t^k,Z_{-T}^k),\ t\in[-T,T])$, so that $(T^k(t)f)(i,z)=\Sigma_j\exp(A^kt)_{j,i}f(j,z)$. The T^k are all Feller semigroups, and $T^k(t)f\to T^0(t)f$ for all $t\in[-T,T]$ and $f\in C(\mathbf{X}\times\mathbb{S}^{n-1};\mathbb{R})$. It thus follows from Theorem 2.5 (p. 167) in [9] that $(X^k,Z_{-T}^k)\Rightarrow(X^0,Z_{-T}^0)$, and from the Skorohod representation theorem that there exists a sequence $((\hat{X}^k,\hat{Z}_{-T}^k),\ k=0,1,2,\ldots)$, defined on a common probability space $(\hat{\Omega},\hat{\mathcal{F}},\hat{\mathbb{P}})$, such that $(\hat{X}^k,\hat{Z}_{-T}^k)$ has the same distribution as (X^k,Z_{-T}^k) for all k, and

$$(\hat{X}^k,\hat{Z}_{-T}^k)\to(\hat{X}^0,\hat{Z}_{-T}^0)\quad\text{pointwise on }\hat{\Omega}.\tag{D.1}$$

Let W^1 be the m-vector shifted Brownian motion of the relaxed problem Φ^1, and, for each k, let $(\hat{Z}_t^k,\ t\in[-T,T])$ be the solution on Ω^1 of the following equation (which is parametrized by $\hat{\omega}$):

$$\hat{Z}_t^k=\hat{Z}_{-T}^k+\int_{-T}^t b^k(\hat{X}_s^k,\hat{Z}_s^k)\,ds+\int_{-T}^t\sigma^k(\hat{Z}_s^k)\,dW_s^1,$$

where b^k and σ^k are the equivalents of b and σ, as defined in (2.5) and (2.10), for the problem Φ^k. (Note that $b^0=b$, and σ^0 is defined in terms of g^0 of (A4).) Jensen's inequality shows that

$$\begin{aligned}\left|\hat{Z}_t^k-\hat{Z}_t^0\right|^2\le 3\left|\hat{Z}_{-T}^k-\hat{Z}_{-T}^0\right|^2+6T\int_{-T}^t\left|b^k(\hat{X}_s^k,\hat{Z}_s^k)-b^0(\hat{X}_s^0,\hat{Z}_s^0)\right|^2ds\\ +3\left|\int_{-T}^t\left(\sigma^k(\hat{Z}_s^k)-\sigma^0(\hat{Z}_s^0)\right)dW_s^1\right|^2,\end{aligned}$$

and Doob's submartingale inequality shows that

$$\mathbf{E}^1\sup_{s\in[-T,t]}\left|\int_{-T}^s\left(\sigma^k(\hat{Z}_r^k)-\sigma^0(\hat{Z}_r^0)\right)dW_r^1\right|^2\le 4\mathbf{E}^1\int_{-T}^t\left\|\sigma^k(\hat{Z}_s^k)-\sigma^0(\hat{Z}_s^0)\right\|^2ds.$$

Now $b^k(i,\cdot)$ and σ^k are Lipschitz continuous on $\mathbb{S}^{n-1}$, uniformly in (i,k), and so, for some $K<\infty$ not depending on k,

$$\mathbf{E}^1\sup_{s\in[-T,t]}\left|\hat{Z}_s^k-\hat{Z}_s^0\right|^2\le R^k+\int_{-T}^t K\mathbf{E}^1\sup_{-T\le r\le s}\left|\hat{Z}_r^k-\hat{Z}_r^0\right|^2dr,$$

where

$$
\begin{aligned}
R^k := 3\left|\hat{Z}^k_{-T} - \hat{Z}^0_{-T}\right|^2 &+ 18T\int_{-T}^{T}\mathbf{E}^1\left|b^k(\hat{X}^k_t,\hat{Z}^k_t) - b^k(\hat{X}^0_t,\hat{Z}^0_t)\right|^2 dt \\
&+ 18T\int_{-T}^{T}\mathbf{E}^1\left|b^k(\hat{X}^0_t,\hat{Z}^0_t) - b^0(\hat{X}^0_t,\hat{Z}^0_t)\right|^2 dt \\
&+ 24T\int_{-T}^{T}\mathbf{E}^1\left\|\sigma^k(\hat{Z}^0_t) - \sigma^0(\hat{Z}^0_t)\right\|^2 dt.
\end{aligned}
\tag{D.2}
$$

Gronwall's lemma now shows that

$$
\mathbf{E}^1\sup_{t\in[-T,T]}\left|\hat{Z}^k_t - \hat{Z}^0_t\right|^2 \le \exp(2KT)R^k. \tag{D.3}
$$

Let $N(\hat{\omega})$ be the number of jumps made by $\hat{X}^0(\hat{\omega})$ in the time interval $[-T,T]$, and suppose that k is sufficiently large that $\|\hat{X}^k(\hat{\omega}) - \hat{X}^0(\hat{\omega})\|_D < \epsilon < 1$, where $\|\cdot\|_D$ is the Skorohod norm; then the Lebesgue measure of the subset of $[-T,T]$ on which $\hat{X}^k(\hat{\omega})$ and $\hat{X}^0(\hat{\omega})$ differ is upper bounded by $N(\hat{\omega})\epsilon$. Thus, since b^k is bounded on $\mathbf{X}\times\mathbb{S}^{n-1}$, uniformly in k, the first integral in (D.2) converges to zero. Since $(b^k,\sigma^k)\to(b^0,\sigma^0)$ uniformly on $\mathbf{X}\times\mathbb{S}^{n-1}$, the second and third integrals in (D.2) also converge to zero. It thus follows from (D.3) that

$$
\mathbf{E}^1\|\hat{Z}^k - \hat{Z}^0\|^2_C \to 0 \quad \text{for all } \hat{\omega}\in\hat{\Omega}, \tag{D.4}
$$

where $\|\cdot\|_C$ is the supremum norm on $C([-T,T];\overline{\mathbb{S}^{n-1}})$. Equations (D.1) and (D.4) now readily show that $(\hat{X}^k,\hat{Z}^k)\to(\hat{X}^0,\hat{Z}^0)$ in probability $(\hat{\mathbb{P}}\otimes\mathbb{P}^1)$, and this proves part (i).

Let f and h be as in part (ii) of the proposition. Then

$$
\begin{aligned}
&\lim_{\epsilon\downarrow 0}\epsilon^{-1}\mathbf{E}\left(f(X_{t-\epsilon},Z_{t-\epsilon}) - f(X_t,Z_t)\right)h(X_t,Z_t) \\
&\quad= \lim_{\epsilon\downarrow 0}\epsilon^{-1}\mathbf{E}\int_{t-\epsilon}^{t}\left(f(X_{t-\epsilon},Z_{t-\epsilon})(\mathcal{L}^{X,Z}h)(X_s,Z_s) - (\mathcal{L}^{X,Z}fh)(X_s,Z_s)\right)ds \\
&\quad\to \mathbf{E}\left(f\mathcal{L}^{X,Z}h - \mathcal{L}^{X,Z}fh\right)(X_t,Z_t),
\end{aligned}
$$

where $\mathcal{L}^{X,Z}$ is the differential generator of (X,Z), defined in (2.9). A similar expression holds with X and Z replaced by X^k and Z^k and $\mathcal{L}^{X,Z}$ replaced by the differential generator of (X^k,Z^k), $\mathcal{L}^{X,Z,k}$. Equation (4.13) with $\pm = -$ now follows from part (i) and the fact that the functions $f\mathcal{L}^{X,Z,k}h - \mathcal{L}^{X,Z,k}fh$ are continuous and converge to $f\mathcal{L}^{X,Z}h - \mathcal{L}^{X,Z}(fh)$ uniformly on $\mathbf{X}\times\mathbb{S}^{n-1}$. Equation (4.13) with $\pm = +$ follows from a similar argument.

The inner lim sup on the right-hand side of (4.12) is the rate of interactive entropy production of the relaxed problem Φ^k, which, according to (4.9), (2.13), and (2.16), takes the value

$$
\dot{S}^k(t) + \dot{D}^k(t) = \frac{1}{2}\mathbf{E}^k\left|g^k(X^k_t) - \bar{g}^k(Z^k_t)\right|^2 - \mathbf{E}^k\sum_{i,j}\log\left(\frac{Z^k_{t,i}}{p^k_X(t)_i}\right)A^k_{i,j}Z^k_{t,j}.
$$

Part (iii) now follows from part (i) and (A3).

Acknowledgments. The author would like to thank Professor Sanjoy Mitter of the Laboratory for Information and Decision Systems (LIDS) at MIT for suggesting this subject area as one fruitful for research, for providing encouragement and support during a number of visits made by the author to LIDS, and for the many engaging and productive discussions that took place during those visits. He would also like to thank one of the anonymous referees for detailed comments that have led to a substantial improvement in the presentation of this material.

REFERENCES

[1] V. Bally, *Lower bounds for the density of locally elliptic Itô processes*, Ann. Probab., 34 (2006), pp. 2406–2440.
[2] V. E. Beneš, *Exact finite-dimensional filters for certain diffusions with nonlinear drift*, Stochastics, 5 (1981), pp. 65–92.
[3] L. Bertini, A. De Sole, D. Gabrielli, G. Jona-Lasinio, and C. Landim, *Macroscopic fluctuation theory for stationary non-equilibrium states*, J. Statist. Phys., 107 (2002), pp. 635–675.
[4] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1999.
[5] P. Brémaud, *Point Processes and Queues. Martingale Dynamics*, Springer, New York, 1981.
[6] M. Chaleyat-Maurel and D. Michel, *Des résultats de non existence de filtre de dimension finie*, Stochastics, 13 (1984), pp. 83–102.
[7] T. E. Duncan, *On the calculation of mutual information*, SIAM J. Appl. Math., 19 (1970), pp. 215–220.
[8] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, New York, 1997.
[9] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterisation and Convergence*, Wiley, New York, 1986.
[10] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.
[11] B. Gaveau and L. S. Schulman, *Creation, dissipation, and recycling of resources in non-equilibrium systems*, J. Statist. Phys., 110 (2003), pp. 1317–1367.
[12] U. G. Haussmann and E. Pardoux, *Time reversal of diffusions*, Ann. Probab., 14 (1986), pp. 1188–1205.
[13] J. L. Lebowitz and H. Spohn, *A Gallavotti–Cohen-type symmetry in the large deviation functional for stochastic dynamics*, J. Statist. Phys., 95 (1999), pp. 333–365.
[14] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes.* I. *General Theory*, Springer, Berlin, 1977.
[15] C. Maes, *The fluctuation theorem as a Gibbs property*, J. Statist. Phys., 95 (1999), pp. 367–392.
[16] E. Mayer-Wolf and M. Zakai, *On a formula relating the Shannon information to the Fisher information for the filtering problem*, in Filtering and Control of Random Processes, H. Korezlioglu, G. Mazziotto, and S. Szpirglas, eds., Lecture Notes in Control and Inform. Sci. 61, Springer, Berlin, 1984, pp. 164–171.
[17] S. K. Mitter and N. J. Newton, *A variational approach to nonlinear estimation*, SIAM J. Control Optim., 42 (2003), pp. 1813–1833.
[18] S. K. Mitter and N. J. Newton, *Information and entropy flow in the Kalman–Bucy filter*, J. Statist. Phys., 118 (2005), pp. 145–176.
[19] N. J. Newton, *Dual Kalman–Bucy filters and interactive entropy production*, SIAM J. Control Optim., 45 (2006), pp. 998–1016.
[20] N. J. Newton, *The interactive statistical mechanics of nonlinear filters*, J. Statist. Phys., submitted.
[21] W. M. Wonham, *Some applications of stochastic differential equations to optimal nonlinear filtering*, J. Soc. Indust. Appl. Math. Ser. A Control, 2 (1965), pp. 347–369.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1664–1682

OPTIMAL DESIGN OF THIN PLATES BY A DIMENSION REDUCTION FOR LINEAR CONSTRAINED PROBLEMS*

GUY BOUCHITTÉ† AND ILARIA FRAGALÀ‡

Abstract. The goal of this paper is to give a rigorous justification for the Hessian-constrained problems introduced in [G. Bouchitté and I. Fragalà, *Arch. Ration. Mech. Anal.*, 184 (2007), pp. 257–284] and to show how they are linked to the optimal design of thin plates. To that aim, we study the asymptotic behavior of a sequence of optimal elastic compliance problems in the double limit when both the maximal height of the design region and the total volume of the material tend to zero. In the vanishing volume limit, a sequence of linear constrained first order vector problems is obtained, which in turn—in the vanishing thickness limit—produces a new linear constrained problem where both first and second order gradients appear. When the load is orthogonal to the plate, only the Hessian constraint is active, and we recover as a particular case the optimization problem studied in [G. Bouchitté and I. Fragalà, *Arch. Ration. Mech. Anal.*, 184 (2007), pp. 257–284] (see also [T. Lewinski and J. J. Telega, *Arch. Mech. (Arch. Mech. Stos.)*, 53 (2001), pp. 457–485]).

Key words. thin plates, optimization, compliance, linear constrained problems, positive measures, Γ-convergence

AMS subject classifications. 49J45, 74K20, 28A25

DOI. 10.1137/060671474

1. Introduction.

Let Ω be an open bounded connected subset of $\mathbb{R}^2$ with a smooth boundary. In [14] we considered the following *mass optimization problem*, which consists in finding the optimal distribution of a given amount of plate-like material in the design region $\overline{\Omega}$ in order to minimize the work made on it by a given system of forces:

$$\mathcal{I} = \inf\big\{\mathcal{C}^{\rm pl}(\mu, j, f) \ : \ \mu \in \mathcal{P}(\overline{\Omega})\big\} \ . \tag{1.1}$$

Here measures μ in the space $\mathcal{P}(\overline{\Omega})$ of probabilities on $\overline{\Omega}$ represent the admissible designs, which are allowed to be diffused as well as concentrated on low-dimensional sets. The cost $\mathcal{C}^{\rm pl}(\mu, j, f)$ that we want to minimize is the *plate compliance*: for any $\mu \in \mathcal{P}(\overline{\Omega})$, for a given stored energy density $j : \mathbb{R}^{2\times 2}_{\rm sym} \to \mathbb{R}$, and for a given real measure $f \in \mathcal{M}(\overline{\Omega}; \mathbb{R})$, it is obtained as

$$\mathcal{C}^{\rm pl}(\mu, j, f) := -\inf\left\{\int j(\nabla^2 u)\, d\mu - \langle f, u\rangle_{\mathbb{R}^2} \ : \ u \in \mathcal{C}^\infty(\mathbb{R}^2; \mathbb{R})\right\} \ . \tag{1.2}$$

In particular, in [14] we established the equality

$$\mathcal{I} = \mathcal{S}^2/2 \ , \tag{1.3}$$

where $\mathcal{S}$ is computed through the following *linear constrained problem*:

$$\mathcal{S} = \sup\Big\{\langle f, u\rangle_{\mathbb{R}^2} \ : \ u \in \mathcal{C}^\infty(\mathbb{R}^2; \mathbb{R}) \text{ such that } \rho(\nabla^2 u) \le 1 \text{ on } \Omega\Big\} \tag{1.4}$$

*Received by the editors October 4, 2006; accepted for publication (in revised form) May 6, 2007; published electronically November 2, 2007.
http://www.siam.org/journals/sicon/46-5/67147.html

†Laboratoire ANLA, Université de Toulon et du Var, 83957 La Garde cedex, France (bouchitte@univ-tln.fr).

‡Dipartimento di Matematica, Politecnico di Milano, Piazza L. da Vinci, 20133 Milano, Italy (ilaria.fragala@polimi.it).

(ρ being related to j by $j(z) = (1/2)\rho^2(z)$). Moreover, we proved that problems (1.1) and (1.4) share the same optimality conditions, which can be explicitly determined.

The goal of this paper is to give a rigorous justification for problems of kind (1.1) or (1.4) and to show how they are linked to the optimal design of thin plates. In fact in [14] these problems were introduced just *formally*, as the second order analogues of their corresponding first order problems. When the design region is a subset of $\mathbb{R}^3$ of the form $Q = \overline{\Omega} \times [-h, h] \subset \mathbb{R}^2 \times \mathbb{R}$, the *elastic compliance* of a mass distribution $\mu \in \mathcal{P}(Q)$ for a given density $j : \mathbb{R}^{3\times 3}_{\rm sym} \to \mathbb{R}$ and a given measure load $F \in \mathcal{M}(Q; \mathbb{R}^3)$ is given by

$$\mathcal{C}^{\rm el}(\mu, j, F) := -\inf\left\{ \int j(e(U))\, d\mu - \langle F, U\rangle_{\mathbb{R}^3} \ : \ U \in \mathcal{C}^\infty(\mathbb{R}^3; \mathbb{R}^3) \right\} , \tag{1.5}$$

where $e(U)$ denotes the symmetric gradient of U. Then the first order three-dimensional (3D) versions of (1.1) and (1.4) read, respectively,

$$\inf\{\mathcal{C}^{\rm el}(\mu, j, F) \ : \ \mu \in \mathcal{P}(Q)\}, \tag{1.6}$$

$$\sup\Big\{ \langle F, U\rangle_{\mathbb{R}^3} \ : \ U \in \mathcal{C}^\infty(\mathbb{R}^3; \mathbb{R}^3) \text{ such that } \rho(e(U)) \le 1 \text{ on } Q \Big\} . \tag{1.7}$$

These problems were studied in detail in [11]; in particular, it turns out that they are related to each other by the condition analogous to (1.3). From a mechanical point of view, they are perfectly justified: when one tries to optimize the compliance of an elastic material under a given load, in the limit of *vanishing volume* microstructures appear—meaning that the material tends to occupy low-dimensional networks—and the limit problem is of type (1.6). This is true both in the case of real materials, due to a common-use result in shape optimization, and in the case of so-called fictitious materials; see section 3.1 for more details.

We now ask,

Do problems of type (1.1) *(or equivalently* (1.4)*) admit any mechanical justification?*

Before explaining the approach we adopt in order to answer this question, let us mention that, prior to [14], problems of type (1.4) had already made their first appearance in the extensive literature on thin plates. Indeed, in the paper [26] by Lewinski and Telega, problems of type (1.4) are obtained in the vanishing volume limit starting from a two-dimensional (2D) compliance model for a plate with *constant* thickness. In the present paper, problems of type (1.1) are obtained through a new, different approach, which in particular enlightens their link with problems of type (1.6). Actually we perform a 3D–2D reduction dimension analysis for problems (1.6): we multiply the maximal height h by an infinitesimal parameter δ, and we take the design region of the form $Q_\delta = \overline{\Omega} \times [-h\delta, h\delta] \subset \mathbb{R}^2 \times \mathbb{R}$. In order to pass to the limit as $\delta \to 0$, a quite natural idea—formerly unexplored to our knowledge—is to start with the 3D vanishing volume model given by (1.7), with Q replaced by Q_δ. One might expect that such suprema remain finite as $\delta \to 0$ and that the convex set of constraints appearing in the limit problem, which is nothing more than the Kuratowski limit of the sets

$$K_\delta := \Big\{ U \in \mathcal{C}^\infty(\mathbb{R}^3; \mathbb{R}^3) \text{ such that } \rho(e(U)) \le 1 \text{ on } \Omega \times (-\delta h, \delta h) \Big\} ,$$

is given by functions whose first order gradient satisfies some suitable relation. Actually, facts come up to these expectations only in the scalar case, namely, when

functions U in K_δ take real values (see Remark 3.8). In spite of this, in the vector case when functions U in K_δ take values in $\mathbb{R}^3$, the situation is dramatically different. First, if the vertical component of the force is of order 1, the suprema in (1.7) blow up to infinity (like δ^{-1}). Then we need to rescale the third component of the force by a factor δ. After such scaling, another crucial difference with respect to the scalar case shows up when studying the Kuratowski limit of K_δ: indeed, due to the role played by a specific strain-displacement relation (of Kirchoff–Love type), two independent constraints appear, each one involving both first and second order derivatives. This analytical fact has an immediate mechanical counterpart: when the load is suitably scaled, a bending effect coupled with membrane energy appears in the limit problem, which can be written as

$$(\mathcal{P}) \qquad \sup\Big\{ \langle \overline{F}, v\rangle_{\mathbb{R}^2} \ : \ v \in \mathcal{C}^\infty(\mathbb{R}^2;\mathbb{R}^3) \text{ such that } \ \overline{\rho}(e(v_1,v_2) \pm h\nabla^2 v_3) \le 1 \text{ on } \Omega \Big\}$$

for a suitably averaged system of forces $\overline{F}$ and a suitably modified function $\overline{\rho}$ (see Theorem 3.3).

Problem $(\mathcal{P})$ reduces to a problem of type (1.4) in the particular case when the unique nonzero component of the load is the vertical one, because in such a case the double constraint imposed on fields v simplifies into one inequality for the Hessian matrix of their third component v_3.

This amounts to saying—see Corollary 3.6—that problems of type (1.1) are recovered as 3D–2D limits of problems of type (1.6) when the load is a vertical one. In particular, for such a kind of load, the optimality conditions found in [14] can be fruitfully employed in order to determine explicit solutions to problem $(\mathcal{P})$. For arbitrary loads, the optimality system has to be suitably generalized in order to cover the case of mixed regimes; see Proposition 3.10 and the examples in section 4.

To conclude this introduction, let us mention that problems of type (1.1) possibly admit further justifications coming from the same background. More precisely, consider a sequence of classical 3D-elasticity problems, where both the maximal height of the design and the total volume of the material are multiplied by infinitesimal parameters, say δ and ε, respectively. The asymptotics of such problems as $\varepsilon, \delta \to 0$ can be studied by adopting one of the two following "strategies (A) or (B)" (notice indeed that δ cannot go to zero for fixed ε). The first possible approach, which we call strategy (A), consists in passing to the limit first in ε—which as already mentioned yields problems of type (1.7)—and then in δ, ending up with problems of type $(\mathcal{P})$ (or (1.1)).

But it is tempting to also look at different ways of performing the double limit in δ and ε. One might follow the alternative strategy (B) of passing to the limit *contemporarily* in ε and δ, keeping the quotient $\eta := \delta/\varepsilon$ fixed, and eventually letting η tend to $+\infty$.

The interest in this alternative strategy (B) is twofold: investigating the commutativity of the double limit process, and linking our approach with the classical thin plates model widely studied in the literature, where a cubic dependence on the profile of the plate appears (without any attempt of being complete, let us refer the reader to [1, 5, 6, 7, 8, 9, 16, 18, 20, 22, 25, 27, 28]).

Here we limit ourselves to address strategy (B) as an interesting open field. We do not accomplish the outcoming results, which are postponed to a forthcoming work.

The paper is organized as follows. In section 2 we fix some notation and the setting of the problem; then we state our main results in section 3. Section 4 is

entirely devoted to exemplifying the application of the results obtained in section 3. Proofs are collected in section 5.

2. Preliminaries and setting of the problem. Let us take a design region in $\mathbb{R}^3$ of the form $Q = \overline{\Omega} \times [-h, h]$, where Ω is an open bounded connected subset of $\mathbb{R}^2$ and h is fixed in $\mathbb{R}^+$; the spatial variable in Q will be denoted by (x', x_3).

Consider a given amount m of elastic material placed in a subset A of the design region: thus A is subject to the constraints

$$A \subseteq Q = \overline{\Omega} \times [-h, h] \ , \qquad \mathrm{vol}(A) = m \ .$$

If the stored energy density is represented by a given integrand $j : \mathbb{R}^{3\times 3}_{\mathrm{sym}} \to \mathbb{R}$ and the material is subject to a given system of forces $F = (F_1, F_2, F_3) \in \mathcal{M}(Q; \mathbb{R}^3)$, the resulting elastic compliance is given by

$$\mathcal{C}^{\mathrm{el}}\big(A, j, F\big) := -\inf \left\{ \int_A j(e(U))\, dx - \langle F, U\rangle_{\mathbb{R}^3} \ : \ U \in \mathcal{C}^\infty(\mathbb{R}^3; \mathbb{R}^3) \right\}$$

(here and in the following, $e(U)$ denotes the symmetric part of the gradient of U).

We assume that j is convex, 2-homogeneous, and coercive, so that it can be written as

$$j(z) = \frac{1}{2}\rho^2(z) \ , \qquad \text{with } \inf_{z\neq 0} \frac{\rho(z)}{|z|} > 0 \ . \tag{2.1}$$

The typical choice of j is the usual quadratic elastic potential of the kind

$$j(z) = \frac{\lambda}{2}(\mathrm{tr}(z))^2 + \mu |z|^2 \ . \tag{2.2}$$

Moreover, for the compliance to be finite, we ask that the system of forces is *balanced*, namely,

$$\langle F, U\rangle_{\mathbb{R}^3} = 0 \qquad \text{whenever } e(U) = 0 \, , \tag{2.3}$$

and also that it belongs to the Sobolev space $H^{-1}(Q; \mathbb{R}^3)$.

We want to now consider the problem of optimizing the compliance when both the maximal height of the design and the total volume of the material become very small. In this situation the maximal height and the total volume will be multiplied by two positive vanishing parameters, say δ and ε, respectively:

$$A \subseteq Q_\delta = \overline{\Omega} \times [-\delta h, \delta h] \ , \qquad \mathrm{vol}(A) = \varepsilon m \ . \tag{2.4}$$

The same optimization problem can also be considered for "fictitious materials," that is, when the set A is replaced by a density θ satisfying

$$\theta \in L^\infty(\mathbb{R}^3; [0, 1]) \ , \qquad \mathrm{spt}(\theta) \subseteq Q_\delta \ , \qquad \int \theta\, dx = \varepsilon m \ , \tag{2.5}$$

and the definition of compliance is extended by setting

$$\mathcal{C}^{\mathrm{el}}\big(\theta, j, F\big) := -\inf \left\{ \int j(e(U))\, \theta\, dx - \langle F, U\rangle_{\mathbb{R}^3} \ : \ U \in \mathcal{C}^\infty(\mathbb{R}^3; \mathbb{R}^3) \right\} \ . \tag{2.6}$$

So we focus attention on the two variational problems

$$\inf\Big\{\mathcal{C}^{\rm el}\big(A,j,F\big)\ :\ A \text{ satisfying (2.4)}\Big\}\,, \tag{2.7}$$

$$\inf\Big\{\mathcal{C}^{\rm el}\big(\theta,j,F\big)\ :\ \theta \text{ satisfying (2.5)}\Big\}\,. \tag{2.8}$$

As announced in the introduction, in this paper we study the asymptotics of the above infima as $\varepsilon,\delta\to 0$ through strategy (A), which seems to be the simpler way leading from the infima in (2.7) or (2.8) to a problem of kind (1.1). Recall that this strategy consists in the following two steps:

- *Step* 1. Keeping δ fixed, let ε tend to zero (so that the quotient $\eta := \delta/\varepsilon$ tends to $+\infty$).
- *Step* 2. Let δ tend to zero.

The first crucial remark is that the infima in (2.7) or (2.8) blow up at each of the two steps. More precisely, if δ is fixed and ε tends to zero, the infima are of order ε^{-1}. Indeed, via the change of variables $V = U/\varepsilon$, it is easy to obtain the identity

$$\mathcal{C}^{\rm el}\left(\frac{\theta}{\varepsilon},j,F\right) = \varepsilon\mathcal{C}^{\rm el}\big(\theta,j,F\big)\ ,$$

whose left-hand side has a finite infimum for θ satisfying (2.5). Therefore, we are led to rescale the system of forces into $\sqrt{\varepsilon}F$; this will ensure that the infimum of the compliance remains finite as ε tends to zero in view of the identity

$$\mathcal{C}^{\rm el}\big(\theta,j,\sqrt{\varepsilon}F\big) = \varepsilon\mathcal{C}^{\rm el}\big(\theta,j,F\big)\ .$$

In turn, the infima obtained through the first step of strategy (A) blow up again when performing the second step, that is, when δ also tends to zero. Thus, we need to rescale the system of forces also with respect to δ. It will be more clear later on (see the proof of Theorem 3.3) that the right scaling of the load in order to keep finite the suprema in (3.3) as $\delta\to 0$ is the following one: set $Q_\delta := \overline{\Omega}\times[-\delta h,\delta h]$, and change F into the element $F^\delta\in H^{-1}(Q_\delta;\mathbb{R}^3)$ which acts on any test function $\varphi\in\mathcal{C}^\infty(\mathbb{R}^3;\mathbb{R}^3)$ as

$$\langle F^\delta,\varphi\rangle_{\mathbb{R}^3} := \sum_{i=1}^{2}\langle F_i(x),\varphi_i(x',\delta x_3)\rangle_{\mathbb{R}^3} + \delta\langle F_3(x),\varphi_3(x',\delta x_3)\rangle_{\mathbb{R}^3}\ .$$

We stress that, in the above definition, the vertical component F_3 is multiplied by δ, as is usual when dealing with plates in a flexion regime.

Summarizing, our rescaled optimization problems read

$$\mathcal{I}_{\varepsilon,\delta} := \inf\Big\{\mathcal{C}^{\rm el}\big(A,j,\sqrt{\varepsilon}F^\delta\big)\ :\ A \text{ satisfying (2.4)}\Big\}\ , \tag{2.9}$$

$$\widetilde{\mathcal{I}}_{\varepsilon,\delta} := \inf\Big\{\mathcal{C}^{\rm el}\big(\theta,j,\sqrt{\varepsilon}F^\delta\big)\ :\ \theta \text{ satisfying (2.5)}\Big\}\ . \tag{2.10}$$

Notice that, for each fixed (ε,δ), $\mathcal{I}_{\varepsilon,\delta}$ and $\widetilde{\mathcal{I}}_{\varepsilon,\delta}$ should remain finite because $\sqrt{\varepsilon}F^\delta$ is still balanced; that is, it fulfills (2.3). Further, in view of the heuristic considerations above, we expect that $\mathcal{I}_{\varepsilon,\delta}$ and $\widetilde{\mathcal{I}}_{\varepsilon,\delta}$ admit finite limits as ε and δ tend to zero. In the remainder of the paper our goal is to identify such limits.

For simplicity of notation, in what follows we take the volume parameter m appearing in (2.4) and (2.5) equal to 1 (this is not restrictive up to a multiplicative factor).

3. Main results. The two steps of strategy (A) are carried out, respectively, in subsections 3.1 and 3.2 below, and the optimality conditions for the limit problem follow in subsection 3.3. All the statements (except the one of Proposition 3.2) will be proved in section 5.

3.1. Vanishing volume limit (fixed thickness). When one performs the limit as $\varepsilon \to 0$ of $\mathcal{I}_{\varepsilon,\delta}$ or of $\widetilde{\mathcal{I}}_{\varepsilon,\delta}$, one obtains a pretty tractable limit infimum problem over the space $\mathcal{P}(Q_\delta)$ of probabilities on Q_δ. The functional to be minimized is of the kind $\mu \mapsto \mathcal{C}^{\rm el}\big(\mu, \mathcal{J}, F^\delta\big)$, where for a given integrand $\mathcal{J}$ we have set

$$\mathcal{C}^{\rm el}\big(\mu, \mathcal{J}, F^\delta\big) := -\inf\left\{ \int \mathcal{J}(e(U))\, d\mu - \langle F^\delta, U\rangle_{\mathbb{R}^3} \ : \ U \in \mathcal{C}^\infty(\mathbb{R}^3, \mathbb{R}^3) \right\} .$$

The only difference between the real and the fictitious case lies in the determination of the integrand $\mathcal{J}$: in the fictitious case one can simply take $\mathcal{J} = j$, while in the real case one has to take $\mathcal{J} = j_0$, with j_0 being deduced from j through a suitable formula. This is stated more precisely in the next two propositions.

Proposition 3.1 (fictitious materials). *There holds*

$$\lim_{\varepsilon\to 0} \widetilde{\mathcal{I}}_{\varepsilon,\delta} = \widetilde{\mathcal{I}}_\delta := \inf\Big\{ \mathcal{C}^{\rm el}\big(\mu, j, F^\delta\big) \ : \ \mu \in \mathcal{P}(Q_\delta) \Big\} .$$

Proposition 3.2 (real materials). *Assume that j is taken of the form* (2.2). *Then there holds*

$$\lim_{\varepsilon\to 0} \mathcal{I}_{\varepsilon,\delta} = \mathcal{I}_\delta := \inf\Big\{ \mathcal{C}^{\rm el}\big(\mu, j_0, F^\delta\big) \ : \ \mu \in \mathcal{P}(Q_\delta) \Big\} ,$$

where $j_0 : \mathbb{R}^{3\times 3}_{\rm sym} \to \mathbb{R}$ denotes the following modified integrand:

$$j_0(z) = \frac{1}{2}\rho_0(z)^2 := \sup\big\{ z\cdot z^* - j^*(z^*) \ : z \in \mathbb{R}^{3\times 3}_{\rm sym} \ , \ \det(z^*) = 0 \big\} . \tag{3.1}$$

Proposition 3.2 is actually a reformulation of the results in [2, 4] (to which we refer for a proof), where the effective stress potential—the Fenchel conjugate $j_0^*(z^*)$ of $j_0(z)$—is characterized explicitly in terms of the eigenvalues of the symmetric tensor z^*. Formula (3.1) is a concise way to recover directly the related effective strain potential j_0; we refer the reader to [3] for some explicit computations in case j is given by (2.2). We believe that Proposition 3.2 remains true even for nonquadratic strain potentials; see [10].

3.2. Vanishing thickness limit. The kind of mass optimization problem given by Propositions 3.1 and 3.2 has been widely studied in [11], where it is proved in particular that

$$\widetilde{\mathcal{I}}_\delta = \widetilde{\mathcal{S}}_\delta^2/2 \ , \qquad \mathcal{I}_\delta = \mathcal{S}_\delta^2/2 \ , \tag{3.2}$$

where $\widetilde{\mathcal{S}}_\delta, \mathcal{S}_\delta$ are given by the following linear constraint problems:

$$\begin{aligned} \widetilde{\mathcal{S}}_\delta &:= \sup\Big\{ \langle F^\delta, U\rangle_{\mathbb{R}^3} \ : \ U \in \mathcal{C}^\infty(\mathbb{R}^3;\mathbb{R}^3) \ \text{ such that } \ \rho(e(U)) \le 1 \text{ on } Q_\delta \Big\} , \\ \mathcal{S}_\delta &:= \sup\Big\{ \langle F^\delta, U\rangle_{\mathbb{R}^3} \ : \ U \in \mathcal{C}^\infty(\mathbb{R}^3;\mathbb{R}^3) \ \text{ such that } \ \rho_0(e(U)) \le 1 \text{ on } Q_\delta \Big\} . \end{aligned} \tag{3.3}$$

Thanks to the crucial equalities (3.2), this second step in strategy (A) is reduced to determining the limit of the sequences $\widetilde{\mathcal{S}}_\delta$ and $\mathcal{S}_\delta$ in (3.3) as $\delta \to 0$. In other words,

our goal is reached if we are able to perform the 3D–2D reduction dimension analysis for such sequences of linear constrained problems.

In order to state our main theorem, we need to introduce an effective system of forces $\overline{F} \in \mathcal{M}(\overline{\Omega};\mathbb{R}^3)$ and an effective integrand $\overline{j} : \mathbb{R}^{2\times 2}_{\rm sym} \to \mathbb{R}$.

For any $\lambda \in \mathcal{M}(Q;\mathbb{R})$, we denote by $[\lambda] \in \mathcal{M}(\overline{\Omega},\mathbb{R})$ the marginal measure defined by the equality

$$\langle [\lambda], \varphi\rangle_{\mathbb{R}^2} := \langle \lambda, \varphi\rangle_{\mathbb{R}^3} \qquad \forall \varphi \in \mathcal{C}^\infty(\mathbb{R}^2;\mathbb{R}) \ ; \tag{3.4}$$

then we define the effective system of forces $\overline{F} = (\overline{F}_1, \overline{F}_2, \overline{F}_3) \in \mathcal{M}(\overline{\Omega};\mathbb{R}^3)$ componentwise by

$$\overline{F}_i := [F_i], \qquad i = 1,2, \qquad \text{and} \qquad \overline{F}_3 := \left[F_3 + x_3 \sum_{i=1}^{2} \frac{\partial F_i}{\partial x_i} \right] \ . \tag{3.5}$$

The effective density $\overline{j} : \mathbb{R}^{2\times 2}_{\rm sym} \to \mathbb{R}$ is obtained from j through the following formula:

$$\overline{j}(z) = \frac{1}{2}\overline{\rho}(z)^2 := \inf \left\{ j\left(z + \sum_{i=1}^{3} \xi_i (e_i \otimes e_3)^* \right) \ : \ \xi_i \in \mathbb{R} \right\} \ . \tag{3.6}$$

THEOREM 3.3. *The limit as $\delta \to 0$ of both the sequences $\{\widetilde{\mathcal{S}}_\delta\}$ and $\{\mathcal{S}_\delta\}$ defined by* (3.3) *is given by*

$$\mathcal{S}_0 := \sup\Big\{ \langle \overline{F}, v\rangle_{\mathbb{R}^2} \ : \ v \in \mathcal{C}^\infty(\mathbb{R}^2;\mathbb{R}^3) \ , \ \overline{\rho}(e(v_1,v_2) \pm h\nabla^2 v_3) \le 1 \text{ on } \Omega \Big\} \ , \tag{3.7}$$

where $\overline{F}$ and $\overline{\rho}$ are given by (3.5) *and* (3.6)*, respectively.*

It is worth noticing that, once Theorem 3.3 is proved for one among the two sequences $\{\widetilde{\mathcal{S}}_\delta\}$ and $\{\mathcal{S}_\delta\}$, the statement for the other one follows immediately from the following algebraic lemma.

LEMMA 3.4. *There holds*

$$\overline{j}(z) = \overline{j_0}(z) \qquad \forall\, z \in \mathbb{R}^{2\times 2}_{\rm sym} \ .$$

Another remark about Theorem 3.3 is that, in general, the limit problem given by (3.7) cannot be "decoupled" into two separate problems, respectively, of first order in (v_1, v_2) and of second order in v_3. Nevertheless, there are special cases when it simplifies into one of them.

COROLLARY 3.5.

(i) *If $\overline{F}_1 = \overline{F}_2 = 0$, then*

$$\mathcal{S}_0 = \sup\Big\{ \langle \overline{F}_3, v_3\rangle_{\mathbb{R}^2} \ : \ v_3 \in \mathcal{C}^\infty(\mathbb{R}^2;\mathbb{R}) \text{ such that } \ \overline{\rho}(\nabla^2 v_3) \le 1/h \text{ on } \Omega \Big\} \ .$$

(ii) *If $\overline{F}_3 = 0$, then*

$$\mathcal{S}_0 = \sup\Bigg\{ \sum_{i=1}^{2} \langle \overline{F}_i, v_i\rangle_{\mathbb{R}^2} \ : \ (v_1,v_2) \in \mathcal{C}^\infty(\mathbb{R}^2;\mathbb{R}^2) \text{ such that } \overline{\rho}(e(v_1,v_2)) \le 1 \text{ on } \Omega \Bigg\} \ .$$

When $\overline{F}_1 = \overline{F}_2 = 0$, combining case (i) of the above corollary with our results in [14], we are finally able to prove that the infima in (2.10) converge to a limit problem of type (1.1).

COROLLARY 3.6. *Let $\mathcal{I}_{\varepsilon,\delta}$ and $\widetilde{\mathcal{I}}_{\varepsilon,\delta}$ be defined, respectively, by* (2.9) *and* (2.10). *If $\overline{F}_1 = \overline{F}_2 = 0$, there holds*

$$\lim_{\delta\to 0}\lim_{\varepsilon\to 0}\mathcal{I}_{\varepsilon,\delta} = \lim_{\delta\to 0}\lim_{\varepsilon\to 0}\widetilde{\mathcal{I}}_{\varepsilon,\delta} = h^{-2}\inf\left\{\mathcal{C}^{\rm pl}\big(\mu,\overline{j},\overline{F}_3\big)\ :\ \mu\in\mathcal{P}(\overline{\Omega})\right\}, \tag{3.8}$$

where the plate compliance $\mathcal{C}^{\rm pl}\big(\mu,\overline{j},\overline{F}_3\big)$ is defined according to (1.2).

Remark 3.7. Let us emphasize that the assumption $F\in H^{-1}(Q;\mathbb{R}^3)$ stated in section 2 is not needed for the well-posedness of the variational problems in (3.7) or (3.8). For instance, it is enough to ask that F is a measure with finite variation. In particular, pointwise applied forces are allowed in our limit problem.

Remark 3.8. The scalar analogue of Theorem 3.3 is simpler, and it can be easily obtained with the same proof. For any $f\in\mathcal{M}(Q;\mathbb{R})$ (with $f\in H^{-1}(Q;\mathbb{R})$ and $\int_Q f = 0$) and any convex, 1-homogeneous, coercive function $\rho:\mathbb{R}^3\to\mathbb{R}$, it can be stated as follows: the limit as $\delta\to 0$ of

$$s_\delta := \sup\Big\{\langle f^\delta, u\rangle_{\mathbb{R}^3}\ :\ u\in\mathcal{C}^\infty(\mathbb{R}^3;\mathbb{R})\ \text{ such that }\ \rho(\nabla u)\le 1 \text{ on } Q_\delta\Big\}$$

is given by

$$s_0 := \sup\Big\{\langle [f], v\rangle_{\mathbb{R}^2}\ :\ v\in\mathcal{C}^\infty(\mathbb{R}^2;\mathbb{R}) \text{ such that }\ \overline{\rho}(\nabla v)\le 1 \text{ on } \Omega\Big\}.$$

Here $f^\delta\in\mathcal{M}(Q;\mathbb{R})$ is the measure which acts on any test function $\varphi\in\mathcal{C}^\infty(\mathbb{R}^3,\mathbb{R})$ as $\langle f^\delta,\varphi\rangle_{\mathbb{R}^3} := \langle f, \varphi(x',\delta x_3)\rangle_{\mathbb{R}^3}$, while $[f]\in\mathcal{M}(\overline{\Omega};\mathbb{R})$ is defined according to (3.4), and $\overline{\rho}:\mathbb{R}^2\to\mathbb{R}$ is given by $\overline{\rho}(z) := \inf\big\{\rho(z+\xi e_3)\ :\ \xi\in\mathbb{R}\big\}$.

3.3. Dual problem and optimality conditions. For the practice computation of $\mathcal{S}_0$, one needs to determine optimality conditions for the infimum problem $(\mathcal{P})$ which defines $\mathcal{S}_0$. Such optimality conditions are obtained in [14], by exploiting the results of [15], in the special situation of Corollary 3.5(i). Let us see how they look in the more general situation of Theorem 3.3. As a preliminary step, we begin by writing the dual problem of $(\mathcal{P})$ (intended in the usual sense of convex analysis; see, e.g., [19]). We denote by $\rho^o:\mathbb{R}^{2\times 2}\to\mathbb{R}$ the *polar function of* ρ, that is,

$$\rho^o(\xi) := \sup\{\xi\cdot z\ :\ \rho(z)\le 1\},$$

where $\xi\cdot z$ indicates the Euclidean scalar product. Then, for λ in the space $\mathcal{M} = \mathcal{M}(\overline{\Omega};\mathbb{R}^{2\times 2}_{\rm sym})$ of $\mathbb{R}^{2\times 2}_{\rm sym}$-valued measures supported on $\overline{\Omega}$ with finite total variation, we use the notation $\int\rho^o(\lambda)$ in the usual sense of convex 1-homogeneous functionals on measures (see, for instance, [21]).

LEMMA 3.9. *The dual problem $(\mathcal{P}^*)$ of $(\mathcal{P})$ is given by*

$$\min\Big\{\int\overline{\rho}^o(\lambda^+)+\int\overline{\rho}^o(\lambda^-)\ :\ \lambda^\pm\in\mathcal{M}\ ,\ -\operatorname{div}(\lambda^+-\lambda^-) = (\overline{F}_1,\overline{F}_2)\ ,\\ h\operatorname{div}^2(\lambda^+-\lambda^-) = \overline{F}_3\Big\}\ ,$$

where the operators div *and* div^2 *are intended in distributional sense.*

PROPOSITION 3.10. *Let v be admissible for $(\mathcal{P})$ and $\lambda^{\pm}$ be admissible for $(\mathcal{P})^*$. They are optimal for the respective problems if and only if the following two equations are satisfied:*

$$\overline{\rho}^o(\lambda^+) = \big\langle \lambda^+,\, e(v_1,v_2) + h\nabla^2 v_3 \big\rangle_{\mathbb{R}^2}\,, \quad \overline{\rho}^o(\lambda^-) = \big\langle \lambda^-,\, e(v_1,v_2) - h\nabla^2 v_3 \big\rangle_{\mathbb{R}^2}\,. \tag{3.9}$$

The application of Proposition 3.10 is exemplified in two concrete cases in section 4.

Remark 3.11. In Lemma 3.9 and Proposition 3.10, the measures $\lambda^{\pm}$ play the role of Lagrange multipliers for the constraint $\overline{\rho}(e(v_1,v_2) \pm h\nabla^2 v_3) \le 1$ in (3.7). From the mechanical point of view, they can be interpreted as follows: for each $\delta > 0$, let σ_δ be the stress tensor in the 3D elastic problem, which after rescaling and change of variables is supported on an optimal structure located in the reference set $Q = \overline{\Omega} \times [-h,h]$. When $\delta \to 0$, σ_δ as a tensor measure concentrates on the top-bottom region $\overline{\Omega} \times \{x_3 = \pm h\}$ and produces $\lambda^{\pm}$ as limit in-plane stress tensors.

Remark 3.12. The optimality condition (3.9) can be extended to the case where v is not smooth by using suitable notions of tangential (first and second order) gradient with respect to measures for which we refer the reader to [12, 15]. Such optimal v are in fact limits of maximizing sequences for $(\mathcal{P})$ and satisfy $v_1, v_2 \in W^{1,q}(\Omega)$ for all $q < +\infty$ and $v_3 \in W^{2,\infty}(\Omega)$.

4. Examples. In the examples we are going to discuss, the systems of loads are discrete (see Remark 3.7). Moreover, they lie in a plane, so that the corresponding optimal structures are supported in that plane. As a consequence, we take a planar design region Q of the form $\overline{\Omega} \times [-h,h]$, with Ω being an open bounded interval of the real line. Thus throughout this section the spatial variable $x' \in \overline{\Omega}$ will become x_1, and the role of the "vertical variable" x_3 will be played by x_2. Clearly, the limit problem will simply reduce to a one-dimensional problem.

We take as a function ρ in (3.3) the Euclidean norm on $\mathbb{R}^{2\times 2}_{\rm sym}$. In this case it is easy to check, by using Lemma 3.4, that the corresponding functions $\overline{\rho}$ and $\overline{\rho_0}$ will simply be equal to the Euclidean norm on $\mathbb{R}$.

Example 4.1 (pure flexion regime). For fixed nonnegative parameters l and h_0, let the points O, A, B have coordinates

$$O := \big(0,0\big)\,, \qquad A := \big(l,0\big)\,, \qquad B := \big(0,h_0\big)\,,$$

and let us consider the following system of forces:

$$F_1 := \delta_O - \delta_B\,, \qquad F_2 = \frac{h_0}{l}(\delta_B - \delta_A)\,.$$

This system of forces is supported on the design region $Q = \overline{\Omega} \times [-h,h]$, provided Ω is an interval containing both O and A, and $h \ge h_0$ (see Figure 4.1). Moreover, it is immediate to check that this system is balanced. Then we can apply Theorem 3.3 to compute $\mathcal{S}_0$, namely the limit as $\delta \to 0$ of the suprema $\mathcal{S}_\delta$ or $\widetilde{\mathcal{S}}_\delta$ in (3.3). The effective system of forces on the x_1-axis is easily obtained:

$$\overline{F}_1 := 0\,, \qquad \overline{F}_2 = \frac{h_0}{l}(\delta_O - \delta_A) - h_0\delta'_O\,.$$

Then according to (3.7) $\mathcal{S}_0$ can be expressed as

$$\sup\left\{ \frac{h_0}{l}\big(v_2(O) - v_2(A)\big) + h_0 v_2'(O) \;:\; v_2 \in \mathcal{C}^\infty(\mathbb{R};\mathbb{R}) \text{ such that } |(v_2)''| \le \frac{1}{h} \text{ on } \Omega \right\}\,.$$

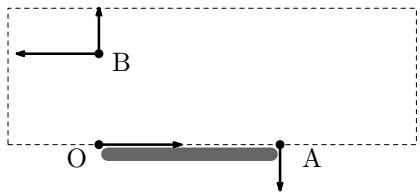


FIG. 4.1. *Loads yielding a pure flexion regime in the whole of* $\overline{OA}$.

In order to compute the explicit value of $\mathcal{S}_0$, we apply Proposition 3.10. Given $v = (v_1, v_2) \in \mathcal{C}^\infty(\mathbb{R}; \mathbb{R}^2)$ and $\lambda^\pm \in \mathcal{M}(\overline{\Omega}; \mathbb{R})$, they are solutions to problem $(\mathcal{P})$ and its dual $(\mathcal{P}^*)$ if the following system is satisfied:

$$\begin{cases} (\lambda^+ + \lambda^-)' = 0\,, \\ h(\lambda^+ - \lambda^-)'' = \dfrac{h_0}{l}(\delta_O - \delta_A) - h_0\delta'_O\,, \\ |(v_2)''| \le \dfrac{1}{h}\,, \\ |\lambda^\pm| = \langle \lambda^\pm, \pm h(v_2)'' \rangle_{\mathbb{R}}\ , \end{cases}$$

where the first two equations select admissible $\lambda^\pm$ in problem $(\mathcal{P})^*$ (see Lemma 3.9), the third equation selects admissible v in problem $(\mathcal{P})$, and the last couple of equations corresponds to the optimality conditions (3.9).

Solutions $\lambda^\pm$ to the first two equations are determined by

$$\lambda^+ = -\lambda^- = \frac{1}{2}\frac{h_0}{hl}(x_1 - l)\chi_{\overline{OA}}(x_1)\mathcal{L}^1 \llcorner \overline{OA}\ ,$$

and the remaining conditions are satisfied if we take

$$v_2(x_1) = -\frac{x_1^2}{2h}\ .$$

Thus we find for the value of the energy

$$\mathcal{S}_0 = \frac{lh_0}{2h}\ .$$

Remark 4.2. (i) Exactly the same result above holds if, in the system of forces, the point A is replaced by any other point of the type (l, h_1), with $|h_1| \le h$ (or even more generally if δ_A is replaced by any probability on the segment $l \times [-h, h]$).

(ii) Exactly the same result above holds if, with the same system of forces, the design region is changed into $\overline{\Omega} \times [0, h]$.

(iii) Note that $\mathcal{S}_0$ is infinitesimal as $h \to +\infty$, as always happens in a pure flexion regime (see Corollary 3.5(i)).

(iv) The role of $\lambda^\pm$ in the reconstruction of 3D-optimal structures will be investigated more deeply in a subsequent work. In the above example we guess that, for any $\delta > 0$, optimal structures are given by two horizontal bars at heights 0 and h, connected by some diagonal bars of vanishing mass.

Example 4.3 (mixed regime). For fixed nonnegative parameters l, h_0, α, let the points O, A, B, C have coordinates

$$O := (0,0)\ , \qquad A := \left(-\frac{l}{2}, 0\right)\ , \qquad B := \left(\frac{l}{2}, 0\right)\ , \qquad C := (0, h_0)\ ,$$

and let us consider the axially symmetric system of forces:

$$F_1 := \alpha(\delta_B - \delta_A)\ , \qquad F_2 = \delta_C - \frac{1}{2}(\delta_A + \delta_B)\ .$$

This system of forces is balanced and is supported on the design region $Q = \overline{\Omega} \times [-h, h]$, provided the interval Ω contains both A and B, and $h \geq h_0$. The effective system of forces is given on the x_1-axis by

$$\overline{F}_1 := \alpha\big(\delta_B - \delta_A\big)\ , \qquad \overline{F}_2 = \delta_O - \frac{1}{2}\big(\delta_A + \delta_B\big)\ .$$

Then according to (3.7) the limit $\mathcal{S}_0$ of the suprema $\mathcal{S}_\delta$ or $\widetilde{\mathcal{S}}_\delta$ in (3.3) can be expressed as

$$\sup\Big\{\alpha\big[v_1(B) - v_1(A)\big] + v_2(O) - \frac{1}{2}\big[v_2(A) + v_2(B)\big]\ :\ v \in \mathcal{C}^\infty(\mathbb{R};\mathbb{R}^2) \text{ such that } |(v_1)' \pm h(v_2)''| \leq 1 \text{ on } \Omega\Big\}\ .$$

Let us compute the explicit value of $\mathcal{S}_0$ in terms of the involved parameters.

By Proposition 3.10, given $v = (v_1, v_2) \in \mathcal{C}^\infty(\mathbb{R}^2;\mathbb{R}^2)$ and $\lambda^\pm \in \mathcal{M}(\overline{\Omega};\mathbb{R})$, they are solutions to problem $(\mathcal{P})$ and its dual $(\mathcal{P}^*)$ if the following system is satisfied:

$$\begin{cases} -(\lambda^+ + \lambda^-)' = \alpha\big(\delta_A - \delta_B\big)\,, \\ h(\lambda^+ - \lambda^-)'' = \delta_O - \dfrac{1}{2}\big(\delta_A + \delta_B\big)\,, \\ |(v_1)' \pm h(v_2)''| \leq 1\,, \\ |\lambda^\pm| = \langle \lambda^\pm, (v_1)' \pm h(v_2)''\rangle_{\mathbb{R}}\ . \end{cases}$$

Solutions $\lambda^\pm$ to the first two equations are determined by

$$\lambda^+ + \lambda^- = \alpha\mathcal{L}^1 \llcorner \overline{AB}\ , \qquad \lambda^+ - \lambda^- = \frac{1}{2h}\left(|x_1| - \frac{l}{2}\right)\mathcal{L}^1 \llcorner \overline{AB}\ , \tag{4.1}$$

and the remaining conditions are satisfied, provided

$$(v_1)' \pm h(v_2)'' = \operatorname{sign}(\lambda^\pm)\ , \tag{4.2}$$

where $\operatorname{sign}(\lambda^\pm)$ denotes the sign of (the densities of) $\lambda^\pm$.

From (4.1), we see in particular that λ^- remains always nonnegative, whereas for λ^+ two cases may occur:

case (1): if $h \geq l/(4\alpha)$, then λ^+ remains nonnegative;

case (2): if $h < l/(4\alpha)$, then

$$\begin{cases} \lambda^+ \geq 0 & \text{if } |x_1| \geq (l/2) - 2h\alpha\,, \\ \lambda^+ < 0 & \text{if } |x_1| < (l/2) - 2h\alpha\ . \end{cases}$$

Accordingly, solutions to (4.2) and the value of $\mathcal{S}_0$ can be easily computed:

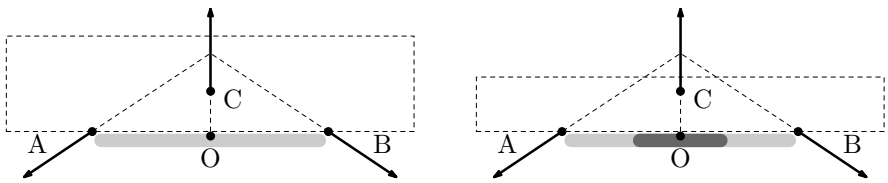


FIG. 4.2. *Loads yielding a membrane/flexion regime in the clear/dark part of* $\overline{AB}$.

case (1): we have $(v_1)' = 1$, $(v_2)'' = 0$ (see the left side of Figure 4.2), and

$$\mathcal{S}_0 = \int \lambda^+ + \int \lambda^- = \alpha l \ ;$$

case (2): we have (see the right side of Figure 4.2)

$$\begin{cases} (v_1)' = 1 \text{ and } (v_2)'' = 0 & \text{if } |x_1| \geq (l/2) - 2h\alpha \,, \\ (v_1)' = 0 \text{ and } (v_2)'' = 1/h & \text{if } |x_1| < (l/2) - 2h\alpha \ , \end{cases}$$

and

$$\mathcal{S}_0 = \int |\lambda^+| + \int \lambda^- = 2h\big[\alpha^2 + l^2/(16h^2)\big] \ .$$

Summing up, we have obtained

$$\mathcal{S}_0 = \begin{cases} \alpha l & \text{if } h \geq l/(4\alpha) \,, \\ 2h\big[\alpha^2 + l^2/(16h^2)\big] & \text{if } h < l/(4\alpha) \ . \end{cases}$$

Remark 4.4. (i) The value found above for $\mathcal{S}_0$ is always independent of the parameter h_0.

(ii) The critical height $h_c := l/(4\alpha)$ is the second coordinate of the intersection point between the straight lines $A + t(-\alpha, -1/2)$ and $B + t(\alpha, -1/2)$ (namely the point where the two forces $(-\alpha, -1/2)\delta_A$ and $(\alpha, -1/2)\delta_B$ concur). If $h \geq h_c$, then the value of $\mathcal{S}_0$ is independent of h. In spite of this, if $h < h_c$, then the dependence of $\mathcal{S}_0$ on h tells that optimal structures for $\mathcal{S}_\delta$ or $\widetilde{\mathcal{S}}_\delta$ do "touch" the bottom of the design region (independently of the choice of h_0).

5. Proofs of the results in section 3.

Proof of Proposition 3.1. Let δ be fixed. We introduce, for every ε, the functional J_ε and the function φ_ε defined, respectively, on $\mathcal{M}(Q_\delta; \mathbb{R}^+)$ and on $\mathbb{R}$ by

$$J_\varepsilon(\mu) := \begin{cases} \mathcal{C}^{\rm el}\big(\mu, j, F^\delta\big) & \text{if } \mu = \theta\, dx \ , \ \theta \in L^\infty(\mathbb{R}^3; [0, \varepsilon^{-1}]) \ , \ \mathrm{spt}(\theta) \subseteq Q_\delta \,, \\ +\infty & \text{otherwise,} \end{cases}$$

$$\varphi_\varepsilon(t) := \begin{cases} \inf\left\{ J_\varepsilon(\mu) \ : \ \mu \in \mathcal{M}(Q_\delta; \mathbb{R}^+) \ , \ \displaystyle\int d\mu = t \right\} & \text{if } 0 < t \leq \varepsilon^{-1} |Q_\delta| \,, \\ +\infty & \text{otherwise.} \end{cases}$$

It is easy to check that J_ε and φ_ε are convex and decrease as ε goes down to zero. In particular, the limit $\varphi_0(t) = \lim_{\varepsilon \to 0} \varphi_\varepsilon(t)$ exists and is convex as a function of t. We claim that, for every $t > 0$, there holds

$$\varphi_0(t) = \frac{(\widetilde{\mathcal{S}}_\delta)^2}{2\, t} \,. \tag{5.1}$$

Recalling (3.2), the proposition will follow by taking $t = 1$, since by (2.10) and (2.6)

$$\widetilde{\mathcal{I}}_{\varepsilon,\delta} = \inf\left\{\mathcal{C}^{\rm el}\left(\frac{\theta}{\varepsilon}, j, F^\delta\right) \ : \ \theta \text{ satisfying (2.5)}\right\} \ = \ \varphi_\varepsilon(1) \ .$$

For proving (5.1), we are going to identify the Fenchel conjugate of φ_0 through the formula

$$\varphi_0^* = \Big(\inf_\varepsilon \varphi_\varepsilon\Big)^* = \sup_\varepsilon \varphi_\varepsilon^* \ . \tag{5.2}$$

To compute φ_ε^*, we begin by noticing that for every $k \in \mathbb{R}$, φ_ε^* computed at $-k$ coincides with the Fenchel conjugate of J_ε computed at the constant function identically equal to $-k$. Indeed,

$$\varphi_\varepsilon^*(-k) = \sup\left\{-\int k\, d\mu - J_\varepsilon(\mu) \ : \ \mu \in \mathcal{M}(Q_\delta;\mathbb{R}^+) \ , \ \int d\mu = t\right\} = J_\varepsilon^*(-k) \ . \tag{5.3}$$

Let us compute $J_\varepsilon^*(-k)$. By definition we have

$$J_\varepsilon^*(-k) = \sup_\mu \inf_U \left\{\int j\big(e(U) - k\big)\ d\mu - \langle F^\delta, U\rangle_{\mathbb{R}^3}\right\} \ ,$$

where the infimum in U is taken over $\mathcal{C}^\infty(\mathbb{R}^3;\mathbb{R}^3)$, while the supremum in μ is taken over the class of measures of the form $\mu = \theta\, dx$ with $\theta \in L^\infty(\mathbb{R}^3;[0,\varepsilon^{-1}])$ and $\mathrm{spt}(\theta) \subseteq Q_\delta$. Since the latter class is compact and since the dependence with respect to (μ, U) is convex-concave, we may exchange the supremum and the infimum (see, e.g., [14, Proposition 2.2]) so that

$$\begin{aligned} J_\varepsilon^*(-k) \ &= \inf_U \left\{-\langle F^\delta, U\rangle_{\mathbb{R}^3} + \sup_\mu \int \big(j(e(U)) - k\big)\ d\mu\right\} \\ &= \inf_U \left\{-\langle F^\delta, U\rangle_{\mathbb{R}^3} + \varepsilon^{-1}\int_{Q_\delta} \big(j(e(U)) - k\big)^+\ dx\right\} \ . \end{aligned}$$

Then, in order to compute the limit as $\varepsilon \to 0$ of $J_\varepsilon^*(-k)$ (which is also their supremum), we are led to consider the functionals G_ε defined on $H^1(\mathbb{R}^3;\mathbb{R}^3)$ by

$$G_\varepsilon(U) := \begin{cases} -\langle F^\delta, U\rangle_{\mathbb{R}^3} + \varepsilon^{-1}\displaystyle\int_{Q_\delta} \big(j(e(U)) - k\big)^+\ dx & \text{if } U \in \mathcal{C}^\infty(\mathbb{R}^3;\mathbb{R}^3)\,, \\ +\infty & \text{otherwise.} \end{cases}$$

It is easy to check that, since $F^\delta \in H^{-1}(Q_\delta;\mathbb{R}^3)$, the functionals G_ε are lower semicontinuous with respect to the weak topology on $H^1(\mathbb{R}^3;\mathbb{R}^3)$. Therefore, since the sequence G_ε is monotone increasing in ε, its Γ-limit with respect to the weak convergence on coincides with the functional G_0 defined by

$$G_0(U) := \begin{cases} -\langle F^\delta, U\rangle_{\mathbb{R}^3} & \text{if } U \in H^1(\mathbb{R}^3;\mathbb{R}^3) \text{ such that } j(e(U)) \le k \text{ a.e. on } Q_\delta\,, \\ +\infty & \text{otherwise} \end{cases}$$

(in particular, $G_0 \equiv +\infty$ for $k < 0$). Moreover, by using the coercivity of j and the Korn inequality, one can easily check that any sequence $\{U^\varepsilon\}$ with $\sup_\varepsilon G_\varepsilon(U^\varepsilon) < +\infty$ is weakly precompact in $H^1(\mathbb{R}^3;\mathbb{R}^3)$ (up to subtracting a rigid displacement, which

is not restrictive thanks to (2.3)). This compactness property, combined with the Γ-convergence of G_ε to G_0, ensures that the infima of G_ε converge to the infimum of G_0. Therefore

$$\begin{aligned}
&-\lim_\varepsilon J_\varepsilon^*(-k)\\
&= -\inf\Big\{-\langle F^\delta, U\rangle_{\mathbb{R}^3} \ : \ U\in H^1(\mathbb{R}^3;\mathbb{R}^3) \text{ such that } j(e(U))\le k \text{ a.e. on } Q_\delta\Big\}\\
&= \sup\Big\{\langle F^\delta, U\rangle_{\mathbb{R}^3} \ : \ U\in \mathcal{C}^\infty(\mathbb{R}^3;\mathbb{R}^3) \text{ such that } j(e(U))\le k \text{ on } Q_\delta\Big\}\ .
\end{aligned}$$

Recalling the definition of $\widetilde{\mathcal{S}}_\delta$ in (3.3) and by the 2-homogeneity of j (see (2.1)), we deduce after an easy computation that

$$-\lim_\varepsilon J_\varepsilon^*(-k) = \sqrt{2k}\,\widetilde{\mathcal{S}}_\delta \qquad \text{for } k\ge 0$$

(and $-\infty$ otherwise). By (5.2) and (5.3), then we arrive at $\varphi_0^*(-k) = -\sqrt{2k}\,\widetilde{\mathcal{S}}_\delta$ for $k\ge 0$ (and $+\infty$ otherwise). Passing to the biconjugate, we infer that, for any $t>0$,

$$\varphi_0^{**}(t) = \sup_{k\ge 0}\Big\{-kt-\varphi_0^*(-k)\Big\} = \sup_{k\ge 0}\Big\{-kt+\sqrt{2k}\,\widetilde{\mathcal{S}}_\delta\Big\} = \frac{1}{2}\frac{(\widetilde{\mathcal{S}}_\delta)^2}{t}\ .$$

Finally, to deduce (5.1), it remains to check that φ_0^{**} coincides with φ_0. This is a consequence of the fact that φ_0 is convex continuous on $\mathbb{R}^+$. Indeed, let μ_0 be the uniform probability density on Q_δ. As F^δ belongs to $H^{-1}(Q_\delta;\mathbb{R}^3)$, we have that $k(\delta) := \mathcal{C}^{\rm el}(\mu_0, j, F^\delta) < +\infty$. Then, for every $t>0$, the measure $t\mu_0$ is admissible for $\varphi_\varepsilon(t)$ whenever $\varepsilon\le t^{-1}|Q_\delta|$. Thus

$$\varphi_0(t) \ \le\ \varphi_\varepsilon(t) \ \le\ \mathcal{C}^{\rm el}(t\mu_0, j, F^\delta) \ =\ \frac{k(\delta)}{t}\ ,$$

where the last equality is obtained performing the rescaling $V=tU$ on the competing strain displacements. The continuity of the convex function φ_0 on $(0,+\infty)$ follows from the latter upper bound, and the proof of Proposition 3.1 is concluded. □

Proof of Theorem 3.3. In view of Lemma 3.4, it is enough to prove that one among the sequences $\widetilde{\mathcal{S}}_\delta$ and $\mathcal{S}_\delta$, say $\widetilde{\mathcal{S}}_\delta$, converges to $\mathcal{S}_0$.

Let us begin by writing $\widetilde{\mathcal{S}}_\delta$ in a more convenient way. We set

$$U(x) = \Big(u_1\big(x',\delta^{-1}x_3\big),\ u_2\big(x',\delta^{-1}x_3\big),\ \delta^{-1}u_3\big(x',\delta^{-1}x_3\big)\Big)\ ,$$

so that

$$e(U)(x) = e_\delta(u)\big(x',\delta^{-1}x_3\big) := \begin{bmatrix} e_{\alpha\beta}(u) & \delta^{-1}e_{\alpha 3}(u)\\ \delta^{-1}e_{\alpha 3}(u) & \delta^{-2}e_{33}(u)\end{bmatrix}(x',\delta^{-1}x_3)\ , \tag{5.4}$$

where the indices α and β take values in $\{1,2\}$. Hence

$$\begin{aligned}
\widetilde{\mathcal{S}}_\delta &= \sup\Big\{\langle F,u\rangle_{\mathbb{R}^3} \ : \ u\in\mathcal{C}^\infty(\mathbb{R}^3;\mathbb{R}^3) \text{ such that } \rho(e_\delta(u)(x',\delta^{-1}x_3))\le 1 \text{ on } Q_\delta\Big\}\\
&= \sup\Big\{\langle F,u\rangle_{\mathbb{R}^3} \ : \ u\in\mathcal{C}^\infty(\mathbb{R}^3;\mathbb{R}^3) \text{ such that } \rho(e_\delta(u))\le 1 \text{ on } Q\Big\}\\
&= \sup\Big\{\langle F,u\rangle_{\mathbb{R}^3} \ : \ u\in K_\delta\Big\}\ ,
\end{aligned}$$

where K_δ denotes the convex set

$$K_\delta := \Big\{ u \in \mathcal{C}^\infty(\mathbb{R}^3;\mathbb{R}^3) \ : \ \rho(e_\delta(u)) \le 1 \text{ on } Q \Big\} \ .$$

As a preliminary remark, we notice that the following compactness property holds: if we take a sequence $\{u^\delta\}$ such that $u^\delta \in K_\delta$, then up to subsequences and up to a rigid motion, it converges uniformly on Q. Indeed, by (2.1) we have that $e_\delta(u^\delta)$ is uniformly bounded in $L^\infty(Q)$; hence, up to subtracting a rigid displacement (which is not restrictive thanks to (2.3)), by the Korn inequality $\{u^\delta\}$ is equibounded in $W^{1,p}(Q;\mathbb{R}^3)$ for every $p \in (1,+\infty)$.

In view of this remark, we are reduced to identifying the Kuratowski limit (if any) of the sequence $\{K_\delta\}$ with respect to the uniform convergence on the compact Q. Indeed, if $\overline{K}$ denotes such a Kuratowski limit, since the linear form $u \mapsto \langle F, u\rangle$ is continuous with respect to the uniform convergence, we will have that

$$\lim_{\delta\to 0} \widetilde{\mathcal{S}}_\delta \ = \ \sup\Big\{ \langle F, u\rangle_{\mathbb{R}^3} \ : \ u \in \overline{K} \Big\} \ . \tag{5.5}$$

We claim that the set $\overline{K}$ can be characterized as follows:

$$\overline{K} = \Big\{ u \in L^\infty(Q;\mathbb{R}^3) \ : \ e(u) \in L^\infty(Q;\mathbb{R}^{3\times 3}_{\rm sym}) \ , \ \overline{\rho}(e_{\alpha\beta}(u)) \le 1 \ , \ e_{i3}(u) = 0 \text{ a.e. on } Q \Big\} \ . \tag{5.6}$$

Let us first show how Theorem 3.3 follows from (5.6) and then give the proof of (5.6).

As a slight variant of Theorem 3.1 in [13], it is easy to check that the right-hand side of (5.6) is the closure in the uniform norm of the set of Kirchoff–Love displacements

$$K = \Big\{ u \in \mathcal{C}^\infty(\mathbb{R}^3;\mathbb{R}^3) \ : \ \overline{\rho}(e_{\alpha\beta}(u)) \le 1 \ , \ e_{i3}(u) = 0 \text{ on } Q \Big\} \ .$$

As is well known, any function $u \in K$ may be written under the form

$$u_i(x) = v_i(x') - \frac{\partial v_3}{\partial x_i}(x') x_3 \qquad \text{for } i = 1,2 \ , \qquad u_3(x) = v_3(x') \ .$$

In terms of the function v, the matrix $e_{\alpha\beta}(u)$ is given by

$$e_{\alpha\beta}(u) = e(v_1, v_2) - x_3 \nabla^2 v_3 \ , \tag{5.7}$$

and hence v must satisfy the inequality

$$\overline{\rho}\big(e(v_1,v_2) - x_3\nabla^2 v_3\big) \le 1 \qquad \forall\, (x', x_3) \in \Omega \times \big(-h, h\big) \ ,$$

which by convexity is equivalent to

$$\overline{\rho}\big(e(v_1,v_2) \pm h\nabla^2 v_3\big) \le 1 \qquad \text{on } \Omega \ . \tag{5.8}$$

On the other hand, we have

$$\langle F, u\rangle_{\mathbb{R}^3} = \sum_{i=1}^3 \langle F_i, v_i\rangle_{\mathbb{R}^3} + \sum_{i=1}^2 \left\langle x_3 \frac{\partial F_i}{\partial x_i}, v_3 \right\rangle_{\mathbb{R}^3} = \langle \overline{F}, v\rangle_{\mathbb{R}^2} \ . \tag{5.9}$$

By (5.5), (5.8), (5.9) and recalling the definition of $\mathcal{S}_0$ in (3.7), we conclude that

$$\lim_{\delta\to 0} \widetilde{\mathcal{S}}_\delta \ = \ \sup\Big\{ \langle F, u\rangle_{\mathbb{R}^3} \ : \ u \in K \Big\} = \mathcal{S}_0 \ .$$

It remains to establish (5.6). Such an equality holds, provided one has

(i) $u^\delta \in K_\delta$, $u^\delta \to u$ uniformly on $Q \implies u \in \overline{K}$,

(ii) $u \in K \implies \exists u^\delta \in K_\delta$ such that $u^\delta \to u$ uniformly on Q.

Proof of (i). Let $u^\delta \in K_\delta$ such that $u^\delta \to u$ uniformly on Q. As already noticed above in this proof, such a sequence $\{u^\delta\}$ is weakly precompact in $W^{1,p}(Q;\mathbb{R}^3)$ for every $p \in (1,+\infty)$, which ensures that u belongs $W^{1,p}(Q;\mathbb{R}^3)$ for every such p. Possibly passing to a subsequence, we may assume that $\{e_\delta(u^\delta)\}$ converges weakly in $L^p(Q;\mathbb{R}^{3\times 3}_{\rm sym})$ to some matrix valued function $M(x)$ which is of the form

$$M = \begin{bmatrix} e_{\alpha\beta}(u) & \xi_{\alpha 3} \\ \xi_{\alpha 3} & \xi_{33} \end{bmatrix} .$$

By the convexity of ρ, one has

$$\|\rho(M)\|_{L^\infty(Q)} \le \liminf_\delta \|\rho(e_\delta(u^\delta))\|_{L^\infty(Q)} \le 1 .$$

Thus, by the definition (3.6) of $\overline{\rho}$, it follows that

$$\|\overline{\rho}(e_{\alpha\beta}(u))\|_{L^\infty(Q)} \le \|\rho(M)\|_{L^\infty(Q)} \le 1 .$$

On the other hand, it is clear that, for $i = 1,2,3$, $e_{i3}(u^\delta)$ does converge strongly to 0 in $L^p(Q)$ and therefore $e_{i3}(u) = 0$. Summarizing we have proved that u belongs to $\overline{K}$.

Proof of (ii). Let $u \in K$. We search for $u^\delta \in K_\delta$ such that $u^\delta \to u$ uniformly on Q. To this end, it not restrictive to assume that the *strict* inequality $\overline{\rho}(e_{\alpha\beta}(u)) < 1$ holds on Q (indeed, for any $u \in K$ the function $\tilde{u} := (1-\delta)u$ satisfies $e_{i3}\tilde{u} = 0$ and $\overline{\rho}(e_{\alpha\beta}\tilde{u}) < 1$). Let $\xi^i = \xi^i(x', x_3)$ be arbitrary smooth functions, and let Φ_i denote their primitives with respect to the x_3 variable:

$$\Phi_i(x', x_3) := \int_0^{x_3} \xi_i(x', s)\, ds .$$

We define the sequence $\{u^\delta\}$ componentwise by

$$u_1^\delta = u_1 + \delta\Phi_1 , \qquad u_2^\delta = u_2 + \delta\Phi_2 , \qquad u_3^\delta = u_3 + \delta^2\Phi_3 .$$

Clearly $\{u^\delta\}$ converges uniformly to u, and, according to definition (5.4), an immediate calculation gives

$$e_\delta(u^\delta) = e_{\alpha\beta}(u) + \sum_{i=1}^{2} \Big(\xi_i + \delta\frac{\partial\Phi_3}{\partial x_i}\Big)(e_i \otimes e_3)^* + \xi_3(e_3 \otimes e_3) ,$$

so that

$$\rho(e_\delta(u^\delta)) \le \rho\left(e_{\alpha\beta}(u) + \sum_{i=1}^{3} \xi^i (e_i \otimes e_3)^*\right) + o(1) .$$

The proof of (ii) is concluded by the arbitrariness of the functions ξ_i. □

Proof of Lemma 3.4. For any given integrand $g : \mathbb{R}^{2\times 2}_{\rm sym} \to \mathbb{R}$, if $\overline{g}$ is defined according to (3.6), one can easily check that the Fenchel conjugates of g and $\overline{g}$ are related by the identity

$$(\overline{g})^*(z^*) = g^*(z^*|0) \qquad \forall\, z^* \in \mathbb{R}^{2\times 2}_{\rm sym} ,$$

where $(z^*|0)$ denotes the 3×3 matrix obtained by adding to z^* a null third line and third column.

Applying this to the convex integrands j and j_0 yields, for every $z\in\mathbb{R}^{2\times 2}_{\rm sym}$,

$$\begin{aligned}\overline{j_0}(z) &= \sup\big\{z\cdot z^* - (j_0)^*(z^*|0)\ :\ z^*\in\mathbb{R}^{2\times 2}_{\rm sym}\big\}\,,\\ \overline{j}(z) &= \sup\big\{z\cdot z^* - j^*(z^*|0)\ :\ z^*\in\mathbb{R}^{2\times 2}_{\rm sym}\big\}\ .\end{aligned}$$

Then the lemma is proved if we can show that

$$(j_0)^*(z^*|0) = j^*(z^*|0)\qquad \forall\, z^*\in\mathbb{R}^{2\times 2}_{\rm sym}\ .$$

Actually, this equality is satisfied since j_0^* and j^* turn out to coincide more generally on the class of degenerated tensors (see [10, Lemma 3.1]). Indeed, the inequality $j_0^*(\tau)\geq j^*(\tau)$ holds for all $\tau\in\mathbb{R}^{3\times 3}_{\rm sym}$ because by definition j_0 satisfies the inequality $j_0\leq j$. On the other hand, let $\tau\in\mathbb{R}^{3\times 3}_{\rm sym}$ be a degenerated tensor. By definition of j_0, for all $z\in\mathbb{R}^{3\times 3}_{\rm sym}$ there holds

$$z\cdot\tau - j_0(z)\leq j^*(\tau)\ ,$$

so that the inequality $j_0^*(\tau)\leq j^*(\tau)$ follows by passing to the supremum over z in the left-hand side. □

Proof of Corollary 3.5. In case (i) it is immediate that

$$\mathcal{S}_0\geq \sup\Big\{\langle\overline{F}_3,v_3\rangle_{\mathbb{R}^2}\ :\ v_3\in\mathcal{C}^\infty(\mathbb{R}^2;\mathbb{R}) \text{ such that } \ \overline{\rho}(\nabla^2 v_3)\leq 1/h \text{ on } \Omega\Big\}\ .$$

The converse inequality is obtained by noticing that, since $\overline{\rho}$ is even and subadditive, the constraint $\overline{\rho}\big(e(v_1,v_2)\pm h\nabla^2 v_3\big)\leq 1$ implies $\overline{\rho}\big(\nabla^2 v_3\big)\leq 1/h$. The proof in case (ii) is analogous. □

Proof of Corollary 3.6. In view of Lemma 3.4, it is enough to prove the statement for one among the sequences $\widetilde{\mathcal{I}}_{\varepsilon,\delta}$ and $\mathcal{I}_{\varepsilon,\delta}$, say $\widetilde{\mathcal{I}}_{\varepsilon,\delta}$.

First, we recall that there holds $\lim_\varepsilon \widetilde{\mathcal{I}}_{\varepsilon,\delta}=\widetilde{\mathcal{I}}_\delta$ (see Proposition 3.1) and that $\widetilde{\mathcal{I}}_\delta=\widetilde{\mathcal{S}}_\delta^2/2$ [11, Theorem 2.3]. Then Theorem 3.3 gives $\lim_\delta\widetilde{\mathcal{S}}_\delta=\mathcal{S}_0$. Finally, we apply Corollary 3.5(i) and [14, Theorem 2.4] to conclude that $\mathcal{S}_0^2/2=h^{-2}\inf\big\{\mathcal{C}^{\rm pl}\big(\mu,\overline{j},\overline{F}_3\big)\ :\ \mu\in\mathcal{P}(\overline{\Omega})\big\}$. □

Proof of Lemma 3.9. Let us rewrite $(\mathcal{P})$ as

$$(\mathcal{P})\qquad -\inf\Big\{-\langle\overline{F},v\rangle_{\mathbb{R}^2}+\chi_{\mathcal{K}}(A^+v)+\chi_{\mathcal{K}}(A^-v)\ :\ v\in\mathcal{C}^\infty(\mathbb{R}^2;\mathbb{R}^3)\Big\}\ ,$$

where $\chi_{\mathcal{K}}$ is the characteristic function of the set

$$\mathcal{K}=\Big\{M\in\mathcal{C}_0(\Omega;\mathbb{R}^{2\times 2}_{\rm sym})\ :\ \overline{\rho}(M)\leq 1\Big\}\ ,$$

and $A:\mathcal{C}_0(\Omega;\mathbb{R}^3)\ni v\mapsto(A^+v,A^-v)\in[\mathcal{C}_0(\Omega;\mathbb{R}^{2\times 2}_{\rm sym})]^2$ is the linear operator densely defined by $A^\pm v:=e(v_1,v_2)\pm h\nabla^2 v_3$ for all smooth functions v.

By standard duality theory (see, for instance, [19]), there holds

$(\mathcal{P}^*)$

$$\min\Big\{\int\overline{\rho}^o(\lambda^+)+\int\overline{\rho}^o(\lambda^-)\ :\ (\lambda^+,\lambda^-)\in[\mathcal{M}(\overline{\Omega};\mathbb{R}^{2\times 2}_{\rm sym})]^2\ ,\ A^*(\lambda^+,\lambda^-)=\overline{F}\Big\}\ ,$$

where $A^* : [\mathcal{M}(\overline{\Omega};\mathbb{R}^{2\times 2}_{\rm sym})]^2 \to \mathcal{M}(\overline{\Omega};\mathbb{R}^3)$ is the adjoint operator of A. It is determined by the following identity (valid for every smooth v):

$$\begin{aligned}\langle A^*(\lambda^+,\lambda^-),v\rangle_{\mathbb{R}^2} &= \langle(\lambda^+,\lambda^-),(A^+v,A^-v)\rangle_{\mathbb{R}^2}\\ &= \langle\lambda^+,e(v_1,v_2)+h\nabla^2 v_3\rangle_{\mathbb{R}^2}+\langle\lambda^-,e(v_1,v_2)-h\nabla^2 v_3\rangle_{\mathbb{R}^2}\\ &= -\langle\operatorname{div}(\lambda^++\lambda^-),(v_1,v_2)\rangle_{\mathbb{R}^2}+\langle h\operatorname{div}^2(\lambda^+-\lambda^-),v_3\rangle_{\mathbb{R}^2}\ .\end{aligned}$$

Therefore, when rewritten componentwise, the constraint $A^*(\lambda^+,\lambda^-)=\overline{F}$ is equivalent to the system of two conditions: $-\operatorname{div}(\lambda^+ + \lambda^-) = (\overline{F}_1,\overline{F}_2)$ and $h\operatorname{div}^2(\lambda^+ - \lambda^-) = \overline{F}_3$. □

Proof of Proposition 3.10. Let v and $\lambda^\pm$ be optimal, respectively, for problems $(\mathcal{P})$ and $(\mathcal{P})^*$. By Lemma 3.9 there holds

$$\int \overline{\rho}^o(\lambda^+) + \int \overline{\rho}^o(\lambda^-) = \langle\overline{F},v\rangle\ . \tag{5.10}$$

On the other hand, if the operator $Av = (A^+v, A^-v)$ is defined as in the proof of Lemma 3.9, we have

$$\overline{\rho}^o(\lambda^\pm) \geq \overline{\rho}^o(\lambda^\pm)\,\overline{\rho}(A^\pm v) \geq \langle\lambda^\pm, A^\pm v\rangle_{\mathbb{R}^2}\ , \tag{5.11}$$

which implies

$$\int \overline{\rho}^o(\lambda^+) + \int \overline{\rho}^o(\lambda^-) \geq \langle(\lambda^+,\lambda^-),Av\rangle_{\mathbb{R}^2} = \langle A^*(\lambda^+,\lambda^-),v\rangle_{\mathbb{R}^2} = \langle\overline{F},v\rangle_{\mathbb{R}^2}\ . \tag{5.12}$$

Combining (5.10) and (5.12), we deduce that the inequalities in (5.11) must turn into equalities, so that the optimality conditions (3.9) hold.

Conversely, any v and $\lambda^\pm$ which are admissible, respectively, for problems $(\mathcal{P})$ and $(\mathcal{P})^*$ satisfy

$$\langle\overline{F},v\rangle_{\mathbb{R}^2} \leq \mathcal{S}_0 \leq \int \overline{\rho}^o(\lambda^+) + \int \overline{\rho}^o(\lambda^-)\ . \tag{5.13}$$

If equations (3.9) hold, we have

$$\langle\overline{F},v\rangle_{\mathbb{R}^2} = \langle A^*(\lambda^+,\lambda^-),v\rangle_{\mathbb{R}^2} = \langle(\lambda^+,\lambda^-),Av\rangle_{\mathbb{R}^2} = \int \overline{\rho}^o(\lambda^+) + \int \overline{\rho}^o(\lambda^-)\ ,$$

and hence the inequalities in (5.13) must turn into equalities, which means that v and $\lambda^\pm$ are optimal. □

Acknowledgments. We are very grateful to Pierre Seppecher for having suggested to us the choice of Example 4.1 and to one of the referees for having drawn the paper [26] to our attention. The second author thanks the University of Toulon for the kind hospitality.

REFERENCES

[1] E. Acerbi, G. Buttazzo, and D. Percivale, *Thin inclusions in linear elasticity: A variational approach*, J. Reine Angew. Math., 386 (1988), pp. 99–115.

[2] G. Allaire, *Shape Optimization by the Homogenization Method*, Springer, Berlin, 2002.

[3] G. Allaire, E. Bonnetier, G. Francfort, and F. Jouve, *Shape optimization by the homogenization method*, Numer. Math., 76 (1997), pp. 27–68.

[4] G. Allaire and R. Kohn, *Optimal design for minimum weight and compliance in plane stress using extremal microstructures*, European J. Mech. A Solids, 12 (1993), pp. 839–878.
[5] N. Antonić and N. Balenović, *Optimal design for plates and relaxation*, Math. Commun., 4 (1999), pp. 111–119.
[6] N. V. Banichuk, *Problems and Methods of Optimal Structural Design*, Plenum Press, New York, 1983.
[7] E. Bonnetier and C. Conca, *Relaxation totale d'un problème d'optimisation de plaques*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 931–936.
[8] E. Bonnetier and C. Conca, *Approximation of Young measures by functions and application to a problem of optimal design for plates with variable thickness*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 399–422.
[9] E. Bonnetier and M. Vogelius, *Relaxation of a compliance functional for a plate optimization problem*, in Applications of Multiple Scaling in Mechanics, P. G. Ciarlet and E. Sanchez-Palencia, eds., Masson, Paris, 1987, pp. 31–53.
[10] G. Bouchitté, *Optimization of light structures: The vanishing mass conjecture*, in Homogenization, 2001 (Naples), GAKUTO Internat. Ser. Math. Sci. Appl. 18, Gakkötosho, Tokyo, 2003, pp. 131–145.
[11] G. Bouchitté and G. Buttazzo, *Characterization of optimal shapes and masses through Monge-Kantorovich equations*, J. Eur. Math. Soc. (JEMS), 3 (2001), pp. 139–168.
[12] G. Bouchitté, G. Buttazzo, and P. Seppecher, *Energies with respect to a measure and applications to low dimensional structures*, Calc. Var. Partial Differential Equations, 5 (1997), pp. 37–54.
[13] G. Bouchitté, G. Buttazzo, and L. De Pascale, *A p-Laplacian approximation for some mass optimization problems*, J. Optim. Theory Appl., 118 (2003), pp. 1–25.
[14] G. Bouchitté and I. Fragalà, *Optimality conditions for mass design problems and applications to thin plates*, Arch. Ration. Mech. Anal., 184 (2007), pp. 257–284.
[15] G. Bouchitté and I. Fragalà, *Second order energies on thin structures: Variational theory and non-local effects*, J. Funct. Anal., 204 (2003), pp. 228–267.
[16] D. Caillerie, *Models of thin or thick plates and membranes derived from linear elasticity*, in Applications of Multiple Scaling in Mechanics, Masson, Paris, 1987, pp. 54–68.
[17] K. T. Cheng and N. Olhoff, *An investigation concerning optimal design of solid elastic plates*, Internat. J. Solids and Structures, 17 (1981), pp. 305–323.
[18] P. Ciarlet, *Mathematical Elasticity, Vol.* 2*, Theory of Plates*, Stud. Math. Appl. 27, North–Holland, Amsterdam, 1997.
[19] I. Ekeland and R. Temam, *Analyse convexe et problèmes variationnels*, Dunod, Gauthier–Villars, Paris, 1976.
[20] L. V. Gibiansky and A. V. Cherkaev, *Design of composite plates of extremal rigidity*, in Topics in the Mathematical Modelling of Composite Materials, Progr. Nonlinear Differential Equations Appl. 31, Birkhäuser, Boston, 1997, pp. 95–137.
[21] C. Goffman and J. Serrin, *Sublinear functions of measures and variational integrals*, Duke Math. J., 31 (1964), pp. 159–178.
[22] R. Kohn and G. Strang, *Optimal design and relaxation of variational problems.* I, Comm. Pure Appl. Math., 39 (1986), pp. 113–137.
[23] R. Kohn and G. Strang, *Optimal design and relaxation of variational problems.* II, Comm. Pure Appl. Math., 39 (1986), pp. 139–182.
[24] R. Kohn and G. Strang, *Optimal design and relaxation of variational problems.* III, Comm. Pure Appl. Math., 39 (1986), pp. 353–377.
[25] R. Kohn and M. Vogelius, *Thin plates with varying thickness, and their relation to structural optimization*, in Homogenization and Effective Moduli of Materials and Media, IMA Vol. Math. Appl. 1, J. Ericksen, D. Kinderleherer, R. Kohn, and J. L. Lions, eds., Springer, New York, 1986, pp. 126–149.
[26] T. Lewinski and J. J. Telega, *Michell-like grillages and structures with locking*, Arch. Mech. (Arch. Mech. Stos.), 53 (2001), pp. 457–485.
[27] J. Muñoz and P. Pedregal, *On the relaxation of an optimal design problem for plates*, Asymptot. Anal., 16 (1998), pp. 125–140.
[28] J. Sprekels and D. Tiba, *A duality approach in the optimization of beams and plates*, SIAM J. Control Optim., 37 (1998), pp. 486–501.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1683–1704

ASYMPTOTIC CONVERGENCE ANALYSIS OF A NEW CLASS OF PROXIMAL POINT METHODS*

WILLIAM W. HAGER† AND HONGCHAO ZHANG‡

Abstract. Finite dimensional local convergence results for self-adaptive proximal point methods and nonlinear functions with multiple minimizers are generalized and extended to a Hilbert space setting. The principle assumption is a local error bound condition which relates the growth in the function to the distance to the set of minimizers. A local convergence result is established for almost exact iterates. Less restrictive acceptance criteria for the proximal iterates are also analyzed. These criteria are expressed in terms of a subdifferential of the proximal function and either a subdifferential of the original function or an iteration difference. If the proximal regularization parameter $\mu(\mathbf{x})$ is sufficiently small and bounded away from zero and f is sufficiently smooth, then there is local linear convergence to the set of minimizers. For a locally convex function, a convergence result similar to that for almost exact iterates is established. For a locally convex solution set and smooth functions, it is shown that if the proximal regularization parameter has the form $\mu(\mathbf{x}) = \beta\|f'[\mathbf{x}]\|^\eta$, where $\eta \in (0,2)$, then the convergence is at least superlinear if $\eta \in (0,1)$ and at least quadratic if $\eta \in [1,2)$.

Key words. proximal point, degenerate optimization, multiple minima, self-adaptive method

AMS subject classifications. 90C06, 90C26, 65Y20

DOI. 10.1137/060666627

1. Introduction.

In this paper, we consider an optimization problem:

$$\min\{f(\mathbf{x}) : \mathbf{x} \in \mathcal{H}\}, \tag{1.1}$$

where $\mathcal{H}$ is a real Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and $f : \mathcal{H} \mapsto \mathcal{R}$. It is assumed that the set of minimizers for (1.1), denoted $\mathbf{X}$, is nonempty and closed. We establish new convergence rate results for proximal point methods for solving (1.1).

Literature connected with the analysis and development of proximal point methods includes [1, 2, 3, 4, 6, 8, 9, 10, 11, 12, 15, 16, 18, 19, 20, 21, 24, 25, 26, 27]. In the proximal point method, iterates $\mathbf{x}_k$, $k \geq 1$, are generated by the following rule:

$$\mathbf{x}_{k+1} \in \arg\ \min\{F_k(\mathbf{x}) : \mathbf{x} \in \mathcal{H}\}, \tag{1.2}$$

where

$$F_k(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2}\mu_k\|\mathbf{x} - \mathbf{x}_k\|^2.$$

Here $\mathbf{x}_0 \in \mathcal{H}$ is an initial guess for a minimizer, the parameters μ_k, $k \geq 0$, are positive scalars, and $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$ is the usual Hilbert space norm. When f is twice continuously differentiable, the eigenvalues of the second derivative operator F_k'' are bounded from below by μ_k at a local minimizer; consequently, the regularization term $\mu_k\|\mathbf{x} - \mathbf{x}_k\|^2$ improves the conditioning of (1.1).

*Received by the editors August 1, 2006; accepted for publication (in revised form) May 9, 2007; published electronically November 2, 2007. This material is based upon work supported by the National Science Foundation under grants 0203270, 0619080, and 0620286.

http://www.siam.org/journals/sicon/46-5/66662.html

†Department of Mathematics, University of Florida, P.O. Box 118105, Gainesville, FL 32611-8105 (hager@math.ufl.edu, http://www.math.ufl.edu/~hager).

‡Institute for Mathematics and Its Applications (IMA), University of Minnesota, 400 Lind Hall, 207 Church Street S.E., Minneapolis, MN 55455-0436 (hozhang@ima.umn.edu).

In [21] Rockafellar shows that if f is convex, then the proximal point method converges linearly when μ_k is bounded away from zero and superlinearly when μ_k tends to zero. For these convergence results, $\mathbf{X}$ is a singleton. Luque [14] studied the case when $\mathbf{X}$ may contain more than one element. By assuming some growth properties for the (multivalued) inverse of the derivative, results analogous to those of Rockafellar were obtained. Kaplan and Tichatschke [11] consider the case where f is convex, μ_k is constant, and $\mathbf{X}$ may contain more than one element. A linear convergence result for the iterates is established under a growth condition for the function which is similar to the growth condition used in our paper (see Assumption 14.4 and Theorem 14.5 in [11]).

In another research direction, Combettes and Pennanen [2], Iusem, Pennanen, and Svaiter [10], and Pennanen [19] replace the monotonicity assumptions appearing in earlier work by a weaker hypomonotonicity condition for the inverse of the derivative, that is, the inverse of the derivative is monotone when a multiple of the identity is added. Additional assumptions, however, enter into the analysis which imply the solution set $\mathbf{X}$ is a singleton.

In [7] we present a new class of self-adaptive proximal point methods for finite dimensional optimization problems. Our analysis employs the following local error bound condition at $\hat{\mathbf{x}} \in \mathbf{X}$: There exist positive constants α and ρ such that

$$f(\mathbf{x}) - f^* \geq \alpha D(\mathbf{x}, \mathbf{X})^2 \quad \text{whenever } \|\mathbf{x} - \hat{\mathbf{x}}\| \leq \rho, \tag{1.3}$$

where f^* is the minimum value in (1.1) and

$$D(\mathbf{x}, \mathbf{X}) = \inf_{\mathbf{y} \in \mathbf{X}} \|\mathbf{x} - \mathbf{y}\|.$$

In other words, $D(\mathbf{x}, \mathbf{X})$ measures the distance to the solution set $\mathbf{X}$. If (1.3) is satisfied, then we say that f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$. For an exact proximal iterate $\mathbf{x}_{k+1}$ satisfying (1.2), we show in [7] that for any starting guess $\mathbf{x}_0$ in a neighborhood of the solution set, the iterates converge to a solution $\mathbf{x}^*$ of (1.1) and the following estimate holds:

$$D(\mathbf{x}_{k+1}, \mathbf{X}) \leq C\mu_k D(\mathbf{x}_k, \mathbf{X}), \tag{1.4}$$

where $C = 2/(2\alpha - \mu_k)$.

In a Hilbert space setting, the exact proximal iterate (1.2) may not exist. In this paper, we establish a similar convergence result using the following acceptance criterion: $\mathbf{x}_{k+1}$ is acceptable when

$$\begin{aligned} F_k(\mathbf{x}_{k+1}) &\leq \inf_{\mathbf{x} \in \mathbf{X}} \left\{ F_k(\mathbf{x}) + \frac{\mu_k^2}{2} \|\mathbf{x} - \mathbf{x}_k\|^2 \right\} \\ &= \inf_{\mathbf{x} \in \mathbf{X}} \left\{ f(\mathbf{x}) + \left(\frac{\mu_k + \mu_k^2}{2} \right) \|\mathbf{x} - \mathbf{x}_k\|^2 \right\}. \end{aligned} \tag{C0}$$

In section 3 we show that there always exists an iterate satisfying (C0), and a convergence result of the form (1.4) holds.

Although (C0) leads to an elegant convergence theory, which can be applied to any function whose set of minimizers is nonempty and closed, the acceptance criterion is not easily implemented since it is expressed in terms of the solution set (which we are trying to compute). Consequently, we now introduce implementable acceptance criteria which are expressed in terms of the (basic) subdifferential of f (see [17, p. 82]

and [22]) denoted $\partial f(\mathbf{x})$. If f is Fréchet differentiable, then $\partial f(\mathbf{x}) = f'[\mathbf{x}]$. If f is convex, then $\partial f(\mathbf{x})$ is the usual subdifferential of convex analysis. The acceptance criteria for (1.2) are

(C1) $F_k(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ and $\|\partial F_k(\mathbf{x}_{k+1})\|_{\inf} \leq \mu_k \|\partial f(\mathbf{x}_k)\|_{\inf}$, and

(C2) $F_k(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ and $\|\partial F_k(\mathbf{x}_{k+1})\|_{\inf} \leq \theta\mu_k \|\mathbf{x}_{k+1} - \mathbf{x}_k\|$.

Here θ is a positive constant smaller than $1/\sqrt{2}$, and $\|\cdot\|_{\inf}$ denotes the distance to the origin; that is, for any set $\mathcal{S} \subset \mathcal{H}$,

$$\|\mathcal{S}\|_{\inf} = \inf_{s \in \mathcal{S}} \|\mathbf{s}\|.$$

If $\mathcal{S} = \emptyset$, then we set $\|\mathcal{S}\|_{\inf} = \infty$.

In [14] and [21], the authors considered the acceptance condition

$$\|\partial F_k(\mathbf{x}_{k+1})\| \leq \epsilon_k \mu_k \|\mathbf{x}_{k+1} - \mathbf{x}_k\|,$$

where $\sum_k \epsilon_k < \infty$. Our criterion (C2) corresponds to the case $\sum_k \epsilon_k = \infty$. In [14] and [21], the authors consider convex functionals, while here we obtain local convergence rates for general nonlinear functionals. Note that our acceptance criteria employ a subdifferential rather than the derivative used in our earlier work.

Slightly different versions of the proximal point method for maximal monotone operators are developed by Solodov and Svaiter in the series of papers [24, 25, 26]. They develop both a hybrid proximal point algorithm where an approximate proximal step is followed by a projection and a hybrid extragradient version in which the original operator is replaced by an ϵ enlargement. In order to compare their analysis to the results in our paper, we focus on the special case where the operator is the subdifferential of a convex function f. In each iteration of the Solodov/Svaiter scheme, they first compute an approximate proximal iterate $\mathbf{y}_k$ satisfying a relaxed version of (C2); they then update the iterate along the negative gradient:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \mathbf{g}_k, \tag{1.5}$$

where $\mathbf{g}_k \in \partial f(\mathbf{y}_k)$, s_k is the scalar stepsize, and $\theta < 1$. In [24], $s_k = 1/\mu_k$ (the reciprocal of the proximal regularization parameter), while in [25], s_k is chosen so that $\mathbf{x}_{k+1}$ is the projection of $\mathbf{x}_k$ onto the half-space

$$\{\mathbf{x} \in \mathcal{H} : \langle \mathbf{g}_k, \mathbf{x} - \mathbf{y}_k \rangle \geq 0\}.$$

Since the Solodov/Svaiter update (1.5) amounts to an extragradient step, their convergence theory for fixed θ (see [26, Thm. 8]) yields linear convergence, even when μ_k tends to 0, unless the accuracy criterion (C2) for the approximate proximal iterate $\mathbf{y}_k$ is strengthened. Ways to improve the accuracy of $\mathbf{y}_k$ so as to obtain superlinear convergence with the Solodov/Svaiter hybrid schemes are the following: (a) Replace θ by θ_k in (C2) and let θ_k tend to 0 (see [26, Rem. 9]). (b) In the case $\mathcal{H} = \mathbb{R}^n$, compute $\mathbf{y}_k$ by a Newton iteration applied to the proximal problem (1.2), with μ_k on the order of $\|\nabla f(\mathbf{x}_k)\|^{1/2}$ (see [26, sect. 5.2]).

In this paper, we obtain superlinear convergence with (C2) by letting μ_k approach 0, and we analyze how the convergence speed depends on the decay rate of μ_k. We consider both a convex cost function analogous to the maximal monotone operator in [24, 25, 26] and the more general case where the solution set $\mathbf{X}$ is locally convex and f is sufficiently smooth. We allow multiple solutions satisfying the local error bound condition (1.3), while in [24, 25, 26] the solution set is unique since the inverse operator is required to be Lipschitz continuous at zero [26, eq. (28)].

We will show that for either (C1) or (C2), for μ_k sufficiently small, and for either f locally convex or $\mathbf{X}$ locally convex and f sufficiently smooth, an estimate of the form (1.4) holds. For μ_k sufficiently small and bounded away from zero, and for smooth functions, there is at least local linear convergence to the set of minimizers. For $\mathbf{X}$ locally convex and f sufficiently smooth, and for $\mu_k = \beta\|f'[\mathbf{x}_k]\|^\eta$, where $\eta \in (0,2)$, the convergence is superlinear when $\eta \in (0,1)$ and at least quadratic when $\eta \in [1,2)$.

Our paper is organized as follows: In section 2 we establish the equivalence, when f is twice continuously differentiable, of our local error bound condition and a gradient-based local error bound condition used in [5, 13, 14, 28, 29, 30]. Note, though, that our local error bound condition can be applied even when f has no derivative. In section 3 we analyze proximal iterates which satisfy (C0). Section 4 studies the criteria (C1) and (C2).

1.1. Notation. Throughout this paper, we use the following notation. If $A : \mathcal{H} \mapsto \mathcal{H}$ is a bounded linear operator, then $\|A\|$ is the operator norm induced by the Hilbert space norm $\|\cdot\|$. The empty set is denoted $\emptyset$. The complement of a set $\mathcal{S} \subset \mathcal{H}$ is denoted $\mathcal{S}^c$. If $\mathbf{x}$ and $\mathbf{y} \in \mathcal{H}$, then $[\mathbf{x},\mathbf{y}]$ is the line segment connecting $\mathbf{x}$ and $\mathbf{y}$. $\mathcal{B}_\rho(\mathbf{x})$ is the ball with center $\mathbf{x}$ and radius ρ. $f'[\mathbf{x}]$ and $f''[\mathbf{x}]$ are the first- and second-order Fréchet derivatives of f at $\mathbf{x}$ when they exist. The derivatives are operators defined on either $\mathcal{H}$ or $\mathcal{H} \times \mathcal{H}$. We also view $f'[\mathbf{x}]$ as an element of $\mathcal{H}$ and write $f'[\mathbf{x}](\mathbf{y}) = \langle f'[\mathbf{x}], \mathbf{y}\rangle$. Similarly, we view $f''[\mathbf{x}]$ as a bounded linear map from $\mathcal{H}$ to itself and write

$$f''[\mathbf{x}](\mathbf{y},\mathbf{z}) = \langle f''[\mathbf{x}]\mathbf{y}, \mathbf{z}\rangle.$$

2. Local error bound based on derivative. In this paper, we utilize the local error bound condition (1.3) based on function value. Earlier work [5, 13, 14, 28, 29, 30] has exploited a local error bound condition based on the derivative. Namely, f' provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ if there exist positive constants α and ρ such that

$$\|f'[\mathbf{x}]\| \geq \alpha D(\mathbf{x}, \mathbf{X}) \quad \text{whenever } \|\mathbf{x} - \hat{\mathbf{x}}\| \leq \rho. \tag{2.1}$$

We now show that when f is smooth enough, these two conditions are equivalent.

LEMMA 2.1. *If f is twice continuously Fréchet differentiable in a neighborhood of $\hat{\mathbf{x}} \in \mathbf{X}$, then f provides a local error bound at $\hat{\mathbf{x}}$ in the sense of* (1.3) *if and only if f' provides a local error bound at $\hat{\mathbf{x}}$ in the sense of* (2.1).

Proof. Suppose f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ with positive scalars α and ρ satisfying (1.3). Choose ρ smaller, if necessary, so that f is twice continuously Fréchet differentiable in $\mathcal{B}_\rho(\hat{\mathbf{x}})$ and

$$\|f''[\mathbf{x}] - f''[\mathbf{y}]\| \leq \alpha/3 \quad \text{for all } \mathbf{x},\mathbf{y} \in \mathcal{B}_\rho(\hat{\mathbf{x}}). \tag{2.2}$$

Define $r = \rho/2$. Given $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$, let $\bar{\mathbf{x}}$ be any element of $\mathbf{X} \cap \mathcal{B}_r(\mathbf{x})$. Since $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$, we have $\hat{\mathbf{x}} \in \mathbf{X} \cap \mathcal{B}_r(\mathbf{x})$, which shows that $\mathbf{X} \cap \mathcal{B}_r(\mathbf{x})$ is nonempty. The triangle inequality implies that

$$\|\bar{\mathbf{x}} - \hat{\mathbf{x}}\| \leq \|\bar{\mathbf{x}} - \mathbf{x}\| + \|\mathbf{x} - \hat{\mathbf{x}}\| \leq 2r = \rho. \tag{2.3}$$

Since both $\mathbf{x}$ and $\bar{\mathbf{x}} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$, f is twice continuously Fréchet differentiable in $\mathcal{B}_\rho(\hat{\mathbf{x}})$, and $f'[\bar{\mathbf{x}}] = \mathbf{0}$, we have

$$f(\mathbf{x}) - f^* = f(\mathbf{x}) - f(\bar{\mathbf{x}}) = \frac{1}{2}\langle \mathbf{x} - \bar{\mathbf{x}}, f''[\bar{\mathbf{x}}](\mathbf{x} - \bar{\mathbf{x}})\rangle + R_2(\mathbf{x}, \bar{\mathbf{x}}), \tag{2.4}$$

where R_2 is the remainder term. The bound (2.2) gives

$$|R_2(\mathbf{x}, \bar{\mathbf{x}})| \leq \frac{\alpha}{3}\|\mathbf{x} - \bar{\mathbf{x}}\|^2$$

whenever $\mathbf{x}$ and $\bar{\mathbf{x}} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$. In this case, (2.4) and the local error bound condition (1.3) give

$$\alpha D(\mathbf{x}, \mathbf{X})^2 \leq f(\mathbf{x}) - f^* \leq \frac{1}{2}\|\mathbf{x} - \bar{\mathbf{x}}\|\|f''[\bar{\mathbf{x}}](\mathbf{x} - \bar{\mathbf{x}})\| + \frac{\alpha}{3}\|\mathbf{x} - \bar{\mathbf{x}}\|^2. \tag{2.5}$$

Again, since f is twice continuously Fréchet differentiable in $\mathcal{B}_\rho(\hat{\mathbf{x}})$ and $f'[\bar{\mathbf{x}}] = \mathbf{0}$, we have

$$f'[\mathbf{x}] = f'[\mathbf{x}] - f'[\bar{\mathbf{x}}] = f''[\bar{\mathbf{x}}](\mathbf{x} - \bar{\mathbf{x}}) + \mathbf{R}_1(\mathbf{x}, \bar{\mathbf{x}}), \tag{2.6}$$

where $\mathbf{R}_1$ is the remainder term. The bound (2.2) gives

$$\|R_1(\mathbf{x}, \bar{\mathbf{x}})\| \leq \frac{\alpha}{3}\|\mathbf{x} - \bar{\mathbf{x}}\| \tag{2.7}$$

whenever $\mathbf{x}$ and $\bar{\mathbf{x}} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$. Combining (2.5)–(2.7) yields

$$\alpha D(\mathbf{x}, \mathbf{X})^2 \leq \frac{1}{2}\left(\|\mathbf{x} - \bar{\mathbf{x}}\|\|f'[\mathbf{x}]\| + \alpha\|\mathbf{x} - \bar{\mathbf{x}}\|^2\right). \tag{2.8}$$

Since $\mathbf{X} \cap \mathcal{B}_r(\mathbf{x})$ is nonempty, we have

$$D(\mathbf{x}, \mathbf{X}) = \inf_{\bar{\mathbf{x}} \in \mathbf{X}} \|\mathbf{x} - \bar{\mathbf{x}}\| = \inf\{\|\mathbf{x} - \bar{\mathbf{x}}\| : \bar{\mathbf{x}} \in \mathbf{X} \cap \mathcal{B}_r(\mathbf{x})\}.$$

Minimizing the right-hand side of (2.8) over $\bar{\mathbf{x}} \in \mathbf{X} \cap \mathcal{B}_r(\mathbf{x})$ gives

$$\alpha D(\mathbf{x}, \mathbf{X})^2 \leq \frac{1}{2}\left(D(\mathbf{x}, \mathbf{X})\|f'[\mathbf{x}]\| + \alpha D(\mathbf{x}, \mathbf{X})^2\right).$$

Rearranging this yields

$$\|f'[\mathbf{x}]\| \geq \alpha D(\mathbf{x}, \mathbf{X}).$$

Hence, $\partial f = f'$ provides a local error bound at $\hat{\mathbf{x}}$ with constants α and r.

Conversely, suppose f' provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ with positive scalars α and ρ satisfying (2.1). Let ρ be as in the first half of the proof. Choose ρ smaller, if necessary, so that

$$\|f''[\mathbf{x}] - f''[\mathbf{y}]\| \leq \frac{7\alpha^2}{18(\lambda + 1)} \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathcal{B}_\rho(\hat{\mathbf{x}}), \tag{2.9}$$

where

$$\lambda = \sup\{\|f''[\mathbf{x}]\| : \mathbf{x} \in \mathcal{B}_\rho(\hat{\mathbf{x}})\}. \tag{2.10}$$

Let $r = \rho/2$, let $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$, and let $\bar{\mathbf{x}} \in \mathbf{X} \cap \mathcal{B}_r(\hat{\mathbf{x}})$.

Since f achieves a minimum at $\bar{\mathbf{x}} \in \mathbf{X}$, $f''[\bar{\mathbf{x}}]$ is positive. Thus, there exists a unique, positive self-adjoint bounded linear operator B, the square root of $f''[\bar{\mathbf{x}}]$,

satisfying $f''[\bar{\mathbf{x}}] = B^2$ [23, Thm. 13.31]. By (2.6), (2.7), and the local error bound condition (2.1), we have

$$\begin{aligned}\alpha D(\mathbf{x},\mathbf{X}) &\le \|f'[\mathbf{x}]\| \\ &\le \|f''[\bar{\mathbf{x}}](\mathbf{x}-\bar{\mathbf{x}})\| + \frac{\alpha}{3}\|\mathbf{x}-\bar{\mathbf{x}}\| \\ &= \|B^2(\mathbf{x}-\bar{\mathbf{x}})\| + \frac{\alpha}{3}\|\mathbf{x}-\bar{\mathbf{x}}\| \\ &\le \|B\|\|B(\mathbf{x}-\bar{\mathbf{x}})\| + \frac{\alpha}{3}\|\mathbf{x}-\bar{\mathbf{x}}\|.\end{aligned}$$

Squaring both sides yields

$$\begin{aligned}\alpha^2 D(\mathbf{x},\mathbf{X})^2 &\le 2\|B\|^2\|B(\mathbf{x}-\bar{\mathbf{x}})\|^2 + \frac{2\alpha^2}{9}\|\mathbf{x}-\bar{\mathbf{x}}\|^2 \\ &= 2\langle \mathbf{x}-\bar{\mathbf{x}}, f''[\bar{\mathbf{x}}](\mathbf{x}-\bar{\mathbf{x}})\rangle\|f''[\bar{\mathbf{x}}]\| + \frac{2\alpha^2}{9}\|\mathbf{x}-\bar{\mathbf{x}}\|^2 \\ &\le 2\langle \mathbf{x}-\bar{\mathbf{x}}, f''[\bar{\mathbf{x}}](\mathbf{x}-\bar{\mathbf{x}})\rangle\lambda + \frac{2\alpha^2}{9}\|\mathbf{x}-\bar{\mathbf{x}}\|^2,\end{aligned}$$

where λ is defined in (2.10). It follows that

$$\langle \mathbf{x}-\bar{\mathbf{x}}, f''[\bar{\mathbf{x}}](\mathbf{x}-\bar{\mathbf{x}})\rangle \ge \frac{\alpha^2[9D(\mathbf{x},\mathbf{X})^2 - 2\|\mathbf{x}-\bar{\mathbf{x}}\|^2]}{18(\lambda+1)}.$$

(1 is added to the denominator to allow for the possibility that $\lambda = 0$.) Using this in (2.4) yields

$$\begin{aligned}f(\mathbf{x}) - f^* &= \frac{1}{2}\langle(\mathbf{x}-\bar{\mathbf{x}}), f''[\bar{\mathbf{x}}](\mathbf{x}-\bar{\mathbf{x}})\rangle + R_2(\mathbf{x},\bar{\mathbf{x}}) \\ &\ge \frac{\alpha^2[9D(\mathbf{x},\mathbf{X})^2 - 2\|\mathbf{x}-\bar{\mathbf{x}}\|^2]}{18(\lambda+1)} + R_2(\mathbf{x},\bar{\mathbf{x}}).\end{aligned} \tag{2.11}$$

By the choice of ρ in (2.9), we have

$$|R_2(\mathbf{x},\bar{\mathbf{x}})| \le \frac{\beta}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|^2, \quad \beta = \frac{7\alpha^2}{18(\lambda+1)},$$

whenever $\mathbf{x}$ and $\bar{\mathbf{x}} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$. By (2.11),

$$f(\mathbf{x}) - f^* \ge \frac{\alpha^2[9D(\mathbf{x},\mathbf{X})^2 - 2\|\mathbf{x}-\bar{\mathbf{x}}\|^2]}{18(\lambda+1)} - \frac{\beta}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|^2.$$

Minimizing $\|\mathbf{x}-\bar{\mathbf{x}}\|$ over $\bar{\mathbf{x}} \in \mathbf{X} \cap \mathcal{B}_r(\mathbf{x})$ gives

$$f(\mathbf{x}) - f^* \ge \left(\frac{\beta}{2}\right) D(\mathbf{x},\mathbf{X})^2,$$

which completes the proof. ∎

3. Convergence analysis for almost exact minimization. We first show that (C0) can always be satisfied.

LEMMA 3.1. *If $\mu_k > 0$, then there exists $\mathbf{x}_{k+1} \in \mathcal{H}$ satisfying* (C0)*; moreover, for any $\mathbf{x}_{k+1}$ satisfying* (C0)*, we have*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \le \sqrt{1+\mu_k}\, D(\mathbf{x}_k,\mathbf{X}). \tag{3.1}$$

Proof. If $\mathbf{x}_k \in \mathbf{X}$, then the lemma holds trivially since $\mathbf{x}_{k+1} = \mathbf{x}_k$. Hence, assume that $D(\mathbf{x}_k, \mathbf{X}) > 0$. Since $\mu_k > 0$, we have

$$\begin{aligned}\inf_{\mathbf{x}\in\mathbf{X}} \left\{ F_k(\mathbf{x}) + \frac{\mu_k^2}{2}\|\mathbf{x} - \mathbf{x}_k\|^2 \right\} &= f^* + \left(\frac{\mu_k + \mu_k^2}{2}\right) \inf_{\mathbf{x}\in\mathbf{X}} \{\|\mathbf{x} - \mathbf{x}_k\|\} \\ &= f^* + \left(\frac{\mu_k + \mu_k^2}{2}\right) D(\mathbf{x}_k, \mathbf{X}) \\ &> f^* + \frac{\mu_k}{2} D(\mathbf{x}_k, \mathbf{X}) \\ &= \inf_{\mathbf{x}\in\mathbf{X}} F_k(\mathbf{x}) \geq \inf_{\mathbf{x}\in\mathcal{H}} F_k(\mathbf{x}).\end{aligned}$$

Since one of these inequalities is strict, there exists $\mathbf{x}_{k+1} \in \mathcal{H}$ satisfying (C0). Moreover, for all $\mathbf{x} \in \mathbf{X}$, (C0) yields

$$\begin{aligned}F_k(\mathbf{x}_{k+1}) = f(\mathbf{x}_{k+1}) + \frac{\mu_k}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 &\leq F_k(\mathbf{x}) + \frac{\mu_k^2}{2}\|\mathbf{x} - \mathbf{x}_k\|^2 \\ &= f^* + \frac{\mu_k + \mu_k^2}{2}\|\mathbf{x} - \mathbf{x}_k\|^2.\end{aligned}$$

Since $f^* \leq f(\mathbf{x}_{k+1})$, we conclude that

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \sqrt{1 + \mu_k}\|\mathbf{x} - \mathbf{x}_k\| \tag{3.2}$$

for all $\mathbf{x} \in \mathbf{X}$. Taking the infimum over $\mathbf{x} \in \mathbf{X}$ gives (3.1). ∎

Iterates which satisfy the criterion (C0) are now analyzed.

THEOREM 3.2. *Assume the following conditions are satisfied:*

(E0) *f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ with positive scalars α and ρ satisfying* (1.3).

(E1) *$\beta > 0$ is small enough that the following inequalities hold:*

$$\frac{\beta + 2\beta^2}{2} \leq \frac{\alpha}{3} \quad \textit{and} \quad \gamma := \frac{\beta\sqrt{3(1+\beta)(3+4\alpha)}}{2\alpha} < 1.$$

(E2) $\mu_k \in (0, \beta]$.

(E3) $\mathbf{x}_0$ *is close enough to* $\hat{\mathbf{x}}$ *that*

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}\|\left(1 + \frac{\sqrt{1+\beta}}{1-\gamma}\right) \leq \rho.$$

Then any proximal iterates $\{\mathbf{x}_k\}$ satisfying (C0) *have the property that $\mathbf{x}_k \in \mathcal{B}_\rho(\hat{\mathbf{x}})$ for each k, and they approach a minimizer $\mathbf{x}^* \in \mathbf{X}$; moreover, for each k, we have*

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq c_1\gamma^k D(\mathbf{x}_0, \mathbf{X}) \quad \textit{and} \quad D(\mathbf{x}_{k+1}, \mathbf{X}) \leq c_2\mu_k D(\mathbf{x}_k, \mathbf{X}), \tag{3.3}$$

where

$$c_1 = \frac{\sqrt{1+\beta}}{1-\gamma} \quad \textit{and} \quad c_2 = \gamma/\beta. \tag{3.4}$$

Proof. For $j = 0$, (E3) implies that

$$\|\mathbf{x}_j - \hat{\mathbf{x}}\| \leq \rho \quad \text{and} \quad D(\mathbf{x}_j, \mathbf{X}) \leq \gamma^j D(\mathbf{x}_0, \mathbf{X}). \tag{3.5}$$

Proceeding by induction, suppose that (3.5) holds for all $j \in [0,k]$ and for some $k \geq 0$. We show that (3.5) also holds for $j = k+1$. By the triangle inequality, Lemma 3.1, (E2), and the induction hypothesis, it follows that

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}_0\| &\leq \sum_{j=0}^{k} \|\mathbf{x}_{j+1} - \mathbf{x}_j\| \leq \sum_{j=0}^{k} \sqrt{1+\mu_j} D(\mathbf{x}_j, \mathbf{X}) \\ &\leq \sqrt{1+\beta} \sum_{j=0}^{k} \gamma^j D(\mathbf{x}_0, \mathbf{X}) \leq \frac{\sqrt{1+\beta}}{1-\gamma} D(\mathbf{x}_0, \mathbf{X}) \leq \frac{\sqrt{1+\beta}}{1-\gamma} \|\mathbf{x}_0 - \hat{\mathbf{x}}\|.\end{aligned}$$

Again, by the triangle inequality and (E3),

$$\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\| \leq \|\mathbf{x}_{k+1} - \mathbf{x}_0\| + \|\mathbf{x}_0 - \hat{\mathbf{x}}\| \leq \left(1 + \frac{\sqrt{1+\beta}}{1-\gamma}\right) \|\mathbf{x}_0 - \hat{\mathbf{x}}\| \leq \rho. \tag{3.6}$$

For any $\mathbf{x} \in \mathcal{H}$, observe that

$$\begin{aligned}&\|\mathbf{x} - \mathbf{x}_k\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &\quad = \langle \mathbf{x} + \mathbf{x}_{k+1} - 2\mathbf{x}_k, \mathbf{x} - \mathbf{x}_{k+1} \rangle \\ &\quad \leq (\|\mathbf{x} - \mathbf{x}_{k+1}\| + 2\|\mathbf{x}_{k+1} - \mathbf{x}_k\|) \|\mathbf{x} - \mathbf{x}_{k+1}\|. \qquad (3.7)\end{aligned}$$

Rearranging (C0) and utilizing (3.7) gives, for all $\mathbf{x} \in \mathbf{X}$,

$$\begin{aligned}&f(\mathbf{x}_{k+1}) - f^* \\ &\quad \leq \frac{\mu_k}{2}(\|\mathbf{x} - \mathbf{x}_k\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2) + \frac{\mu_k^2}{2}\|\mathbf{x} - \mathbf{x}_k\|^2 \\ &\quad \leq \frac{\mu_k}{2}(\|\mathbf{x} - \mathbf{x}_{k+1}\| + 2\|\mathbf{x}_{k+1} - \mathbf{x}_k\|)\|\mathbf{x} - \mathbf{x}_{k+1}\| + \frac{\mu_k^2}{2}\|\mathbf{x} - \mathbf{x}_k\|^2 \\ &\quad \leq \frac{\mu_k}{2}(\|\mathbf{x} - \mathbf{x}_{k+1}\| + 2\|\mathbf{x}_{k+1} - \mathbf{x}_k\|)\|\mathbf{x} - \mathbf{x}_{k+1}\| + \frac{\mu_k^2}{2}(\|\mathbf{x} - \mathbf{x}_{k+1}\| + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|)^2 \\ &\quad \leq \frac{\mu_k}{2}\|\mathbf{x} - \mathbf{x}_{k+1}\|^2 + \mu_k\|\mathbf{x}_{k+1} - \mathbf{x}_k\|\|\mathbf{x} - \mathbf{x}_{k+1}\| \\ &\qquad\quad + \mu_k^2(\|\mathbf{x} - \mathbf{x}_{k+1}\|^2 + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2).\end{aligned}$$

Utilizing the inequalities

$$\mu_k\|\mathbf{x}_{k+1} - \mathbf{x}_k\|\|\mathbf{x} - \mathbf{x}_{k+1}\| \leq \frac{\alpha}{3}\|\mathbf{x} - \mathbf{x}_{k+1}\|^2 + \frac{3\mu_k^2}{4\alpha}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$$

and $\mu_k \leq \beta$, we obtain

$$f(\mathbf{x}_{k+1}) - f^* \leq \left(\frac{\beta + 2\beta^2}{2} + \frac{\alpha}{3}\right)\|\mathbf{x} - \mathbf{x}_{k+1}\|^2 + \left(\frac{3+4\alpha}{4\alpha}\right)\mu_k^2\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2. \tag{3.8}$$

Taking the infimum over $\mathbf{x} \in \mathbf{X}$ on the right-hand side of (3.8) gives

$$f(\mathbf{x}_{k+1}) - f^* \leq \left(\frac{\beta + 2\beta^2}{2} + \frac{\alpha}{3}\right)D(\mathbf{x}_{k+1}, \mathbf{X})^2 + \left(\frac{3+4\alpha}{4\alpha}\right)\mu_k^2\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2. \tag{3.9}$$

By (3.6), $\mathbf{x}_{k+1} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$. Since f provides a local error bound at $\hat{\mathbf{x}}$,

$$\alpha D(\mathbf{x}_{k+1}, \mathbf{X})^2 \leq f(\mathbf{x}_{k+1}) - f^*. \tag{3.10}$$

Combining this with (3.9) gives

$$\left(\frac{2\alpha}{3} - \frac{\beta + 2\beta^2}{2}\right) D(\mathbf{x}_{k+1}, \mathbf{X})^2 \le \left(\frac{3+4\alpha}{4\alpha}\right) \mu_k^2 \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2. \tag{3.11}$$

By (E1), the coefficient of D in (3.11) is bounded from below by $\alpha/3$. Hence, (3.11), Lemma 3.1, and (3.5), with $j = k$, yield

$$\begin{aligned} D(\mathbf{x}_{k+1}, \mathbf{X}) &\le \mu_k \frac{\sqrt{3(3+4\alpha)}}{2\alpha} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \\ &\le \mu_k \sqrt{1+\mu_k} \frac{\sqrt{3(3+4\alpha)}}{2\alpha} D(\mathbf{x}_k, \mathbf{X}) \\ &\le \mu_k \frac{\sqrt{3(1+\beta)(3+4\alpha)}}{2\alpha} D(\mathbf{x}_k, \mathbf{X}) \qquad (3.12) \\ &\le \gamma D(\mathbf{x}_k, \mathbf{X}) \le \gamma^{k+1} D(\mathbf{x}_0, \mathbf{X}). \qquad (3.13) \end{aligned}$$

Relations (3.6) and (3.13) complete the proof of the induction step. Relations (3.12) and (3.13) give the estimate (3.3).

By Lemma 3.1 and (3.5), the proximal iterates $\mathbf{x}_k$ form a Cauchy sequence in $\mathcal{H}$, which has a limit denoted $\mathbf{x}^*$. By (3.5), Lemma 3.1, and the bound $\mu_k \le \beta$, we have

$$\begin{aligned} \|\mathbf{x}_k - \mathbf{x}^*\| &\le \sum_{j=k}^{\infty} \|\mathbf{x}_{j+1} - \mathbf{x}_j\| \le \sum_{j=k}^{\infty} \sqrt{1+\mu_k} D(\mathbf{x}_j, \mathbf{X}) \\ &\le \sqrt{1+\beta} \sum_{j=k}^{\infty} \gamma^j D(\mathbf{x}_0, \mathbf{X}) = \gamma^j \frac{\sqrt{1+\beta}}{1-\gamma} D(\mathbf{x}_0, \mathbf{X}). \qquad (3.14) \end{aligned}$$

By (3.5) and (3.14), for any $k \ge 0$ we have

$$D(\mathbf{x}^*, \mathbf{X}) \le D(\mathbf{x}_k, \mathbf{X}) + \|\mathbf{x}_k - \mathbf{x}^*\| \le \left(\gamma^k + \gamma^k \frac{\sqrt{1+\beta}}{1-\gamma}\right) D(\mathbf{x}_0, \mathbf{X}).$$

Thus $D(\mathbf{x}^*, \mathbf{X}) = 0$. Since $\mathbf{X}$ is closed, the limit $\mathbf{x}^* \in \mathbf{X}$. ∎

We now give a choice for μ_k which leads to a quadratic convergence rate for the proximal point iteration.

COROLLARY 3.3. *Assume that conditions* (E0), (E1), *and* (E3) *of Theorem* 3.2 *are satisfied. In addition, let* $e : \mathcal{H} \mapsto \mathbb{R}$ *be any nonnegative function with the property that*

$$e(\mathbf{x}) \le \beta \quad \text{and} \quad e(\mathbf{x}) \le L D(\mathbf{x}, \mathbf{X}) \tag{3.15}$$

for all $\mathbf{x} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$ *and for some* $L \in \mathbb{R}$. *Then for the choice* $\mu_k = e(\mathbf{x}_k)$, *any proximal iterates* $\{\mathbf{x}_k\}$ *satisfying* (C0) *have the property that* $\mathbf{x}_k \in \mathcal{B}_\rho(\hat{\mathbf{x}})$ *for each* k, *the iterates approach a minimizer* $\mathbf{x}^* \in \mathbf{X}$, *and for each* k, *we have*

$$D(\mathbf{x}_{k+1}, \mathbf{X}) \le c_2 L D(\mathbf{x}_k, \mathbf{X})^2, \tag{3.16}$$

where c_2 *is given in* (3.4).

Proof. This follows directly from the proof of Theorem 3.2; simply append the condition $\mu_j \le \beta$ for each $j \in [0, k]$ to the induction hypothesis (3.5):

$$\|\mathbf{x}_j - \hat{\mathbf{x}}\| \le \rho, \quad D(\mathbf{x}_j, \mathbf{X}) \le \gamma^j D(\mathbf{x}_0, \mathbf{X}), \quad \text{and} \quad \mu_j \le \beta. \tag{3.17}$$

Since $\mathbf{x}_0 \in \mathcal{B}_\rho(\hat{\mathbf{x}})$, (3.15) implies that $\mu_0 = e(\mathbf{x}_0) \le \beta$. Hence, (3.17) is satisfied for $j = 0$. In the proof of Theorem 3.2, we show that if $\mu_j \le \beta$ for $j \in [0, k]$, then the first two conditions in (3.17) hold for $j = k+1$. In (3.6), we show that $\mathbf{x}_{k+1} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$. Consequently, $\mu_{k+1} = e(\mathbf{x}_{k+1}) \le \beta$, and (3.17) holds for $j = k+1$. Replacing μ_k by $e(\mathbf{x}_k) \le LD(\mathbf{x}_k, \mathbf{X})$ in (3.3) gives (3.16). □

If f is Lipschitz continuously differentiable, then the function $e(\mathbf{x}) = \|f'[\mathbf{x}]\|$ satisfies the hypotheses of Corollary 3.3 when ρ is sufficiently small since $f'[\bar{\mathbf{x}}] = \mathbf{0}$ for all $\bar{\mathbf{x}} \in \mathbf{X}$.

4. Convergence analysis for approximate minimization. We now analyze the situation where the proximal point iteration (1.2) need only satisfy (C1) or (C2). The following property of a convex function is used in the analysis.

PROPOSITION 4.1. *If $\mathbf{x}^*$ is a local minimizer of F_k and f is convex in $\mathcal{B}_\rho(\mathbf{x}^*)$ for some $\rho > 0$, then*

$$F_k(\mathbf{x}) \le F_k(\mathbf{x}^*) + \frac{\|\partial F_k(\mathbf{x})\|_{\inf}^2}{\mu_k} \tag{4.1}$$

for all $\mathbf{x} \in \mathcal{B}_\rho(\mathbf{x}^*)$.

Proof. If $\partial F_k(\mathbf{x})$ is empty, then $\|\partial F_k(\mathbf{x})\|_{\inf} = \infty$, and there is nothing to prove. Hence, we assume that $\partial F_k(\mathbf{x}) \ne \emptyset$. Since f is convex in $\mathcal{B}_\rho(\mathbf{x}^*)$, we have

$$F_k(\mathbf{x}^*) \ge F_k(\mathbf{x}) + \langle \mathbf{y}, \mathbf{x}^* - \mathbf{x} \rangle \quad \text{for all } \mathbf{y} \in \partial F_k(\mathbf{x}). \tag{4.2}$$

For a convex functional, the subdifferentials satisfy the monotonicity condition [17, Thm. 3.56]

$$\langle \bar{\mathbf{y}} - \mathbf{y}^*, \mathbf{x} - \mathbf{x}^* \rangle \ge 0 \quad \text{for all } \bar{\mathbf{y}} \in \partial f(\mathbf{x}) \quad \text{and} \quad \mathbf{y}^* \in \partial f(\mathbf{x}^*). \tag{4.3}$$

Given $\bar{\mathbf{y}} \in \partial f(\mathbf{x})$, define

$$\mathbf{y} = \bar{\mathbf{y}} + \mu_k(\mathbf{x} - \mathbf{x}_k) \in \partial F_k(\mathbf{x}). \tag{4.4}$$

Since $\mathbf{x}^*$ is a local minimizer of F_k, $\mathbf{0} \in \partial F_k(\mathbf{x}^*)$, or equivalently, there exists $\mathbf{y}^* \in \partial f(\mathbf{x}^*)$ such that

$$\mathbf{0} = \mathbf{y}^* + \mu_k(\mathbf{x}^* - \mathbf{x}_k). \tag{4.5}$$

By (4.3), (4.4), and (4.5), we have

$$\begin{aligned} \langle \mathbf{y}, (\mathbf{x} - \mathbf{x}^*) \rangle &= \langle \bar{\mathbf{y}} + \mu_k(\mathbf{x} - \mathbf{x}_k) - (\mathbf{y}^* + \mu_k(\mathbf{x}^* - \mathbf{x}_k)), \mathbf{x} - \mathbf{x}^* \rangle \\ &= \langle \bar{\mathbf{y}} - \mathbf{y}^*, \mathbf{x} - \mathbf{x}^* \rangle + \mu_k \|\mathbf{x} - \mathbf{x}^*\|^2 \\ &\ge \mu_k \|\mathbf{x} - \mathbf{x}^*\|^2 \end{aligned} \tag{4.6}$$

for any $\mathbf{y} \in \partial F_k(\mathbf{x})$. The Schwarz inequality yields

$$\|\mathbf{x} - \mathbf{x}^*\| \le \frac{\|\mathbf{y}\|}{\mu_k}. \tag{4.7}$$

Thus, it follows from (4.2) and (4.7) that for any $\mathbf{y} \in \partial F_k(\mathbf{x})$,

$$F_k(\mathbf{x}) \le F_k(\mathbf{x}^*) + \frac{\|\mathbf{y}\|^2}{\mu_k}.$$

Minimizing over $\mathbf{y} \in \partial F_k(\mathbf{x})$ gives (4.1). □

In our first convergence result, we focus on the case where f is convex over the level set defined by the starting guess. We also employ a subdifferential generalization of the gradient-based local error bound condition (2.1): For some $\hat{\mathbf{x}} \in \mathbf{X}$, there exist positive constant α and ρ such that

$$\|\partial f(\mathbf{x})\|_{\inf} \geq \alpha D(\mathbf{x}, \mathbf{X}) \quad \text{whenever } \|\mathbf{x} - \hat{\mathbf{x}}\| \leq \rho. \tag{4.8}$$

If (4.8) is satisfied, then we say that ∂f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$.

THEOREM 4.2. *Assume that the following conditions are satisfied:*

(A0) *∂f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ with positive scalars α and ρ satisfying* (4.8).

(A1) *If $\mathcal{L}$ is the level set $\{\mathbf{x} \in \mathcal{H} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$, then f is convex and lower semicontinuous on $\mathcal{L}$, and there exists a constant L such that $\|\partial f(\mathbf{x})\|_{\inf} \leq LD(\mathbf{x}, \mathbf{X})$ for all $\mathbf{x} \in \mathcal{L}$.*

(A2) *Define the parameters*

$$\Lambda = L + \tau \quad \text{and} \quad \tau^2 = 1 + 2L^2 \quad \text{if acceptance criterion (C1) is used,}$$

while

$$\Lambda = \tau(1 + \theta) \quad \text{and} \quad \tau^2 = \frac{1}{1 - 2\theta^2} \quad \text{if acceptance criterion (C2) is used.}$$

$\beta > 0$ is small enough that the following inequality holds:

$$\gamma := \frac{\beta\Lambda}{\alpha} < 1. \tag{4.9}$$

(A3) *$\mu_k \in (0, \beta]$ and $\theta < 1/\sqrt{2}$.*

(A4) *$\mathbf{x}_0$ is close enough to $\hat{\mathbf{x}}$ that*

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}\| \left(1 + \frac{\tau}{1 - \gamma}\right) \leq \rho.$$

If the approximate proximal iterates $\mathbf{x}_k$ satisfy either (C1) *or* (C2)*, then the iterates are all contained in $\mathcal{B}_\rho(\hat{\mathbf{x}})$, and they approach a minimizer $\mathbf{x}^* \in \mathbf{X}$; moreover, for each k, we have*

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq c_1 \gamma^k D(\mathbf{x}_0, \mathbf{X}) \quad \text{and} \quad D(\mathbf{x}_{k+1}, \mathbf{X}) \leq c_2 \mu_k D(\mathbf{x}_k, \mathbf{X}), \tag{4.10}$$

where

$$c_1 = \frac{\tau}{1 - \gamma} \quad \text{and} \quad c_2 = \gamma/\beta.$$

Proof. For $j = 0$, we have

$$\|\mathbf{x}_j - \hat{\mathbf{x}}\| \leq \rho, \quad \mathbf{x}_j \in \mathcal{L}, \quad \text{and} \quad D(\mathbf{x}_j, \mathbf{X}) \leq \gamma^j D(\mathbf{x}_0, \mathbf{X}). \tag{4.11}$$

Proceeding by induction, suppose that (4.11) holds for all $j \in [0, k]$ and for some $k \geq 0$. We show that (4.11) also holds for $j = k + 1$.

Due to the convexity and lower semicontinuity of f on $\mathcal{L}$, this level set is closed and convex. Suppose $j \in [0, k]$. By (C1) or (C2), we have

$$f(\mathbf{x}_{j+1}) \leq F_j(\mathbf{x}_{j+1}) \leq f(\mathbf{x}_j). \tag{4.12}$$

We conclude that $f(\mathbf{x}_j) \le f(\mathbf{x}_0)$ for each j. Since $F_j(\mathbf{x}_j) = f(\mathbf{x}_j) \le f(\mathbf{x}_0)$, minimizing F_j over $\mathcal{H}$ is equivalent to minimizing F_j over $\mathcal{L}$. Since F_j is strongly convex and lower semicontinuous on $\mathcal{L}$, F_j is weakly lower semicontinuous on $\mathcal{L}$, and there exists an exact proximal point iterate $\mathbf{x}_j^*$ defined by

$$\mathbf{x}_j^* \in \arg\ \min \{F_j(\mathbf{x}) : \mathbf{x} \in \mathcal{H}\}.$$

Moreover, $f(\mathbf{x}_j^*) \le f(\mathbf{x}_j) \le f(\mathbf{x}_0)$. Combining this with (4.12), both $\mathbf{x}_{j+1}$ and $\mathbf{x}_j^*$ lie in $\mathcal{L}$. By (A2) and Proposition 4.1, we have

$$\begin{aligned} F_j(\mathbf{x}_{j+1}) &= F_j(\mathbf{x}_j^*) + (F_j(\mathbf{x}_{j+1}) - F_j(\mathbf{x}_j^*)) \\ &\le F_j(\mathbf{x}_j^*) + \frac{\|\partial F_j(\mathbf{x}_{j+1})\|_{\inf}^2}{\mu_j} \\ &\le f^* + \frac{\mu_j}{2} D(\mathbf{x}_j, \mathbf{X})^2 + \frac{\|\partial F_j(\mathbf{x}_{j+1})\|_{\inf}^2}{\mu_j}. \end{aligned}$$

Since $f^* \le f(\mathbf{x}_{j+1})$, it follows that

$$\frac{\mu_j}{2}\|\mathbf{x}_{j+1} - \mathbf{x}_j\|^2 \le \frac{\mu_j}{2} D(\mathbf{x}_j, \mathbf{X})^2 + \frac{\|\partial F_j(\mathbf{x}_{j+1})\|_{\inf}^2}{\mu_j}. \tag{4.13}$$

By (C1) and (A1),

$$\|\partial F_j(\mathbf{x}_{j+1})\|_{\inf} \le \mu_j \|\partial f(\mathbf{x}_j)\|_{\inf} \le \mu_j L D(\mathbf{x}_j, \mathbf{X}). \tag{4.14}$$

Combining this with (4.13), we have

$$\|\mathbf{x}_{j+1} - \mathbf{x}_j\|^2 \le (1 + 2L^2) D(\mathbf{x}_j, \mathbf{X})^2. \tag{4.15}$$

Similarly, if criterion (C2) is used, then $\|\partial F_j(\mathbf{x}_{j+1})\|_{\inf} \le \theta\mu_j \|\mathbf{x}_{j+1} - \mathbf{x}_j\|$, and by (4.13), we have

$$\|\mathbf{x}_{j+1} - \mathbf{x}_j\|^2 \le \frac{1}{1 - 2\theta^2} D(\mathbf{x}_j, \mathbf{X})^2. \tag{4.16}$$

Together, (4.15) and (4.16) yield

$$\|\mathbf{x}_{j+1} - \mathbf{x}_j\| \le \tau D(\mathbf{x}_j, \mathbf{X}), \tag{4.17}$$

where τ is defined in (A2); this holds for any $j \in [0, k]$.

By (4.11), we have

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}_0\| &\le \sum_{j=0}^{k} \|\mathbf{x}_{j+1} - \mathbf{x}_j\| \le \sum_{j=0}^{k} \tau D(\mathbf{x}_j, \mathbf{X}) \\ &\le \tau \sum_{j=0}^{k} \gamma^j D(\mathbf{x}_0, \mathbf{X}) \le \frac{\tau}{1-\gamma} D(\mathbf{x}_0, \mathbf{X}) \le \frac{\tau}{1-\gamma} \|\mathbf{x}_0 - \hat{\mathbf{x}}\|. \end{aligned}$$

Again, by the triangle inequality and (A4),

$$\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\| \le \|\mathbf{x}_{k+1} - \mathbf{x}_0\| + \|\mathbf{x}_0 - \hat{\mathbf{x}}\| \le \left(1 + \frac{\tau}{1-\gamma}\right) \|\mathbf{x}_0 - \hat{\mathbf{x}}\| \le \rho.$$

Hence, $\mathbf{x}_{k+1} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$, which establishes the first relation in (4.11).

By (A0), we have

$$\alpha D(\mathbf{x}_{k+1}, \mathbf{X}) \le \|\partial f(\mathbf{x}_{k+1})\|_{\inf} \le \|\partial F_k(\mathbf{x}_{k+1})\|_{\inf} + \mu_k \|\mathbf{x}_{k+1} - \mathbf{x}_k\|. \tag{4.18}$$

If (C1) is used, then $\|\partial F_k(\mathbf{x}_{k+1})\|_{\inf} \le \mu_k \|\partial f(\mathbf{x}_k)\|_{\inf}$; hence, (A1) and (4.18) imply that

$$\begin{aligned} \alpha D(\mathbf{x}_{k+1}, \mathbf{X}) &\le \mu_k(\|\partial f(\mathbf{x}_k)\|_{\inf} + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|) \\ &\le \mu_k(L D(\mathbf{x}_k, \mathbf{X}) + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|). \end{aligned} \tag{4.19}$$

If (C2) is used, then $\|\partial F_k(\mathbf{x}_{k+1})\|_{\inf} \le \theta \mu_k \|\mathbf{x}_{k+1} - \mathbf{x}_k\|$, and by (4.18), we have

$$\alpha D(\mathbf{x}_{k+1}, \mathbf{X}) \le \mu_k(1+\theta)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|. \tag{4.20}$$

Inserting the bound (4.17) into (4.19) or (4.20) yields the second half of (4.10). By the second half of (4.10) and (A3), we have

$$D(\mathbf{x}_{k+1}, \mathbf{X}) \le \left(\frac{\Lambda \mu_k}{\alpha}\right) D(\mathbf{x}_k, \mathbf{X}) \le \gamma D(\mathbf{x}_k, \mathbf{X}) \le \gamma^{k+1} D(\mathbf{x}_0, \mathbf{X}).$$

This establishes the last relation in (4.11) for $j = k+1$, and the proof of the induction step is complete. The proof that the $\mathbf{x}_k$ form a Cauchy sequence converging to a limit $\mathbf{x}^* \in \mathbf{X}$ and the first part of (4.10) are exactly as in Theorem 3.2. □

Suppose that $\mathbf{x}^*$ is a local minimizer of F_k, f is convex in $\mathcal{B}_\rho(\mathbf{x}^*)$ for some $\rho > 0$, and the following inequality holds:

$$f(\mathbf{x}_1) \ge f(\mathbf{x}_2) + \frac{1}{2}\langle \mathbf{y}_1 + \mathbf{y}_2, \mathbf{x}_1 - \mathbf{x}_2 \rangle \tag{4.21}$$

whenever $\mathbf{y_i} \in \partial f(\mathbf{x}_i)$, $i = 1, 2$. For example, when f is a quadratic, (4.21) is satisfied with equality. When (4.21) holds, Proposition 4.1 can be strengthened to

$$F_k(\mathbf{x}) \le F_k(\mathbf{x}^*) + \frac{\|\partial F_k(\mathbf{x})\|_{\inf}^2}{2\mu_k} \tag{4.22}$$

for all $\mathbf{x} \in \mathcal{B}_\rho(\mathbf{x}^*)$. In Theorem 4.2, we require that $\theta < 1/\sqrt{2}$ in (A3); this requirement arises at inequality (4.16) since we need to ensure that $1 - 2\theta^2 > 0$. If f satisfies (4.21), then by exploiting the stronger inequality (4.22), the restriction on θ for stopping criterion (C2) can be relaxed to $\theta < 1$.

We now relax the convexity requirement for f while strengthening the smoothness condition. We require only that $\mathbf{X}$ is locally convex, while f is locally, twice continuously differentiable. If the set $\mathcal{B}_\rho(\hat{\mathbf{x}}) \cap \mathbf{X}$ is convex for some $\rho > 0$, then the projection $\bar{\mathbf{x}}$ of $\mathbf{x}$ onto $\mathcal{B}_\rho(\hat{\mathbf{x}}) \cap \mathbf{X}$ exists. For $\mathbf{x} \in \mathcal{B}_{\rho/2}(\hat{\mathbf{x}})$, it follows that $\|\mathbf{x} - \mathbf{y}\| \ge \rho/2$ when $\mathbf{y} \in \mathcal{B}_\rho(\hat{\mathbf{x}})^c$, where c denotes complement. Hence, the distance from $\mathbf{x} \in \mathcal{B}_{\rho/2}(\hat{\mathbf{x}})$ to $\mathbf{X}$ is the same as the distance from $\mathbf{x}$ to $\mathbf{X} \cap \mathcal{B}_\rho(\hat{\mathbf{x}})$:

$$\|\mathbf{x} - \bar{\mathbf{x}}\| = \min\{\|\mathbf{x} - \mathbf{y}\| : \mathbf{y} \in \mathcal{B}_\rho(\hat{\mathbf{x}}) \cap \mathbf{X}\} = \min\ \{\|\mathbf{x} - \mathbf{y}\| : \mathbf{y} \in \mathbf{X}\} = D(\mathbf{x}, \mathbf{X}).$$

The following lemma plays the role of Proposition 4.1 when we remove the convexity requirement for f.

LEMMA 4.3. *Suppose f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ with positive scalars α and ρ satisfying* (1.3), *$\mathbf{X} \cap B_\rho(\hat{\mathbf{x}})$ is convex, and f is twice Lipschitz continuously*

Fréchet differentiable on $\mathcal{B}_\rho(\hat{\mathbf{x}})$. *If* $\mu_k = \beta\|f'[\mathbf{x}_k]\|^\eta$, *where* $\eta \geq 0$ *and* $\beta > 0$, *then there exist* $r \in (0, \rho/2]$ *and positive constants* C_1 *and* C_2 *with the following property: For each* $\mathbf{x}_k \in \mathcal{B}_r(\hat{\mathbf{x}})$, *we have*

$$F_k(\mathbf{x}) - F_k(\bar{\mathbf{x}}) \leq \frac{C_1}{\mu_k}\|F_k'(\mathbf{x})\|^2 \tag{4.23}$$

whenever $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$ *and* $\|\mathbf{x} - \bar{\mathbf{x}}\| \leq C_2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta$.

Proof. If $\mathbf{x}_k = \bar{\mathbf{x}}_k$, then (4.23) is trivial. Hence, we assume that $\mathbf{x}_k \neq \bar{\mathbf{x}}_k$. By Lemma 2.1, f' provides a local error bound with constants α and $r \in (0, \rho/2]$. Hence, for any $\mathbf{x}_k \in \mathcal{B}_r(\hat{\mathbf{x}})$, we have

$$\alpha\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| = \alpha D(\mathbf{x}_k, \mathbf{X}) \leq \|f'[\mathbf{x}_k]\|.$$

Raising this inequality to the η power and utilizing the definition of μ_k gives

$$\mu_k \geq \beta\alpha^\eta\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta. \tag{4.24}$$

By the second-order necessary optimality condition, the second derivative operator $f''[\mathbf{x}]$ is positive for any $\mathbf{x} \in \mathbf{X}$. Hence, given any $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$ and $\mathbf{y} \in \mathcal{H}$, we deduce from (4.24) that

$$\begin{aligned}\langle \mathbf{y}, F_k''(\mathbf{x})\mathbf{y}\rangle &= \langle \mathbf{y}, (f''[\bar{\mathbf{x}}] + \mu_k\mathbf{I} + f''[\mathbf{x}] - f''[\bar{\mathbf{x}}])\mathbf{y}\rangle \\ &\geq \mu_k\|\mathbf{y}\|^2 + \langle \mathbf{y}, (f''[\mathbf{x}] - f''[\bar{\mathbf{x}}])\mathbf{y}\rangle \\ &\geq \Big(\beta\alpha^\eta\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta - L_2\|\mathbf{x} - \bar{\mathbf{x}}\|\Big)\|\mathbf{y}\|^2,\end{aligned}$$

where L_2 is a Lipschitz constant for f'' on $\mathcal{B}_\rho(\hat{\mathbf{x}})$. Hence, if $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$ satisfies

$$\|\mathbf{x} - \bar{\mathbf{x}}\| \leq C_2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta, \quad C_2 = \frac{\beta\alpha^\eta}{2L_2}, \tag{4.25}$$

then we have

$$\langle \mathbf{y}, F_k''(\mathbf{x})\mathbf{y}\rangle \geq \frac{\beta\alpha^\eta}{2}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta\|\mathbf{y}\|^2. \tag{4.26}$$

Let $\mathcal{A}$ be the collection of $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$ which satisfies (4.25):

$$\mathcal{A} = \{\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}}) : \|\mathbf{x} - \bar{\mathbf{x}}\| \leq C_2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta\}.$$

We now show that $\mathcal{A}$ is closed and convex. Since $r \leq \rho/2$, it follows from the discussion preceding the lemma that for each $\mathbf{y}$ and $\mathbf{z} \in \mathcal{B}_r(\hat{\mathbf{x}})$, the projections $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$ onto $\mathbf{X}$ exist in $\mathcal{B}_\rho(\hat{\mathbf{x}})$. By the convexity of $\mathcal{B}_\rho(\hat{\mathbf{x}}) \cap \mathbf{X}$, the line segment $[\bar{\mathbf{y}}, \bar{\mathbf{z}}]$ is contained in $\mathcal{B}_\rho(\hat{\mathbf{x}}) \cap \mathbf{X}$. Thus, if $\mathbf{y}$ and $\mathbf{z} \in \mathcal{A}$, then each $\mathbf{x} \in [\mathbf{y}, \mathbf{z}]$ lies in $\mathcal{A}$. $\mathcal{A}$ is closed since the projection onto a convex set is Lipschitz continuous.

By (4.26), F_k is strongly convex over the closed, convex set $\mathcal{A}$. Consequently, there exists a unique minimizer $\mathbf{x}_k^*$:

$$\mathbf{x}_k^* = \arg\min\{F_k(\mathbf{x}) : \mathbf{x} \in \mathcal{A}\}. \tag{4.27}$$

Given $\mathbf{x} \in \mathcal{A}$ and $t \in [0, 1]$, we define $\mathbf{x}(t) = \mathbf{x}_k^* + t(\mathbf{x} - \mathbf{x}_k^*)$. Since $\mathcal{A}$ is convex and both $\mathbf{x}$ and $\mathbf{x}_k^* \in \mathcal{A}$, it follows that $\mathbf{x}(t) \in \mathcal{A}$ for all $t \in [0, 1]$. Since $\mathbf{x}(0) = \mathbf{x}_k^*$

achieves the minimum in (4.27), we have $F_k(\mathbf{x}_k^*) \le F_k(\mathbf{x}(t))$ for all $t \in [0,1]$. Thus for $\phi(t) := F_k(\mathbf{x}(t))$, we have $\phi'(0) \ge 0$, and by (4.26),

$$\begin{aligned}\|F_k'(\mathbf{x})\|\|\mathbf{x}-\mathbf{x}_k^*\| &\ge F_k'(\mathbf{x})(\mathbf{x}-\mathbf{x}_k^*) = \phi'(1) \ge \phi'(1)-\phi'(0) = \phi''(s)\\ &= \langle F_k''(\mathbf{x}(s))(\mathbf{x}-\mathbf{x}_k^*), \mathbf{x}-\mathbf{x}_k^*\rangle\\ &\ge \frac{\beta\alpha^\eta}{2}\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^\eta\|\mathbf{x}-\mathbf{x}_k^*\|^2\end{aligned}$$

for some $s \in [0,1]$. Hence, we have

$$\|\mathbf{x}-\mathbf{x}_k^*\| \le \frac{2}{\beta\alpha^\eta\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^\eta}\|F_k'(\mathbf{x})\|. \tag{4.28}$$

By (4.26), $F_k(\mathbf{x}(t))$ is convex as a function of $t \in [0,1]$. This implies that $\phi'(t)$ is an increasing function of $t \in [0,1]$. We combine this observation with (4.27) and (4.28) to obtain

$$\begin{aligned}F_k(\mathbf{x}) - F_k(\bar{\mathbf{x}}_k) &\le F_k(\mathbf{x}) - F_k(\mathbf{x}_k^*)\\ &= F_k(\mathbf{x}(1)) - F_k(\mathbf{x}(0))\\ &= \int_0^1 \phi'(t)dt \le \phi'(1) = F_k'(\mathbf{x})(\mathbf{x}-\mathbf{x}_k^*)\\ &\le \frac{2}{\beta\alpha^\eta\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^\eta}\|F_k'(\mathbf{x})\|^2 \qquad (4.29)\end{aligned}$$

whenever $\mathbf{x} \in \mathcal{A}$ and $\mathbf{x}_k \in \mathcal{B}_r(\hat{\mathbf{x}})$.

If L_1 is a Lipschitz constant for f' over $\mathcal{B}_\rho(\hat{\mathbf{x}})$, then we have

$$\|f'[\mathbf{x}_k]\| = \|f'[\mathbf{x}_k] - f'[\bar{\mathbf{x}}_k]\| \le L_1\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|$$

whenever $\mathbf{x}_k \in \mathcal{B}_r(\hat{\mathbf{x}})$. By the definition of μ_k, it follows that

$$\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta \ge \mu_k/(\beta L_1^\eta).$$

Combining this with (4.29) gives (4.23) with $C_1 = 2L_1^\eta/\alpha^\eta$. □

THEOREM 4.4. *Assume that the following conditions are satisfied:*

(B0) *f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ with positive scalars α and ρ satisfying* (1.3).

(B1) *The set $\mathcal{B}_\rho(\hat{\mathbf{x}}) \cap \mathbf{X}$ is convex.*

(B2) *f is twice Lipschitz continuously Fréchet differentiable in $\mathcal{B}_\rho(\hat{\mathbf{x}})$.*

(B3) *The proximal iterates $\mathbf{x}_k$ satisfy either* (C1) *or* (C2) *with $\mu_k = \beta\|f'[\mathbf{x}_k]\|^\eta$, where $\eta \in (0,2)$ and β is positive. If* (C2) *is used, then θ is small enough that $2C_1\theta^2 < 1$, where C_1 is the constant in* (4.23).

Then for ϵ sufficiently small and for each $\mathbf{x}_0 \in \mathcal{B}_\epsilon(\hat{\mathbf{x}})$, the proximal iterates $\mathbf{x}_k$ are all contained in $\mathcal{B}_\rho(\hat{\mathbf{x}})$, and they approach a minimizer $\mathbf{x}^* \in \mathbf{X}$; moreover, for each k, we have

$$\begin{aligned}&\|\mathbf{x}_k - \mathbf{x}^*\| \le c_1\gamma^k D(\mathbf{x}_0, \mathbf{X}) \quad \textit{and}\\ &D(\mathbf{x}_{k+1}, \mathbf{X}) \le c_2\mu_k D(\mathbf{x}_k, \mathbf{X}) \le \beta c_2 L_1^\eta D(\mathbf{x}_k, \mathbf{X})^{1+\eta}, \qquad (4.30)\end{aligned}$$

where $\gamma < 1$, c_1, and c_2 are constants independent of k.

Proof. We start by explaining how to choose ϵ so that the theorem holds. Define the parameters

$$\Lambda = L_1 + \tau \quad \text{and} \quad \tau^2 = 1 + 2C_1L_1^2 \quad \text{if acceptance criterion (C1) is used,}$$

while

$$\Lambda = \tau(1+\theta) \quad \text{and} \quad \tau^2 = \frac{1}{1-2C_1\theta^2} \quad \text{if acceptance criterion (C2) is used,}$$

where C_1 is the constant in (4.23) and L_1 is a Lipschitz constant for f' over $\mathcal{B}_\rho(\hat{\mathbf{x}})$. Notice that the hypotheses of the theorem are satisfied if ρ is decreased. Choose ρ small enough that

$$\gamma := \sup_{\mathbf{x}\in\mathcal{B}_\rho(\hat{\mathbf{x}})} \frac{\beta\|f'[\mathbf{x}]\|^\eta\Lambda}{\alpha} < 1. \tag{4.31}$$

By Lemma 2.1, we can choose ρ smaller, if necessary, so that f' provides a local error bound with constants α and $\rho/2$. By Lemma 4.3, we can choose ρ smaller, if necessary, so that (4.23) holds whenever $\mathbf{x} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$ and $\|\mathbf{x}-\bar{\mathbf{x}}\| \le C_2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta$. Choose $\epsilon > 0$ small enough that

$$\left(\epsilon + \frac{E\epsilon^{1-\eta/2}}{1-\gamma^{1-\eta/2}}\right) \le \frac{\rho}{2}, \quad \text{where} \quad E = \sqrt{\frac{L_1}{2\beta\alpha^\eta}}, \tag{4.32}$$

and

$$\epsilon(L_1 + E\epsilon^{-\eta/2})\left(\frac{\beta L_1^\eta}{\alpha}\right) \le C_2 \quad \text{if stopping criterion (C1) is used,} \tag{4.33}$$

$$E\epsilon^{1-\eta/2}(1+\theta)\left(\frac{\beta L_1^\eta}{\alpha}\right) \le C_2 \quad \text{if stopping criterion (C2) is used.} \tag{4.34}$$

Since $\eta \in (0,2)$, (4.33) and (4.34) are satisfied for ϵ sufficiently small.

We now prove the theorem. Again, let $\bar{\mathbf{x}}$ be the projection of $\mathbf{x}$ onto $\mathbf{X}$. For $j = 0$, we have

$$\|\mathbf{x}_j - \hat{\mathbf{x}}\| \le \rho/2 \quad \text{and} \quad \|\mathbf{x}_j - \bar{\mathbf{x}}_j\| \le \gamma^j\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\| \tag{4.35}$$

since $\mathbf{x}_0 \in \mathcal{B}_\epsilon(\hat{\mathbf{x}}) \subset \mathcal{B}_{\rho/2}(\hat{\mathbf{x}})$. Proceeding by induction, suppose that (4.35) holds for all $j \in [0,k]$ and for some $k \ge 0$.

For any $j \in [0,k]$, the condition $F_j(\mathbf{x}_{j+1}) \le f(\mathbf{x}_j)$ in (C1) or (C2) implies that

$$\mu_j\|\mathbf{x}_{j+1} - \mathbf{x}_j\|^2 \le f(\mathbf{x}_j) - f(\mathbf{x}_{j+1}) \le f(\mathbf{x}_j) - f(\bar{\mathbf{x}}_j). \tag{4.36}$$

By the induction hypothesis, $\mathbf{x}_j \in \mathcal{B}_{\rho/2}(\hat{\mathbf{x}})$, and by the triangle inequality, we have

$$\|\bar{\mathbf{x}}_j - \hat{\mathbf{x}}\| \le \|\bar{\mathbf{x}}_j - \mathbf{x}_j\| + \|\mathbf{x}_j - \hat{\mathbf{x}}\| \le \|\bar{\mathbf{x}}_0 - \mathbf{x}_0\| + \frac{\rho}{2} \le \rho. \tag{4.37}$$

Hence, $\bar{\mathbf{x}}_j \in \mathcal{B}_\rho(\hat{\mathbf{x}})$. We expand f in (4.36) in a Taylor series around $\bar{\mathbf{x}}_j$ and use the fact that $f'[\bar{\mathbf{x}}_j] = \mathbf{0}$ to obtain

$$\mu_j\|\mathbf{x}_{j+1} - \mathbf{x}_j\|^2 \le \frac{L_1}{2}\|\mathbf{x}_j - \bar{\mathbf{x}}_j\|^2, \tag{4.38}$$

where L_1 is a Lipschitz constant for f'. Since f' provides a local error bound with constants α and $\rho/2$ and $\mathbf{x}_j \in \mathcal{B}_{\rho/2}(\hat{\mathbf{x}})$, it follows that

$$\mu_j = \beta \|f'[\mathbf{x}_j]\|^\eta \geq \beta\alpha^\eta \|\mathbf{x}_j - \bar{\mathbf{x}}_j\|^\eta.$$

Combining this with (4.38) gives

$$\|\mathbf{x}_{j+1} - \mathbf{x}_j\| \leq E\|\mathbf{x}_j - \bar{\mathbf{x}}_j\|^{1-\eta/2} \tag{4.39}$$

for $j \in [0, k]$, where E is defined in (4.32). By the triangle inequality, (4.32), (4.35), and (4.39), we have

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\| &\leq \|\mathbf{x}_0 - \hat{\mathbf{x}}\| + \sum_{j=0}^{k} \|\mathbf{x}_{j+1} - \mathbf{x}_j\| \\
&\leq \|\mathbf{x}_0 - \hat{\mathbf{x}}\| + E\sum_{j=0}^{k} \|\mathbf{x}_j - \bar{\mathbf{x}}_j\|^{1-\eta/2} \\
&\leq \|\mathbf{x}_0 - \hat{\mathbf{x}}\| + E\|\mathbf{x}_0 - \hat{\mathbf{x}}\|^{1-\eta/2} \sum_{j=0}^{k} (\gamma^{1-\eta/2})^j \\
&\leq \|\mathbf{x}_0 - \hat{\mathbf{x}}\| + \frac{E\|\mathbf{x}_0 - \hat{\mathbf{x}}\|^{1-\eta/2}}{1 - \gamma^{1-\eta/2}} \\
&\leq \epsilon + \frac{E\epsilon^{1-\eta/2}}{1 - \gamma^{1-\eta/2}} \leq \rho/2.
\end{aligned}$$

This establishes the first half of (4.35) for $j = k + 1$.

To establish the second half of (4.35), we will apply Lemma 4.3 to $\mathbf{x} = \mathbf{x}_{k+1}$. Since $\mathbf{x}_{k+1} \in \mathcal{B}_{\rho/2}(\hat{\mathbf{x}})$, we need only show that $\mathbf{x} = \mathbf{x}_{k+1}$ satisfies the qualification $\|\mathbf{x} - \bar{\mathbf{x}}\| \leq C_2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta$ for (4.23). Since $\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\| \leq \rho/2$ and since f' provides a local error bound at $\hat{\mathbf{x}}$ with constants α and $\rho/2$,

$$\alpha\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \leq \|f'[\mathbf{x}_{k+1}]\| \leq \|F_k'(\mathbf{x}_{k+1})\| + \mu_k\|\mathbf{x}_{k+1} - \mathbf{x}_k\|. \tag{4.40}$$

Since f' is Lipschitz continuous over $\mathcal{B}_\rho(\hat{\mathbf{x}})$, it follows from (4.39), (4.40), and the definition of μ_k that for stopping criterion (C1),

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| &\leq \frac{\mu_k}{\alpha}(\|\nabla f(\mathbf{x}_k)\| + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|) \\
&\leq \left(\frac{\beta L_1^\eta}{\alpha}(L_1\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|)\right)\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta \\
&\leq \left(\frac{\beta L_1^\eta}{\alpha}(L_1\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\| + E\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^{1-\eta/2})\right)\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta \\
&\leq \left(\frac{\beta L_1^\eta}{\alpha}(L_1\|\mathbf{x}_0 - \hat{\mathbf{x}}\| + E\|\mathbf{x}_0 - \hat{\mathbf{x}}\|^{1-\eta/2})\right)\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta \\
&\leq C_2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^\eta. \qquad (4.41)
\end{aligned}$$

The first inequality is due to (4.40) and (C1); the second inequality is based on the definition of μ_k and the Lipschitz continuity of f'; the third inequality utilizes the induction hypothesis (4.35), the bound (4.39) for $j = k$, and the fact that $\mathbf{x}_k$ and

$\bar{\mathbf{x}}_k \in \mathcal{B}_\rho(\hat{\mathbf{x}})$ (see (4.37)); and the fourth inequality is a consequence of the induction hypothesis and the fact that ϵ satisfies (4.33).

If stopping criterion (C2) is used, then in the same fashion, (4.34) gives

$$\begin{aligned}\|\mathbf{x}_{k+1}-\bar{\mathbf{x}}_{k+1}\| &\le \frac{\mu_k}{\alpha}(1+\theta)\|\mathbf{x}_{k+1}-\mathbf{x}_k\| \\ &\le \left(\frac{\beta L_1^\eta}{\alpha}(1+\theta)\|\mathbf{x}_{k+1}-\mathbf{x}_k\|\right)\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^\eta \\ &\le \left(\frac{\beta L_1^\eta}{\alpha}(1+\theta)E\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^{1-\eta/2}\right)\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^\eta \\ &\le \left(\frac{\beta L_1^\eta}{\alpha}(1+\theta)E\|\mathbf{x}_0-\hat{\mathbf{x}}\|^{1-\eta/2}\right)\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^\eta \\ &\le C_2\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^\eta. \qquad (4.42)\end{aligned}$$

Thus, if either stopping criterion (C1) or (C2) is used, then $\mathbf{x}=\mathbf{x}_{k+1}$ satisfies the qualifications of Lemma 4.3.

We now give another bound for the change $\|\mathbf{x}_{k+1}-\mathbf{x}_k\|$. Lemma 4.3 yields

$$F_k(\mathbf{x}_{k+1}) \le F_k(\bar{\mathbf{x}}_k) + \frac{C_1}{\mu_k}\|F_k'(\mathbf{x}_{k+1})\|^2. \qquad (4.43)$$

Since $f(\bar{\mathbf{x}}_k) \le f(\mathbf{x}_{k+1})$, we conclude that

$$\frac{\mu_k}{2}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 \le \frac{\mu_k}{2}\|\bar{\mathbf{x}}_k-\mathbf{x}_k\|^2 + \frac{C_1}{\mu_k}\|F_k'(\mathbf{x}_{k+1})\|^2. \qquad (4.44)$$

If (C1) is used, then we have

$$\begin{aligned}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 &\le \|\bar{\mathbf{x}}_k-\mathbf{x}_k\|^2 + 2C_1\|f'[\mathbf{x}_k]\|^2 \\ &\le \|\bar{\mathbf{x}}_k-\mathbf{x}_k\|^2 + 2C_1L_1^2\|\bar{\mathbf{x}}_k-\mathbf{x}_k\|^2.\end{aligned}$$

If (C2) is used, then (4.44) gives

$$\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 \le \frac{1}{1-2C_1\theta^2}\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^2.$$

In either case,

$$\|\mathbf{x}_{k+1}-\mathbf{x}_k\| \le \tau\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|, \qquad (4.45)$$

where τ is defined at the start of the proof.

If (C1) is used, then

$$\|F_k'(\mathbf{x}_{k+1})\| \le \mu_k\|f'[\mathbf{x}_k]\| = \mu_k\|f'[\mathbf{x}_k]-f'[\bar{\mathbf{x}}_k]\| \le \mu_k L_1\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|.$$

By (4.40) and (4.45), we have

$$\|\mathbf{x}_{k+1}-\bar{\mathbf{x}}_{k+1}\| \le \frac{\mu_k}{\alpha}(L_1+\tau)\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|. \qquad (4.46)$$

If (C2) is used, then $\|F_k'(\mathbf{x}_{k+1})\| \le \theta\mu_k\|\mathbf{x}_{k+1}-\mathbf{x}_k\|$, and by (4.40) and (4.45), we have

$$\|\mathbf{x}_{k+1}-\bar{\mathbf{x}}_{k+1}\| \le \frac{\tau\mu_k(1+\theta)}{\alpha}\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|. \qquad (4.47)$$

In either case, since $\mathbf{x}_k \in \mathcal{B}_{\rho/2}(\hat{\mathbf{x}})$, it follows from the definition of γ and μ_k that

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \leq \gamma \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|.$$

This establishes the second half of (4.35) for $j = k+1$. Hence, the proof of the induction step is complete. The proof that the $\mathbf{x}_k$ form a Cauchy sequence converging to a limit $\mathbf{x}^* \in \mathbf{X}$ is exactly as in Theorem 3.2. The first half of (4.30) follows from (4.46) and (4.47). In the second half of (4.30), we replace μ_k by $\beta \|f'[\mathbf{x}_k]\|^\eta$ and exploit the Lipschitz continuity of f'. □

Remark. During the proof of Theorem 4.4, in (4.41) and (4.42), we show that each iterate $\mathbf{x}_{k+1}$ lies in the region where F_k is convex (4.26).

When $\eta = 0$, we can drop the requirement of Theorem 4.4 that $\mathbf{X}$ is locally convex. This convexity requirement arose since we use Lemma 4.3, which assumes that $\mathbf{X}$ is locally convex. We now show that Lemma 4.3 can be established without local convexity when μ_k is bounded away from 0.

LEMMA 4.5. *Let $\beta > 0$ and suppose that $\mu_k \geq \beta$ for each k. Given $\hat{\mathbf{x}} \in \mathbf{X}$ and $\delta \in (0,1)$, suppose f is twice Lipschitz continuously Fréchet differentiable on $\mathcal{B}_\rho(\hat{\mathbf{x}})$, let L_2 be a Lipschitz constant for f'' on $\mathcal{B}_\rho(\hat{\mathbf{x}})$, and let $r = \min\{\rho, \delta\beta/L_2\}$. Then we have*

$$F_k(\mathbf{x}) - F_k^* \leq \left(\frac{1}{(1-\delta)\mu_k}\right) \|F_k'(\mathbf{x})\|^2 \tag{4.48}$$

for all $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$, where

$$F_k^* = \min_{\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})} F_k(\mathbf{x}).$$

Proof. Suppose $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$. We start with the identity

$$\langle F_k''[\mathbf{x}]\mathbf{y}, \mathbf{y}\rangle = \langle F_k''[\hat{\mathbf{x}}]\mathbf{y}, \mathbf{y}\rangle + \langle (F_k''[\mathbf{x}] - F_k''[\hat{\mathbf{x}}])\mathbf{y}, \mathbf{y}\rangle.$$

By the second-order optimality condition, $f''[\hat{\mathbf{x}}]$ is positive. Consequently, we have

$$\langle F_k''[\hat{\mathbf{x}}]\mathbf{y}, \mathbf{y}\rangle \geq \mu_k \|\mathbf{y}\|^2.$$

Since $F_k''[\mathbf{x}] - F_k''[\hat{\mathbf{x}}] = f''[\mathbf{x}] - f''[\hat{\mathbf{x}}]$, it follows from the Lipschitz continuity of f'' that

$$\langle (F_k''[\mathbf{x}] - F_k''[\hat{\mathbf{x}}])\mathbf{y}, \mathbf{y}\rangle \leq L_2 \|\mathbf{x} - \hat{\mathbf{x}}\| \, \|\mathbf{y}\|^2 \leq \delta\beta \|\mathbf{y}\|^2 \leq \delta\mu_k \|\mathbf{y}\|^2$$

since $r \leq \delta\beta/L_2$. Hence, if $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$, we have

$$\langle F_k''[\mathbf{x}]\mathbf{y}, \mathbf{y}\rangle \geq (1-\delta)\mu_k \|\mathbf{y}\|^2. \tag{4.49}$$

By (4.49), F_k is convex on $\mathcal{B}_r(\hat{\mathbf{x}})$. Consequently, the minimizer $\mathbf{x}_k^*$ over $\mathcal{B}_r(\hat{\mathbf{x}})$ exists:

$$\mathbf{x}_k^* = \arg\min\{F_k(\mathbf{x}) : \mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})\}.$$

Since $\mathcal{B}_r(\hat{\mathbf{x}})$ is a convex set, the first-order optimality condition

$$\langle F_k'[\mathbf{x}_k^*], \mathbf{x} - \mathbf{x}_k^*\rangle \geq 0$$

holds for all $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$. It follows that

$$\langle F_k'[\mathbf{x}], \mathbf{x} - \mathbf{x}_k^* \rangle \geq \langle F_k'[\mathbf{x}] - F_k'[\mathbf{x}_k^*], \mathbf{x} - \mathbf{x}_k^* \rangle.$$

We utilize the strong convexity property (4.49) to obtain

$$\langle F_k'[\mathbf{x}], \mathbf{x} - \mathbf{x}_k^* \rangle \geq (1-\delta)\mu_k \|\mathbf{x} - \mathbf{x}_k^*\|^2,$$

which gives

$$\|\mathbf{x} - \mathbf{x}_k^*\| \leq \frac{1}{(1-\delta)\mu_k} \|F_k'[\mathbf{x}]\|. \tag{4.50}$$

The convexity of F_k on $\mathcal{B}_r(\hat{\mathbf{x}})$ implies that

$$F_k(\mathbf{x}^*) \geq F_k(\mathbf{x}) + F_k'[\mathbf{x}](\mathbf{x}_k^* - \mathbf{x}).$$

Combining this with (4.50) completes the proof. ∎

Theorem 4.4 holds with the following modifications: (i) The assumption (B0) that $\mathcal{B}_\rho(\hat{\mathbf{x}}) \cap \mathbf{X}$ is convex is dropped; and (ii) $\mu_k \in [\beta_0, \beta_1]$, where β_1 is chosen small enough that the constant $\gamma = \beta_2 \Lambda/\alpha$ is less than 1. For completeness, we state the modified result.

THEOREM 4.6. *Assume that the following conditions are satisfied:*

(b0) *f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ with positive scalars α and ρ satisfying* (1.3).

(b1) *f is twice Lipschitz continuously Fréchet differentiable in $\mathcal{B}_\rho(\hat{\mathbf{x}})$.*

(b2) *The proximal iterates $\mathbf{x}_k$ satisfy either* (C1) *or* (C2) *with $\mu_k \in [\beta_0, \beta_1]$, where $\beta_0 > 0$. If* (C2) *is used, then $\theta < 1/\sqrt{2}$. $\delta \in (0,1)$ is chosen small enough that $\theta^2 < (1-\delta)/2$.*

(b3) *Define the parameters*

$$\Lambda = L_1 + \tau \quad \text{and} \quad \tau^2 = 1 + 2L_1^2/(1-\delta) \quad \text{if acceptance criterion (C1) is used,}$$

while

$$\Lambda = \tau(1+\theta) \quad \text{and} \quad \tau^2 = \frac{1}{1 - 2\theta^2/(1-\delta)} \quad \text{if acceptance criterion (C2) is used,}$$

where L_1 is a Lipschitz constant for f' over $\mathcal{B}_\rho(\hat{\mathbf{x}})$. β_1 is small enough that

$$\gamma := \frac{\beta_1 \Lambda}{\alpha} < 1.$$

Then for ϵ sufficiently small and for each $\mathbf{x}_0 \in \mathcal{B}_\epsilon(\hat{\mathbf{x}})$, the proximal iterates $\mathbf{x}_k$ are all contained in $\mathcal{B}_\rho(\hat{\mathbf{x}})$, and they approach a minimizer $\mathbf{x}^* \in \mathbf{X}$; moreover, for each k, we have

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq c_1 \gamma^k D(\mathbf{x}_0, \mathbf{X}),$$

where c_1 is a constant independent of k.

The proof of Theorem 4.6 is the same as the proof of Theorem 4.4, except that we use Lemma 4.5 instead of Lemma 4.3 and we replace expressions like $\|\mathbf{x} - \bar{\mathbf{x}}\|$ by $D(\mathbf{x}, \mathbf{X})$.

5. Final discussion. The local convergence results obtained in [7] for a new class of self-adaptive proximal point methods have been extended from a finite dimensional setting to a Hilbert space setting. In particular the local convergence estimates obtained for exact iterates in [7] are now established for approximate iterates satisfying (C0). Our analysis, which permits multiple minimizers, employs a local error bound condition (1.3) relating the growth in f to the distance from the set of minimizers. The gradient-based acceptance criteria in [7] have been replaced by subdifferential-based criteria (C1) and (C2). The local convergence results for the subdifferential-based stopping criteria are similar to the convergence results for iterates satisfying (C0). Three types of assumptions were considered in our analysis connected with (C1) and (C2): (a) f is convex and lower semicontinuous on a level set; (b) the set $\mathbf{X} \cap \mathcal{B}_\rho(\hat{\mathbf{x}})$ is convex for some $\rho > 0$, and f is twice continuously differentiable on $\mathcal{B}_\rho(\hat{\mathbf{x}})$; and (c) $\mu_k \in [\beta_0, \beta_1]$, with β_1 sufficiently small, $\beta_0 > 0$, and f twice continuously differentiable on $\mathcal{B}_\rho(\hat{\mathbf{x}})$. The conditions (b) and (c) are weaker than the local convexity requirement for f in [7].

The analysis in this paper has focused on local convergence. Global convergence issues are studied in section 6 of [7], where we also present computational results which show that for a class of ill-conditioned nonlinear optimization problem, a proximal point approach could reduce the computing time.

REFERENCES

[1] A. Auslender and M. Teboulle, *Lagrangian duality and related multiplier methods for variational inequality problems*, SIAM J. Optim., 10 (2000), pp. 1097–1115.

[2] P. L. Combettes and T. Pennanen, *Proximal methods for cohypomonotone operators*, SIAM J. Control Optim., 43 (2004), pp. 731–742.

[3] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Math. Oper. Res., 18 (1993), pp. 203–226.

[4] J. Eckstein, *Approximate iterations in Bregman-function-based proximal algorithms*, Math. Programming, 83 (1998), pp. 113–123.

[5] J. Fan and Y. Yuan, *On the convergence of the Levenberg-Marquardt method without nonsingularity assumption*, Computing, 74 (2005), pp. 23–39.

[6] C. D. Ha, *A generalization of the proximal point algorithm*, SIAM J. Control Optim., 28 (1990), pp. 503–512.

[7] W. W. Hager and H. Zhang, *Self-adaptive inexact proximal point methods*, Comput. Optim. Appl., to appear.

[8] C. Humes and P. Silva, *Inexact proximal point algorithms and descent methods in optimization*, Optim. Eng., 6 (2005), pp. 257–271.

[9] A. Iusem, B. Svaiter, and M. Teboulle, *Entropy-like proximal methods in convex programming*, Math. Oper. Res., 19 (1994), pp. 790–814.

[10] A. N. Iusem, T. Pennanen, and B. F. Svaiter, *Inexact variants of the proximal point algorithm without monotonicity*, SIAM J. Optim., 13 (2003), pp. 1080–1097.

[11] A. Kaplan and R. Tichatschke, *Stable Methods for Ill-Posed Variational Problems*, Akademie Verlag, Berlin, 1994.

[12] A. Kaplan and R. Tichatschke, *Proximal point methods and nonconvex optimization*, J. Global Optim., 13 (1998), pp. 389–406.

[13] D. Li, M. Fukushima, L. Qi, and N. Yamashita, *Regularized Newton methods for convex minimization problems with singular solutions*, Comput. Optim. Appl., 28 (2004), pp. 131–147.

[14] F. J. Luque, *Asymptotic convergence analysis of the proximal point algorithm*, SIAM J. Control Optim., 22 (1984), pp. 277–293.

[15] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle, 4 (1970), pp. 154–158.

[16] B. Martinet, *Determination approachée d'un point fixe d'une application pseudo-contractante*, C. R. Acad. Sci. Paris Sér. A-B, 274 (1972), pp. 163–165.

[17] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation* I, Springer-Verlag,

Berlin, 2006.
[18] J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[19] T. Pennanen, *Local convergence of the proximal point algorithm and multiplier methods without monotonicity*, Math. Oper. Res., 27 (2002), pp. 170–191.
[20] R. T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 2 (1976), pp. 97–116.
[21] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[22] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[23] W. Rudin, *Functional Analysis*, McGraw–Hill, New York, 1973.
[24] M. V. Solodov and B. F. Svaiter, *A hybrid extragradient-proximal point algorithm using the enlargement of a maximal monotone operator*, Set-Valued Anal., 7 (1999), pp. 323–345.
[25] M. V. Solodov and B. F. Svaiter, *A hybrid projection-proximal point algorithm*, J. Convex Anal., 6 (1999), pp. 59–70.
[26] M. V. Solodov and B. F. Svaiter, *A unified framework for some inexact proximal point algorithms*, Numer. Funct. Anal. Optim., 22 (2001), pp. 1013–1035.
[27] J. E. Spingarn, *Submonotone mappings and the proximal point algorithm*, Numer. Funct. Anal. Optim., 4 (1981), pp. 123–150.
[28] P. Tseng, *Error bounds and superlinear convergence analysis of some Newton-type methods in optimization*, in Nonlinear Optimization and Related Topics, G. Di Pillo and F. Giannessi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 445–462.
[29] N. Yamashita and M. Fukushima, *The proximal point algorithm with genuine superlinear convergence for the monotone complementarity problem*, SIAM J. Optim., 11 (2000), pp. 364–379.
[30] N. Yamashita and M. Fukushima, *On the rate of convergence of the Levenberg-Marquardt method*, in Topics in Numerical Analysis, Comput. Suppl. 15, Springer-Verlag, Vienna, 2001, pp. 239–249.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1705–1725

OPTIMAL ENERGY CONTROL IN FINITE TIME BY VARYING THE LENGTH OF THE STRING*

MARTIN GUGAT†

Abstract. We consider a finite string where, at both end points, a homogeneous Dirichlet boundary condition holds. One boundary point is fixed, and the other is moving; hence the length of the string is changing in time. The string is controlled through the movement of this boundary point. We consider movements of the boundary that are Lipschitz continuous. Only movements for which at the given finite terminal time the string has the same length as at the beginning are admissible. Moreover, we impose an upper bound for the Lipschitz constant of the movement that is smaller than the speed of wave propagation. We consider the optimal control problem to find an admissible movement for which at the given terminal time the energy of the string is minimal. We give a sufficient condition for the existence and uniqueness of an optimal movement and construct an optimal control movement.

Key words. PDE-constrained optimization, optimal control of PDEs, optimal boundary control, wave equation, optimal energy control, moving boundary, control constraint, optimal shape

AMS subject classifications. 49K20, 35L05

DOI. 10.1137/06065636x

1. Introduction.

We consider a string of finite length that is governed by the wave equation with homogeneous Dirichlet boundary conditions. The left boundary point is fixed, and the other boundary point of the string is moving. This system has already been studied in [3].

The boundary control of the wave equation has been studied by many authors (see, e.g., [16], [13], [11], [12], [1], [20], [8], and the references therein). In most studies both boundary points of the string are fixed, and the string is controlled through prescribed function values at the fixed boundary points.

In contrast to this approach, in this paper we control the system through the movement of the boundary point, that is, through the length of the string as a function of time. For the case of dimensions greater than one, this corresponds to the control of the shape as a function of time. Thus our problem is a one-dimensional (1-d) case for optimal shape control of a hyperbolic partial differential equation. The monograph [5] gives an overview about problems where moving boundaries play an essential role. Controllability of the wave equation with point control where the point is moving in the system's fixed spatial domain is studied in [9]. Observability and stabilizability of this system are considered in [10]. A problem with moving control of the heat equation is studied in [4].

The well-posedness of the wave equation in a noncylindrical, time periodical domain in $R \times R^N$ has been studied in [17]. Distributed control of the wave equation in a domain with a moving boundary has been studied in [2] for dimensions unequal to two. For these dimensions, a contraction of the domain always leads to nondecreasing energy, and an expansion always leads to nonincreasing energy (see Theorem 2.1 in [2]).

*Received by the editors April 5, 2006; accepted for publication (in revised form) July 27, 2007; published electronically November 2, 2007. This work has been supported by the PROCOPE program of DAAD, D/0427464.

http://www.siam.org/journals/sicon/46-5/65636.html

†Lehrstuhl für angewandte Mathematik AM2, Martensstr. 3, 91058 Erlangen, Germany (gugat@am.uni-erlangen.de).

The situation in dimension two is completely different. For certain expansions, the energy is nonincreasing, but also, if the interval is contracting sufficiently fast, a decay of the energy can be achieved. In [18], [19] the stabilization of the wave equation through the movement of the boundary is studied in dimension two. A movement is constructed that assures exponential decay of the energy.

In this paper, we study the 1-d-case. In [6] the stability of the string with varying length has been studied, and it has been shown that the energy cannot become arbitrarily small unless it is zero from the start. This follows from inequality (5.10), where a lower bound for the energy at time t is given that depends on the initial state. A sufficient condition for exponential growth of the energy is given in [7]. Interestingly, this situation of exponential energy growth could have applications in photon creation from the vacuum; see [14].

We consider the following optimal boundary movement control problem: Let at time $t = 0$ the length L of the string and an initial state be given. Let a terminal time T and a positive real number D that is strictly less than the speed of wave propagation be given. We admit movements of the right boundary point of the string that are given by a Lipschitz continuous function with a Lipschitz constant that is less than or equal to the number D. Moreover, we assume that at the terminal time T the string has the same length as at the beginning. For the set of Lipschitz continuous functions on the time interval $[0, T]$, we use the notation

$$Lip = \{\phi \,|\, \phi : [0,T] \to (0,\infty) \text{ such that } \phi \text{ is Lipschitz continuous } \}.$$

We consider the boundary movements in the admissible set

$$\Phi = \{\phi \in Lip \text{ with Lipschitz constant } \leq D \text{ and } \phi(0) = L,\ \phi(T) = L\}\,. \tag{1.1}$$

Our problem is to find an admissible boundary movement for which the energy of the string at the terminal time T is miminal. We present an explicit formula for optimal movements that solve this optimal control problem.

This paper has the following structure: In section 2, we define the problem of energy minimizing boundary movement control for a vibrating string with homogeneous Dirichlet boundary conditions at both ends. We assume that at the terminal time the string has the same length as at the beginning. As a control constraint, we prescribe an upper bound for the Lipschitz constants of the admissible boundary movements.

Section 3 contains our main results. In Theorem 3.2, we present an optimal boundary movement that depends on the initial state in a robust way. The energy decay that can be achieved depends on the initial state. For certain initial states, it is not possible to achieve an energy decay by boundary movement control. Theorem 3.1 contains sufficient conditions for the existence and uniqueness of optimal controls.

The proof of the main result uses the solution of the initial-boundary-value problem for a given boundary movement that is given in section 4. The construction is based upon the method of characteristics. It is necessary to make sure that the Lipschitz constants of the admissible boundary movements are strictly less than the speed of wave propagation, because, otherwise, the solution of the corresponding initial boundary value problem is, in general, not well-defined. The reason is that the information travels on the characteristic curves, and, if the length of the string increases too fast, it may happen that there exist points that are not reached by characteristic curves of both families.

In section 5, we introduce a set of functions that depend in a bijective way on the boundary movements (see Lemma 5.1). These functions are the unknowns in our

transformed optimization problem. To obtain the transformed problem, we write our objective function, that is, the energy of the string at time T in terms of the new unknowns. The solution of the transformed problem is given in Lemma 5.6.

2. The problem. Let the wave speed $c > 0$ be given. At the initial time $t = 0$, the string has the length $L > 0$. Define the control time $T = 2L/c$. Let $p \in [1, \infty)$ be given. Now we define the set B of admissible initial states. In the definition of the set B and in what follows, we use generalized derivatives. Let

$$B = \{(y_0, y_1) : \ y_0' \in L^p(0, L), \ y_1 \in L^p(0, L), \ y_0(0) = y_0(L) = 0\}. \tag{2.1}$$

Let a real number $D \in (0, c)$ be given. Define the set Φ of admissible boundary motions as in (1.1). Let $(y_0, y_1) \in B$ be given. We consider the problem

$$P: \qquad \text{Find } \ \phi \in \Phi \text{ such that}$$

$$W(T) = \int_0^L \left| v_x(x,T) + \frac{1}{c} v_t(x,T) \right|^p + \left| v_x(x,T) - \frac{1}{c} v_t(x,T) \right|^p \, dx$$

is minimized, where $v(x,t)$ is the solution of the initial boundary value problem

$$v(x,0) = y_0(x), \ v_t(x,0) = y_1(x), \ x \in (0, L), \tag{2.2}$$

$$v(0,t) = 0, \ v(\phi(t), t) = 0, \ t \in (0, T), \tag{2.3}$$

$$v_{tt}(x,t) = c^2 v_{xx}(x,t), \ (x,t) \in \Omega = \{(x,t) : \ t \in (0,T), \ x \in (0, \ \phi(t))\}. \tag{2.4}$$

For $p = 2$, we have

$$W(T) = 2 \int_0^L \left(v_x(x,T)^2 + \frac{1}{c^2} v_t(x,T)^2 \right) dx;$$

hence, $W(T)$ is equivalent to the classical energy. In the general case, we further refer to $W(T)$ as a generalized energy function. For any $t \in (0,T)$, the definition of $W(t)$ is given in (5.8), where the integration interval is $(0, \phi(t))$. If both boundary points are fixed (that is, $\phi(t) \equiv L$), the generalized energy defined in (5.8) is conserved; see Remark 5.1.

3. Main result: Energy minimizing movement.

THEOREM 3.1 (existence and uniqueness of the solution of problem P). *There exists a movement $\phi \in \Phi$ that solves problem P.*

Let $p \in (1, \infty)$ and $(y_0, y_1) \in B$ be given. Define the L^p-function A as

$$A(x) = \begin{cases} y_0'(-x) & -(1/c) y_1(-x) \quad \text{if } x \in [-L, 0), \\ y_0'(x) & +(1/c) y_1(x) \quad \text{if } x \in [0, L]. \end{cases} \tag{3.1}$$

Define the set $M_z = \{x \in [-L, L] : A(x) = 0\}$. If the set M_z has measure zero, the solution of P is uniquely determined.

For a point z on the real axis define the projection on the interval $[\frac{c-D}{c+D}, \ \frac{c+D}{c-D}]$ in the usual way as

$$\Pi_{[\frac{c-D}{c+D}, \ \frac{c+D}{c-D}]}(z) = \max\{(c-D)/(c+D), \min\{(c+D)/(c-D), \ z\}\}.$$

Define $W(0)$ as follows:

$$W(0) = \int_0^L \left| y_0'(x) + \frac{1}{c} y_1(x) \right|^p + \left| y_0'(x) - \frac{1}{c} y_1(x) \right|^p \, dx.$$

THEOREM 3.2 (solution of problem P). *For $p = 1$, for all $\phi \in \Phi$ we have $W(T) = W(0)$.*

Let $p \in (1, \infty)$ and an initial state $(y_0, y_1) \in B$ be given. Define the L^p-function A as in (3.1). *If $\int_{-L}^{L} |A(y)| \, dy > 0$, there exists a real number $\lambda > 0$ such that the moment equation*

$$\int_{-L}^{L} \Pi_{[\frac{c-D}{c+D}, \, \frac{c+D}{c-D}]}(\lambda \, |A(y)| \,) \, dy = 2L \tag{3.2}$$

holds. With this choice of λ, define the function $h : [-L, L] \to [L, 3L]$ by

$$h(x) = L + \int_{-L}^{x} \Pi_{[\frac{c-D}{c+D}, \, \frac{c+D}{c-D}]}(\lambda \, |A(y)| \,) \, dy$$

and the functions $H_1 : [-L, L] \to (0, \infty)$ and $H_2 : [-L, L] \to [0, 2L]$ by

$$H_1(x) = \frac{h(x) - x}{2}, \; H_2(x) = \frac{h(x) + x}{2}.$$

Then an optimal control movement that is a solution of problem P is given by the function $\phi \in \Phi$ defined by

$$\phi(t) = H_1(H_2^{-1}(ct)), \; t \in (0, T), \tag{3.3}$$

that yields the minimal value for

$$W(T) = \int_{-L}^{L} \frac{|A(s)|^p}{h'(s)^{p-1}} \, ds. \tag{3.4}$$

If $\int_{-L}^{L} |A(y)| \, dy = 0$, we have $W(T) = 0$ for all $\phi \in \Phi$.

The proofs of Theorems 3.1 and 3.2 will be given in section 5.4. They are based upon an explicit representation of the solution of the initial-boundary-value problem (2.2), (2.3), (2.4) for a given boundary motion. This representation of the solution is given in section 4.2. It is used to transform the optimal control problem to a convex optimization problem in a function space that we can solve. The transformed set of admissible functions is defined in section 5.1 by positive pointwise bounds for the function values and a moment equation that prescribes the integral of the admissible functions. For the solution of the transformed problem, the convexity of the transformed objective function on the transformed set of admissible functions is essential.

Remark 3.1. The (non)movement $\phi(t) = L$, $t \in [0, T]$ is admissible. For this movement, the energy is conserved for all p (see Remark 5.1). Therefore the optimal control movement does *not* lead to energy increase.

Remark 3.2. Assume that there exists a real constant $r > 0$ such that $|A(x)| = r$ for all $x \in [-L, L]$. Then for the minimal value of $W(T)$ we have $W(T) = W(0)$. This follows from Theorem 3.2, since to satisfy the moment equation (3.2), the number λ has to be chosen such that $\Pi_{[\frac{c-D}{c+D}, \, \frac{c+D}{c-D}]}(\lambda \, |A(y)| \,) = 1$. This yields $h(x) = x + 2L$,

$H_1(x) = L$, $\phi(t) = L$; hence, in this case it is optimal not to move the boundary point, and thus the energy is conserved.

Example 3.1. Assume that $c = 1$ and $D \in [1/2, 1)$. Then $(c + D)/(c - D) \geq 3$. Let $\gamma > 0$ be given. Assume that

$$\gamma\, y_0(x) = \begin{cases} \frac{1}{3}x, & x \in [0, L/2], \\ (L/6) - (4/3)(x - (L/2)), & x \in (L/2, (5L/6)], \\ -(5L/18) + (5/3)(x - (5L/6)), & x \in (5L/6), L], \end{cases}$$

$$\gamma\, y_1(x) = \begin{cases} 0, & x \in [0, L/2], \\ -5/3, & x \in (L/2, (5L/6)], \\ 4/3, & x \in (5L/6, L]. \end{cases}$$

Then we have

$$\gamma\, A(x) = \begin{cases} \frac{1}{3}, & x \in [-L, L/2], \\ -3, & x \in (L/2, (5L/6)], \\ 3, & x \in ((5L/6), L]. \end{cases}$$

Equation (3.2) holds with $\lambda = \gamma$. For the function h defined in Theorem 3.2 we have

$$h(t) = \begin{cases} \frac{4}{3}L + \frac{1}{3}t, & t \in [-L, L/2], \\ 3t, & t \in (L/2, L]. \end{cases}$$

This yields the optimal movement

$$\phi(t) = \frac{L}{2} + \frac{1}{2}|t - L|,\ t \in [0, 2L].$$

For the energy we have

$$W(0) = \frac{1}{\gamma^p}\,\frac{1}{2}\left(3^p + \frac{1}{3^{p-1}}\right)L,\ W(T) = \frac{1}{\gamma^p}\,2L.$$

This yields the ratio

$$\frac{W(T)}{W(0)} = \frac{4}{3^p + \frac{1}{3^{p-1}}};$$

thus, for $p = 2$ we have $W(T)/W(0) = 3/7$. Hence the optimal movement absorbs more than half of the initial energy. Since the set M_z has measure zero, Theorem 3.1 implies that the solution ϕ of P is uniquely determined.

Example 3.2. Let $y_0(x) = |x - (L/2)| - (L/2)$, $y_1(x) = 0$, $x \in [0, L]$. We have $|A| = 1$ almost everywhere; thus, the set M_z has measure zero. Theorem 3.1 implies the existence of a solution of P. Since M_z has measure zero, Theorem 3.1 implies that the solution of P is unique. Theorem 3.2 implies that the unique optimal movement is the nonmovement $\phi(t) = L$, $t \in [0, 2L]$ that corresponds to the function $h(x) = 2L + x$. Remark 5.1 implies that $W(T)/W(0) = 1$, so by boundary movement an energy decrease cannot be achieved. Since the solution of P is unique, this implies that for every other admissible boundary movement an energy growth is produced at the terminal time.

Example 3.3. Assume that $2L < D < c$ and $(y_0, y_1) \in B$ are such that for some $\lambda > 0$ we have

$$\lambda |A(x)| = \frac{2c}{\sqrt{c^2(1-T)^2 + 4c(L+x)}} - 1.$$

Then M_z has measure zero; thus, Theorem 3.1 implies that P has a unique solution. For all $x \in [-L, L]$: $\lambda|A(x)| \in [(c-D)/(c+D),\ (c+D)/(c-D)]$, and (3.2) holds. We have $h(x) = -x + \sqrt{c^2(1-T)^2 + 4c(L+x)} - c(1-T)$, and the optimal movement is given by $\phi(x) = L + cx(T-x)$.

Example 3.4. Assume that $L(c+D) \le 2c$. Let $\gamma = 2c/(L(c-D))$. Then $\gamma \ge (c+D)/(c-D)$. Let $y_1(x) = 0$ and

$$y_0(x) = \begin{cases} \gamma x, & x \in [0, \frac{1}{2\gamma}], \\ 1 - \gamma x, & x \in (\frac{1}{2\gamma}, \frac{1}{\gamma}], \\ 0, & x > \frac{1}{\gamma}. \end{cases}$$

With $\lambda = 1$, (3.2) holds. Theorem 3.2 implies that an optimal movement is given by the function ϕ corresponding to

$$h'(t) = \begin{cases} \frac{c+D}{c-D}, & t \in [-\frac{1}{\gamma}, \frac{1}{\gamma}], \\ \frac{c-D}{c+D}, & t \notin [-\frac{1}{\gamma}, \frac{1}{\gamma}]. \end{cases}$$

For the energy, this yields

$$\frac{W(T)}{W(0)} = \left(\frac{c-D}{c+D}\right)^{p-1}.$$

For the corresponding optimal movement we have $\phi'(H_2(x)/c) = cH_1'(x)/H_2'(x)$, and thus $\phi'(H_2(x)/c) = D$, if $x \in [-1/\gamma, 1/\gamma]$ and $\phi'(H_2(x)/c) = -D$ otherwise. Hence

$$\phi'(t) = \begin{cases} D, & t \in [\frac{H_2(-1/\gamma)}{c},\ \frac{H_2(1/\gamma)}{c}] = [\frac{L}{2c},\ \frac{L}{c}\left(\frac{3}{2} + \frac{D}{c}\right)], \\ -D, & \text{otherwise.} \end{cases}$$

Example 3.5. Assume that $\varepsilon \in (0, 2D/(c+D))$ and $|A(x)| = 1 + \varepsilon \sin(x)$. Then M_z has measure zero; thus, P has a unique solution. Equation (3.2) holds with $\lambda = 1$, $H_1(x) = L + \varepsilon[\cos(L) - \cos(x)]/2$, $H_2(x) = H_1(x) + x$. For the graph of ϕ, by (3.3) we obtain the curve $G = \{(t, \phi(t)) : t \in [0, T]\} = \{(H_2(x)/c,\ H_1(x)), x \in [-L, L]\}$.

4. Transformation of the problem. For a given boundary movement ϕ from the set Φ defined in (1.1), we want to find a representation of the solution of the initial-boundary-value problem (2.2), (2.3), (2.4) with one moving boundary point in terms of traveling waves; that is, we want to derive d'Alembert's solution for our problem. In particular, we have to show that such a solution exists.

4.1. Wave propagation auxiliary functions. In this section we define some auxiliary functions that we need to derive d'Alembert's solution for our problem. Let $\phi \in \Phi$ be given. Since ϕ is Lipschitz, ϕ is absolutely continuous. For $t \in [0, T]$, define

$$\psi_1(t) = \phi(t) - ct,\ \psi_2(t) = \phi(t) + ct. \tag{4.1}$$

Then $\psi_1'(t) = \phi'(t) - c$. The definition of the set Φ implies the inequality

$$-D \le \phi'(t) \le D$$

and thus $\psi_1'(t) \leq D - c < 0$; hence, ψ_1 is strictly decreasing on $[0,T]$ and thus invertible. We have $\psi_1(0) = L$ and $\psi_1(T) = -L$; therefore, $\psi_1^{-1}(s)$ is defined for $s \in [-L, L]$.

We have $\psi_2'(t) = \phi'(t) + c \geq -D + c > 0$; hence, ψ_2 is strictly increasing on $[0,T]$ and thus invertible. We have $\psi_2(0) = L$ and $\psi_2(T) = 3L$; therefore, $\psi_2^{-1}(s)$ is defined for $s \in [L, 3L]$.

Since $\phi(t) \geq 0$, for all $t \in [0,T]$ we have

$$-\psi_1(t) = ct - \phi(t) < ct + \phi(t) = \psi_2(t). \tag{4.2}$$

For $x \in [-L, L]$ define

$$h(x) = \psi_2(\psi_1^{-1}(-x)). \tag{4.3}$$

Then h is strictly increasing and

$$h'(x) = -\frac{\psi_2'(\psi_1^{-1}(-x))}{\psi_1'(\psi_1^{-1}(-x))} > 0. \tag{4.4}$$

On account of (4.2) for all $x \in [-L, L]$ we have

$$x = -\psi_1(\psi_1^{-1}(-x)) < \psi_2(\psi_1^{-1}(-x)) = h(x).$$

For the inverse of h we have

$$h^{-1}(x) = -\psi_1(\psi_2^{-1}(x)). \tag{4.5}$$

We have

$$h^{-1}(L) = -L,\ h^{-1}(3L) = L. \tag{4.6}$$

Note that

$$L = \psi_2(0) < h(0) = \psi_2(\psi_1^{-1}(0)), \tag{4.7}$$

since

$$0 < \psi_1^{-1}(0),$$

which is true on account of

$$L = \psi_1(0) > 0.$$

Let $t_1 = \psi_1^{-1}(0)$. Then $\phi(t_1) - ct_1 = 0$, and thus $\phi(t_1) = ct_1$. Therefore

$$\psi_2(t_1) = \phi(t_1) + ct_1 = 2ct_1.$$

Hence

$$h(0) = \psi_2(\psi_1^{-1}(0)) = \psi_2(t_1) = 2ct_1 = 2c\psi_1^{-1}(0).$$

In fact, t_1 is the time that a characteristic curve starting at time zero at the left end point of the string needs to reach the moving end of the string.

Later in Lemma 5.1 we will show that there is a bijection between the maps h as defined in (4.3) and the corresponding maps ϕ, which allows one to transform our optimal control problem to an optimization problem in terms of the function h.

4.2. Solution of the initial-boundary-value problem. In this section, we give a representation of the solution of the initial-boundary-value problem for a given fixed boundary movement $\phi \in \Phi$.

THEOREM 4.1. *Let $\phi \in \Phi$ and $(y_0, y_1) \in B$ be given. With h^{-1} as in* (4.5)*, define the functions*

$$\alpha(x) = \begin{cases} -y_0(-x) + (1/c) \int_0^{-x} y_1(s)\,ds, & x \in (-L, 0), \\ y_0(x) + (1/c) \int_0^{x} y_1(s)\,ds, & x \in [0, L), \\ -y_0(-h^{-1}(x)) + (1/c) \int_0^{-h^{-1}(x)} y_1(s)\,ds, & x \in [L, h(0)), \\ y_0(h^{-1}(x)) + (1/c) \int_0^{h^{-1}(x)} y_1(s)\,ds, & x \in [h(0), 3L) \end{cases} \tag{4.8}$$

and

$$v(x,t) = [\alpha(x+ct) - \alpha(-x+ct)]/2, \quad (x,t) \in \Omega. \tag{4.9}$$

We have $\alpha' \in L^p(-L, 3L)$. The function v is continuous on Ω and v_t, $v_x \in L^1(\Omega)$.

Define the family of test functions $\mathcal{T}$ as

$\mathcal{T} = \{\varphi \in C^2(\Omega)$: There exists a set $Q = [x_1, x_2] \times [t_1, t_2] \subset \Omega$ such that the support of φ is contained in the interior of $Q\}$.

The function v satisfies the wave equation (2.4) *in the following weak sense:*

$$\int_\Omega v_t(x,t)\varphi_t(x,t)\,d(x,t) = c^2 \int_\Omega v_x(x,t)\varphi_x(x,t)\,d(x,t) \text{ for all } \varphi \in \mathcal{T}. \tag{4.10}$$

The function v satisfies (2.2) *and* (2.3)*. In this sense, v is the solution of the initial-boundary-value problem* (2.2)*,* (2.3)*,* (2.4)*.*

Proof. Since $y_0' \in L^p(0, L)$, the Sobolev imbedding theorem implies that y_0 is continuous. Moreover, y_1 is in $L^p(0, L)$, and thus α is well-defined. Now we discuss the regularity of α. On the intervals $(-L, 0)$, $[0, L)$, $[L, h(0))$, and $[h(0), 3L)$ the function α is continuous. Due to the definition of the set B we have

$$\lim_{x\to 0-} \alpha(x) = -y_0(0) = 0 = y_0(0) = \lim_{x\to 0+} \alpha(x),$$

$$\lim_{x\to L-} \alpha(x) = y_0(L) + \frac{1}{c}\int_0^L y_1(s)\,ds = -y_0(L) + \frac{1}{c}\int_0^L y_1(s)\,ds = \lim_{x\to L+} \alpha(x),$$

$$\lim_{x\to h(0)-} \alpha(x) = -y_0(0) = y_0(0) = \lim_{x\to h(0)+} \alpha(x),$$

and hence α is continuous on the interval $(-L, 3L)$. The derivative α' in the sense of distributions exists on the intervals $(-L, 0)$, $(0, L)$, $(L, h(0))$, and $(h(0), 3L)$ as a L^p-function. Since α is continuous, this implies that α is absolutely continuous on $(-L, 3L)$. Hence $\alpha' \in L^1(-L, 3L)$, and the L^p regularity on the subintervals $(-L, 0)$, $(0, L)$, $(L, h(0))$, and $(h(0), 3L)$ implies that $\alpha' \in L^p(-L, 3L)$. The continuity of v follows from the continuity of α. For $t = 0$ and $x \in (0, L)$ we have

$$v(x, 0) = [\alpha(x) - \alpha(-x)]/2 = y_0(x).$$

For $(x, t) \in \Omega$ almost everywhere, we have

$$v_t(x,t) = c[\alpha'(x+ct) - \alpha'(-x+ct)]/2. \tag{4.11}$$

Thus the definition of α implies the equation $v_t(x,0) = y_1(x)$. Hence the initial conditions (2.2) are valid.

For $(x,t) \in \Omega$ almost everywhere, we have

$$v_x(x,t) = [\alpha'(x+ct) + \alpha'(-x+ct)]/2. \tag{4.12}$$

By Tonelli's theorem (see, e.g., [15]), (4.12) implies $v_x \in L^1(\Omega)$, and (4.11) implies $v_t \in L^1(\Omega)$.

For all $\varphi \in \mathcal{T}$, integration by parts, (4.12) and (4.11) yield

$$\begin{aligned}
\int_\Omega v_x(x,t)\varphi_x(x,t)\,d(x,t) &= \int_{x_1}^{x_2}\int_{t_1}^{t_2} \varphi_x(x,t)[\alpha'(x+ct) + \alpha'(-x+ct)]/2\,dt\,dx \\
&= -\int_{x_1}^{x_2}\int_{t_1}^{t_2} \varphi_{xt}(x,t)[\alpha(x+ct) + \alpha(-x+ct)]/(2\,c)\,dt\,dx \\
&= -\int_{t_1}^{t_2}\int_{x_1}^{x_2} \varphi_{tx}(x,t)[\alpha(x+ct) + \alpha(-x+ct)]/(2\,c)\,dx\,dt \\
&= \int_{t_1}^{t_2}\int_{x_1}^{x_2} \varphi_t(x,t)[\alpha'(x+ct) - \alpha'(-x+ct)]/(2\,c)\,dx\,dt \\
&= \int_\Omega \varphi_t(x,t)\,v_t(x,t)/c^2\,d(x,t),
\end{aligned}$$

and hence (4.10) holds.

For $x = 0$ we have $v(0,t) = [\alpha(ct) - \alpha(ct)]/2 = 0$; hence, at $x = 0$ the boundary condition $v(0,t) = 0$ holds for all $t \in (0,T)$.

We have $v(\phi(t),t) = [\alpha(\phi(t)+ct) - \alpha(-\phi(t)+ct)]/2 = [\alpha(\psi_2(t)) - \alpha(-\psi_1(t))]/2$. The definition of the set Φ implies that, on the interval $[0,T]$, h is strictly increasing and invertible (see section 4.1). By (4.6), for all $s \in [L,3L] = [\psi_2(0), \psi_2(T)]$ we have $h^{-1}(s) \in [-L,L]$. Hence the definition of α implies the equation

$$\alpha(h^{-1}(s)) = \begin{cases} -y_0(-h^{-1}(s)) + (1/c)\int_0^{-h^{-1}(s)} y_1(t)\,dt, & h^{-1}(s) \in (-L,0), \\ y_0(h^{-1}(s)) + (1/c)\int_0^{h^{-1}(s)} y_1(t)\,dt, & h^{-1}(s) \in [0,L). \end{cases} \tag{4.13}$$

On the other hand, for $s \in (L, h(0))$ the definition of α implies that $\alpha(s) = -y_0(-h^{-1}(s)) + (1/c)\int_0^{-h^{-1}(s)} y_1(t)\,dt$, and we have $h^{-1}(s) \in (h^{-1}(L),\, 0) = (-L, 0)$. Thus (4.13) implies that $\alpha(h^{-1}(s)) = \alpha(s)$ for $s \in (L, h(0))$.

For $s \in (h(0), 3L)$, the definition of α yields $\alpha(s) = y_0(h^{-1}(s)) + \frac{1}{c}\int_0^{h^{-1}(s)} y_1(t)\,dt$, and we have $h^{-1}(s) \in (0,\, h^{-1}(3L)) = (0,\, L)$. Thus also in this case (4.13) implies that $\alpha(h^{-1}(s)) = \alpha(s)$.

Hence for all $s \in (L, 3L)$ the following equation holds:

$$\alpha(h^{-1}(s)) = \alpha(s). \tag{4.14}$$

Thus for all $t \in (0,T)$ we have

$$\alpha(\psi_2(t)) = \alpha(h^{-1}(\psi_2(t)) = \alpha(-\psi_1(\psi_2^{-1}(\psi_2(t)))) = \alpha(-\psi_1(t)).$$

Therefore, the boundary condition $v(\phi(t),t) = 0$ holds for all $t \in (0,T)$. ∎

The following lemma contains regularity conditions for the initial state and the boundary movement ϕ that guarantee the existence of a solution of the initial-boundary-value problem (2.2), (2.3), (2.4) that satisfies the wave equation pointwise almost everywhere.

LEMMA 4.2. *If $y_0 \in C^2[0,L]$, $y_1 \in C^1[0,L]$, $y_0(0) = 0$, $y_0(L) = y_0'(L) = 0$, $y_1(0) = 0 = y_1(L)$, and $\phi \in \Phi \cap C^2[0,T]$, then α defined by* (4.8) *is in $C^1(-L,3L)$, α' is absolutely continuous, $\alpha'' \in L^\infty(-L,3L)$, and v defined by* (4.9) *satisfies the wave equation* (2.4) *in the following weak sense:*

$$\int_\Omega v_{tt}(x,t)\varphi(x,t)\,d(x,t) = c^2 \int_\Omega v_{xx}(x,t)\varphi(x,t)\,d(x,t) \text{ for all } \varphi \in \mathcal{T}. \tag{4.15}$$

Moreover, v satisfies (2.4) *pointwise almost everywhere in Ω.*

Proof. Since $\phi \in C^2[0,T]$, definition (4.1) implies that ψ_1 and ψ_2 are in $C^2[0,T]$. Moreover, ψ_1^{-1} and ψ_2^{-1} are in $C^2[0,T]$. Hence the definition (4.3) of h implies that h is in $C^2[-L,L]$, and (4.5) implies that h^{-1} is also two times continuously differentiable.

In the proof of Theorem 4.1, we have seen that α is absolutely continuous. Since $y_0 \in C^2[0,L]$ and $y_1 \in C^1[0,L]$, the definition (4.8) of α implies that on the intervals $(-L,0)$, $(0,L)$, $(L,h(0))$, and $(h(0),3L)$ the function α is two times continuously differentiable. Due to the conditions $y_0(0) = 0$, $y_0(L) = y_0'(L) = 0$, $y_1(0) = 0 = y_1(L)$ in the end point of these intervals, we have the one-sided derivatives

$$\begin{aligned}
\alpha_-'(0) &= y_0'(0) - (1/c)y_1(0) = y_0'(0) = y_0'(0) + (1/c)y_1(0) = \alpha_+'(0),\\
\alpha_-'(L) &= y_0'(L) + (1/c)y_1(L) = 0 = y_0'(L)\ (h^{-1})'(L) - (1/c)y_1(L)\ (h^{-1})'(L)\\
&= y_0'(-h^{-1}(L))\ (h^{-1})'(L) - (1/c)y_1(-h^{-1}(L))\ (h^{-1})'(L)\\
&= \alpha_+'(L),\\
\alpha_-'(h(0)) &= y_0'(0)\ (h^{-1})'(h(0)) - (1/c)y_1(0)\ (h^{-1})'(h(0)) = y_0'(0)\ (h^{-1})'(h(0))\\
&= y_0'(0)\ (h^{-1})'(h(0)) + (1/c)y_1(0)\ (h^{-1})'(h(0)) = \alpha_+'(h(0)).
\end{aligned}$$

Since the one-sided derivatives are equal, α' is continuous, and hence $\alpha \in C^1(-L,3L)$. Since α'' exists on the intervals $(-L,0)$, $(0,L)$, $(L,h(0))$, and $(h(0),3L)$ as a bounded continuous function, this implies that α' is absolutely continuous on $(-L,3L)$ and $\alpha'' \in L^\infty(-L,3L)$. For (x,t) in Ω almost everywhere we have

$$\begin{aligned}
v_{tt}(x,t) &= c^2[\alpha''(x+ct) - \alpha''(-x+ct)]/2,\\
v_{xx}(x,t) &= [\alpha''(x+ct) - \alpha''(-x+ct)]/2,
\end{aligned}$$

and hence (2.4) holds almost everywhere in Ω. For all $\varphi \in \mathcal{T}$, integration by parts yields

$$\int_\Omega \varphi_t(x,t)\,v_t(x,t)/c^2\,d(x,t) = -\int_\Omega \varphi(x,t)\,v_{tt}(x,t)/c^2\,d(x,t)$$

and

$$\int_\Omega \varphi_x(x,t)\,v_x(x,t)\,d(x,t) = -\int_\Omega \varphi(x,t)\,v_{xx}(x,t)\,d(x,t).$$

Hence (4.10) implies that (4.15) holds. ∎

Remark 4.1. For x in the interval $(0,L)$ we have

$$\alpha'(x)^2 + \alpha'(-x)^2 = \frac{[\alpha'(x) + \alpha'(-x)]^2}{2} + \frac{[\alpha'(x) - \alpha'(-x)]^2}{2} = 2\,y_0'(x)^2 + \frac{2}{c^2}y_1(x)^2.$$

5. Solution of the optimal shape control problem.

5.1. The transformed set of admissible controls. Definition (4.3) states how h can be obtained from a given movement ϕ. The following lemma shows that, if h is known, the corresponding movement ϕ is uniquely determined and can be computed.

LEMMA 5.1. *Let $\phi \in \Phi$ be given. Define the function h by* (4.3). *For $x \in [-L, L]$, define*

$$H_1(x) = \frac{h(x) - x}{2},\ H_2(x) = \frac{h(x) + x}{2}. \tag{5.1}$$

Then for all $t \in [0, T]$, we have

$$\phi(t) = H_1(H_2^{-1}(ct)). \tag{5.2}$$

Proof. In section 4.1, we have seen that h is strictly increasing, and hence H_2 is strictly increasing. Since $h(-L) = L$, we have $H_2(-L) = 0$. Since $h(L) = 3L$, $H_2(L) = 2L$. Thus the assertion (5.2) is equivalent to the statement that for all $x \in [-L, L]$ we have

$$\phi\left(\frac{H_2(x)}{c}\right) = H_1(x).$$

From (4.1), we have for all $t \in [0, T]$ the equation $\psi_1(t) + 2ct = \psi_2(t)$. For $t = \psi_1^{-1}(x)$ with $x \in [-L, L]$, this yields

$$x + 2c\psi_1^{-1}(x) = \psi_2(\psi_1^{-1}(x)),$$

and by (4.3) this implies

$$h(x) = \psi_2(\psi_1^{-1}(-x)) = 2c\psi_1^{-1}(-x) - x.$$

Therefore, the following equation holds: $H_2(x) = [h(x) + x]/2 = c\psi_1^{-1}(-x)$. This implies the equation

$$\psi_1\left(\frac{H_2(x)}{c}\right) = \psi_1(\psi_1^{-1}(-x)) = -x = \phi\left(\frac{H_2(x)}{c}\right) - H_2(x),$$

where we have again used the definition of ψ_1. Hence

$$\phi\left(\frac{H_2(x)}{c}\right) = H_2(x) - x = H_1(x),$$

where the last equation follows from definition (5.1). ∎

Define the set H of functions defined on the interval $[-L, L]$ as follows:

$$H = \{h : h(x) = \psi_2(\psi_1^{-1}(-x)),\ x \in [-L, L], \text{ with } \psi_1, \psi_2 \text{ as in (4.1) for some } \phi \in \Phi\ \}.$$

LEMMA 5.2. *Define the map $\theta : \Phi \to H$ by $\theta(\phi) = h$, where $h(x) = \psi_2(\psi_1^{-1}(-x))$, $x \in [-L, L]$ with ψ_1, ψ_2 as in* (4.1). *Then θ is bijective.*

Proof. First we show that θ is injective. Let ϕ_1, $\phi_2 \in \Phi$ be given such that $h_1 = \theta(\phi_1) = \theta(\phi_2) = h_2$. Lemma 5.1 implies that, with

$$H_1^i(x) = \frac{h_i(x) - x}{2},\ H_2^i(x) = \frac{h_i(x) + x}{2},\ i \in \{1, 2\},$$

we have

$$\phi_i(t) = H_1^i((H_2^i)^{-1}(ct)),\ i \in \{1, 2\}.$$

Since $h_1 = h_2$, we have $H_1^1 = H_1^2$ and $H_2^1 = H_2^2$, and thus $\phi_1(t) = \phi_2(t)$. Therefore θ is injective. The definition of the set H implies that $H = \theta(\Phi)$, and thus θ is surjective. □

Later we will need the following result in a proof.

LEMMA 5.3. *Let $I = [\frac{c-D}{c+D},\ \frac{c+D}{c-D}]$. The following equation holds:*

$$c \max_{z \in I} \left\{ \frac{z-1}{z+1},\ \frac{1-z}{1+z} \right\} = D. \tag{5.3}$$

Proof. For $z \in I$, let $\eta_1(z) = (z-1)/(z+1)$. Since $\eta_1'(z) > 0$, for all $z \in I$ we have $\eta_1(z) \le \eta_1(\frac{c+D}{c-D}) = D/c$.

For $z \in I$, let $\eta_2(z) = (1-z)/(1+z)$. Since $\eta_2'(z) < 0$, for all $z \in I$ we have $\eta_2(z) \le \eta_2(\frac{c-D}{c+D}) = D/c$. □

LEMMA 5.4. *Define the set*

(5.4)

$$M = \left\{ h | h : [-L, L] \to [L, 3L] \text{ such that } h(x) = L + \int_{-L}^{x} f(s)\, ds \text{ with } f \in F \right\},$$

where the set F is defined as

$$F = \left\{ f | f : [-L, L] \to \left[\frac{c-D}{c+D},\ \frac{c+D}{c-D} \right] \text{ such that } f \text{ is Lebesgue integrable and } \int_{-L}^{L} f(x)\, dx = 2L \right\}.$$

Then $M = H$.

Proof. First we show that $H \subset M$. Let $h \in H = \theta(\Phi)$, $h(x) = \psi_2(\psi_1^{-1}(-x))$, $x \in [-L, L]$. Since ψ_2 is Lipschitz continuous with Lipschitz constant $c + D$, for all $x_1,\ x_2 \in [-L, L]$ we have

$$|h(x_1) - h(x_2)| = |\psi_2(\psi_1^{-1}(-x_1)) - \psi_2(\psi_1^{-1}(-x_2))| \le (c+D)|\psi_1^{-1}(-x_1) - \psi_1^{-1}(-x_2)|.$$

For ψ_1 and $t_1,\ t_2 \in [0, T]$ we have the inequality

$$|\psi_1(t_1) - \psi_1(t_2)| = |\phi(t_1) - \phi(t_2) - c(t_1 - t_2)| \ge c|t_1 - t_2| - |\phi(t_1) - \phi(t_2)| \ge (c-D)|t_1 - t_2|.$$

With $t_1 = \psi_1^{-1}(-x_1)$, $t_2 = \psi_1^{-1}(-x_2)$, this yields

$$|\psi_1^{-1}(-x_1) - \psi_1^{-1}(-x_2)| \le \frac{1}{c-D}|x_1 - x_2|.$$

This implies the inequality

$$|h(x_1) - h(x_2)| \le \frac{c+D}{c-D}\,|x_1 - x_2|,$$

and thus h is Lipschitz continuous with Lipschitz constant $\frac{c+D}{c-D}$. Since h is Lipschitz, h is absolutely continuous, and, for the derivative, we have the inequality $|h'(x)| \le (c+D)/(c-D)$ almost everywhere.

For all t_1, $t_2 \in [0,T]$, we have

$$|\psi_2(t_1) - \psi_2(t_2)| \geq (c-D)|t_1 - t_2|,\ |\psi_1(t_1) - \psi_1(t_2)| \leq (c+D)|t_1 - t_2|.$$

With $t_1 = \psi_1^{-1}(-x_1)$, $t_2 = \psi_1^{-1}(-x_2)$, this yields

$$|x_1 - x_2| \leq (c+D)|\psi_1^{-1}(-x_1) - \psi_1^{-1}(-x_2)|.$$

Hence

$$|h(x_1) - h(x_2)| \geq (c-D)|\psi_1^{-1}(-x_1) - \psi_1^{-1}(-x_2)| \geq \frac{c-D}{c+D}|x_1 - x_2|.$$

This implies that $|h'(x)| \geq (c-D)/(c+D)$ almost everywhere. In section 4.1, we have shown that h is strictly increasing. Hence we have the inequality

$$\frac{c-D}{c+D} \leq h'(x) \leq \frac{c+D}{c-D}. \tag{5.5}$$

We can write

$$h(x) = h(-L) + \int_{-L}^{x} h'(s)\,ds = L + \int_{-L}^{x} h'(s)\,ds.$$

Since $h(L) = L + \int_{-L}^{L} h'(s)\,ds = 3L$, we have $\int_{-L}^{L} h'(s)\,ds = 2L$. So we have shown $h' \in F$, which implies $h \in M$, and thus $H \subset M$.

Now we show that $M \subset H$. Let $m \in M$ be given, $m(x) = L + \int_{-L}^{x} f(s)\,ds$, with $f \in F$ and $x \in [-L,L]$. Then m is strictly increasing. For $x \in [-L,L]$, define $H_1(x) = [m(x) - x]/2$, $H_2(x) = [m(x) + x]/2$. Then H_2 is strictly increasing, and $\{H_2(x) :\ x \in [-L,L)\} = [H_2(-L),\ H_2(L)] = [\frac{m(-L)-L}{2},\ \frac{m(L)+L}{2}] = [0,2L]$. For $t \in [0,T]$, we have $ct \in [0,2L] = H_2[-L,L]$. Hence $H_2^{-1}(ct)$ is well-defined and in $[-L,L]$. So we can define

$$\phi(t) = H_1(H_2^{-1}(ct)). \tag{5.6}$$

Since $H_2(-L) = 0$ and $H_2(L) = 2L$ we have

$$\phi(0) = H_1(H_2^{-1}(0)) = H_1(-L) = L,$$
$$\phi(T) = H_1(H_2^{-1}(2L)) = H_1(L) = L.$$

Our assumptions imply that H_1 is Lipschitz continuous with Lipschitz constant $c/(c-D)$ and H_2^{-1} is Lipschitz continuous with Lipschitz constant $(c+D)/c$. Hence ϕ is Lipschitz continuous and thus absolutely continuous. Since $\phi(H_2(x)/c) = H_1(x)$, the chain rule implies that for the derivative we have the equation $\phi'(H_2(x)/c) = cH_1'(x)/H_2'(x)$, which implies the inequality

$$\left|\phi'\left(\frac{H_2(x)}{c}\right)\right| = c\,\frac{|H_1'(x)|}{H_2'(x)} = c\frac{|f(x)-1|}{f(x)+1} \leq c\max\left\{\frac{f(x)-1}{f(x)+1},\ \frac{1-f(x)}{1+f(x)}\right\} \leq D,$$

where the last inequality follows from Lemma 5.3 since $f \in F$. Hence $\phi \in \Phi$. Therefore we can compute $\theta(\phi)$.

For $x \in [-L,L]$, we have $H_2(x)/c \in [0,2L/c] = [0,T]$, and the definition of ϕ implies the equation $\phi(H_2(x)/c) = H_1(x) = [m(x) - x]/2$, and hence

$$\psi_2\left(\frac{H_2(x)}{c}\right) = \phi\left(\frac{H_2(x)}{c}\right) + H_2(x) = \frac{m(x)-x}{2} + \frac{m(x)+x}{2} = m(x).$$

Moreover, we have

$$-\psi_1\left(\frac{H_2(x)}{c}\right) = H_2(x) - \phi\left(\frac{H_2(x)}{c}\right) = H_2(x) - H_1(x) = \frac{m(x)+x}{2} - \frac{m(x)-x}{2} = x.$$

This implies the equation $\psi_1^{-1}(-x) = H_2(x)/c$. Hence $\theta(\phi)(x) = \psi_2(\psi_1^{-1}(-x)) = \psi_2(H_2(x)/c) = m(x)$. So we have shown that $\theta(\phi) = m$. Hence $m \in \theta(\Phi) = H$. Since $m \in M$ was arbitrary, this yields $M \subset H$. Since we have already shown the inclusion $H \subset M$, this yields the equation $H = M$. Moreover, this implies the equation

$$\phi = \theta^{-1}(m). \qquad \square \tag{5.7}$$

We see that the mapping θ defined in Lemma 5.2 is a bijection between the admissible motions $\phi \in \Phi$ and the functions $h \in M$. Moreover, for all $m \in M$, (5.7) implies that $\phi = \theta^{-1}(m)$ is given by (5.6).

5.2. The objective function: Computation of the energy. Let $t \in [0,T]$ be given. Define the integrals

$$I_1(t) = \int_0^{\phi(t)} \left| v_x(x,t) + \frac{1}{c} v_t(x,t) \right|^p \, dx, \; I_2(t) = \int_0^{\phi(t)} \left| v_x(x,t) - \frac{1}{c} v_t(x,t) \right|^p \, dx$$

and the generalized energy by

$$W(t) = I_1(t) + I_2(t). \tag{5.8}$$

Equation (4.9) implies that

$$I_1(t) = \int_{ct}^{\psi_2(t)} |\alpha'(x)|^p \, dx, \; I_2(t) = \int_{-\psi_1(t)}^{ct} |\alpha'(x)|^p \, dx.$$

Thus for all $t \in [0,T]$, we have

$$W(t) = \int_{-\psi_1(t)}^{ct} |\alpha'(x)|^p \, dx + \int_{ct}^{\psi_2(t)} |\alpha'(x)|^p \, dx = \int_{-\psi_1(t)}^{\psi_2(t)} |\alpha'(x)|^p \, dx.$$

For our terminal time T this implies that

$$W(T) = \int_L^{3L} |\alpha'(x)|^p \, dx = \int_{h^{-1}(L)}^{h^{-1}(3L)} |\alpha'(h(s))|^p \, h'(s) \, ds = \int_{-L}^{L} |\alpha'(h(s))|^p \, h'(s) \, ds.$$

By (4.14), for all $u \in [L,3L]$ we have $h^{-1}(u) \in [-L,L]$ and $\alpha(h^{-1}(u)) = \alpha(u)$. Thus for all $s \in [-L,L]$ we have $\alpha(s) = \alpha(h(s))$ and thus

$$\alpha'(s) = \alpha'(h(s)) \, h'(s).$$

This yields the equation

$$W(T) = \int_{-L}^{L} \frac{|\alpha'(x)|^p}{h'(x)^{p-1}} \, dx. \tag{5.9}$$

We see that W as a function of t on the interval $[0,T]$ is absolutely continuous and the derivative is given by the L^1 function

$$\begin{aligned} W'(t) &= |\alpha'(\psi_2(t))|^p\,\psi_2'(t) + |\alpha'(-\psi_1(t))|^p\,\psi_1'(t) \\ &= |\alpha'(\psi_2(t))|^p\,\psi_2'(t) + |\alpha'(h(-\psi_1(t)))|^p\,[h'(-\psi_1(t))]^p\,\psi_1'(t) \\ &= |\alpha'(\psi_2(t))|^p\,\psi_2'(t) + |\alpha'(\psi_2(t))|^p\left[-\frac{\psi_2'(t)}{\psi_1'(t)}\right]^p\psi_1'(t) \\ &= |\alpha'(\psi_2(t))|^p\,\psi_2'(t)\frac{|\psi_1'(t)|^{p-1} - |\psi_2'(t)|^{p-1}}{|\psi_1'(t)|^{p-1}}. \end{aligned}$$

Here we have used (4.4) to evaluate $h'(-\psi_1(t))$. Since $|\phi'(t)| < c$, this implies that the sign of $W'(t)$ is equal to the sign of

$$|\psi_1'(t)|^{p-1} - |\psi_2'(t)|^{p-1} = (c - \phi'(t))^{p-1} - (c + \phi'(t))^{p-1}.$$

If, for almost all $t \in (0,t_1)$, $\phi'(t) > 0$, this implies that $W'(t) < 0$ on $(0,t_1)$, and thus $W(0) > W(t_1)$. This means that an expansion causes a decrease in energy.

On the other hand, if, for almost all $t \in (0,t_1)$, $\phi'(t) < 0$, we have $W'(t) > 0$ on $(0,t_1)$, and therefore $W(0) < W(t_1)$. Thus a contraction causes an increase in energy.

This means that the results given in Theorem 2.1 in [2] for the case $p = 2$ and dimension unequal to two are also valid for $p \neq 2$, $p \in (1,\infty)$ in the 1-d case.

Remark 5.1 (conservation of the energy for $\phi(t) \equiv L$). Let $\phi(t) \equiv L$. Then for all $t \in (0,T)$, $\phi'(t) = 0$; hence, $W'(t) = 0$ and therefore $W(t) = W(0)$; that is, in the case of two fixed boundary points the generalized energy $W(t)$ is conserved.

Remark 5.2 (conservation of the energy for $p = 1$). For $p = 1$ we have for all $\phi \in \Phi$ and for $t \in (0,T)$

$$W'(t) = |\alpha'(\psi_2(t))|\,\psi_2'(t)\frac{|\psi_1'(t)|^{1-1} - |\psi_2'(t)|^{1-1}}{|\psi_1'(t)|^{1-1}} = 0,$$

and thus the integral

$$W(t) = \int_{-L}^{L} |\alpha'(x)|\,dx = W(T)$$

is conserved for the limit case $p = 1$ regardless of ϕ. In other words, for $p = 1$ the control problem P is meaningless.

Remark 5.3 (sharp lower bounds for the energy). Let q be such that $\frac{1}{p} + \frac{1}{q} = 1$. Hölder's inequality implies that

$$\begin{aligned} \int_{-L}^{L} |\alpha'(x)|\,dx = \int_{-\psi_1(t)}^{\psi_2(t)} |\alpha'(x)|\,dx &\le \left(\int_{-\psi_1(t)}^{\psi_2(t)} |\alpha'(x)|^p\,dx\right)^{1/p}\left(\int_{-\psi_1(t)}^{\psi_2(t)} 1^q\,dx\right)^{1/q} \\ &= (2\,\phi(t))^{1-1/p} W(t)^{1/p}. \end{aligned}$$

Thus for all $t > 0$ we have the following lower bound for the energy:

$$W(t) \ge \frac{1}{(2\,\phi(t))^{p-1}}\left(\int_{-L}^{L} |\alpha'(x)|\,dx\right)^p. \tag{5.10}$$

Assume now that $\alpha(x)$ is such that there exists a real number r such that for all $x \in [-\psi_1(t), \psi_2(t)]$ we have $|\alpha'(x)| = r$. Such a situation occurs in Example 3.2 with $r = 1$. Then $\int_{-L}^{L} |\alpha'(x)|\,dx = \int_{-\psi_1(t)}^{\psi_2(t)} |\alpha'(x)|\,dx = 2\phi(t)\,r$ and

$$W(t) = 2\phi(t)\,r^p = \frac{1}{(2\,\phi(t))^{p-1}} \left(\int_{-L}^{L} |\alpha'(x)|\,dx \right)^p,$$

which shows that the lower bound (5.10) is sharp.

Equation (5.9) yields a lower bound for the energy $W(T)$ at the terminal time T. With (5.5), (5.9) implies the inequality

$$W(T) \geq \left(\frac{c-D}{c+D} \right)^{p-1} \int_{-L}^{L} |\alpha'(x)|^p = \left(\frac{c-D}{c+D} \right)^{p-1} W(0).$$

This lower bound is attained in Example 3.4.

5.3. The transformed optimization problem. Later in this section, we will need the following result in a proof.

LEMMA 5.5. *Assume that $p \in (1, \infty)$. Let $r \geq 0$ be given. For $z \in (0, [c+D]/[c-D]]$, define the function $\tau(z) = r/z^{p-1}$, and let $\kappa = \frac{1}{2}\,(p-1)p(c-D)^{p+1}/(c+D)^{p+1} > 0$. Then for all z_1, $z_2 \in (0, [c+D]/[c-D]]$, the following inequality holds:*

$$\tau(z_2) - \tau(z_1) \geq r(1-p)\frac{1}{z_1^p}(z_2 - z_1) + r\kappa(z_2 - z_1)^2.$$

Proof. We have $\tau'(z) = r(1-p)/z^p$ and $\tau''(z) = r(p-1)p/z^{p+1}$. We consider the Taylor expansion for τ which yields the existence of a point μ between z_1 and z_2 such that

$$\begin{aligned} \tau(z_2) &= \tau(z_1) + \tau'(z_1)(z_2 - z_1) + \frac{\tau''(\mu)}{2}(z_2 - z_1)^2 \\ &\geq \tau(z_1) + \tau'(z_1)(z_2 - z_1) + \frac{1}{2} r(p-1)p(z_2 - z_1)^2 / \max\{z_1, z_2\}^{p+1} \\ &\geq \tau(z_1) + \tau'(z_1)(z_2 - z_1) + r\,\kappa\,(z_2 - z_1)^2, \end{aligned}$$

and the assertion follows. ∎

Now we come to the transformed optimization problem. Let the set F be defined as in Lemma 5.4. For $f \in F$, define

$$J(f) = \int_{-L}^{L} \frac{|\alpha'(x)|^p}{f(x)^{p-1}}\,dx.$$

In Theorem 4.1, we have seen that $\alpha' \in L^p(-L, L)$. Due to the bounds for f in the definition of the set F, this implies that the number $J(f)$ is well-defined. Lemma 5.4 states that $M = H$. Due to (5.9), for all $h \in M$, we have $W(T) = J(h')$. Hence the definition (5.4) of the set M and the bijection θ between the sets Φ and M given in Lemma 5.2 imply that our problem P is equivalent to the problem

$$Q: \qquad \text{Find } f \in F \text{ such that } J(f) \text{ is minimized.} \tag{5.11}$$

Problem Q is a convex optimization problem. The necessary optimality conditions lead us to the following solution of problem Q.

LEMMA 5.6. *Let $p \in (1,\infty)$, and let α be defined as in* (4.8). *For $\lambda > 0$, define the function f on the interval $[-L,L]$ by*

$$f(x) = \begin{cases} \frac{c-D}{c+D} & \text{if} \quad x \in [-L,L] \quad \text{and} \quad \lambda|\alpha'(x)| \le \frac{c-D}{c+D}, \\ \lambda|\alpha'(x)| & \text{if} \quad x \in [-L,L] \quad \text{and} \quad \lambda|\alpha'(x)| \in (\frac{c-D}{c+D},\ \frac{c+D}{c-D}), \\ \frac{c+D}{c-D} & \text{if} \quad x \in [-L,L] \quad \text{and} \quad \lambda|\alpha'(x)| \ge \frac{c+D}{c-D}. \end{cases}$$

If $\int_L^L |\alpha'(y)|\,dy > 0$, there exists a real number $\lambda > 0$ such that

$$\int_{-L}^{L} f(x)\,dx = 2L, \tag{5.12}$$

and, with this choice of λ, we have $f \in F$, and f is a solution of problem Q.

Define the set $M_z = \{x \in [-L,L] : \alpha'(x) = 0\}$. If M_z has measure zero, the solution of Q is uniquely determined.

If $\int_L^L |\alpha'(y)|\,dy = 0$, we have $J(f) = 0$ for all $f \in F$.

Proof. **A special case.** For the special case that $\int_L^L |\alpha'(y)|\,dy > 0$ and for

$$\lambda = \frac{2L}{\int_{-L}^{L} |\alpha'(x)|\,dx},$$

we have $\lambda|\alpha'(x)| \in [\frac{c-D}{c+D},\ \frac{c+D}{c-D}]$ for all $x \in [-L,L]$, and we give a proof for the optimality of f that is based upon Hölder's inequality. We have

$$J(f) = \frac{1}{\lambda^{p-1}} \int_{-L}^{L} \frac{|\alpha'(x)|^p}{|\alpha'(x)|^{p-1}}\,dx = \frac{(\int_{-L}^{L} |\alpha'(x)|\,dx)^p}{(2L)^{p-1}}.$$

Let q be such that $(1/p) + (1/q) = 1$, and let $g \in F$. Then we have

$$\begin{aligned} \int_{-L}^{L} |\alpha'(x)|\,dx &= \int_{-L}^{L} \frac{|\alpha'(x)|}{g(x)^{1/q}}\, g(x)^{1/q}\,dx \\ &\le \left(\int_{-L}^{L} \frac{|\alpha'(x)|^p}{g(x)^{p/q}}\,dx\right)^{1/p} \left(\int_{-L}^{L} g(x)\,dx\right)^{1/q} \\ &\le \left(\int_{-L}^{L} \frac{|\alpha'(x)|^p}{g(x)^{p-1}}\,dx\right)^{1/p} (2L)^{1/q} \\ &= J(g)^{1/p}(2L)^{1/q}. \end{aligned}$$

This implies the desired inequality

$$J(g) \ge \frac{(\int_{-L}^{L} |\alpha'(x)|\,dx)^p}{(2L)^{p/q}} = \frac{(\int_{-L}^{L} |\alpha'(x)|\,dx)^p}{(2L)^{p-1}} = J(f).$$

The general case. For the general case where $\int_L^L |\alpha'(y)|\,dy > 0$, we give a proof that is based upon the convexity of the objective function. Assume that f is given as defined in Lemma 5.6 and that (5.12) holds. Then $f \in F$. Let $g \in F$ be given. Define

the difference function $\Delta = g - f$. Then $\int_{-L}^{L} \Delta(x)\,dx = 0$. Define the sets

$$M_{\leq} = \left\{ x \in [-L, L] :\ |\alpha'(x)| \leq \frac{c-D}{c+D}\ \frac{1}{\lambda} \right\},$$
$$M_0 = \left\{ x \in [-L, L] :\ |\alpha'(x)| \in \left(\frac{c-D}{c+D}\ \frac{1}{\lambda},\ \frac{c+D}{c-D}\ \frac{1}{\lambda} \right) \right\},$$
$$M_{\geq} = \left\{ x \in [-L, L] :\ |\alpha'(x)| \geq \frac{c+D}{c-D}\ \frac{1}{\lambda} \right\}.$$

For all $x \in M_{\leq}$, we have $\Delta(x) = g(x) - \frac{c-D}{c+D} \geq 0$. Hence due to the definition of f the following inequality holds:

$$\int_{M_{\leq}} \frac{|\alpha'(x)|^p}{f(x)^p}\ \Delta(x)\,dx = \left(\frac{c+D}{c-D} \right)^p \int_{M_{\leq}} |\alpha'(x)|^p\ \Delta(x)\,dx$$
$$\leq \left(\frac{c+D}{c-D} \right)^p \left(\frac{c-D}{c+D} \right)^p \frac{1}{\lambda^p} \int_{M_{\leq}} \Delta(x)\,dx = \frac{1}{\lambda^p} \int_{M_{\leq}} \Delta(x)\,dx.$$

For all $x \in M_{\geq}$, we have $\Delta(x) = g(x) - \frac{c+D}{c-D} \leq 0$. Hence due to the definition of f the following inequality holds:

$$\int_{M_{\geq}} \frac{|\alpha'(x)|^p}{f(x)^p}\ \Delta(x)\,dx = \left(\frac{c-D}{c+D} \right)^p \int_{M_{\geq}} |\alpha'(x)|^p\ \Delta(x)\,dx$$
$$\leq \left(\frac{c-D}{c+D} \right)^p \left(\frac{c+D}{c-D} \right)^p \frac{1}{\lambda^p} \int_{M_{\geq}} \Delta(x)\,dx = \frac{1}{\lambda^p} \int_{M_{\geq}} \Delta(x)\,dx.$$

Moreover, the definition of f implies the equation

$$\int_{M_0} \frac{|\alpha'(x)|^p}{f(x)^p}\ \Delta(x)\,dx = \int_{M_0} \frac{|\alpha'(x)|^p}{\lambda^p\,|\alpha'(x)|^p}\ \Delta(x)\,dx = \frac{1}{\lambda^p} \int_{M_0} \Delta(x)\,dx.$$

We have $f(x),\ g(x) \in [\frac{c-D}{c+D},\ \frac{c+D}{c-D}]$ almost everywhere on the interval $[-L, L]$. Hence, for our objective function, the application of Lemma 5.5 pointwise for all $x \in [-L, L]$ with $r = |\alpha'(x)|^p$ and $z_2 = g(x)$, $z_1 = f(x)$ yields

$$\begin{aligned} J(g) - J(f) &= \int_{-L}^{L} \frac{|\alpha'(x)|^p}{g(x)^{p-1}} - \frac{|\alpha'(x)|^p}{f(x)^{p-1}}\,dx \\ &\geq \int_{-L}^{L} (1-p) \frac{|\alpha'(x)|^p}{f(x)^p}\ \Delta(x)\,dx + \int_{-L}^{L} |\alpha'(x)|^p\,\kappa\,\Delta(x)^2\,dx \\ &= (1-p) \int_{M_{\leq}} \frac{|\alpha'(x)|^p}{f(x)^p}\ \Delta(x)\,dx + (1-p) \int_{M_0} \frac{|\alpha'(x)|^p}{f(x)^p}\ \Delta(x)\,dx \\ &\quad + (1-p) \int_{M_{\geq}} \frac{|\alpha'(x)|^p}{f(x)^p}\ \Delta(x)\,dx + \kappa \int_{-L}^{L} |\alpha'(x)|^p\,\Delta(x)^2\,dx. \end{aligned}$$

Since $(1-p) < 0$, with the inequalities for the integrals on $M_{\leq}$, $M_{\geq}$, and M_0 derived

above, this implies the inequality

$$\begin{aligned} J(g) - J(f) &\geq (1-p)\int_{M_{\leq}} \frac{1}{\lambda^p}\,\Delta(x)\,dx + (1-p)\int_{M_0} \frac{1}{\lambda^p}\,\Delta(x)\,dx \\ &\quad + (1-p)\int_{M_{\geq}} \frac{1}{\lambda^p}\,\Delta(x)\,dx + \kappa\int_{-L}^{L} |\alpha'(x)|^p\,\Delta(x)^2\,dx \\ &= \frac{1-p}{\lambda^p}\int_{-L}^{L} \Delta(x)\,dx + \kappa\int_{-L}^{L} |\alpha'(x)|^p\,\Delta(x)^2\,dx \\ &= \kappa\int_{-L}^{L} |\alpha'(x)|^p\,\Delta(x)^2\,dx \geq 0. \end{aligned}$$

Thus $J(g) \geq J(f)$. Hence f is a solution of the optimization problem Q.

Define the set $M_1 = \{x \in [-L, L] :\ \Delta(x) \neq 0\}$. If M_z has measure zero and M_1 has strictly positive measure, that is, $g \neq f$, we have the inequality

$$J(g) - J(f) \geq \kappa\int_{-L}^{L} |\alpha'(x)|^p\,\Delta(x)^2\,dx > 0.$$

Hence in this case, the solution of Q is uniquely determined.

Define the function U on the interval $(0, \infty)$ by the equation

$$U(\lambda) = \int_{-L}^{L} \Pi_{[\frac{c-D}{c+D},\ \frac{c+D}{c-D}]}(\lambda\,|\alpha'(y)|\,)\,dy,\ \lambda \in (0, \infty).$$

For all z_1, $z_2 \in (0, \infty)$, we have the inequality

$$\left|\Pi_{[\frac{c-D}{c+D},\ \frac{c+D}{c-D}]}(z_1) - \Pi_{[\frac{c-D}{c+D},\ \frac{c+D}{c-D}]}(z_2)\right| \leq |z_1 - z_2|.$$

Hence for all λ_1, $\lambda_2 \in (0, \infty)$ we have

$$\begin{aligned} |U(\lambda_1) - U(\lambda_2)| &\leq \int_{-L}^{L} \left|\Pi_{[\frac{c-D}{c+D},\ \frac{c+D}{c-D}]}(\lambda_1\,|\alpha'(y)|) - \Pi_{[\frac{c-D}{c+D},\ \frac{c+D}{c-D}]}(\lambda_2|\alpha'(y)|)\right|\,dy \\ &\leq \int_{-L}^{L} |\ \lambda_1\,|\alpha'(y)| - \lambda_2|\alpha'(y)|\ |\,dy \\ &= |\lambda_1 - \lambda_2|\int_{-L}^{L} |\alpha'(y)|\,dy. \end{aligned}$$

Thus U is Lipschitz continuous. We have $\lim_{\lambda\to 0} U(\lambda) = 2L\frac{c-D}{c+D} < 2L$.

Since $\int_{L}^{L} |\alpha'(y)|\,dy > 0$ we have $\lim_{\lambda\to\infty} U(\lambda) = 2L\frac{c+D}{c-D} > 2L$. Hence there exists a number $\lambda > 0$ such that $U(\lambda) = 2L$, which means that (5.12) is valid.

The case where $\int_{L}^{L} |\alpha'(y)|\,dy = 0$ is trivial. □

5.4. Proofs of the main results.

Proof of Theorem 3.1. For $p = 1$, $W(0) = W(T)$ for all $\phi \in \Phi$ (see Remark 5.2), so $\phi(t) = L \in \Phi$ is a solution of P.

Assume that $p > 1$. Due to (5.9) and the transformation of the set of admissible controls described in section 5.1, problem P is equivalent to the problem

$$P_1 : \text{Find } h \in H \text{ such that } J(h') = \int_{-L}^{L} \frac{|\alpha'(x)|^p}{h'(x)^{p-1}}\,dx \text{ is minimized.}$$

For all $h \in H$, we have $h'(x) \in F$, and, for all $f \in F$, we have $h(x) = -L + \int_{-L}^{x} f(x)\,ds \in H$. Moreover, $J(h') = J(f)$. Due to the representation (5.4) of the set H, this implies that P_1 is equivalent to Q.

Lemma 5.6 implies the existence of a solution of Q, which yields in turn the existence of a solution of P.

By Theorem 4.1, we have $\alpha' \in L^p(-L, 3L)$, and, due to the compatibility conditions $y_0(0) = y_0(L) = 0$ in the definition of the set B, the definition of the function α implies that, for all $x \in [-L, L]$, we have $\alpha'(x) = A(x)$.

Hence the definition of the set M_z in Theorem 3.1 is equivalent to the definition in Lemma 5.6. If the set M_z has measure zero, Lemma 5.6 implies the uniqueness of the solution of Q. We have seen that each function $f \in F$ corresponds to an admissible function $\phi \in \Phi$ with $J(f) = W(T)$. This implies the uniqueness of the solution of P. □

Proof of Theorem 3.2. If $\int_{L}^{L} |A(y)|\,dy > 0$, Lemma 5.6 implies the existence of a number $\lambda > 0$ such that (5.12) holds. If (5.12) holds for $\lambda > 0$, then (3.2) is valid with this value of λ, and, for the function f defined in Lemma 5.6, we have

$$f(x) = \Pi_{[\frac{c-D}{c+D},\ \frac{c+D}{c-D}]}(\lambda\,|A(x)|\,).$$

For the function h defined in Theorem 3.2, we have $h(x) = -L + \int_{-L}^{x} f(x)\,ds$. Since f solves Q, h solves problem P_1 defined above. In Lemma 5.2, we have shown that the solution ϕ of P can then be obtained by $\phi = \theta^{-1}(h)$. Equations (5.7) and (5.6) show that ϕ is given by (3.3). Equation (5.9) yields the minimal value of $W(T)$ and the result for the case $\int_{L}^{L} |A(y)|\,dy = 0$. □

6. Conclusion. We study a system that is controlled through the movement of the boundary and where the boundary movements are described by Lipschitz continuous functions. To obtain a well-posed problem, we require that the Lipschitz constants for the admissible controls are less than or equal to a given number D that is strictly less than the speed of wave propagation. We give a representation of a boundary movement that generates a maximal decrease of the energy in the given finite time interval. In particular, we give sufficient conditions for the existence and uniqueness of an optimal movement. Due to the nature of our system it is impossible to drive the energy arbitrarily close to zero unless it is zero from the start. For some initial states, it is even impossible to achieve any energy decrease by boundary movement control. The optimal energy decrease depends on the initial state. Depending on the initial state a considerable reduction of the energy can be achieved.

Acknowledgment. I want to thank the anonymous referees for their comments.

REFERENCES

[1] S. A. Avdonin and S. S. Ivanov, *Families of Exponentials*, Cambridge University Press, Cambridge, 1995.

[2] C. Bardos and G. Chen, *Control and stabilization for the wave equation, part* III: *Domain with moving boundary*, SIAM J. Control Optim., 19 (1981), pp. 123–138.

[3] N. Balazs, *On the solution of the wave equation with moving boundaries*, J. Math. Anal. Appl., 3 (1961), pp. 472–484.

[4] C. Castro and E. Zuazua, *Unique continuation and control for the heat equation from an oscillating lower dimensional manifold*, SIAM J. Control Optim., 43 (2005), pp. 1400–1434.

[5] M. C. Delfour and J. P. Zolesio, *Shapes and Geometries: Analysis, Differential Calculus and Optimization*, SIAM, Philadelphia, 2001.

[6] J. Dittrich, P. Duclos, and N. Gonzalez, *Stability and instability of the wave equation solution in a pulsating domain*, Rev. Math. Phys., 10 (1998), pp. 925–962.
[7] J. Dittrich, P. Duclos, and P. Seba, *Instability in a classical periodically driven string*, Phys. Rev. E, 49 (1994), pp. 3535–3538.
[8] M. Gugat, *Optimal boundary control of a string to rest in finite time with continuous state*, ZAMM Z. Angew. Math. Mech., 86 (2006), pp. 134–150.
[9] A. Khapalov, *Controllability of the wave equation with moving point control*, Appl. Math. Optim., 31 (1995), pp. 155–175.
[10] A. Khapalov, *Observability and stabilization of the vibrating string equipped with bouncing point sensors and actuators*, Math. Methods Appl. Sci., 24 (2001), pp. 1055–1072.
[11] W. Krabs, *Optimal control of processes governed by partial differential equations part* ii: *Vibrations*, Z. Oper. Res., 26 (1982), pp. 63–86.
[12] W. Krabs, *On moment theory and controllability of one–dimensional vibrating systems and heating processes*, Lecture Notes in Control and Inform. Sci. 173, Springer-Verlag, Heidelberg, 1992.
[13] J. L. Lions, *Exact controllability, stabilization and perturbations of distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.
[14] O. Meplan and C. Gignoux, *Exponential growth of the energy of a wave in a* $1D$ *vibrating cavity: Application to the quantum vacuum*, Phys. Rev. Lett., 76 (1996), pp. 408–410.
[15] G. K. Pedersen, *Analysis Now*, Springer-Verlag, New York, 1989.
[16] D. L. Russell, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–560.
[17] C. Truchi and J. P. Zolesio, *Wave equation in time periodical domain*, in Stabilization of Flexible Structures, Proceedings of the ComCon Workshop, Montpellier/France, 1987, A. V. Balakrishnan and J. P. Zolesio, eds., ISBN 0-911575-37-5, Springer-Verlag, Berlin, 1988, pp. 282–294.
[18] J. P. Zolesio and C. Truchi, *Shape stabilization of wave equation*, in Boundary Control and Boundary Variations, Proceedings of the IFIP WG 7.2 Conference, Nice/France, 1987, Lecture Notes in Control and Inform. Sci. 100, J. P. Zolesio, ed., Springer-Verlag, Berlin, 1988, pp. 372–398.
[19] J. P. Zolesio, *Shape stabilization of flexible structure*, in Distributed Parameter Systems, Proceedings of the 2nd International Conference, Vorau, 1984, Lecture Notes in Control and Inform. Sci. 75, F. Kappel and K. Kunisch, eds., Springer-Verlag, Berlin, 1985, pp. 446–460.
[20] E. Zuazua, *Optimal and approximate control of finite-difference approximation schemes for the* $1d$ *wave equation*, Rend. Mat. Appl., 7 (2004), pp. 201–237.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1726–1753

CONSTRAINED DIRICHLET BOUNDARY CONTROL IN L^2 FOR A CLASS OF EVOLUTION EQUATIONS*

K. KUNISCH† AND B. VEXLER‡

Abstract. Optimal Dirichlet boundary control based on the very weak solution of a parabolic state equation is analyzed. This approach allows us to consider the boundary controls in L^2, which has advantages over approaches which consider control in Sobolev spaces involving (fractional) derivatives. Pointwise constraints on the boundary are incorporated by the primal-dual active set strategy. Its global and local superlinear convergences are shown. A discretization based on space-time finite elements is proposed and numerical examples are included.

Key words. Dirichlet boundary control, inequality constraints, parabolic equations, very weak solution

AMS subject classifications. 49J20, 35K05, 49K20, 49K05, 49M29

DOI. 10.1137/060670110

1. Introduction.

In this work we focus on the Dirichlet boundary optimal control problem with pointwise constraints on the boundary, formally given by

$$\begin{cases} \min & J(y,u) \\ \text{subject to} & \partial_t y - \kappa\Delta y + b\cdot\nabla y = f \quad \text{in } Q, \\ & y = u, \quad u \le \psi \quad \text{on } \Sigma, \\ & y(0) = y_0 \quad \text{in } \Omega, \end{cases} \tag{1.1}$$

where $Q = (0,T]\times\Omega$, $\Sigma = (0,T]\times\partial\Omega$, and κ, b, f, y_0, ψ, and $T > 0$ are fixed. We propose and analyze a function space formulation which is amenable to efficient numerical realizations. To incorporate the constraints numerically the primal-dual active set (PDAS) strategy is used and its convergence is investigated. We also propose a space-time Galerkin approximation and provide numerical examples.

The specific difficulties involved in Dirichlet control problems result from the fact that they are not of variational type. In the literature several treatments of Dirichlet boundary control problems can be found, where the function space for the controls is H^s with $s \ge \frac{1}{2}$. As a consequence, the numerical realization by finite elements or finite differences is more involved than if the control space were L^2. Our approach will be based on the concept of very weak solutions to the state equation. This allows the use of L^2 as the control space.

Let us briefly describe possible approaches to treating Dirichlet boundary optimal control problems. While in our work we shall treat the time-dependent case, it will be convenient for the present purpose to restrict our attention to a tracking-type optimal

*Received by the editors September 18, 2006; accepted for publication (in revised form) March 20, 2007; published electronically November 9, 2007.
http://www.siam.org/journals/sicon/46-5/67011.html

†Institute for Mathematics and Scientific Computing, University of Graz, Heinrichstraße 36, A-8010 Graz, Austria (karl.kunisch@uni-graz.at).

‡Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenberger Straße 69, 4040 Linz, Austria (boris.vexler@oeaw.ac.at).

control problem with the most simple stationary elliptic equation as constraint:

$$
(1.2)\qquad \begin{cases} \min & \frac{1}{2}|y-z|^2_{L^2(\Omega)} + \frac{\beta}{2}|u|^2_{L^2(\partial\Omega)} \\ \text{over} & (y,u) \in L^2(\Omega)\times L^2(\partial\Omega) \\ \text{subject to} & -(y,\Delta v)_{L^2(\Omega)} = -(u, \frac{\partial}{\partial n} v)_{L^2(\partial\Omega)} \quad \text{for all } v \in H^2(\Omega)\cap H^1_0(\Omega) \\ \text{and} & u \le \psi \quad \text{on } \partial\Omega, \end{cases}
$$

where $z \in L^2(\Omega)$ and $\partial\Omega$ denotes the boundary of the domain Ω. The variational equation in (1.2) is the very weak form of

$$
\begin{cases} -\Delta y = 0 \quad \text{in } \Omega, \\ y = u \quad \text{on } \partial\Omega; \end{cases}
$$

see [36]. In our work we shall use the analogue of (1.2). If the state variable y is considered in $H^1(\Omega)$, then a proper formulation is given by

$$
(1.3)\qquad \begin{cases} \min & \frac{1}{2}|y-z|^2_{L^2(\Omega)} + \frac{\beta}{2}|u|^2_{H^{\frac{1}{2}}(\partial\Omega)} \\ \text{over} & (y,u) \in H^1(\Omega)\times H^{\frac{1}{2}}(\partial\Omega) \\ \text{subject to} & (\nabla y, \nabla v)_{L^2(\Omega)} = 0 \quad \text{for all } v \in H^1_0(\Omega) \text{ and } y = u \text{ on } \partial\Omega \\ \text{and} & u \le \psi \quad \text{on } \partial\Omega. \end{cases}
$$

For both formulations (1.2) and (1.3) it is classical to argue existence of a unique solution; see, e.g., [36]. Numerically realizing the $H^{1/2}$-norm in (1.3) is more involved than realizing the L^2-norm in (1.2). To avoid difficulties with implementing the $H^{1/2}$-norm it was replaced in several publications by the H^1-norm. As a consequence the Laplace–Beltrami operator appears in the optimality condition. This formulation, properly modified for the specific application and without control constraints, was used in the context of optimal boundary control of the Navier–Stokes equations and the Boussinesq equations; see, e.g., [23, 24, 33]. For a numerical wavelet–based realization of H^s-norms in the context of Dirichlet control of elliptic equations, we refer to [30].

A third alternative is given by

$$
(1.4)\qquad \begin{cases} \min & \frac{1}{2}|y-z|^2_{H^1(\Omega)} + \frac{\beta}{2}|u|^2_{L^2(\partial\Omega)} \\ \text{over} & (y,u) \in H^1(\Omega)\times H^{1/2}(\partial\Omega) \\ \text{subject to} & (\nabla y, \nabla v)_{(\Omega)} = 0 \quad \text{for all } v \in H^1_0(\Omega) \text{ and } y = u \quad \text{on } \partial\Omega \\ \text{and} & u \le \psi \quad \text{on } \partial\Omega. \end{cases}
$$

Again existence can be argued by standard arguments, but for (1.4), differently from (1.2) and (1.3), the essential term for obtaining coercivity is the H^1-norm of the tracking functional. Just like (1.2) this formulation also avoids having to deal with fractional order Sobolev spaces. It was used in the context of boundary control of the stationary Navier–Stokes equations in [15], for example. In the adjoint equation, however, a Laplacian now appears in the source term acting on the defect $y - z$.

Besides the difficulties already mentioned with (1.3) and (1.4) there is yet another, possibly more essential, reason to favor the formulation in (1.2). For (1.2) the Lagrange multiplier associated to the constraint $u \le \psi$ is an L^2-function, whereas

it is only a measure for the formulations in (1.3) and (1.4). As a consequence the complementarity conditions related to the inequality constraint can be expressed in a pointwise a.e. manner by the common pointwise complementarity functions like the max or the Fischer–Burmeister functions only for formulation (1.2). Such a pointwise formulation is a basis for efficient optimization algorithms such as the PDAS strategy or the semismooth Newton method.

Let us also recall the possibility of approximating Dirichlet boundary control problems by regularization based on Robin boundary controls of the form $\delta \frac{\partial y}{\partial n} + y = u$ for $\delta \to 0^+$. This results in the variational formulation

$$
\left\{
\begin{array}{ll}
\text{min} & \frac{1}{2}|y-z|^2_{L^2(\Omega)} + \frac{\beta}{2}\,|u|^2_{L^2(\partial\Omega)} \\
\text{over} & (y,u) \in H^1(\Omega) \times L^2(\partial\Omega) \\
\text{subject to} & (\nabla y, \nabla v)_{L^2(\Omega)} = \frac{1}{\delta}(y-u,v)_{L^2(\partial\Omega)} \quad \text{for all } v \in H^1(\Omega) \\
\text{and} & u \le \psi \quad \text{on } \partial\Omega.
\end{array}
\right. \tag{1.5}
$$

The choice of δ remains a delicate matter. This approach was used for stationary and nonstationary problems in [6] and [2], respectively. In [5] a numerical approach to Dirichlet boundary control based on a discretization using the Nitsche method was proposed.

We next point out some additional features of this paper. As already mentioned, the pointwise inequality constraint $u \le \psi$ will be treated by the PDAS algorithm. Its global, as well as local, superlinear convergence will be analyzed. Here it is essential that the Lagrange multiplier is an L^2-function and that the resulting complementarity condition involving the max operation is Newton differentiable. This is the case for (1.2), whereas this is not true for the other two formulations. Newton differentiability will be shown for (1.2) for time-dependent problems in the present paper. For stationary problems it easily follows as well.

Discretization of infinite-dimensional problems will be carried out by a space-time finite element method. This approach guarantees that the algorithm is invariant with respect to the ordering of discretization of the problem and gradient computations.

In spite of the fact that we use the very weak solution concept as our functional analytic setting for Dirichlet boundary control, the numerical discretization is based on standard space-time Galerkin finite-dimensional spaces. This will be justified by the fact that the solutions of the optimal control problems are more regular than those required by (1.2).

In our numerical implementation we use piecewise (bi)linear elements for spatial discretization of the primal and adjoint states as well as for the controls. This may appear to be incompatible at first, since the optimality condition involves $\frac{\partial p}{\partial n}$ and u in an additive manner, where p denotes the adjoint state. However, we replace $\frac{\partial p}{\partial n}$ by a variational expression in such a way that the resulting discretization is well balanced.

In section 2 we gather well-posedness results and a priori estimates for a class of evolution equations with Dirichlet boundary conditions in L^2. We include a convection term, due to future interest in considering similar problems for the Boussinesq systems, with specific nonconvex cost functionals, motivated by fluid mechanics considerations. In this case the convection coefficient is the velocity field of the fluid. Section 3 is devoted to the statements and analysis of the optimal control problems under consideration. In particular, we describe regularity properties of the optimal solutions. These are not only of interest in their own right, but are essential for superlinear convergence of the PDAS strategy, as explained in section 4. Section 5 contains a

description of the finite element discretization, and section 6 is devoted to selected numerical examples.

2. On the state equation. In this section we provide the necessary existence and a priori estimates for very weak solutions to

$$\begin{cases} \partial_t y - \kappa\Delta y + b\cdot\nabla y = f \quad \text{in } Q,\\ y = u \quad \text{on } \Sigma,\\ y(0) = y_0 \quad \text{in } \Omega, \end{cases} \tag{2.1}$$

where $Q = (0,T]\times\Omega$, $\Sigma = (0,T]\times\partial\Omega$, and Ω is a bounded domain in $\mathbb{R}^n$, $n\geq 2$, with C^2 boundary $\partial\Omega$. This boundary regularity of Ω guarantees that the Laplacian with homogeneous Dirichlet boundary conditions, denoted by Δ_0, is an isomorphism form $H^2(\Omega)\cap H_0^1(\Omega)$ to $L^2(\Omega)$. We shall denote the adjoint of Δ_0, mapping from $L^2(\Omega)$ to $H^{-2}(\Omega) = (H^2(\Omega)\cap H_0^1(\Omega))^*$ by Δ_0 as well. For any Banach space Y, we use the abbreviations $L^2(Y) = L^2(0,T;Y)$, $H^s(Y) = H^s(0,T;Y)$, $s\in[0,\infty)$, and $C(Y) = C([0,T];Y)$.

Further $\kappa>0$, $y_0\in H^{-1}(\Omega)$, $f\in L^2(H^{-2}(\Omega))$, $u\in L^2(\Sigma)$ and $b\in\mathbb{L}^\infty(Q)$, div b $\in L^\infty(L^{\hat n}(\Omega))$, where $\hat n = \max(n,3)$, and $\mathbb{L}^\infty(Q) = \bigotimes_{i=1}^n L^\infty(Q)$. At times we shall simply write $L^p(Q)$ for $\mathbb{L}^p(Q)$.

The very weak form of (2.1) that we shall utilize is given by

$$\begin{cases} \langle\partial_t y(t), v\rangle - \kappa(y(t),\Delta v) - (y(t), \operatorname{div}(b(t))\, v) - (y(t), b(t)\nabla v)\\ \qquad = \langle f(t), v\rangle - \kappa(u(t), \frac{\partial}{\partial n} v)_{\partial\Omega} \quad \text{for all } v\in H^2(\Omega)\cap H_0^1(\Omega)\\ \qquad\qquad\qquad\qquad\qquad\qquad \text{and a.e. } t\in(0,T),\\ y(0) = y_0, \end{cases} \tag{2.2}$$

where $\langle\cdot,\cdot\rangle = \langle\cdot,\cdot\rangle_{H^{-2}(\Omega),H^2(\Omega)\cap H_0^1(\Omega)}$ and denotes the canonical duality pairing, and $(\cdot,\cdot)$ and $(\cdot,\cdot)_{\partial\Omega}$ stand for the inner products in $L^2(\Omega)$ and $L^2(\partial\Omega)$, respectively. The last equality in (2.2) is understood in $H^{-1}(\Omega)$. The existence and uniqueness of a very weak solution in the space $L^2(Q)\cap H^1(H^{-2}(\Omega))\cap C(H^{-1}(\Omega))$ is shown in the following theorem.

THEOREM 2.1. *For every $\kappa>0$, $b\in L^\infty(Q)$, with* div $b\in L^\infty(L^{\hat n}(\Omega))$, $y_0\in H^{-1}(\Omega)$, $f\in L^2(H^{-2}(\Omega))$, *and* $u\in L^2(\Sigma)$, *there exists a unique very weak solution* $y\in L^2(Q)\cap H^1(H^{-2}(\Omega))\cap C(H^{-1}(\Omega))$ *satisfying*

$$|y|_{L^2(Q)\cap H^1(H^{-2}(\Omega))\cap C(H^{-1}(\Omega))} \leq C(|y_0|_{H^{-1}(\Omega)} + |f|_{L^2(H^{-2}(\Omega))} + |u|_{L^2(\Sigma)}), \tag{2.3}$$

where C depends continuously on $\kappa>0$, $|b|_{L^\infty(Q)}$, and $|\operatorname{div} b|_{L^\infty(L^{\hat n}(\Omega))}$, *and is independent of $f\in L^2(H^{-2}(\Omega))$, $u\in L^2(\Sigma)$, and $y_0\in H^{-1}(\Omega)$.*

Proof. Let us first assume existence of y with the claimed regularity and verify the a priori estimate (2.3). Throughout, k will denote a generic embedding constant. Let us introduce the transformed state-variable $\hat y(t) = y(t)e^{-ct}$, $c\geq 0$, and note that if y is a very weak solution of (2.1), then $\hat y\in L^2(Q)\cap H^1(H^{-2}(\Omega))$ is a very weak solution of

$$\begin{cases} \partial_t\hat y + c\hat y - \kappa\Delta\hat y + b\cdot\nabla\hat y = \hat f \quad \text{in } Q,\\ \hat y = \hat u \quad \text{on } \Sigma,\\ \hat y(0) = y_0 \quad \text{in } \Omega, \end{cases}$$

where $\hat f = f e^{-ct}$, $\hat u = u e^{-ct}$. The constant c will be fixed below. We further introduce $\omega = (-\Delta_0)^{-1}\,\hat y \in L^2(H^2(\Omega)\cap H_0^1(\Omega))\cap H^1(L^2(\Omega))$ and note that ω satisfies for all $v \in H^2(\Omega)\cap H_0^1(\Omega)$

$$\begin{aligned}\langle(-\Delta_0)\,\partial_t\omega(t),\,v\rangle + \kappa(\Delta_0\,\omega(t),\,\Delta v) + c(-\Delta_0\,\omega(t),\,v)\\ + (\Delta_0\,\omega(t),\,\operatorname{div} b(t)\,v) + (\Delta_0\,\omega(t),\,b(t)\nabla v) = \langle\hat f(t),\,v\rangle - \kappa\left(\hat u(t),\,\frac{\partial}{\partial n}\,v\right)_{\partial\Omega}\end{aligned}$$

for all $t\in(0,T)$. Setting $v=\omega(t)$ and integrating over $(0,t)$, we find

$$\begin{aligned}\frac12|\nabla\omega(t)|^2 - \frac12|\nabla\omega(0)|^2 + \kappa\int_0^t|\Delta_0\;\omega(s)|^2\,ds + c\int_0^t|\nabla\omega(s)|^2\,ds\\ + \int_0^t(\Delta_0\,\omega(s),\,\operatorname{div} b(s)\,\omega(s))\,ds + \int_0^t(\Delta_0\,\omega(s),\,b(s)\nabla\omega(s))\,ds\\ = \int_0^t\langle\hat f(s),\omega(s)\rangle\,ds - \kappa\int_0^t\left(\hat u(s),\frac{\partial}{\partial n}w(s)\right)_{\partial\Omega},\end{aligned}$$

and consequently using $\|\frac{\partial}{\partial n}w(s)\|_{L^2(\partial\Omega)}\le k\,\|\Delta_0 w(s)\|_{L^2(\Omega)}$ we obtain

$$\begin{aligned}&\frac12|\nabla\omega(t)|^2\,ds + \kappa\int_0^t|\Delta_0\,\omega(s)|^2ds + c\int_0^t|\nabla\omega(s)|^2\,ds\\ &\le \frac12|\nabla\omega(0)|^2 + \frac{\kappa}{8}\int_0^t|\Delta_0\,\omega(s)|^2\,ds + \frac{2k}{\kappa}\,|\operatorname{div} b|^2_{L^\infty(L^{\hat n}(\Omega)}\int_0^t|\nabla\omega(s)|^2\,ds\\ &\quad+\frac{\kappa}{8}\int_0^t|\Delta_0\,\omega(s)|^2 + \frac{2|b|^2_{L^\infty(Q)}}{\kappa}\int_0^t|\nabla\omega(s)|^2\,ds + \frac{2k^2}{\kappa}\int_0^t|\hat f(s)|^2_{H^{-2}} + \frac{\kappa}{8}\int_0^t|\Delta_0\,\omega|^2\,ds\\ &\quad+2\kappa^2\int_0^t|\hat u(s)|^2_{L^2(\partial\Omega)}\,ds + \frac{\kappa}{8}\int_0^t|\Delta_0\,\omega(s)|^2\,ds\\ &\le \frac12|\nabla\omega(0)|^2 + \frac{4\kappa}{8}\int_0^t|\Delta_0\,\omega(s)|^2\,ds + \left(\frac{2k}{\kappa}\,|\operatorname{div} b|^2_{L^\infty(L^{\hat n}(\Omega))} + \frac{2|b|^2_{L^\infty(Q)}}{\kappa}\right)\int_0^t|\nabla\omega(s)|^2\,ds\\ &\quad+\frac{2k^2}{\kappa}\int_0^t|\hat f(s)|^2_{H^{-2}(\Omega)}\,ds + 2k^2\int_0^t|\hat u(s)|^2_{L^2(\partial\Omega)}\,ds\,.\end{aligned}$$

If we choose c such that

$$\frac{2k}{\kappa}\,|\operatorname{div} b|^2_{L^\infty(L^{\hat n}(\Omega))} + \frac{2\,|b|^2_{L^\infty(Q)}}{\kappa} \le \frac{c}{2}\,, \tag{2.4}$$

then

$$\begin{aligned}\frac12\,|\nabla\omega(t)|^2 + \frac{k}{2}\int_0^t|\Delta_0\,\omega(s)|^2\,ds + \frac{c}{2}\int_0^t|\nabla\omega(s)|^2\,ds\\ \le \frac12\,|\nabla\omega(0)|^2 + \frac{2k^2}{\kappa}\int_0^t|\hat f(s)|^2_{H^{-2}(\Omega)}\,ds + 2k^2\int_0^t|\hat u(s)|^2_{L^2(\partial\Omega)}\,ds\,.\end{aligned} \tag{2.5}$$

From (2.5) we deduce the existence of a constant C with the specified properties such that for all $t\in[0,T]$

$$|\hat y(t)|_{H^{-1}(\Omega)} + \int_0^t|\hat y(s)|^2_{L^2(\Omega)}\,ds \le C(|y_0|_{H^{-1}(\Omega)} + |f|_{L^2(H^{-2}(\Omega))} + |u|_{L^2(\Sigma)})\,,$$

and, since $\hat{y}(t) = y(t)e^{-ct}$, we find for a possibly modified C:

$$|y(t)|_{H^{-1}(\Omega)} + \int_0^t |y(s)|^2_{L^2(\Omega)} ds \le C(|y_0|_{H^{-1}(\Omega)} + |f|_{L^2(H^{-2}(\Omega))} + |u|_{L^2(\Sigma)}) . \tag{2.6}$$

Finally using (2.2) we obtain

$$\begin{aligned}\int_0^T |\partial_t y(t)|^2_{H^{-2}(\Omega)}\, dt &= \int_0^T \sup_{\substack{v\in H^2(\Omega)\cap H^1_0(\Omega),\\ |\Delta_0 v|\le 1}} \langle \partial_t y(t), v\rangle^2\, dt \\ &\le \kappa^2 \int_0^T |y(t)|^2\, dt + \int_0^T (y(t), \operatorname{div} b\, v)^2_{L^2(\Omega)}\, dt \\ &\quad + |b|^2_{L^\infty(Q)} \int_0^T |y(t)|^2\, dt + |f|^2_{L^2(H^{-2}(\Omega))} + k\,|u|^2_{L^2(\Sigma)} .\end{aligned}$$

For the second term on the right-hand side we estimate for $n > 4$

$$\begin{aligned}\int_0^T (y(t), \operatorname{div} b\, v)^2_{L^2(\Omega)}\, dt &\le \int_0^T |y(t)|^2_{L^2(\Omega)}\, |\operatorname{div} b|^2_{L^{2p}(\Omega)}\, |v|^2_{L^{2q}(\Omega)}\, dt \\ &\le k \int_0^T |y(t)|^2_{L^2(\Omega)}\, |\operatorname{div} b|^2_{L^{\hat{n}}(\Omega)}\, dt ,\end{aligned}$$

where $q = \frac{n}{n-4}$, $p = \frac{n}{4}$, and we used that $H^2(\Omega) \hookrightarrow L^{\frac{2n}{n-4}}(\Omega)$ and $\hat{n} > 2p = \frac{n}{2}$. The same estimate for dimensions $n = 2, 3, 4$ follows quite easily.

We obtain

$$\begin{aligned}&\int_0^T |\partial_t y|^2_{H^{-2}(\Omega)}\, dt \\ &\quad \le (\kappa^2 + k\,|\operatorname{div} b|_{L^\infty(L^{\hat{n}}(\Omega))} + |b|_{L^\infty(Q)}) \int_0^T |y(t)|^2\, dt + |f|^2_{L^2(H^{-2}(\Omega))} + k\,|u|^2_{L^2(\Sigma)} .\end{aligned}$$

Together with (2.6) this gives the desired estimate (2.3), which in particular also implies the uniqueness of the very weak solution to (2.1). Existence follows, for example, by combining this a priori estimate with a Galerkin procedure; see, e.g., [14, Chapter 18]. Alternatively analytic semigroup theory as in [32] can be used, noting that $-\kappa\Delta - b\cdot\nabla + cI$ generates an analytic semigroup in $L^2(\Omega)$. □

From the proof it follows that the solution y to (2.2) also satisfies the variational equation in Q given by

$$\begin{aligned}&\int_0^T \big(\langle \partial_t y(t), v(t)\rangle - \kappa(y(t), \Delta v(t)) - (y(t), \operatorname{div}(b(t))v(t)\,) - (y(t), b(t)\nabla v(t))\big)\, dt \\ &= \int_0^T \langle f(t), v(t)\rangle dt - \kappa \int_0^T \left(u(t), \frac{\partial}{\partial n} v(t)\right)_{L^2(\Omega)} dt \quad \text{for all } v \in L^2(H^2(\Omega)\cap H^1_0(\Omega)).\end{aligned} \tag{2.7}$$

The following result will allow us to consider cost functionals with pointwise-in-time evaluation of the trajectory.

COROLLARY 2.2. *If, in addition to the assumptions of Theorem* 2.1, $y_0 \in L^2(\Omega), f \in L^2(Q)$, *and* $u \in L^\infty(L^2(\partial\Omega))$, *then the very weak solution satisfies* $y \in$

$L^\infty(L^2(\Omega))$ and $y(\bar t)$ is a well-defined element in $L^2(\Omega)$ for every fixed $\bar t \in (0,T]$. Moreover, there exists a constant C independent of y_0, f, and u, such that for the corresponding solution $y = y(u)$ we have

$$|y(\bar t)|_{L^2(\Omega)} \le C(|y_0|_{L^2(\Omega)} + |f|_{L^2(Q)} + |u|_{L^\infty(L^2(\partial\Omega))}). \tag{2.8}$$

Proof. Fix $\kappa > 0$ and $b \in L^\infty(Q)$ with $\operatorname{div} b \in L^\infty(L^{\hat n}(\Omega))$. Without loss of generality we can assume that $A = -\kappa\Delta - b\cdot\nabla$ is uniformly elliptic. If not, we add a multiple c of the identity operator and accordingly multiply the constant C by the factor e^{cT}. Then A generates an analytic semigroup in $L^2(\Omega)$. For the equation with $u = 0$, estimate (2.8) follows by standard semigroup arguments. Using the superposition principle for (2.1) it therefore suffices to consider the case $y_0 = 0$, $f = 0$, and $u \in L^\infty(L^2(\partial\Omega))$. From [32] (see also [2]), we have the existence of $C > 0$ such that

$$|y|_{L^\infty(L^2(\Omega))} \le C|u|_{L^\infty(L^2(\partial\Omega))}\,. \tag{2.9}$$

From Theorem 2.1 we deduce $y \in C(H^{-1}(\Omega))$ and therefore

$$y(\bar t) = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\int_{-\varepsilon}^{0} y(\bar t + \tau)\,d\tau, \tag{2.10}$$

where the integral and the equality are interpreted in $H^{-1}(\Omega)$. Denoting

$$g_\varepsilon = \frac{1}{\varepsilon}\int_{-\varepsilon}^{0} y(\bar t + \tau)\,d\tau\,,$$

we obtain using (2.9) that

$$|g_\varepsilon|_{L^2(\Omega)} \le C|u|_{L^\infty(L^2(\partial\Omega))}.$$

Therefore, there is a subsequence converging weakly in $L^2(\Omega)$ to $\tilde g$ with

$$|\tilde g|_{L^2(\Omega)} \le C|u|_{L^\infty(L^2(\partial\Omega))}.$$

Using (2.10) we obtain that $y(\bar t) = \tilde g$. The desired conclusion follows. ∎

3. The optimal control problems and regularity of optimal controls. We consider the following two optimal control problems:

$$\text{(P1)}\qquad \begin{cases} \min & J(y,u) = G(y) + \frac{\beta}{2}\,|u|^2_{L^2(\Sigma)} \\ \text{over} & (y,u) \in L^2(Q)\times L^2(\Sigma) \\ \text{subject to} & (2.1) \text{ and } \ u \le \psi \text{ on } \ \Sigma, \end{cases}$$

where $\beta > 0$, $\psi \in L^2(\Sigma)$, and $G : L^2(Q) \to \mathbb{R}$ is bounded below, C^1, and weakly lower semicontinuous. The second problem under consideration is

$$\text{(P2)}\qquad \begin{cases} \min & J(y,u) = G\,(y(T)) + \frac{\beta}{2}\,|u|^2_{L^2(\Sigma)} \\ \text{over} & (y,u) \in L^2(Q)\times L^2_{T_1}(\Sigma) \\ \text{subject to} & (2.1), \quad \varphi \le u \le \psi \text{ on } \ \Sigma, \end{cases}$$

where $\beta > 0$, $\varphi, \psi \in L^\infty(L^2(\partial\Omega))$, $\varphi(x) < \psi(x)$ a.e. on Σ, and $G : L^2(\Omega) \to \mathbb{R}$ is bounded below, weakly lower semicontinuous, and C^1. Here

$$L^2_{T_1}(\Sigma) = \{u \in L^2(\Sigma) : u(t,x) = 0 \quad \text{for } t \in (T_1, T)\},$$

with $T_1 \in [0,T]$. For (P2) we require that $\varphi \le 0 \le \psi$ a.e. on (T_1, T). In section 3.2 we shall require that $T_1 < T$. The practical interpretation of setting $u = 0$ in a neighborhood of T is that the controller and the observer are not acting simultaneously. This choice will be important for obtaining better regularity results for $y(T)$.

We refer to (y,u) as a solution of (2.1) if that equation is satisfied in the very weak sense (2.2). Throughout this section the regularity assumptions of Theorem 2.1 for b are supposed to hold, and

$$f \in L^2(Q), \quad y_0 \in L^2(\Omega).$$

Then we have the following result.

PROPOSITION 3.1. *There exist solutions* $(y^*, u^*) = (y(u^*), u^*)$ *to* (P1) *as well as* (P2)*, which are unique if* G *is convex.*

This follows from weak sequential limit arguments (see, e.g., [36]) utilizing Theorem 2.1, respectively, Corollary 2.2. For (P1) a lower bound $\varphi \le u$ can be added and treated as we do for (P2). In (P2) the simultaneous use of upper and lower bounds for the control is essential to guarantee the $L^\infty(L^2(\partial\Omega))$ bound for the controls, which is required by Corollary 2.2.

The above theorem is valid for all $T_1 \le T$. If one additionally assumes that $T_1 < T$, then the condition $\varphi, \psi \in L^\infty(L^2(\partial\Omega))$ can be weakened to $\varphi, \psi \in L^2(\Sigma)$, and the statement of the theorem follows from additional regularity of $y(T)$ in this case.

3.1. Problem (P1). To argue the existence of Lagrange multipliers for the inequality constraint in (P1), we introduce

$$\begin{aligned} e = (e_1, e_2) : (L^2(Q) \cap H^1(H^{-2})) \times L^2(\Sigma) &\to L^2(H^{-2}(\Omega)) \times H^{-1}(\Omega)\,, \\ g : L^2(\Sigma) &\to L^2(\Sigma) \end{aligned}$$

by

$$\begin{aligned} \langle e_1(y,u), v\rangle &= \int_0^T \left(\langle \partial_t y - f, v\rangle - (y \operatorname{div} b, v) - \kappa(y, \Delta v) - (y, b\cdot\nabla v) + \kappa \left(u, \frac{\partial}{\partial n} v \right)_{\partial\Omega} \right) dt\,, \\ e_2(y,u) &= y(0) - y_0\,, \\ g(u) &= u - \psi \end{aligned}$$

for arbitrary $v \in L^2(H^2(\Omega) \cap H^1_0(\Omega))$. Recall that $L^2(Q) \cap H^1(H^{-2}) \subset C(H^{-1}(\Omega))$, so that e_2 is well defined. The linearizations e' of e and g' of g are obtained from e and g by deleting the affine terms y_0, f, and ψ, respectively. We introduce the Lagrangian

$$\mathcal{L}(y,u,p,p_0,\lambda) = G(y) + \frac{\beta}{2}|u|^2_{L^2(\Sigma)} + \langle (p,p_0), e(y,u)\rangle + (\lambda, g(u)).$$

From Theorem 2.1 it follows that (e', g') is surjective, and hence there exists a Lagrange multiplier $(p, p_0, \lambda) \in L^2(H^2(\Omega) \cap H^1_0(\Omega)) \times H^1_0(\Omega) \times L^2(\Sigma)$ associated to the constraints (e, g); see, e.g., [37]. It follows that the optimality system satisfied

by an optimal pair (y^*, u^*) is obtained by setting $\nabla_{y,u,p,p_0}\mathcal{L}(y,u,p,p_0,\lambda) = 0$, and $\lambda \geq 0$, $g(u) \leq 0$, $\lambda\, g(u) = 0$. Consequently the optimality system for (P1) is given by

$$\begin{cases} \partial_t y - \kappa\Delta y + b\cdot\nabla y = f \quad \text{in } Q, \\ y = u \text{ on } \Sigma,\ y(0) = y_0 \quad \text{in } \Omega, \\ -\partial_t p - \kappa\Delta p - \operatorname{div} b\, p - b\cdot\nabla p = -G'(y) \quad \text{in } Q, \\ p = 0 \text{ on } \Sigma, \quad p(T) = 0 \quad \text{in } \Omega, \\ \kappa\frac{\partial p}{\partial n} + \beta u + \lambda = 0 \quad \text{on } \Sigma, \\ \lambda = \max(0, \lambda + c(u-\psi)) \quad \text{on } \Sigma \end{cases} \tag{3.1}$$

for any $c > 0$. Moreover, $p(0) = p_0$. Note that the last equation in (3.1) is equivalent to $\lambda \geq 0$, $u \leq \psi$, and $\lambda(u-\psi) = 0$. The equations in the last two lines of (3.1) are equivalent to

$$u = \min\left(\psi, -\frac{\kappa}{\beta}\frac{\partial p}{\partial n}\right).$$

The equations in the first two lines of (3.1) are understood in the sense of very weak solutions. The time derivative in $\partial_t p$ must first be interpreted in variational form, but from the third equation in (3.1) it immediately follows that $p \in L^2(H^2(\Omega)\cap H_0^1(\Omega))\cap H^1(L^2(\Omega))$. This is consistent with the regularity results for parabolic equations, since $G'(y) \in L^2(Q)$; see, e.g., [31, p. 342]. If G is convex, then (3.1) is a necessary and sufficient optimal condition for (P1).

We now turn to regularity properties of the optimal solution on Σ. This result is essential for superlinear convergence of the PDAS method; see section 4. Henceforth let (y,u,p,λ) denote a solution to (3.1). The active and inactive sets at a solution are denoted by

$$\mathcal{A} = \{\,(t,x)\in\Sigma\,:\,u(t,x) = \psi\,\}, \quad \mathcal{I} = \{\,(t,x)\in\Sigma\,:\,u(t,x) < \psi\,\}.$$

THEOREM 3.2. *On the inactive set $\mathcal{I}$ we have for the optimal solution $u|\mathcal{I} \in L^{q_n}(\mathcal{I})$ with*

$$q_n = \begin{cases} \frac{2(n+1)}{n} & \text{if } n \geq 3, \\ 3-\varepsilon & \text{if } n = 2. \end{cases} \tag{3.2}$$

On the active set the regularity of u is determined by ψ. Moreover,

$$\frac{\partial p}{\partial n} \in L^{q_n}(\Sigma) \quad \text{and} \quad \left\|\frac{\partial p}{\partial n}\right\|_{L^{q_n}(\Sigma)} \leq C\,\|p\|_{L^2(H^2(\Omega))\cap H^1(L^2(\Omega))}$$

with an embedding constant C.

Proof. As already noted, $p \in L^2(H^2(\Omega))\cap H^1(L^2(\Omega))$. This implies that

$$\frac{\partial p}{\partial n} \in L^2(H^{\frac12}(\partial\Omega))\cap H^{\frac14}(L^2(\partial\Omega));$$

see [21], or [31, Chapter II and p. 342]. Since $H^{\frac14}(L^2(\partial\Omega)) \hookrightarrow L^4(L^2(\partial\Omega))$ (see [1]), we find

$$\frac{\partial p}{\partial n} \in L^2(H^{\frac12}(\partial\Omega))\cap L^4(L^2(\partial\Omega)), \tag{3.3}$$

and hence interpolation [42, Chapter 1] implies that

$$\frac{\partial p}{\partial n} \in L^{r_s}([H^{\frac{1}{2}}(\partial\Omega), L^2(\partial\Omega)]_s), \quad \text{where } \frac{1}{r_s} = \frac{1-s}{2} + \frac{s}{4}.$$

For $n \geq 3$ we use the fact that for $H^{\frac{1}{2}}(\partial\Omega) \hookrightarrow L^{\frac{2n-2}{n-2}}(\partial\Omega)$, and we obtain

$$[H^{\frac{1}{2}}(\partial\Omega), L^2(\partial\Omega)]_s \hookrightarrow L^{q_s}(\partial\Omega), \quad \text{where } \frac{1}{q_s} = \frac{(1-s)(n-2)}{2n-2} + \frac{s}{2}.$$

Next we choose s such that $r_s = q_s$, i.e.,

$$r_s = \frac{8}{4-2s} = \frac{2n-2}{n+s-2} = q_s.$$

This implies that $s = \frac{2}{n+1}$ and hence $q_s = \frac{2(n+1)}{n}$. Consequently for $n \geq 3$ we obtain $\frac{\partial p}{\partial n} \in L^{\frac{2(n+1)}{n}}(\Sigma)$.

For $n = 2$ we have that $H^{\frac{1}{2}}(\partial\Omega) \hookrightarrow L^r(\partial\Omega)$ for all $r < \infty$. Using similar arguments to those before, we deduce that $\frac{\partial p}{\partial n} \in L^{3-\frac{1}{r-1}}(\Sigma)$.

From (3.1) we have that $\frac{\partial p}{\partial n} = -\beta u$ on $\mathcal{I}$, and the asserted regularity of u follows. The desired estimate for $\|\frac{\partial p}{\partial n}\|_{L^{q_n}(\Sigma)}$ holds due to the continuity of all embeddings involved. ∎

Our next objective is to show that for the optimal solution u the corresponding very weak solution y to the state equation is in fact a variational solution in the sense that $y \in L^2(H^1(\Omega)) \cap H^1(H^{-1}(\Omega))$, $y = u$ a.e. on Σ, and

$$\int_Q \partial_t y\, v\, dxdt = \int_Q (-\kappa \nabla y\, \nabla v - b \cdot \nabla y\, v + fv) dxdt$$

for all $v \in L^2(H^2(\Omega) \cap H_0^1(\Omega))$. This is important for numerical realizations which are conveniently based on this formulation. We shall require the following lemma, which uses the notion of uniform 1-smooth regularity property of the boundary, for which we refer to [1].

LEMMA 3.3. *Let D be a domain in $\mathbb{R}^n$, having the uniform 1-smooth regularity property and a bounded boundary, and let $s \in [0,1]$.*

(a) *If $v \in H^s(D)$, then $\max(0,v) \in H^s(D)$ and*

$$|\max(0,v)|_{H^s(D)} \leq |v|_{H^s(D)}.$$

(b) *If $v \in H^s(0,T;L^2(D))$, then $\max(0,v) \in H^s(0,T;L^2(D))$ and*

$$|\max(0,v)|_{H^s(0,T;L^2(D))} \leq |v|_{H^s(0,T;L^2(D))}.$$

Proof. (a) For $s = 0$ the claim is trivial and for $s = 1$ it is well known; see [42]. Thus let us consider the case $0 < s < 1$. Under the stated regularity properties for ∂D, the interpolation norm on $H^s(D)$ is equivalent to the intrinsic $H^s(D)$-norm on D given by

$$|v|^2_{L^2(D)} + \int_D \int_D \frac{|v(x) - v(y)|^2}{|x-y|^{n+2s}}\, dxdy; \tag{3.4}$$

see [1]. Let $S_i \subset D \times D$ be given by

$$S_1 = \{(x,y) : v(x) \geq 0,\, v(y) \geq 0\},\ S_2 = \{(x,y) : v(x) \geq 0,\, v(y) < 0\},$$
$$S_3 = \{(x,y) : v(x) < 0,\, v(y) \geq 0\},\ S_4 = \{(x,y) : v(x) < 0,\, v(y) < 0\}.$$

Then with $v^+ = \max(0, v)$

$$\int_D \int_D \frac{|v^+(x) - v^+(y)|^2}{|x-y|^{n+2s}}\, dxdy \le \int_{s_1 \cup s_2 \cup s_3} \int_{s_1 \cup s_2 \cup s_3} \frac{|v(x) - v(y)|^2}{|x-y|^{n+2s}}\, dxdy \le \int_D \int_D \frac{|v(x) - v(y)|^2}{|x-y|^{n+2s}}\, dxdy,$$

and (a) follows. Turning to (b), from [29, Theorem 1.7], it is known that for $s \in (0,1)$ up to equivalence of norms we have

$$|v|^2_{H^s(L^2(D))} = |v|^2_{L^2(L^2(D))} + 2\int_0^T \int_0^{T-t} t^{-1-2s} \, |v(\tau) - v(t+\tau)|^2_{L^2(D)} d\tau dt.$$

Setting $t + \tau = r$ the last term can equivalently be expressed as

$$\int_0^T \int_\tau^T |r-\tau|^{-1-2s} \, |v(\tau) - v(r)|^2 \, dr d\tau,$$

and using the symmetry of this expression with respect to s and τ, we find

$$|v|^2_{H^s(L^2(D))} = |v|^2_{L^2(L^2(D))} + \int_0^T \int_0^T \frac{|v(\tau) - v(r)|^2_{L^2(D)}}{|\tau - r|^{1+2s}}\, dr d\tau,$$

which is analogous to (3.4). The integral term can be expressed as

$$\int_0^T \int_0^T \int_\Omega \frac{|v(\tau, x) - v(r, x)|^2}{|\tau - r|^{1+2s}}\, dx\, dr\, d\tau,$$

and hence the proof can be completed as in (a). ∎

THEOREM 3.4. *Let (y, u) denote a solution to* (P1) *and assume that $\psi \in L^2(H^{\frac{1}{2}}(\partial\Omega)) \cap H^{\frac{1}{4}}(L^2(\partial\Omega))$. Then y is a variational solution of the state equation with*

$$u \in L^2(H^{\frac{1}{2}}(\partial\Omega)) \cap H^{\frac{1}{4}}(L^2(\partial\Omega)) \quad \text{and} \quad y \in L^2(H^1(\Omega)) \cap H^{\frac{1}{2}}(L^2(\Omega)) \cap H^1(H^{-1}(\Omega)).$$

If, moreover, $G'(y) \in L^2(H^1(\Omega)) \cap H^{\frac{1}{2}}(L^2(\Omega))$, $y_0 \in H^{\frac{1}{2}-\epsilon}(\Omega)$, and $\psi \in L^2(H^1(\partial\Omega)) \cap H^{\frac{1}{2}}(L^2(\partial\Omega))$, then

$$u \in L^2(H^1(\partial\Omega)) \cap H^{\frac{1}{2}}(L^2(\partial\Omega)) \quad \text{and} \quad y \in L^2(H^{\frac{3}{2}-\epsilon}(\Omega)) \cap H^{\frac{3-2\epsilon}{4}}(L^2(\Omega))$$

for every $\epsilon \in (0, \frac{1}{2}]$. In addition $u = 0$ on $\mathcal{I} \cap (\{T\} \times \partial\Omega)$.

Proof. From the proof of Theorem 3.2 we have that

$$\frac{\partial p}{\partial n} \in L^2(H^{\frac{1}{2}}(\partial\Omega)) \cap H^{\frac{1}{4}}(L^2(\partial\Omega)).$$

From (3.1) with $\beta = c$ we deduce that $u = \min(0, -\frac{1}{\beta}\frac{\partial p}{\partial n} - \psi) + \psi$, and hence Lemma 3.3 implies that $u \in L^2(H^{\frac{1}{2}}(\partial\Omega)) \cap H^{\frac{1}{4}}(L^2(\partial\Omega))$. By regularity results for parabolic equations based on interpolation theory [35, p. 78] (with $s = -\frac{1}{2}$), we obtain that $y \in L^2(H^1(\Omega)) \cap H^{\frac{1}{2}}(L^2(\Omega))$. Therefore

$$\int_0^T \langle \partial_t y, v\rangle \, dt = \int_Q (-\kappa \nabla y \, \nabla v - b \cdot \nabla y \, v + f v) dx dt$$

for all $v \in L^2(H^2(\Omega) \cap H_0^1(\Omega))$. Since the right-hand side can uniquely be extended to a continuous functional on $L^2(H_0^1(\Omega))$, it follows that $\partial_t y \in L^2(H^{-1}(\Omega))$. From (2.7), moreover, $y = u$ in $L^2(H^{\frac{1}{2}}(\partial\Omega))$. We conclude that y is a variational solution to (2.2).

If $G'(y) \in L^2(H^1(\Omega)) \cap H^{\frac{1}{2}}(L^2(\Omega))$, then $p \in L^2(H^3(\Omega)) \cap H^{\frac{3}{2}}(L^2(\Omega))$ [35, p. 32] (with $k = 1$). It follows that $\frac{\partial p}{\partial n} \in L^2(H^{\frac{3}{2}}(\partial\Omega)) \cap H^{\frac{3}{4}}(L^2(\partial\Omega))$; see, e.g., [20, p. 9]. Due to the regularity assumption on ψ and Lemma 3.3, we find that $u \in L^2(H^1(\partial\Omega)) \cap H^{\frac{1}{2}}(L^2(\partial\Omega))$. This implies that $y \in L^2(H^{\frac{3}{2}-\epsilon}(\Omega)) \cap H^{\frac{3}{4}-\frac{\epsilon}{2}}(L^2(\Omega))$ for every $\epsilon > 0$ [35, p. 78] (with $s = -\frac{1}{4} - \frac{\epsilon}{2}$). Regularity of p implies that $p(T) \in H^{2-\epsilon}(\Omega)$ and hence $\frac{\partial p}{\partial n}(T) \in H^{\frac{1}{2}-\epsilon}(\partial\Omega)$. Since $p(T) = 0$ on Ω we find that $\frac{\partial p}{\partial n}(T) = 0$ on $\partial\Omega$. Hence from the fifth equation in (3.1) we deduce that $u = 0$ on $\mathcal{I} \cap (\{T\} \times \partial\Omega)$. □

Remark 3.1. For $G(y) = \frac{1}{2}|y - y_d|^2$ the condition $G'(y) \in L^2(H^1(\Omega)) \cap H^{\frac{1}{2}}(L^2(\Omega))$ is satisfied if $y_d \in L^2(H^1(\Omega)) \cap H^{\frac{1}{2}}(L^2(\Omega))$ and $\psi \in L^2(H^{\frac{1}{2}}(\partial\Omega)) \cap H^{\frac{1}{4}}(L^2(\partial\Omega))$.

COROLLARY 3.5 (extra L^p regularity). *By interpolation one can show that if* $u \in L^2(H^1(\partial\Omega)) \cap H^{\frac{1}{2}}(L^2(\partial\Omega))$, *then* $u \in L^{q_\epsilon}(\Sigma)$, *where* $q_\epsilon = \frac{2(n+1)}{n-1} - \epsilon$, *for every* $\epsilon > 0$.

3.2. Problem (P2). We first derive the optimality system for (P2). This requires more care than for (P1) since G in this case is not defined on the space of trajectories $L^2(Q)$.

Let (y, u) denote an optimal solution to (P2). We shall require that $G'(y(T)) \in H_0^1(\Omega)$. This will guarantee the required regularity of the adjoint state. In case $G(y(T)) = \frac{1}{2}|y(T) - z|^2$, this is the case if $y(T) - z \in H_0^1(\Omega)$, i.e., we require regularity of $y(T)$ (and z) beyond that which is guaranteed by Corollary 2.2, as well as boundary conditions for $y(T) - z$. The required regularity of y at T can be achieved by restricting u to be a function of t only in a neighborhood of T. To take into consideration the additional boundary condition, we require that $u = 0$ in a neighborhood of $T = 0$. Then by semigroup theory $y(T) \in H_0^1(\Omega) \cap H^2(\Omega)$ and, if $z \in H_0^1(\Omega)$, we have $y(T) - z \in H_0^1(\Omega)$. Thus for tracking-type functionals the requirement that $G'(y(T)) \in H_0^1(\Omega)$ holds if $u \in L^2_{T_1}(\Sigma)$ and $z \in H_0^1(\Omega)$. This motivates the use of $L^2_{T_1}(\Sigma)$ in (P2).

THEOREM 3.6. *Let* (y, u) *denote a solution to* (P2) *with* $T_1 < T$ *and assume that* $G'(y(T)) \in H_0^1(\Omega)$. *Then there exist* $p \in L^2(H^2(\Omega) \cap H_0^1(\Omega)) \cap H^1(L^2(\Omega))$ *and* $\lambda \in L^2(\Sigma_{T_1})$ *such that for all* $c > 0$

$$
\text{(3.5)} \qquad \begin{cases} \partial_t y - \kappa\Delta y + b \cdot \nabla y = f \quad \text{in } Q, \\ y = u \quad \text{on } \Sigma, \quad y = y_0 \quad \text{in } \Omega, \\ -\partial_t p - \kappa\Delta p - \operatorname{div} b\, p - b \cdot \nabla p = 0 \quad \text{in } Q, \\ p = 0 \quad \text{on } \Sigma, \quad p(T) = -G'(y(T)) \quad \text{in } \Omega, \\ \kappa \frac{\partial p}{\partial n} + \beta u + \lambda = 0 \quad \text{on } \Sigma_{T_1}, \\ \lambda = \max(0, \lambda + c(u - \psi)) + \min(0, \lambda + c(u - \varphi)) \quad \text{on } \Sigma_{T_1} \end{cases}
$$

holds, where $\Sigma_{T_1} = (0, T_1) \times \partial\Omega$.

Proof. From Theorem 2.1 the affine mapping $u \to y(u)$ is continuous from $L^2(\Sigma)$ to $L^2(Q) \cap H^1(H^{-2}(\Omega))$. The linearization $\dot{y}$ at u in direction h satisfies

$$
\text{(3.6)} \quad \langle \partial_t \dot{y}(t),\, v\rangle - \kappa(\dot{y}(t),\, \Delta v) - (\dot{y}(t),\, \operatorname{div}(b(t))v) - (\dot{y}(t),\, b(t)\nabla v) = \kappa \left(h(t), \frac{\partial}{\partial n} v \right)_{\partial\Omega} \quad \text{for all } v \in H^2(\Omega) \cap H_0^1(\Omega) \text{ and a.e. } t \in (0, T).
$$

Moreover, by Corollary 2.2, the affine mapping $u \to y(T;u)$ is continuous from $L^\infty(\Sigma)$ to $L^2(\Omega)$, and hence it is differentiable at u in directions $h \in L^\infty(\Sigma)$. By assumption, $G'(y(T)) \in H_0^1(\Omega))$, and hence the solution to the adjoint equation satisfies $p \in L^2(H^2(\Omega) \cap H_0^1(\Omega)) \cap H^1(L^2(\Omega))$ [31]. Let $j(u) = J(y(u),u)$ denote the reduced cost functional corresponding to (P2). For the derivative at $u \in L^\infty(\Sigma)$ in direction $h \in L^2(\Sigma)$ we obtain by (3.6)

$$\begin{aligned}(j'(u),h)_{L^2(\Sigma)} &= (G'(y(T)),\, \dot y(T))_{L^2(\Omega)} + \beta(u,h)_{L^2(\Sigma)} \\ &= -(p(T),\, \dot y(T))_{L^2(\Omega)} + \beta(u,h)_{L^2(\Sigma)} = -\int_0^T \frac{d}{dt}(p(t),\dot y(t))_{L^2(\Omega)}dt + \beta(u,h)_{L^2(\Sigma)} \\ &= \left(\kappa\frac{\partial p}{\partial n} + \beta u,\, h\right)_{L^2(\Sigma)}.\end{aligned}$$

At the solution we therefore have

$$(j'(u),\, h-u) \geq 0 \quad \text{for all } \ h \in L^2_{T_1}(\Sigma), \ \text{with } \ \varphi \leq h \leq \psi. \tag{3.7}$$

Note that the directions h in (3.7) are in $L^\infty_{T_1}(\Sigma)$ as well. Define

$$\mathcal{A}_\varphi = \{(t,x) \in \Sigma_{T_1} : u = \varphi\}, \quad \mathcal{A}_\psi = \{(t,x) \in \Sigma_{T_1} : u = \psi\}, \quad \mathcal{I} = \Sigma_{T_1} \setminus (\mathcal{A}_\varphi \cup \mathcal{A}_\psi),$$

where $\Sigma_1 = (0,T_1)\times\partial\Omega$. Set $\mathcal{S} = \{(t,x) \in \mathcal{I} : j'(u) \geq 0\}$ and define $\bar h = \varphi\,\chi_{\mathcal{S}} + u\,\chi_{\Sigma\setminus\mathcal{S}}$, which satisfies $\varphi \leq \bar h \leq \psi$ on Σ_{T_1}. By (3.7)

$$0 \leq (j'(u),\, \bar h - u)_{L^2(\Sigma_{T_1})} = (j'(u),\, \varphi - u)_{L^2(\mathcal{S})} \leq 0,$$

and hence $j'(u) = 0$ on $\mathcal{S}$, since $\varphi < u < \psi$ on $\mathcal{S}$. Analogously one shows that $j'(u) = 0$ on $\{(t,x) \in \mathcal{I} : j'(u) \leq 0\}$ and hence $j'(u) = 0$ on $\mathcal{I}$. Next set $\mathcal{S}_\psi = \{(t,x) \in \mathcal{A}_\psi : j'(u) \geq 0\}$, and define $\bar h = \varphi\,\chi_{\mathcal{S}_\psi} + u\,\chi_{\Sigma\,\setminus\,\mathcal{S}_\psi}$. Then by (3.7)

$$0 \leq (j'(u),\, \bar h - u)_{L^2(\Sigma_{T_1})} = (j'(u),\, \varphi - \psi)_{L^2(\Sigma_{T_1})} \leq 0.$$

Since $\varphi < \psi$ a.e. on Σ_{T_1} this implies that $j'(u) = 0$ on $\mathcal{S}_\psi$ and hence $j'(u) \leq 0$ on $\mathcal{A}_\psi$. Analogously one shows that $j'(u) \geq 0$ on $\mathcal{A}_\varphi$.

Setting

$$\lambda = \begin{cases} -\kappa\frac{\partial p}{\partial n} - \beta u & \text{on } \ \Sigma_{T_1} \setminus \mathcal{I}, \\ 0 & \text{on } \ \mathcal{I}, \end{cases}$$

the last two equations of (3.5) follow and the optimality system is verified. □

COROLLARY 3.7. *Under the assumptions of Theorem* 3.4 *we have* $\frac{\partial p}{\partial n} \in L^{q_n}(\Sigma)$ *and* $u|\mathcal{I} \in L^{q_n}(\mathcal{I})$ *with* q_n *defined in* (3.2).

This is a direct consequence of Theorem 3.6, which asserts that $p \in L^2(H^2(\Omega)) \cap H^1(L^2(\Omega))$, and of the proof of Theorem 3.2.

COROLLARY 3.8. *Under the assumptions of Theorem* 3.6 *and if* $\varphi,\psi \in L^2(H^{\frac12}(\partial\Omega)) \cap H^{\frac14}(L^2(\partial\Omega))$, *then* y *is a variational solution of the state equation with*

$$u \in L^2(H^{\frac12}(\partial\Omega)) \cap H^{\frac14}(L^2(\partial\Omega)) \ \ \textit{and} \ \ y \in L^2(H^1(\Omega)) \cap H^{\frac12}(L^2(\Omega)) \cap H^1(H^{-1}(\Omega)).$$

If, moreover, $G'(y(T)) \in H^2(\Omega)\cap H_0^1(\Omega)$, $y_0 \in H^{\frac12-\epsilon}(\Omega)$, *and* $\varphi,\psi \in L^2(H^1(\partial\Omega))\cap H^{\frac12}(L^2(\partial\Omega))$, *then*

$$u \in L^2(H^1(\partial\Omega)) \cap H^{\frac{1-\varepsilon}{2}}(L^2(\partial\Omega)) \ \ \textit{and} \ \ y \in L^2(H^{\frac32-\epsilon}(\Omega)) \cap H^{\frac{3-2\epsilon}{4}}(L^2(\Omega))$$

for every $\epsilon \in (0, \frac{1}{2}]$.

Proof. The proof of the first part is similar to that of Theorem 3.4. By the last two equations of (3.5) we find

$$u = \max\left(\varphi, \min\left(\psi, -\frac{\kappa}{\beta}\frac{\partial p}{\partial n}\right)\right) \quad \text{a.e. on } \Sigma_{T_1}, \tag{3.8}$$

which is equivalent to $u = \max(0, \min(0, -\frac{\kappa}{\beta}\frac{\partial p}{\partial n} - \psi) + \psi - \varphi) + \varphi$. Since $\frac{\partial p}{\partial n} \in L^2(H^{\frac{1}{2}}(\partial\Omega)) \cap H^{\frac{1}{4}}(L^2(\partial\Omega))$ this implies by Lemma 3.3 that

$$u|(0,T_1) \in L^2(0,T_1; H^{\frac{1}{2}}(\partial\Omega)) \cap H^{\frac{1}{4}}(0,T_1; L^2(\partial\Omega)),$$

and by concatenation of functions in $H^{\frac{1}{4}}$ this implies that

$$u \in L^2(0,T; H^{\frac{1}{2}}(\partial\Omega)) \cap H^{\frac{1}{4}}(0,T; L^2(\partial\Omega))$$

(see [29, Proposition 1.13]), and hence $y \in L^2(H^1(\Omega)) \cap H^{\frac{1}{2}}(L^2(\Omega))$. Turning to the second part of the proof, we assume that $G'(y(T)) \in H^2(\Omega) \cap H_0^1(\Omega)$. Then $p \in L^2(H^3(\Omega)) \cap H^{\frac{3}{2}}(L^2(\Omega))$ [35, p. 32], and $\frac{\partial p}{\partial n} \in L^2(H^{\frac{3}{2}}(\partial\Omega)) \cap H^{\frac{3}{4}}(L^2(\partial\Omega))$. By (3.8) and concatenation of H^s-functions with $s \in [0, \frac{1}{2})$, we find that $u \in L^2(H^1(\partial\Omega)) \cap H^{\frac{1-\varepsilon}{2}}(L^2(\partial\Omega))$ for every $\varepsilon \in (0,1)$. This implies that $y \in L^2(H^{\frac{3}{2}-\varepsilon}(\Omega)) \cap H^{\frac{3-2\varepsilon}{4}}(L^2(\Omega))$. ∎

4. The PDAS strategy and its convergence properties. The PDAS strategy has proved to be very efficient for solving constrained optimal control problems [8]. We describe it here for (P1) and defer the necessary modifications for (P2) to Remark 4.3.

In addition to the assumptions on $G : L^2(Q) \to \mathbb{R}$ made in section 3, we assume that G is convex so that all auxiliary optimal control problems that arise in this section have unique solutions.

The PDAS strategy is an iterative algorithm which, based on the current iterate (u_k, λ_k), defines the active set

$$\mathcal{A}_k = \{\, x \in \Omega \,:\, \lambda_k(x) + c(u_k - \psi)(x) > 0 \,\}$$

and the inactive set

$$\mathcal{I}_k = \Omega \setminus \mathcal{A}_k.$$

The subsequent step consists in solving the optimal control problem with equality constraints only:

$$(\mathrm{P_k}) \qquad \begin{cases} \min & J(y,u) = G(y) + \frac{\beta}{2}|u|^2_{L^2(\Sigma)} \\ \text{over} & (y,u) \in L^2(Q) \times L^2(\Sigma) \\ \text{subject to} & (2.1) \ \text{ and } u = \psi \ \text{ on } \mathcal{A}_k. \end{cases}$$

This leads to the following iterative algorithm, in which step (iii) is the necessary and sufficient optimality condition for (P_k).

PDAS ALGORITHM.

(i) Choose $(u_1, \lambda_1) \in L^2(\Sigma) \times L^2(\Sigma),\ c > 0$.

(ii) Define $\mathcal{A}_k = \{\, x \in \Omega \,:\, \lambda_k(x) + c(u_k - \psi)(x) > 0 \,\},\ \mathcal{I}_k = \Omega \setminus \mathcal{A}_k$.

(iii) Solve for $(y_{k+1}, u_{k+1}, p_{k+1}) \in L^2(Q) \cap H^1(H^{-2}(\Omega)) \cap C(H^{-1}(\Omega)) \times L^2(\Sigma) \times L^2(H^2(\Omega) \cap H_0^1(\Omega))$:

$$\text{(4.1)} \qquad \begin{cases} \partial_t y - \kappa \Delta y + b \cdot \nabla y = f \quad \text{in } Q, \\ y = u \quad \text{on } \Sigma, \quad y(0) = y_0 \quad \text{in } \Omega, \\ -\partial_t p - \kappa \Delta p - \operatorname{div} b\, p - b \cdot \nabla p = -G'(y) \quad \text{in } Q, \\ p = 0 \quad \text{on } \Sigma, \quad p(T) = 0 \quad \text{in } \Omega, \\ u = \psi \text{ on } \mathcal{A}_k, \quad \kappa \frac{\partial p}{\partial n} + \beta u = 0 \text{ on } \mathcal{I}_k. \end{cases}$$

(iv) Set

$$\lambda_{k+1} = \begin{cases} 0 & \text{on } \mathcal{I}_k, \\ -\kappa \frac{\partial p_{k+1}}{\partial n} - \beta \psi & \text{on } \mathcal{A}_k. \end{cases}$$

(v) Stop or return to (ii).

Remark 4.1. For practical features of this algorithm, we refer to [8] and [9], for example. Here, we mention only that

1. for $k \geq 2$ the iterates of the algorithm are independent of the choice of c, and
2. if the algorithm finds two successive active sets, for which $\mathcal{A}_k = \mathcal{A}_{k+1}$, then $(y(u_k), u_k)$ is the solution of the problem.

The latter will be used as a stopping criterion for numerical examples in section 6.

Remark 4.2. The equality-constrained optimization problem (P_k) is solved using the Newton method for the reduced cost functional $j(u) = G(y(u)) + \frac{\beta}{2} |u|^2_{L^2(\Sigma)}$. The required first and second derivatives of j are computed using solutions of the adjoint problems; see, e.g., [3]. In section 5 we describe the computation of these derivatives on the discrete level.

For the following result it will be convenient to choose a specific initialization for λ, given by

$$\text{(4.2)} \qquad \begin{cases} \text{choose} & u_1 \in L^2(\Sigma), \\ \text{set} & \lambda_1 = -\kappa \frac{\partial p(u_1)}{\partial n} - \beta u_1, \\ \text{and set} & c = \beta \text{ for the first iteration.} \end{cases}$$

THEOREM 4.1. *If the PDAS algorithm is initialized by* (4.2)*, and if further* $\psi \in L^{\frac{2(n+1)}{n}}(\Sigma)$, $G' : L^2(Q) \to L^2(Q)$ *is locally Lipschitz, and* $|u_1 - u^*|_{L^2(\Sigma)}$ *is sufficiently small, then the iterates* $(y_k, u_k, p_k, \lambda_k)$ *converge superlinearly in* $L^2(Q) \cap H^1(H^{-2}(\Omega)) \cap C(H^{-1}(\Omega)) \times L^2(\Sigma) \times L^2(H^2(\Omega) \cap H_0^1(\Omega)) \times L^2(\Sigma)$ *to* $(y^*, u^*, p^*, \lambda^*)$.

Proof. Let us consider λ in the last equation of (3.1) as a function of u. Then (3.1) can equivalently be expressed as

$$\text{(4.3)} \qquad F(u) = \lambda(u) - \max(0, \lambda(u) + \beta(u - \psi)) = 0, \text{ where } F : L^2(\Sigma) \to L^2(\Sigma).$$

Note that (4.3) is equivalent to

$$\text{(4.4)} \qquad F(u) = \beta u - \beta \psi + \max\left(0, \kappa \frac{\partial p}{\partial n} + \beta \psi\right) = 0,$$

due to the fifth equation in (3.1). By Theorem 3.2 and the assumption that $\psi \in L^{\frac{2(n+1)}{n}}(\Sigma)$ we have that $\kappa\frac{\partial p}{\partial n} + \beta\psi \in L^{q_n}(\Sigma)$ with q_n defined in (3.2). Due to the fact that $q_n > 2$ we obtain that

$$u \to F(u)$$

is Newton differentiable, as introduced in Definition 1 of [25] (see Proposition 4.1 of [25]), with the generalized derivate of F at u in direction $h \in L^2(\Sigma)$ given by

$$G_F(u)h = \beta h + G_{\max}\left(\kappa\frac{\partial p}{\partial n} + \beta\psi\right)\frac{\partial p(h)}{\partial n},$$

where

$$G_{\max}(u)(x) = \begin{cases} 1 & \text{if } u(x) > 0, \\ 0 & \text{if } u(x) \le 0. \end{cases}$$

It was proved in general terms in [25, Theorem 4.1] that $G_F(u)$ has a bounded inverse from $L^2(\Sigma)$ to itself for every $u \in L^2(\Sigma)$. Hence it follows that the semismooth Newton algorithm applied to $F(u) = 0$ is locally superlinearly convergent. The semismooth Newton iteration consists of the iteration

$$\begin{cases} G_F(u_\kappa)\delta u = -F(u_\kappa), \\ u_{k+1} = u_k + \delta u. \end{cases} \tag{4.5}$$

In the following arguments we show that the semismooth Newton iteration and the PDAS strategy coincide. In principle this argument can be extracted from [25], but we believe that it is instructive to carry it out for the present case. A short consideration shows that a semismooth Newton step (4.5) is equivalent to

$$\begin{cases} \partial_t y_{k+1} - \kappa\Delta y_{k+1} + b\cdot\nabla y_{k+1} = f \quad \text{in } Q, \\ y_{k+1} = u_{k+1} \quad \text{on } \Sigma,\ y(0) = y_0 \quad \text{in } \Omega, \\ -\partial_t p_{k+1} - \kappa\Delta p_{k+1} - \operatorname{div} b\, p_{k+1} - b\cdot\nabla p_{k+1} = -G'(y_{k+1}) \quad \text{in } Q, \\ p_{k+1} = 0 \quad \text{on } \Sigma, \quad p_{k+1}(T) = 0 \quad \text{in } \Omega, \\ u_{k+1} = \psi \quad \text{on } \mathcal{A}_k^{SN},\ \kappa\frac{\partial p_{k+1}}{\partial n} + \beta u_{k+1} = 0 \quad \text{on } \mathcal{I}_k^{SN}, \end{cases} \tag{4.6}$$

where

$$\mathcal{A}_k^{SN} = \left\{x : \left(-\kappa\frac{\partial p_k}{\partial n} - \beta\psi\right)(x) > 0\right\}, \qquad \mathcal{I}_k^{SN} = \Omega \setminus \mathcal{A}_k^{SN}.$$

We further set

$$\lambda_{k+1} = \begin{cases} 0 & \text{on } \mathcal{I}_k^{SN}, \\ -\kappa\frac{\partial p_{k+1}}{\partial n} - \beta\psi & \text{on } \mathcal{A}_k^{SN} \end{cases} \tag{4.7}$$

and observe that

$$\lambda_k + \beta(u_k - \psi) = -\kappa\frac{\partial p_{k+1}}{\partial n} - \beta\psi \quad \text{for } k = 2, 3, \dots. \tag{4.8}$$

Note that

$$\lambda_k(u_k - \psi) = 0 \quad \text{for } k = 2, 3, \dots. \tag{4.9}$$

Hence $\lambda_k + \beta(u_k - \psi) > 0$ if and only if $\lambda_k + c(u_k - \psi) > 0$ for any $c > 0$. From (4.8) we have that

$$\mathcal{A}_k = \mathcal{A}_k^{SN} \quad \text{and} \quad \mathcal{I}_k = \mathcal{I}_k^{SN} \quad \text{for } k = 2, 3, \dots.$$

Therefore the PDAS strategy and the semismooth Newton iteration coincide, provided that their initialization phases coincide. For that it suffices to check that $\mathcal{A}_1 = \mathcal{A}_1^{SN}$. This is the case since for λ_1 as in (4.2) we have

$$\lambda_1 + \beta(u_1 - \psi) = -\kappa \frac{\partial p(u_1)}{\partial n} - \beta\psi_1.$$

Superlinear convergence of y_k to y^* in $L^2(Q) \cap H^1(H^{-2}(\Omega)) \cap C(H^{-1}(\Omega))$ follows from Theorem 2.1. Moreover, superlinear convergence of (p_k, λ_k) to (p^*, λ^*) in $L^2(H^2(\Omega) \cap H_0^1(\Omega)) \times L^2(\Sigma)$ is a consequence of (3.1) and (4.1),

$$\lambda^* - \lambda_k = -\beta(u^* - u_k) - \kappa \left(\frac{\partial p^*}{\partial n} - \frac{\partial p_k}{\partial n} \right),$$

and of Theorem 3.1. ∎

In Theorem 4.1 we addressed local convergence of the PDAS algorithm. We now turn to global convergence, i.e., to convergence from arbitrary initializations $(u_1, \lambda_1) \in L^2(\Sigma) \times L^2(\Sigma)$.

THEOREM 4.2. *If β is sufficiently large and $G(y) = \frac{1}{2}|y - z|^2_{L^2(Q)}$ for some $z \in L^2(Q)$, then the iterates $(y_k, u_k, p_k, \lambda_k) \to (y^*, u^*, p^*, \lambda^*)$ in $L^2(Q) \cap H^1(H^{-2}(\Omega)) \cap C(H^{-1}(\Omega)) \times L^2(\Sigma) \times L^2(H^2(\Omega) \cap H_0^1(\Omega)) \times L^2(\Sigma)$.*

Proof. Convergence of $(u_k, \lambda_k) \to (u^*, \lambda^*)$ in $L^2(\Sigma) \times L^2(\Sigma)$ follows from a general result in [27, Theorem 4.1]. It requires that $\beta > \|T\|_{\mathcal{L}(L^2(\Sigma), L^2(Q))}$, where $Tu = y(u)$. Convergence of (y_k, u_k) in the specified norms is a consequence of the governing equations for y_k and p_k. ∎

Remark 4.3. For (P2), under the assumptions of Theorem 3.6, identical assertions to Theorems 4.1 and 4.2 hold. (P2) differs from (P1) in that it involves a terminal observation and bilateral constraints. We again have, provided by Corollary 3.7, the necessary additional regularity $\frac{\partial p}{\partial n} \in L^{q_n}(\Sigma)$. Global convergence and local superlinear convergence for bilaterally constrained problems were treated in [27, Theorems 4.1 and 6.1].

5. Finite element discretization. In this section we discuss the space-time finite element discretization of the optimization problem under consideration. The space discretization of the state equation is based on the usual H^1-conforming finite elements, whereas the time discretization is done by a discontinuous Galerkin method as proposed in [16, 17]. We refer to [3, 38] for a detailed description of the space-time finite element methods for parabolic optimization problems including adaptivity. We emphasize that space-time Galerkin discretizations of optimal control problems allow a natural translation of the optimality system and the optimization algorithms from the continuous to the discrete level: in fact, the approaches "discretize-then-optimize" and "optimize-then-discretize" coincide. We return to this aspect in Remark 6.2 below.

Since the state equation (2.2) is considered in the very weak sense, it may appear at first that its approximation by finite elements based on the standard variational

formulation may not be appropriate. However, such an approach is justified since the optimal state and control which need to be approximated possess the common regularity of a variational solution; see Theorem 3.4. For an interesting discussion of finite element discretizations of equations with rough boundary data, we refer to [7] in the elliptic case and to [19] in the parabolic case. Finite element approximation of Dirichlet optimal control problems governed by elliptic equations are discussed in [11, 43].

For this section it is convenient to introduce the following notation: $V = H^1(\Omega)$, $V_0 = H_0^1(\Omega)$, $H = L^2(\Omega)$, and $X = L^2(0,T;V) \cap H^1(0,T;V^*)$. We introduce a bilinear form $a\colon X \times X \to \mathbb{R}$ corresponding to the standard variational formulation of the state equation:

$$a(y,v) = \int_0^T \{(\partial_t y, v) + \kappa(\nabla y, \nabla v) + (b \cdot \nabla y, v)\}\, dt\,.$$

To define the discretization in time, let us partition the time interval $\bar{I} = [0,T]$ as

$$\bar{I} = \{0\} \cup I_1 \cup I_2 \cup \cdots \cup I_M$$

with subintervals $I_m = (t_{m-1}, t_m]$ of size k_m and time points

$$0 = t_0 < t_1 < \cdots < t_{M-1} < t_M = T.$$

We define the discretization parameter k as a piecewise constant function by setting $k|_{I_m} = k_m$ for $m = 1, \ldots, M$.

By means of the subintervals I_m, we define for $r \in \mathbb{N}_0$ a semidiscrete space X_k^r consisting of discontinuous-in-time piecewise polynomial functions:

$$X_k^r = \{\, v_k \in L^2(I, V_0) : v_k\big|_{I_m} \in \mathcal{P}^r(I_m, V_0) \text{ and } v_k(0) \in H \,\}\,.$$

Here, $\mathcal{P}^r(I_m, V_0)$ denotes the space of polynomials up to order r defined on I_m with values in V_0. For the definition of the discontinuous Galerkin method we introduce the following notation for a function $v_k \in X_k^r$:

$$v_{k,m}^+ := \lim_{t \to 0^+} v_k(t_m + t), \quad v_{k,m}^- := \lim_{t \to 0^+} v_k(t_m - t) = v_k(t_m), \quad [v_k]_m := v_{k,m}^+ - v_{k,m}^-\,.$$

Using this notation we define a discretized version of the bilinear form a:

$$a_k(y_k, v_k) = \sum_{m=1}^{M} \int_{I_m} \{(\partial_t y_k, v_k) + \kappa(\nabla y_k, \nabla v_k) + (b \cdot \nabla y_k, v_k)\}\, dt + \sum_{m=0}^{M-1} ([y_k]_{m-1}, v_{k,m-1}^+) + (y_{k,0}^-, v_{k,0}^-)\,.$$

For the space discretization, we consider two- or three-dimensional shape-regular meshes; see, e.g., [12]. A mesh consists of quadrilateral or hexahedral cells K, which constitute a nonoverlapping cover of the computational domain Ω. The corresponding mesh is denoted by $\mathcal{T}_h = \{K\}$, where we define the discretization parameter h as a cellwise constant function by setting $h\big|_K = h_K$ with the diameter h_K of the cell K.

On the mesh $\mathcal{T}_h$ we construct a conforming finite element space $V_h \subset V$ in a standard way:

$$V_h^s = \{\, v \in V : v\big|_K \in \mathcal{Q}^s(K) \text{ for } K \in \mathcal{T}_h \,\}.$$

Here, $\mathcal{Q}^s(K)$ consists of shape functions obtained via bi- or trilinear transformations of polynomials in $\widehat{\mathcal{Q}^s}(\widehat{K})$ defined on the reference cell $\widehat{K} = (0,1)^n$, where

$$\widehat{\mathcal{Q}^s}(\widehat{K}) = \operatorname{span}\left\{ \prod_{j=1}^{n} x_j^{k_j} :\ k_j \in \mathbb{N}_0,\ k_j \le s \right\}.$$

Remark 5.1. The definition of V_h^s can be extended to the case of triangular meshes in the obvious way.

The discrete space with homogeneous Dirichlet boundary conditions is denoted by $V_{h,0}^s = V_h^s \cap H_0^1(\Omega)$. Moreover, we introduce the space of traces of function in V_h^s:

$$W_h^s = \{\, w_h \in H^{1/2}(\partial\Omega) : w_h = \gamma(v_h), v_h \in V_h^s \,\},$$

where $\gamma\colon H^1(\Omega) \to H^{1/2}(\partial\Omega)$ is the trace operator.

With these preliminaries, we define the discrete spaces for the control and state variables:

$$X_{k,h}^{r,s} = \{\, v_{kh} \in L^2(I, V_{h,0}^s) : v_{kh}\big|_{I_m} \in \mathcal{P}^r(I_m, V_{h,0}^s) \text{ and } v_{kh}(0) \in V_h^s \,\} \subset X_k^r,$$
$$U_{k,h}^{r,s} = \{\, u_{kh} \in L^2(I, W_h^s) : u_{kh}\big|_{I_m} \in \mathcal{P}^r(I_m, W_h^s) \,\}.$$

Remark 5.2. In the above definition of the discrete spaces $X_{k,h}^{r,s}$ and $U_{k,h}^{r,s}$, we assumed that the spatial discretization is fixed for all time intervals. However, in many situations the use of different meshes $\mathcal{T}_h^m$ for each of the subintervals I_m is required for efficient adaptive discretizations. The consideration of such dynamically changing meshes can be included in the above definitions in a natural way [41].

For a function $u_{kh} \in U_{k,h}^{r,s}$ we define an extension $\widehat{u}_{kh} \in X_{k,h}^{r,s}$ such that

$$\gamma(\widehat{u}_{kh}(t,\cdot)) = u_{kh}(t,\cdot) \ \text{ and } \ \widehat{u}_{kh}(t,x_i) = 0 \ \text{ on all interior nodes } x_i \text{ of } \mathcal{T}_h. \tag{5.1}$$

The optimization problem on the discrete level is then formulated as follows:

$$\min J(y_{kh}, u_{kh}), \quad u_{kh} \in U_{k,h}^{r,s} \cap U_{ad} \tag{5.2}$$

subject to

$$y_{kh} \in \widehat{u}_{kh} + X_{k,h}^{r,s}, \quad a_k(y_{kh}, v_{kh}) = \int_0^T (f, v_{kh})\, dt + (y_0, v_{kh,0}^-) \quad \text{for all } v_{kh} \in X_{k,h}^{r,s}. \tag{5.3}$$

The discrete state equation (5.3) defines a discrete solution operator S_{kh} which maps a given discrete control u_{kh} to the (unique) solution of (5.3). As on the continuous level we introduce a discrete reduced cost functional

$$j_{kh}(u_{kh}) = J(S_{kh}(u_{kh}), u_{kh}). \tag{5.4}$$

The discrete optimization problem (5.2)–(5.3) is solved by the PDAS strategy described in the previous section. In each step an equality-constrained optimization

problem is solved by the Newton method for the discrete reduced cost functional j_{kh}; see Remark 4.2. For the realization of the Newton method, we need representations of the first and second directional derivatives of j_{kh}.

PROPOSITION 5.1. *Let the discrete reduced cost functional j_{kh} be defined as in* (5.4). *Then the following hold:*

(a) *The first directional derivative in direction $\delta u_{kh} \in U_{k,h}^{r,s}$ can be expressed as*

$$j'_{kh}(u_{kh})(\delta u_{kh}) = J'_y(y_{kh}, u_{kh})(\widehat{\delta u}_{kh}) - a_k(\widehat{\delta u}_{kh}, p_{kh}) + J'_u(y_{kh}, u_{kh})(\delta u_{kh}), \tag{5.5}$$

where $y_{kh} = S_{kh}(u_{kh})$, the extension $\widehat{\delta u}_{kh}$ is defined in (5.1), *and $p_{kh} \in X_{k,h}^{r,s}$ is the solution of the discrete adjoint equation*

$$a_k(v_{kh}, p_{kh}) = J'_y(y_{kh}, u_{kh})(v_{kh}) \quad \text{for all } v_{kh} \in X_{k,h}^{r,s}. \tag{5.6}$$

(b) *The second derivatives of j_{kh} in directions $\delta u_{kh}, \tau u_{kh} \in U_{k,h}^{r,s}$ can be expressed as*

$$\begin{aligned} j''_{kh}(u_{kh})(\delta u_{kh}, \tau u_{kh}) = J''_{yy}(y_{kh}, u_{kh})(\delta y_{kh}, \widehat{\tau u}_{kh}) - a_k(\widehat{\tau u}_{kh}, \delta p_{kh}) \\ + J''_{uu}(y_{kh}, u_{kh})(\delta u_{kh}, \tau u_{kh}), \end{aligned} \tag{5.7}$$

where δy_{kh} is the solution of the discrete tangent equation

$$\delta y_{kh} \in \widehat{\delta u}_{kh} + X_{k,h}^{r,s} \ : \ a_k(\delta y_{kh}, v_{kh}) = 0 \quad \text{for all } v_{kh} \in X_{k,h}^{r,s}\,, \tag{5.8}$$

$\delta p_{kh} \in X_{k,h}^{r,s}$ is given by

$$a_k(v_{kh}, \delta p_{kh}) = J''_{yy}(y_{kh}, u_{kh})(\delta y_{kh}, v_{kh}) \quad \text{for all } v_{kh} \in X_{k,h}^{r,s}\,, \tag{5.9}$$

and $\widehat{\delta u}_{kh}, \widehat{\tau u}_{kh}$ are the extensions of $\delta u_{kh}, \tau u_{kh}$ defined as in (5.1).

Proof. Using the solution $\delta y_{kh} = S'_{kh}(u_{kh})(\delta u_{kh})$ of the discretized tangent equation (5.8), we obtain

$$j'_{kh}(u_{kh})(\delta u_{kh}) = J'_y(y_{kh}, u_{kh})(\delta y_{kh}) + J'_u(y_{kh}, u_{kh})(\delta u_{kh}).$$

We rewrite the first term using (5.8) and (5.6):

$$\begin{aligned} &J'_y(y_{kh}, u_{kh})(\delta y_{kh}) = J'_y(y_{kh}, u_{kh})(\delta y_{kh} - \widehat{\delta u}_{kh}) + J'_y(y_{kh}, u_{kh})(\widehat{\delta u}_{kh}) \\ &= a_k(\delta y_{kh} - \widehat{\delta u}_{kh}, p_{kh}) + J'_y(y_{kh}, u_{kh})(\widehat{\delta u}_{kh}) = -a_k(\widehat{\delta u}_{kh}, p_{kh}) + J'_y(y_{kh}, u_{kh})(\widehat{\delta u}_{kh}). \end{aligned}$$

This gives the desired representation (5.5). The representation of the second derivatives is obtained in a similar way. ∎

Remark 5.3. On the continuous level, similar representations of the derivatives hold. They can be equivalently expressed using the normal derivatives of the adjoint state on the boundary; see (3.1). A direct discretization of the representation involving normal fluxes is in general not equivalent to (5.5) and would not lead to the exact representation of the derivatives of j_{kh} due to the lack of appropriate formulas for integration by parts of the discretized solutions.

Remark 5.4. In the convection dominated case, i.e., if $|b| \gg \kappa$, the finite element discretization may lead to strongly oscillatory solutions. Several stabilization methods are known to improve the approximation properties of the pure Galerkin discretization and to reduce the oscillatory behavior; see, e.g., [10, 22, 28, 39, 40]. For the stabilized finite elements in the context of stationary optimal control problems, we refer the reader to [13, 4].

6. Numerical examples. In this section we discuss numerical examples illustrating our results and give some details on the numerical realization.

Due to the fact that the trial and the test spaces in the formulation of the discrete state equation (5.3) are discontinuous in time, this formulation results in a time stepping scheme. In our numerical realization we use bilinear finite elements for the space discretization and piecewise constant functions in time resulting in spaces $X_{k,h}^{0,1}$ and $U_{k,h}^{0,1}$. In the following we describe the state equation (5.3), the adjoint equation (5.6), equations (5.8) and (5.9), and the evaluation of the derivatives of the discrete reduced cost functional for this choice of discretization. We define

$$U_m = u_{kh}\big|_{I_m},\ Y_m = y_{kh}\big|_{I_m},\ P_m = p_{kh}\big|_{I_m}, \quad i = 1, \dots, M,$$

$$Y_0 = y_{kh,0}^-,\ P_0 = p_{kh,0}^- .$$

The discrete state equation reads as follows for $Y_0 \in V_h$ and $Y_m \in U_m + V_{h,0}$:

$$(Y_0, \phi) = (y_0, \phi) \quad \text{for all } \phi \in V_h ,$$

$$\begin{aligned}(Y_m, \phi) + k_m\,(\nabla Y_m, \nabla\phi) + k_m\left(\int_{I_m} b(s)\,ds \cdot \nabla Y_m, \phi\right) &= (Y_{m-1}, \phi) \\ + k_m\left(\int_{I_m} f(s)\,ds, \phi\right) \quad &\text{for all } \phi \in V_{h,0},\ m = 1, \dots, M .\end{aligned}$$

Remark 6.1. If the time integrals are approximated by the box rule, then the resulting scheme is equivalent to the implicit Euler method. However, a better approximation of these time integrals leads to a scheme which allows for better error estimates with respect to the required smoothness of the solution and to long time integration ($T \gg 1$); see, e.g., [18]. For the numerical examples which follow, the trapezoidal rule is used, which guarantees this improved convergence behavior.

In order to cover both problem (P1) with a time-distributed cost functional and the problem (P2) with a terminal time functional, we write the cost functional in the form

$$J(y, u) = \int_0^T I(y(s))\,ds + K(y(T)) + \frac{\beta}{2}|u|^2_{L^2(\Sigma)} .$$

The discrete adjoint equation reads as follows for $P_0 \in V_h$ and $P_m \in V_{h,0}$:

$$\begin{aligned}(\phi, P_M) + k_M\,(\nabla\phi, \nabla P_M) + k_M\left(\int_{I_M} b(s)\,ds \cdot \nabla\phi, P_M\right) &= K'(Y_M)(\phi) \\ + k_M\,I'(Y_M)(\phi) \quad &\text{for all } \phi \in V_{h,0},\end{aligned}$$

$$\begin{aligned}(\phi, P_m) + k_m\,(\nabla\phi, \nabla P_m) + k_m\left(\int_{I_m} b(s)\,ds \cdot \nabla\phi, P_m\right) &= (\phi, P_{m+1}) \\ + k_m\,I'(Y_m)(\phi) \quad &\text{for all } \phi \in V_{h,0},\ m = M-1, \dots, 1 ,\end{aligned}$$

$$(\phi, P_0) = (\phi, P_1) \quad \text{for all } \phi \in V_h .$$

Remark 6.2. There are two possible ways to obtain the above equations for P_m, $m = 0, \dots, M$:

- discretization of the continuous adjoint equation with dG(0) in time and with H^1-conforming finite elements in space (optimize-then-discretize approach);
- application of the Lagrange formalism on the discrete level for the optimization problem with the state equation discretized by dG(0) in time and H^1-conforming finite elements in space (discretize-then-optimize approach).

The resulting schemes for P_m coincide independent of the temporal grid. This fact relies on the space-time Galerkin discretization.

For a standard formulation of the implicit Euler scheme, i.e.,

$$\frac{1}{k_m}\,(Y_m - Y_{m-1}, \phi) + (\nabla Y_m, \nabla\phi) + (b(t_m)\nabla Y_m, \phi) = (f(t_m), \phi) \quad \text{for all } \phi \in V_{h,0},$$

the optimize-then-discretize approach leads to the discrete adjoint

$$\frac{1}{k_{m+1}}\,(\phi, P_m - P_{m+1}) + (\nabla\phi, \nabla P_m) + (b(t_m)\nabla\phi, P_m) = (I'(Y_m), \phi) \quad \text{for all } \phi \in V_{h,0},$$

whereas the discretize-then-optimize approach produces

$$\frac{1}{k_m}\,(\phi, P_m) - \frac{1}{k_{m+1}}\,(\phi, P_{m+1}) + (\nabla\phi, \nabla P_m) + (b(t_m)\nabla\phi, P_m) = (I'(Y_m), \phi) \quad \text{for all } \phi \in V_{h,0}.$$

These schemes are different for nonconstant time steps k_m.

For the optimization algorithm we need the evaluation of the derivatives of j_{kh} for basis functions in $U_{k,h}^{0,1}$. We consider the following basis of $U_{k,h}^{0,1}$:

$$w_{i,m}(t,x) = \begin{cases} \phi_i(x), & t \in I_m, \\ 0 & \text{otherwise}, \end{cases} \tag{6.1}$$

where $\phi_i = \gamma(\widehat{\phi}_i)$ and $\widehat{\phi}_i \in V_h$ is a finite element nodal basis function for a boundary node i. We obtain the following corollary from Proposition 5.1.

COROLLARY 6.1. *The following representation holds:*

$$\begin{aligned} j'_{kh}(u_{kh})(w_{i,M}) = {} & \beta(U_M, \phi_i)_{\partial\Omega} + K'(Y_M)(\widehat{\phi}_i) + k_M\, I'(Y_M)(\widehat{\phi}_i) \\ & - (\widehat{\phi}_i, P_M) - k_M\,(\nabla\widehat{\phi}_i, \nabla P_M) - k_M \left(\int_{I_M} b(s)\,ds \cdot \nabla\widehat{\phi}_i, P_M\right), \end{aligned}$$

$$\begin{aligned} j'_{kh}(u_{kh})(w_{i,m}) = {} & \beta(U_m, \phi_i)_{\partial\Omega} + k_m\, I'(Y_m)(\widehat{\phi}_i) + (\widehat{\phi}_i, P_{m+1}) \\ & - (\widehat{\phi}_i, P_m) - k_m\,(\nabla\widehat{\phi}_i, \nabla P_m) - k_m \left(\int_{I_m} b(s)\,ds \cdot \nabla\widehat{\phi}_i, P_m\right), \\ & m = M-1, \ldots, 1\,. \end{aligned}$$

Remark 6.3. Due to the fact that $\widehat{\phi}_i$ has local support, the spatial integration in the representations above is done only over cells adjacent to the boundary.

Next, we describe (5.8) and (5.9) and the evaluation of the second derivatives. We define

$$\delta U_m = \delta u_{kh}\big|_{I_m},\ \delta Y_m = \delta y_{kh}\big|_{I_m},\ \delta P_m = \delta p_{kh}\big|_{I_m}, \quad i = 1, \ldots, M,$$

$$\delta Y_0 = \delta y^-_{kh,0},\ \delta P_0 = \delta p^-_{kh,0}\,.$$

The discrete tangent equation reads as follows for $\delta Y_0 \in V_h$ and $\delta Y_m \in \delta U_m + V_{h,0}$:

$$\delta Y_0 = 0\,,$$

$$(\delta Y_m, \phi) + k_m\,(\nabla \delta Y_m, \nabla \phi) + k_m\,\left(\int_{I_m} b(s)\,ds \cdot \nabla \delta Y_m, \phi\right) = (\delta Y_{m-1}, \phi)$$
$$\text{for all } \phi \in V_h\,,\ m = 1, \dots, M\,.$$

The discrete equation (5.9) reads as follows for $\delta P_0 \in V_h$ and $\delta P_m \in V_{h,0}$:

$$(\phi, \delta P_M) + k_M\,(\nabla \phi, \nabla \delta P_M) + k_M\,\left(\int_{I_M} b(s)\,ds \cdot \nabla \phi, \delta P_M\right) = K''(Y_M)(\delta Y_M, \phi)$$
$$+\ k_M\,I''(Y_M)(\delta Y_M, \phi) \quad \text{for all } \phi \in V_h\,,$$

$$(\phi, \delta P_m) + k_m\,(\nabla \phi, \nabla \delta P_m) + k_m\,\left(\int_{I_m} b(s)\,ds \cdot \nabla \phi, \delta P_m\right) = (\phi, \delta P_{m+1})$$
$$+\ k_m\,I''(Y_m)(\delta Y_m, \phi) \quad \text{for all } \phi \in V_h\,,\ m = M-1, \dots, 1\,,$$

$$(\phi, \delta P_0) = (\phi, \delta P_1) \quad \text{for all } \phi \in V_h\,.$$

Using the basis (6.1) we obtain the following representation of $j''_{kh}(u_{kh})(\delta u_{kh}, w_{i,m})$ as a corollary to Proposition 5.1.

COROLLARY 6.2. *The following representation holds:*

$$j''_{kh}(u_{kh})(\delta u_{kh}, w_{i,M}) = \beta(\delta U_M, \phi_i)_{\partial\Omega} + K''(Y_M)(\delta Y_M, \widehat{\phi}_i) + k_M\,I''(Y_M)(\delta Y_M, \widehat{\phi}_i)$$
$$-\ (\widehat{\phi}_i, \delta P_M) - k_M\,(\nabla \widehat{\phi}_i, \nabla \delta P_M) - k_M\,\left(\int_{I_M} b(s)\,ds \cdot \nabla \widehat{\phi}_i, \delta P_M\right)$$

$$j''_{kh}(u_{kh})(\delta u_{kh}, w_{i,m}) = \beta(\delta U_m, \phi_i)_{\partial\Omega} + k_m\,I''(Y_m)(\delta Y_M, \widehat{\phi}_i) + (\widehat{\phi}_i, \delta P_{m+1})$$
$$-\ (\widehat{\phi}_i, \delta P_m) - k_m\,(\nabla \widehat{\phi}_i, \nabla \delta P_m) - k_m\,\left(\int_{I_m} b(s)\,ds \cdot \nabla \widehat{\phi}_i, \delta P_m\right),$$
$$m = M-1, \dots, 1\,.$$

We close the paper with three numerical examples. The first two examples correspond to problems (P1) and (P2), respectively, and examine the behavior of the PDAS method if the dimension of the discrete problem increases due to the refinement of spatial and time meshes. The third example is devoted to the superlinear convergence of the PDAS method.

Our special interest in considering the behavior of the algorithm as the mesh is refined results from previous experience with constrained optimal control problems with distributed controls. Pointwise control, respectively, state constraints, result in a very different behavior of the algorithm in the sense that it is mesh-independent for the former but strongly mesh-dependent for the latter; see [8] and [26]. Analytically this is reflected in the fact that for the former the Lagrange multipliers are L^2-functions, whereas they are only measures in the case of state constraints. Regularization or nested iteration can be used in the latter case to nearly restore mesh-independence.

For the case of Dirichlet boundary control with pointwise constraints on the controls the practical performance of the algorithm and specifically its behavior with respect to mesh refinement cannot easily be predicted from previous experience. On the one hand, as in the case of distributed control, the associated Lagrange multipliers are L^2-regular and we can prove superlinear convergence. However, at least formally, inequality constraints on the control along the boundary are equivalent to inequality constraints on the state on the boundary, and, second, the extra regularity on the adjoint states and the optimal controls obtained in section 3 is rather less than that in the case of distributed controls.

In section 4, we have shown the superlinear convergence of the PDAS method on the continuous level for Dirichlet optimal control problems. On the discrete level, we typically have finite step convergence (cf. the stopping criterion discussed in Remark 4.1), which is, of course, better than superlinear convergence. In our last numerical example, presented in section 6.3, we observe the behavior of the PDAS method corresponding to superlinear convergence also before the method stops finding the optimal discrete solution.

6.1. Example 1: Time-distributed functional. We consider the following Dirichlet optimal control problem on $\Omega \times (0,T)$ with $\Omega = (0,1)^2 \subset \mathbb{R}^2$ and $T = 1$:

$$\begin{aligned} \min \quad & J(u,y) = \frac{1}{2}\|y - y_d\|^2_{L^2(Q)} + \frac{\beta}{2}\|u\|^2_{L^2(\Sigma)}\,, \\ \text{subject to} \quad & y_t - \kappa\Delta y + b\cdot\nabla u = f \quad \text{in } \Omega\times(0,T), \\ & y = u \qquad\qquad\qquad\quad \text{on } \partial\Omega\times(0,T), \\ & y(0) = y_0 \qquad\qquad\qquad\quad \text{in } \Omega \end{aligned}$$

and control constraints

$$u \geq \phi\,.$$

The data are given as follows:

$$f = 0, \quad \kappa = 1, \quad b(t,x) = 15\,(\sin(2\pi t), \cos(2\pi t)), \quad y_0 = 0, \quad \beta = 10^{-4}\,,$$

$$y_d(t,x) = x_1x_2(\cos(\pi t) - x_1)(\sin(\pi t) - x_2), \quad \phi = -0.25\,.$$

This optimal control problem is discretized by space-time finite elements as described above. The resulting finite-dimensional problem is solved by the PDAS method. In Table 6.1 the number of iterations of the method is shown for a sequence of uniformly refined discretizations. Here, M denotes the number of time-steps and N is the number of nodes in the space discretization. In all cases the algorithm terminated with two consecutive active sets coinciding, so that the exact solution of the discretized problem is found.

We present the results for two choices of initial guesses for the control variable: the same choice for all discretization levels ($u_0 = 1$), and an interpolated solution from the previous discretization level (nested iteration). The goal consists in obtaining practical experience as to which degree the weak additional regularity established in Theorem 3.2 and Corollary 3.7 is sufficient for near mesh-independent behavior. The results indicate that the additional regularity is sufficient for nearly mesh-independent behavior and that nested iterations provide only a relatively moderate improvement. The algorithm was also tested with other initial guesses and led to very similar results.

TABLE 6.1
PDAS method on the sequence of uniformly refined discretizations.

N	M	$\dim X_h = M \cdot N$	$\dim U_h$	PDAS iterations	PDAS nested iterations
25	2	50	32	2	2
81	4	324	128	3	3
289	8	2312	512	4	3
1089	16	17424	2048	4	3
4225	32	135200	8192	5	4
16641	64	1065024	32768	6	4

6.2. Example 2: Terminal functional. In this example we consider a Dirichlet optimal control problem with a terminal cost functional:

$$\begin{aligned} \min \quad & J(u,y) = \frac{1}{2}\|y(T) - y_d^T\|^2_{L^2(\Omega)} + \frac{\beta}{2}\|u\|^2_{L^2(\Sigma)}\,, \\ \text{subject to} \quad & y_t - \kappa\Delta y + b\cdot\nabla u = f \quad \text{in } \Omega\times(0,T), \\ & y = u \quad \text{on } \partial\Omega\times(0,T), \\ & y(0) = y_0 \quad \text{in } \Omega\,, \end{aligned}$$

and control constraints

$$\phi \le u \le \psi, \quad u = 0 \text{ on } \partial\Omega\times(T_1,T).$$

The data are given as follows:

$$f = 0, \quad \kappa = 1, \quad b(t,x) = 15\,(\sin(2\pi t), \cos(2\pi t)), \quad y_0 = 0, \quad \beta = 10^{-4}, \quad T_1 = 0.75\,,$$

$$y_d^T(x) = 3\,\left(x_1x_2 + \sin(12\pi x_1^2(1-x_1)^2)\,\sin(12\pi x_2^2(1-x_2)^2)\right), \quad \phi = -0.1, \quad \psi = 2.5\,.$$

In Table 6.2 we present the corresponding results.

TABLE 6.2
PDAS method on the sequence of uniformly refined discretizations.

N	M	$\dim X_h = M \cdot N$	$\dim U_h$	PDAS iterations	PDAS nested iterations
25	2	50	32	3	3
81	4	324	128	3	3
289	8	2312	512	4	4
1089	16	17424	2048	5	4
4225	32	135200	8192	5	5
16641	64	1065024	32768	6	5

6.3. Example 3. In this example, we consider the following Dirichlet optimal control problem on $\Omega\times(0,T)$ with $\Omega = (0,1)^2 \subset \mathbb{R}^2$ and $T = 1$:

$$\begin{aligned} \min \quad & J(u,y) = \frac{1}{2}\|y - y_d\|^2_{L^2(Q)} + \frac{\beta}{2}\|u\|^2_{L^2(\Sigma)}\,, \\ \text{subject to} \quad & y_t - \kappa\Delta y + b\cdot\nabla u = f \quad \text{in } \Omega\times(0,T), \\ & y = u \quad \text{on } \partial\Omega\times(0,T), \\ & y(0) = y_0 \quad \text{in } \Omega \end{aligned}$$

and control constraints

$$\phi \le u \le \psi\,.$$

The data are given as follows:

$$f = \begin{cases} 2, & x_1 \le 0.25, \\ -35 & \text{else}, \end{cases} \quad \kappa = 1, \;\; b(t,x) = 10\,(\sin(2\pi t), \cos(2\pi t)), \;\; y_0 = 0, \;\; \beta = 10^{-5}\,,$$

$$y_d(t,x) = \begin{cases} 2 - 2x_1, & x_1 \le 0.5, \\ 2 - 2x_2 & \text{else}, \end{cases} \qquad \phi = -1, \quad \psi = 2\,.$$

For a fixed discretization with $M = 64$ time-steps and $N = 1089$ nodes in the spacial mesh, we consider the iteration error

$$e_i = \|u_{kh}^{(i)} - u_{kh}\|_{L^2(\Sigma)},$$

where $u_{kh}^{(i)}$ is the ith iterate, and u_{kh} is the optimal discrete solution. The results presented in Table 6.3 demonstrate superlinear convergence of the algorithm.

TABLE 6.3
Iteration error for the PDAS method.

i	0	1	2	3	4	5	6	7
e_i	2.3e-1	1.7e-1	4.1e-2	1.9e-2	6.5e-3	3.8e-4	4.5e-6	0
e_{i+1}/e_i	7.4e-1	2.4e-1	4.6e-1	3.4e-1	5.8e-2	1.2e-2	0	–

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, Amsterdam, 2005.

[2] N. ARADA AND J.-P. RAYMOND, *Dirichlet boundary control of semilinear parabolic equations, Part* 1: *Problems with no state constraints*, Appl. Math. Optim., 45 (2002), pp. 125–143.

[3] R. BECKER, D. MEIDNER, AND B. VEXLER, *Efficient numerical solution of parabolic optimization problems by finite element methods*, Optim. Methods Softw., to appear.

[4] R. BECKER AND B. VEXLER, *Optimal control of the convection-diffusion equation using stabilized finite element methods*, Numer. Math., 106 (2007), pp. 349–367.

[5] R. BECKER, *Mesh adaptation for Dirichlet flow control via Nitsche's method*, Comm. Numer. Methods Engrg., 18 (2002), pp. 669–680.

[6] F. B. BELGACEM, H. E. FEKIH, AND H. METOUI, *Singular perturbation for the Dirichlet boundary control of elliptic problems*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 883–850.

[7] M. BERGGREN, *Approximations of very weak solutions to boundary-value problems*, SIAM J. Numer. Anal., 42 (2004), pp. 860–877.

[8] M. BERGOUNIOUX, K. ITO, AND K. KUNISCH, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[9] M. BERGOUNIOUX, M. HADDOU, M. HINTERMÜLLER, AND K. KUNISCH, *A comparison of a Moreau–Yosida-based active set strategy and interior point methods for constrained optimal control problems*, SIAM J. Optim., 11 (2000), pp. 495–521.

[10] E. BURMAN AND P. HANSBO, *Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 1437–1453.

[11] E. CASAS AND J.-P. RAYMOND, *Error estimates for the numerical approximation of Dirichlet boundary control for semilinear elliptic equations*, SIAM J. Control Optim., 45 (2006), pp. 1586–1611.

[12] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.
[13] S. S. Collis and M. Heinkenschloss, *Analysis of the Streamline Upwind/Petrov Galerkin Method Applied to the Solution of Optimal Control Problems*, CAAM TR02-01, Rice University, Houston, TX, 2002.
[14] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology: Evolution Problems* I, Vol. 5, Springer-Verlag, Berlin, 1992.
[15] J. De Los Reyes and K. Kunisch, *A semi-smooth Newton method for control constrained boundary optimal control of the Navier-Stokes equations*, Nonlinear Anal., 62 (2005), pp. 1289–1316.
[16] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.
[17] K. Eriksson, C. Johnson, and V. Thomée, *Time discretization of parabolic problems by the discontinuous Galerkin method*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 611–643.
[18] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* I: *A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.
[19] D. French and J. King, *Analysis of a robust finite element approximation for a parabolic equation with rough boundary data*, Math. Comp., 60 (1993), pp. 79–104.
[20] A. Fursikov, *Optimal Control of Distributed Systems: Theory and Applications*, Transl. Math. Monogr. 187, AMS, Providence, RI, 1999.
[21] P. Grisvard, *Commutativité de deux foncteurs d'interpolation et applications*, J. Math. Pures Appl., 45 (1966), pp. 143–206.
[22] J.-L. Guermond, *Stabilization of Galerkin approximations of transport equations by subgrid modeling*, Modél. Math. Anal. Numér., 36 (1999), pp. 1293–1316.
[23] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with Dirichlet controls*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 711–748.
[24] M. D. Gunzburger and S. Manservisi, *The velocity tracking problem for Navier–Stokes flows with boundary control*, SIAM J. Control Optim., 39 (2000), pp. 594–634.
[25] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semi-smooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.
[26] M. Hintermüller and K. Kunisch, *Feasible and noninterior path-following in constrained minimization with low multiplier regularity*, SIAM J. Control Optim., 45 (2006), pp. 1198–1221.
[27] K. Ito and K. Kunisch, *The primal-dual active set method for nonlinear optimal control problems with bilateral constraints*, SIAM J. Control Optim., 43 (2004), pp. 357–376.
[28] C. Johnson, *Numerical Solution of Partial Differential Equations by Finite Element Method*, Cambridge University Press, Cambridge, UK, 1987.
[29] F. Kappel and K. Kunisch, *Invariance results for delay and Volterra equations in fractional order Sobolev spaces*, Trans. Amer. Math. Soc., 304 (1987), pp. 1–51.
[30] A. Kunoth, *Adaptive wavelet schemes for an elliptic control problem with Dirichlet boundary control*, Numer. Algorithms, 39 (2005), pp. 199–200.
[31] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Uralćeva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.
[32] I. Lasiecka, *Galerkin approximation of abstract parabolic boundary value problems with rough boundary data-l^p theory*, Math. Comp., 175 (1986), pp. 55–75.
[33] H.-C. Lee, *Analysis and computational methods of Dirichlet boundary optimal control problems for $2d$ Boussinesq equations*, Adv. Comput. Math., 19 (2003), pp. 255–275.
[34] J. L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Vol. I, Springer-Verlag, Berlin, 1972.
[35] J. L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Vol. II, Springer-Verlag, Berlin, 1972.
[36] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Grundlehren Math. Wiss. 170, Springer-Verlag, Berlin, 1971.
[37] H. Maurer and J. Zowe, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Program., 16 (1979), pp. 98–110.
[38] D. Meidner and B. Vexler, *Adaptive space-time finite element methods for parabolic optimization problems*, SIAM J. Control Optim., 46 (2007), pp. 116–142.
[39] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.
[40] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Dif-*

ferential Equations, Springer Ser. Comput. Math. 24, Springer-Verlag, Berlin, 1996.

[41] M. Schmich and B. Vexler, *Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations*, SIAM J. Sci. Comput., to appear.

[42] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, J. A. Barth Verlag, Heidelberg, 1995.

[43] B. Vexler, *Finite element approximation of elliptic Dirichlet optimal control problems*, Numer. Funct. Anal. Optim., to appear.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1754–1778

OPTIMAL CONTROL OF SEMILINEAR PARABOLIC EQUATIONS WITH $\mathcal{K}$-APPROXIMATE PERIODIC SOLUTIONS*

LING LEI† AND GENGSHENG WANG‡

Abstract. In this paper, we study some optimal control problems governed by certain semilinear parabolic equations with $\mathcal{K}$-approximate periodic solutions. We first prove the existence and uniqueness theorems for $\mathcal{K}$-approximate periodic solutions of the equations. We then use these results to establish the qualified Pontryagin maximum principle. The existence for such optimal controls is also investigated in the paper.

Key words. $\mathcal{K}$-approximate periodic solution, optimal control, non-well-posed parabolic equation, Pontryagin's maximum principle

AMS subject classifications. 49K20, 35J65

DOI. 10.1137/060665063

1. Introduction.

Let Ω be a bounded domain in $\mathbf{R}^n$, $n \geq 3$, with a C^2-smooth boundary $\partial\Omega$. Write Q and Σ for the sets $\Omega \times (0,T)$ and $\partial\Omega \times (0,T)$, respectively. Denote by χ_ω the characteristic function of ω, where ω is a subdomain of Ω. Consider the following controlled parabolic equation:

$$\begin{cases} y_t(x,t) - \Delta y(x,t) + f(x,t,y(x,t)) = \chi_\omega(x)u(x,t) & \text{in } Q, \\ y(x,t) = 0 & \text{on } \Sigma, \end{cases} \tag{1.1}$$

where $f : \Omega \times (0,T) \times \mathbf{R} \to \mathbf{R}$ is a given function satisfying certain conditions to be specified later and the control function $u(x,t)$ is taken from the space $L^2(Q)$. For simplicity, we shall call the aforementioned function $u(x,t)$ a control. This paper is concerned with the optimal control problem governed by (1.1) with $\mathcal{K}$-approximate periodic solutions.

To start, we recall the concept of $\mathcal{K}$-approximate periodic solutions for a parabolic equation, first considered by the first author in [6], [7]. Let $H = L^2(\Omega)$ and $V = H_0^1(\Omega)$. Write $|\cdot|_2$ and $\|\cdot\|$ for the norms of H and V, respectively. Denote by $\langle\cdot,\cdot\rangle$ the inner product in H. Let A be the linear operator from H to H such that $Ay = -\triangle y$, with $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$. Let $\{X_i(x)\}_{i=1}^\infty$ be a complete set of eigenfunctions of A, which serves as an orthonormal basis of H, and let $\{\lambda_i\}_{i=1}^\infty$, with $0 < \lambda_1 < \lambda_2 \leq \cdots < +\infty$, be the corresponding eigenvalues. For each nonnegative integer $\mathcal{K}$, we denote by $H^{I,\mathcal{K}}$ and $H^{II,\mathcal{K}}$ the subspaces spanned by $\{X_i(x)\}_{i=1}^{\mathcal{K}}$ and $\{X_i(x)\}_{i=\mathcal{K}+1}^\infty$, respectively. Then $H = H^{I,\mathcal{K}} \oplus H^{II,\mathcal{K}}$, and each element $h \in H$ can be written as $h = h^{I,\mathcal{K}} + h^{II,\mathcal{K}}$, where $h^{I,\mathcal{K}} \in H^{I,\mathcal{K}}$ and $h^{II,\mathcal{K}} \in H^{II,\mathcal{K}}$. Moreover, each function $\varphi(x,t) \in L^2(0,T;H)$ can be expressed as $\varphi(x,t) = \varphi^{I,\mathcal{K}}(x,t) + \varphi^{II,\mathcal{K}}(x,t) \equiv \sum_{i=1}^{\mathcal{K}} \varphi_i(t)X_i(x) + \sum_{i=\mathcal{K}+1}^\infty \varphi_i(t)X_i(x)$, where $\varphi_i(t) = \langle\varphi(\cdot,t), X_i(\cdot)\rangle \in L^2(0,T)$. In what follows, we shall omit variables x and t for functions in $L^2(0,T;H)$ and omit x for functions of

*Received by the editors July 13, 2006; accepted for publication (in revised form) June 30, 2007; published electronically November 9, 2007. This work was supported by National Natural Science Foundation of China under grants 10471053 and 60574017 and by the key project of Chinese Ministry of Education, No. 106118.

http://www.siam.org/journals/sicon/46-5/66506.html

†Department of Mathematics, Zhejiang University, Hangzhou, Zhejiang, 310027, People's Republic of China (leiling0810@yahoo.com.cn)

‡School of Mathematics, Wuhan University, Wuhan, Hubei, 430072, People's Republic of China (wanggs@public.wh.hb.cn).

x, if there is no risk of causing confusion. Write $Y = \{y \in L^2(0,T;D(A));\ y_t \in L^2(0,T;H)\}$ with norm $\|y\|_Y = \{\|y\|^2_{L^2(0,T;D(A))} + \|y_t\|^2_{L^2(0,T;H)}\}^{\frac{1}{2}}$. Then Y is a Hilbert space. We say that y is a $\mathcal{K}$-approximate periodic solution of (1.1) if $y \in Y$ satisfies (1.1) and $y^{II,\mathcal{K}}(0) = y^{II,\mathcal{K}}(T)$. When $\mathcal{K} = 0$, we will always regard $\sum_{i=1}^{0} = 0$. Hence, a 0-approximate periodic solution of (1.1) is a usual periodic solution.

A $\mathcal{K}$-approximate solution of (1.1) can be explained mathematically as a special type of solution for the equation like periodic solutions, steady state solutions, and others. It can also be viewed as a perturbation of a periodic solution in the low frequency part. In physics and applied sciences, it can be regarded as a solution with a periodic higher frequency part.

In this paper, we shall study the following optimal control problem:

$$(\mathbf{P}_{\mathcal{K},r}):\quad \mathrm{Inf}J(y,u) \equiv \mathrm{Inf}\frac{1}{2}\int_Q (y^2+u^2)dxdt \text{ over all } (y,u) \in Y \times L^2(Q)$$

satisfying (1.1) and the following state constraint:

$$\begin{cases} y^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}},\ \ y^{I,\mathcal{K}}(T) \in B^{I,\mathcal{K}}(0,r), \\ y^{II,\mathcal{K}}(0) = y^{II,\mathcal{K}}(T), \end{cases} \tag{1.2}$$

where $a^{I,\mathcal{K}} \in H^{I,\mathcal{K}}$ is a given element and $B^{I,\mathcal{K}}(0,r)$ denotes the closed ball in $H^{I,\mathcal{K}}$ centered at 0 and of radius $r > 0$. We shall establish the Pontryagin maximum principle and the existence of optimal controls for the problem $(\mathbf{P}_{\mathcal{K},r})$. We define the following certain properties for the function f, which will be used in our later discussions.

Property (A_f). The function $f : \Omega \times (0,T) \times \mathbf{R} \to \mathbf{R}$ is measurable in $(x,t) \in \Omega \times (0,T)$ and continuously differentiable in the third variable. Moreover, there exist positive constants L and α, with $1 < \alpha \le \frac{n}{n-2}$, such that

$$|f(x,t,\xi)| \le L(|\xi|^\alpha + 1)\ \forall\ \xi \in \mathbf{R} \text{ and for almost all } (x,t) \text{ in } Q, \tag{1.3}$$

$$|f'_\xi(x,t,\xi)| \le L(|\xi|^{\alpha-1} + 1)\ \forall\ \xi \in \mathbf{R} \text{ and for almost all } (x,t) \text{ in } Q. \tag{1.4}$$

Property (H_f). The function $f : \Omega \times (0,T) \times \mathbf{R} \to \mathbf{R}$ is measurable in $(x,t) \in \Omega \times (0,T)$ and continuously differentiable in the third variable. Moreover,

(i) there exists a positive constant μ such that

$$f(x,t,\xi)\xi + \lambda_1\xi^2 \ge \mu\xi^2,\ \forall\ \xi \in \mathbf{R},\ \text{a.e. } (x,t) \in Q,$$

where λ_1 is the first eigenvalue of operator A;

(ii) there exist positive constants $\widetilde{L}$ and β with $\beta \le \frac{2}{n}$, such that

$$|f(x,t,\xi)| \le \widetilde{L}|\xi|(1+|\xi|^\beta),\ \forall\ \xi \in \mathbf{R},\ \text{a.e. } (x,t) \in Q.$$

We shall derive the Pontryagin maximum principle and the existence of optimal controls for the problem $(\mathbf{P}_{\mathcal{K},r})$ under property $(\mathbf{A_f})$ and $(\mathbf{H_f})$, respectively. We now are ready to state our main results.

THEOREM 1.1. *Suppose that f satisfies property $(\mathbf{A_f})$. Let $(y^*,u^*) \in Y \times L^2(Q)$ be optimal for problem $(\mathbf{P}_{\mathcal{K},r})$. Then there exist a number μ_0, with $\mu_0 \ne 0$, a function p in the space Y, and an element $a_0^{I,\mathcal{K}}$ in the space $H^{I,\mathcal{K}}$ such that*

$$\begin{cases} \chi_\omega p = \mu_0 u^* & a.e.\ in\ \ Q, \\ p_t + \triangle p - f'_y(x,t,y^*)p = \mu_0 y^* & in\ \ Q, \\ p = 0 & on\ \ \Sigma, \\ p^{I,\mathcal{K}}(T) = -a_0^{I,\mathcal{K}},\ p^{II,\mathcal{K}}(T) = p^{II,\mathcal{K}}(0) & in\ \ \Omega. \end{cases}$$

THEOREM 1.2. *Suppose that f satisfies property $(\mathbf{H_f})$. Then, for each $M > 0$, there exist an integer $\mathcal{K} \equiv \mathcal{K}(M) \geq 0$ and a positive constant $r \equiv r(M)$ such that for each $a^{I,\mathcal{K}} \in H^{I,\mathcal{K}}$, with $(1 + \lambda_{\mathcal{K}})|a^{I,\mathcal{K}}|_2^2 \leq M^2$, where $\lambda_{\mathcal{K}}$ is the $\mathcal{K}$th eigenvalue of the operator A, problem $(\mathbf{P}_{\mathcal{K},r})$ has at least one solution.*

Under the assumption $(\mathbf{A_f})$, the controlled system (1.1) with given initial data $y(x,0) = y_0(x)$ (in $L^2(\Omega)$ or in $H_0^1(\Omega)$) may have no solution or may have many solutions in (0,T) for some control $u \in L^2(Q)$. Thus the corresponding optimal control problem $(\mathbf{P}_{\mathcal{K},r})$ cannot be treated by the traditional methods provided in [1], [2], [4], [8], [9], where one regards the state function y as a well-defined functional of the control u. Hence, our optimal control problem is not well-posed. (See [13]). The non-well-posed optimal control problem was first mentioned in [10], where an example of the optimal control problem governed by a semilinear parabolic equation, admitting no solution for some controls, was studied. In [13], a more general non-well-posed optimal control problem governed by some semilinear parabolic equations involving some kinds of state constraints was further investigated. However, in the aforementioned work, due to the technical difficulties, the control is allowed to be acted only in the whole domain Ω, and the state is viewed as the weak solution of the controlled equation, namely, the solution in the space $Y_1 = \{y \in L^2(0,T;H_0^1(\Omega)); y_t - \triangle y \in L^2(Q)\}$. The novelty of the present work compared with the others dealing with non-well-posed optimal control problems mentioned above is as follows: First, we consider a new type of state constraints—$\mathcal{K}$-approximate periodic state constraints (which include periodic state constraints). Moreover, the controls are applied internally into the equations. Second, we consider the strong solution y of the controlled equation (1.1), namely, $y \in Y$, as the state of the optimal control problem. This, together with the state constraint (1.2) and the growth condition $(\mathbf{A_f})$ on the nonlinear term f in (1.1), motivates us to develop a new method to construct an approximate problem $(\mathbf{P}^{\varepsilon}_{\mathcal{K},r})$ to approach the original problem $(\mathbf{P}_{\mathcal{K},r})$. The approximate problems in the works [10], [13] are to ask the infimum of a penalty functional over the whole product space of the state space with the control space, namely, the space $Y_1 \times L^2(Q)$. This is also a general method to construct an approximate problem when people use a penalization method to derive the Pontryagin maximum principle. However, for the current problem, it is difficult for us to construct a suitable penalty functional such that the corresponding infimum problem over the space $Y \times L^2(Q)$ approximates the original problem well. The difficulty is mainly caused by the state constraint (1.2) and the facts that the state is the strong solution and a weak solution of (1.1) satisfying the $\mathcal{K}$-approximate periodic condition may be not a strong solution due to the growth condition $(\mathbf{A_f})$. In this work, we construct a new type of approximate problem $(\mathbf{P}^{\varepsilon}_{\mathcal{K},r})$ which is to ask the infimum of a penalty functional over a suitable subset of the product space $Y \times L^2(Q)$. This approximate problem approaches the original problem well because of the existence and uniqueness result of $\mathcal{K}_0$-approximate solution for the linearized equation of (1.1) at the optimal state y^* for sufficiently large $\mathcal{K}_0$. (See Theorem 2.1.)

Problem $(\mathbf{P}_{\mathcal{K},r})$ studied here is an optimal control problem without a control constraint if one views χ_ω as a linear bounded operator from the space $L^2(\Omega)$ into itself. But, on the other hand, if we set $V(\omega) = \{v(x,t) \in L^2(Q);\ \text{supp}v(x,t) \subset \overline{\omega} \times [0,T]\}$ and write $v(x,t)$ for the term $\chi_\omega(x)u(x,t)$ on the right-hand side of (1.1), then the problem turns to an optimal control problem with control constraint $v \in V(\omega)$. Both the control constraint and the state constraint may cause difficulties for us to get the Pontryagin maximum principle in a qualified form. In this problem, the state constraint is a special kind of two endpoint state constraint, and the control constraint also has a special form. Thus, the unique continuation property for linear

parabolic equations leads us to the Pontryagin maximum principle in a qualified form. If problem $(\mathbf{P}_{\mathcal{K},r})$ involves another type of control constraints, then how to derive the Pontryagin maximum principle in a qualified form is still open for us due to some technical problems.

We emphasize that, even though the cost functional studied in this paper is a typical quadratic form, our main results (Theorems 1.1 and 1.2) can be extended in an identical way to the case where the cost functional has a more general form than that appearing in [1], [13].

Existence and uniqueness problems for periodic solutions for parabolic equations have been extensively studied in the literature. There have also been many works done, in the past years, on the optimal control problems (including the Pontryagin maximum principle and the existence of the optimal controls) governed by semilinear parabolic equations with state constraints. These constraints include two ending point state constraints in the time variable (see [1], [8], [9], [12], [13], [14], and the references therein), the integral type of state constraints (see [5] and the references therein), and the pointwise state constraints (see [4] and the references therein). One of the most important two endpoint state constraints is the periodic state constraints, i.e., $y(x,0) = y(x,T)$ (see [1], [12], and the references therein). The state constraint considered in this paper has the periodic state constraint as a special case.

The paper is organized as follows. In section 2, we establish the existence and uniqueness of $\mathcal{K}$-approximate periodic solutions in the space Y for certain linear parabolic equations which is important for us to get the Pontryagin maximum of optimal controls for problem $(\mathbf{P}_{\mathcal{K},r})$. We also obtain the existence of $\mathcal{K}$-approximate periodic solutions for certain semilinear parabolic equations which plays a key role in the proof of Theorem 1.2. In section 3, we prove Theorem 1.1 by constructing a suitable approximate problem $(\mathbf{P}^{\varepsilon}_{\mathcal{K},r})$. In the last section, we give the proof of Theorem 1.2.

2. Approximate periodic solution for parabolic equations. In this section, we shall give the existence and uniqueness of $\mathcal{K}$-approximate periodic solutions in space Y for some linear parabolic equations and the existence of $\mathcal{K}$-approximate periodic solutions in Y for some semilinear parabolic equations. In the linear case, our method is basically the same as that used in [6], where the existence and uniqueness of $\mathcal{K}$-approximate periodic solutions in space $V_2(Q) \equiv L^\infty(0,T;H) \cap L^2(0,T;V)$ was established. In the semilinear case, we will apply the fixed point theorem and the existence result obtained in the linear problem.

THEOREM 2.1. *Let M be a positive number. Then there exists a nonnegative integer $\mathcal{K}_0 \equiv \mathcal{K}_0(-\triangle, \Omega, T, M)$ depending only on Ω, T, M, and $(-\Delta)$, where the dependence on the operator $-\Delta$ is only through the eigenvalues of the operator $-\Delta$ such that, for each integer $\mathcal{K}$ with $\mathcal{K} \geq \mathcal{K}_0$, each function e in the space $L^\infty(0,T;L^n(\Omega))$, with $\|e\|_{L^\infty(0,T;L^n(\Omega))} \leq M$, each element $a^{I,\mathcal{K}}$ in the space $H^{I,\mathcal{K}}$, and each function v in the space $L^2(0,T;H)$, the following problem has a unique solution y in the space Y:*

$$\begin{cases} y_t - \triangle y - ey = v & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}},\ y^{II,\mathcal{K}}(0) = y^{II,\mathcal{K}}(T) & \text{in } \Omega. \end{cases} \tag{2.1}$$

Moreover, this solution y satisfies the following estimate:

$$\|y\|_Y^2 \leq C(\lambda_{\mathcal{K}} |a^{I,\mathcal{K}}|_2^2 + \|v\|^2_{L^2(0,T;H)}), \tag{2.2}$$

where $C \equiv C(-\triangle, T, \Omega, M)$ is a positive constant depending only on $-\Delta, T, \Omega$, and M and where $\lambda_{\mathcal{K}}$ is the $\mathcal{K}$th eigenvalue of the operator A.

Proof of Theorem 2.1. By the theory of linear evolution equations (see Corollary 4.3, Chapter 1 of [2]), one can easily get that for each function e in the space $L^\infty(0,T;L^n(\Omega))$, with $\|e\|_{L^\infty(0,T;L^n(\Omega))} \le M$, each element z_0 in the space V, and each function v in the space $L^2(0,T;H)$, the following equation:

$$\begin{cases} z_t - \triangle z - ez = v & \text{in } Q, \\ z = 0 & \text{on } \Sigma, \\ z(0) = z_0 & \text{in } \Omega \end{cases} \tag{2.3}$$

has a unique solution z in the space Y. Moreover, there exists a positive constant C such that

$$\|z\|_Y^2 \le C(\|v\|^2_{L^2(0,T;H)} + \|z_0\|^2). \tag{2.4}$$

Here and in what follows, C stands for a positive constant depending only on $-\triangle, T, \Omega$, and M, which may be different in different contexts. Let e be a function in the space $L^\infty(0,T;L^n(\Omega))$ satisfying $\|e\|_{L^\infty(0,T;L^n(\Omega))} \le M$. Then, for each nonnegative integer $\mathcal{K}$ and for each element $a^{II,\mathcal{K}}$ in the space $H^{II,\mathcal{K}} \cap V$, there exists a unique solution $\widetilde{y}(t) \equiv \widetilde{y}(t; a^{II,\mathcal{K}})$ in the space Y satisfying

$$\begin{cases} \widetilde{y}_t - \triangle \widetilde{y} - e\widetilde{y} = 0 & \text{in } Q, \\ \widetilde{y} = 0 & \text{on } \Sigma, \\ \widetilde{y}^{I,\mathcal{K}}(0) = 0,\ \widetilde{y}^{II,\mathcal{K}}(0) = a^{II,\mathcal{K}} & \text{in } \Omega, \end{cases} \tag{2.5}$$

and

$$\|\widetilde{y}\|_Y^2 \le C\|a^{II,\mathcal{K}}\|^2. \tag{2.6}$$

Since $Y \hookrightarrow C([0,T];V)$ (see Chapter 1 of [11]), we have

$$\|\widetilde{y}\|^2_{C([0,T];V)} \le C\|a^{II,\mathcal{K}}\|^2. \tag{2.7}$$

Multiplying the first equation of (2.5) by $(-\triangle \widetilde{y}^{II,\mathcal{K}}(t))$ and integrating over Ω, we get

$$\frac{1}{2}\frac{d}{dt}\|\widetilde{y}^{II,\mathcal{K}}(t)\|^2 + \lambda_{\mathcal{K}+1}\|\widetilde{y}^{II,\mathcal{K}}(t)\|^2 \le |\langle e\widetilde{y}(t), \triangle \widetilde{y}^{II,\mathcal{K}}(t)\rangle|.$$

By the Hölder inequality and the Sobolev embedding theorem, we obtain

$$\begin{aligned} 2|\langle e\widetilde{y}(t), \triangle \widetilde{y}^{II,\mathcal{K}}(t)\rangle| &\le 2\int_\Omega |e\widetilde{y}(t)\triangle \widetilde{y}^{II,\mathcal{K}}(t)|dx \\ &\le 2\left(\int_\Omega |e(x,t)|^n dx\right)^{\frac{1}{n}} \left(\int_\Omega |\widetilde{y}(x,t)|^{\frac{2n}{n-2}} dx\right)^{\frac{n-2}{2n}} \left(\int_\Omega |\triangle \widetilde{y}^{II,\mathcal{K}}(x,t)|^2 dx\right)^{\frac{1}{2}} \\ &\le 2MC_1\|\widetilde{y}(t)\| \cdot |\triangle \widetilde{y}^{II,\mathcal{K}}(t)|_2 \\ &\le \varepsilon |\triangle \widetilde{y}^{II,\mathcal{K}}(t)|_2^2 + C(\varepsilon, M, C_1)\|\widetilde{y}(t)\|^2 \quad \forall\, \varepsilon > 0, \end{aligned}$$

where C_1 is the constant of the embedding from $H_0^1(\Omega)$ into $L^{\frac{2n}{n-2}}(\Omega)$ and $C(\varepsilon, M, C_1)$ is a positive constant depending only on ε, M, and C_1. Thus, it holds that

$$\frac{d}{dt}(e^{2\lambda_{\mathcal{K}+1}t}\|\widetilde{y}^{II,\mathcal{K}}(t)\|^2) \le (\varepsilon|\triangle \widetilde{y}^{II,\mathcal{K}}(t)|_2^2 + C(\varepsilon, M, C_1)\|\widetilde{y}(t)\|^2)e^{2\lambda_{\mathcal{K}+1}t}.$$

Integrating the above over $(0,T)$ and making use of (2.7), we get

$$e^{2\lambda_{\mathcal{K}+1}T}\|\widetilde{y}^{II,\mathcal{K}}(T)\|^2 \le \|a^{II,\mathcal{K}}\|^2 + \varepsilon\int_0^T e^{2\lambda_{\mathcal{K}+1}t}|\triangle\widetilde{y}^{II,\mathcal{K}}(t)|_2^2 dt + C\cdot C(\varepsilon,M,C_1)\|a^{II,\mathcal{K}}\|^2\int_0^T e^{2\lambda_{\mathcal{K}+1}t}dt,$$

which, together with (2.6), implies

$$\begin{aligned}\|\widetilde{y}^{II,\mathcal{K}}(T)\|^2 &\le e^{-2\lambda_{\mathcal{K}+1}T}\|a^{II,\mathcal{K}}\|^2 + C\cdot C(\varepsilon,M,C_1)\left(\frac{1}{2\lambda_{\mathcal{K}+1}} - \frac{e^{-2\lambda_{\mathcal{K}+1}T}}{2\lambda_{\mathcal{K}+1}}\right)\|a^{II,\mathcal{K}}\|^2\\ &\quad +\varepsilon\int_0^T e^{-2\lambda_{\mathcal{K}+1}(T-t)}|\Delta\widetilde{y}^{II,\mathcal{K}}(t)|_2^2 dt\\ &\le \left\{e^{-2\lambda_{\mathcal{K}+1}T} + C\cdot C(\varepsilon,M,C_1)\left(\frac{1}{2\lambda_{\mathcal{K}+1}} - \frac{e^{-2\lambda_{\mathcal{K}+1}T}}{2\lambda_{\mathcal{K}+1}}\right)\right\}\cdot\|a^{II,\mathcal{K}}\|^2\\ &\quad +\varepsilon\cdot C\|a^{II,\mathcal{K}}\|^2.\end{aligned} \tag{2.8}$$

We now first choose an ε sufficiently small such that $\varepsilon\cdot C < \frac{1}{4}$. Then we fix such an ε and an integer $\mathcal{K}_0 \gg 1$ such that $e^{-2\lambda_{\mathcal{K}_0+1}T} < \frac{1}{4}$ and $C(\varepsilon,M,C_1)\cdot C(\frac{1}{2\lambda_{\mathcal{K}_0+1}} - \frac{e^-2\lambda_{\mathcal{K}_0+1}T}{2\lambda_{\mathcal{K}_0+1}}) < \frac{1}{4}$. (Apparently, the choice of such a $\mathcal{K}_0$ depends only on the operators $-\Delta, M, \Omega$, and T.) Then, for each integer $\mathcal{K}$ with $\mathcal{K}\ge\mathcal{K}_0$, it holds that

$$\|\widetilde{y}^{II,\mathcal{K}}(T)\|^2 \le \frac{3}{4}\|a^{II,\mathcal{K}}\|^2 \text{ for each } a^{II,\mathcal{K}}\in H^{II,\mathcal{K}}\cap V. \tag{2.9}$$

Now, for each integer $\mathcal{K}$ with $\mathcal{K}\ge\mathcal{K}_0$, each element $a^{I,\mathcal{K}}$ in the space $H^{I,\mathcal{K}}$, each function v in the space $L^2(0,T;H)$, and each function e in the space $L^\infty(0,T;L^n(\Omega))$ with the estimate $\|e\|_{L^\infty(0,T;L^n(\Omega))}\le M$, we define a map $J:\ H^{II,\mathcal{K}}\cap V\to H^{II,\mathcal{K}}\cap V$ by setting $J(a^{II,\mathcal{K}}) = y^{II,\mathcal{K}}(T;a^{I,\mathcal{K}}+a^{II,\mathcal{K}})$, where $y(t;a^{I,\mathcal{K}}+a^{II,\mathcal{K}})$ is the unique solution of (2.3) with z_0 being replaced by $(a^{I,\mathcal{K}}+a^{II,\mathcal{K}})$. Then, for any $a_1^{II,\mathcal{K}}$, $a_2^{II,\mathcal{K}}$ in the space $H^{II,\mathcal{K}}\cap V$, we have

$$\begin{aligned}\|J(a_1^{II,\mathcal{K}}) - J(a_2^{II,\mathcal{K}})\|^2 &= \|y^{II,\mathcal{K}}(T;a^{I,\mathcal{K}}+a_1^{II,\mathcal{K}}) - y^{II,\mathcal{K}}(T;a^{I,\mathcal{K}}+a_2^{II,\mathcal{K}})\|^2\\ &= \|\widetilde{y}(T;a_1^{II,\mathcal{K}} - a_2^{II,\mathcal{K}})\|^2,\end{aligned}$$

where $\widetilde{y}(t;a_1^{II,\mathcal{K}} - a_2^{II,\mathcal{K}})$ is the unique solution of (2.5) with $a^{II,\mathcal{K}}$ being replaced by $a_1^{II,\mathcal{K}} - a_2^{II,\mathcal{K}}$. Thus by (2.9), we get

$$\|J(a_1^{II,\mathcal{K}}) - J(a_2^{II,\mathcal{K}})\|^2 \le \frac{3}{4}\|a_1^{II,\mathcal{K}} - a_2^{II,\mathcal{K}}\|^2.$$

Hence, the map J has a unique fixed point $\widetilde{a}^{II,\mathcal{K}}$ in the space $H^{II,\mathcal{K}}\cap V$, and thus $y(t;a^{I,\mathcal{K}}+\widetilde{a}^{II,\mathcal{K}})\in Y$ is the unique solution of (2.1).

Next we shall prove the estimate (2.2). Let $\mathcal{K}$ be an integer, with $\mathcal{K}\ge\mathcal{K}_0$, and let e be a function in the space $L^\infty(0,T;L^n(\Omega))$, with $\|e\|_{L^\infty(0,T;L^n(\Omega))}\le M$. We write y for the unique solution of (2.1) corresponding to $a^{I,\mathcal{K}}\in H^{I,\mathcal{K}}$ and $v\in L^2(0,T;H)$. We can write $y = y_1 + y_2$, where the functions y_1 and y_2 satisfy

$$\begin{cases}(y_1)_t - \triangle y_1 - ey_1 = v & \text{in } Q,\\ y_1 = 0 & \text{on } \Sigma,\\ y_1^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}},\ y_1^{II,\mathcal{K}}(0) = 0 & \text{in } \Omega\end{cases}$$

and

$$\begin{cases} (y_2)_t - \triangle y_2 - e y_2 = 0 & \text{in } Q, \\ y_2 = 0 & \text{on } \Sigma, \\ y_2^{I,\mathcal{K}}(0) = 0,\ y_2^{II,\mathcal{K}}(0) = y^{II,\mathcal{K}}(0) & \text{in } \Omega, \end{cases}$$

respectively. Then by (2.4), we have

$$\|y_1^{II,\mathcal{K}}(T)\|^2 \le \|y_1(T)\|^2 \le C\|y_1\|_Y^2 \le C(\|v\|^2_{L^2(0,T;H)} + \|a^{I,\mathcal{K}}\|^2).$$

It follows from (2.9) that

$$\|y_2^{II,\mathcal{K}}(T)\|^2 \le \frac{3}{4}\|y_2^{II,\mathcal{K}}(0)\|^2 = \frac{3}{4}\|y^{II,\mathcal{K}}(0)\|^2.$$

Thus it holds that

$$\begin{aligned} \|y^{II,\mathcal{K}}(0)\|^2 &= \|y^{II,\mathcal{K}}(T)\|^2 = \|y_1^{II,\mathcal{K}}(T) + y_2^{II,\mathcal{K}}(T)\|^2 \\ &\le C(\varepsilon)\|y_1^{II,\mathcal{K}}(T)\|^2 + (1+\varepsilon)\|y_2^{II,\mathcal{K}}(T)\|^2 \\ &\le C(\varepsilon)\cdot C(\|v\|^2_{L^2(0,T;H)} + \|a^{I,\mathcal{K}}\|^2) + \frac{3}{4}(1+\varepsilon)\|y^{II,\mathcal{K}}(0)\|^2 \ \forall\, \varepsilon > 0, \end{aligned}$$

where $C(\varepsilon)$ is a positive constant depending only on ε. Then by taking ε small enough such that $\frac{3}{4}(1+\varepsilon) < 1$, we get

$$\begin{aligned} \|y^{II,\mathcal{K}}(0)\|^2 &\le C(\|v\|^2_{L^2(0,T;H)} + \|a^{I,\mathcal{K}}\|^2) \\ &\le C(\|v\|^2_{L^2(0,T;H)} + \lambda_{\mathcal{K}}|a^{I,\mathcal{K}}|_2^2), \end{aligned}$$

from which and by (2.4), we obtain

$$\|y\|_Y^2 \le C(\|v\|^2_{L^2(0,T;H)} + \|a^{I,\mathcal{K}}\|^2 + \|y^{II,\mathcal{K}}(0)\|^2) \le C(\|v\|^2_{L^2(0,T;H)} + \lambda_{\mathcal{K}}|a^{I,\mathcal{K}}|_2^2).$$

This completes the proof of Theorem 2.1. □

Next, we shall study the existence of $\mathcal{K}$-approximate periodic solutions for some semilinear parabolic equations. Consider the following semilinear parabolic equation:

$$\begin{cases} y_t - \Delta y + f(x,t,y) = v & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}},\ y^{II,\mathcal{K}}(0) = y^{II,\mathcal{K}}(T) & \text{in } \Omega, \end{cases} \tag{2.10}$$

where $\mathcal{K}$ is a nonnegative integer, $a^{I,\mathcal{K}}$ is an element in the space $H^{I,\mathcal{K}}$, and v is a function in the space $L^2(0,T;H)$.

THEOREM 2.2. *Suppose that* ($\mathbf{H_f}$) *holds. Then, for each positive number M, there exists a nonnegative integer $\mathcal{K} \equiv \mathcal{K}(M)$ such that, for each function v in the space $L^2(0,T;H)$ and each element $a^{I,\mathcal{K}}$ in the space $H^{I,\mathcal{K}}$ satisfying the estimate*

$$\|v\|^2_{L^2(0,T;H)} + (1+\lambda_{\mathcal{K}})|a^{I,\mathcal{K}}|_2^2 \le 2M^2,$$

(2.10) *has at least one solution y in Y satisfying $\|y\|_Y^2 \le C(M)$, where $C(M)$ is a positive constant depending on M.*

Proof of Theorem 2.2. Let ρ be a positive number. We write $\mathcal{A}_\rho$ for the set $\{z \in Y;\ \|z\|_Y \le \rho\}$. It is clear that $\mathcal{A}_\rho$ is a compact and convex subset in $L^2(Q)$. We define a function $g : \Omega \times (0,T) \times \mathbf{R} \to \mathbf{R}$ by setting

$$g(x,t,\xi) = \begin{cases} \frac{f(x,t,\xi)}{\xi}, & \xi \neq 0, \\ \lim_{\xi \to 0} \frac{f(x,t,\xi)}{\xi}, & \xi = 0. \end{cases}$$

Since $Y \hookrightarrow L^\infty(0,T;V) \hookrightarrow L^\infty(0,T;L^{\frac{2n}{n-2}}(\Omega))$, it follows from the assumption $(\mathbf{H_f})$ that, for each element z in the set $\mathcal{A}_\rho$, we have

$$\begin{aligned} \operatorname{esssup}_{0\le t\le T} \int_\Omega |g(x,t,z)|^n dx &\le C \operatorname{esssup}_{0\le t\le T} \int_\Omega (1+|z|^\beta)^n dx \\ &\le C \operatorname{esssup}_{0\le t\le T} \int_\Omega (1+|z|^2) dx \le C(1+\rho^2). \end{aligned}$$

Here and throughout the proof of this theorem, C stands for a positive constant independent of ρ, $\mathcal{K}$, v, and $a^{I,\mathcal{K}}$, which may be different in different contexts. Then by Theorem 2.1, there exists a nonnegative integer $\mathcal{K}_0 \equiv \mathcal{K}_0(\rho, -\Delta, \Omega, T)$ such that, for each integer $\mathcal{K}$ with $\mathcal{K} \ge \mathcal{K}_0$, each element z in the set $\mathcal{A}_\rho$, each function v in the space $L^2(0,T;H)$, and each element $a^{I,\mathcal{K}}$ in the space $H^{I,\mathcal{K}}$, the equation

$$\begin{cases} y_t - \Delta y + g(x,t,z)y = v & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}},\ y^{II,\mathcal{K}}(0) = y^{II,\mathcal{K}}(T) & \text{in } \Omega \end{cases} \tag{2.11}$$

has a unique solution $y \equiv y(t; z, v, a^{I,\mathcal{K}}) \in Y$ satisfying

$$\|y\|_Y^2 \le C(\rho, -\Delta, \Omega, T)(\|v\|^2_{L^2(Q)} + \lambda_{\mathcal{K}} |a^{I,\mathcal{K}}|_2^2).$$

Now we fix such an integer $\mathcal{K}$ and such a pair $(a^{I,\mathcal{K}},\ v)$ in the space $H^{I,\mathcal{K}} \times L^2(Q)$ and then define a map $\Phi \equiv \Phi_{\rho,\mathcal{K},v,a^{I,\mathcal{K}}} : \mathcal{A}_\rho \to L^2(Q)$ by setting $\Phi(z) = y$.

We shall prove that the map Φ has a fixed point, and then (2.10) has a solution in Y. To this end, we first claim that the map Φ is continuous from $\mathcal{A}_\rho$ with the topology induced by the $L^2(Q)$-norm to $L^2(Q)$. Let $\{z_m\}$ be a sequence in the set $\mathcal{A}_\rho$ such that $z_m \to \widetilde{z}$ strongly in $L^2(Q)$ as $m \to \infty$. Then one can easily show that $\widetilde{z} \in \mathcal{A}_\rho$. Write $y_m,\ m = 1, 2, \ldots$, for the solutions of (2.11) with z being replaced by z_m. It suffices to show that $y_m \to \widetilde{y}$ strongly in $L^2(Q)$, as $m \to \infty$, and that $\widetilde{y}$ is the solution of (2.11) with z being replaced by $\widetilde{z}$. Here is the argument: Since $\{y_m\}$ and $\{z_m\}$ are bounded in Y, there exist subsequences $\{y_{m_k}\} \subset \{y_m\}$ and $\{z_{m_k}\} \subset \{z_m\}$ such that, as $m_k \to \infty$,

$$\begin{aligned} y_{m_k} \to \overline{y} \quad & \text{weakly in } Y, \\ & \text{strongly in } L^2(Q), \\ & \text{a.e. in } Q \end{aligned} \tag{2.12}$$

and

$$\begin{aligned} z_{m_k} \to \widetilde{z} \quad & \text{weakly in } Y, \\ & \text{strongly in } L^2(Q), \\ & \text{a.e. in } Q. \end{aligned} \tag{2.13}$$

Moreover,

$$(y_{m_k}(0), y_{m_k}(T)) \to (\overline{y}(0), \overline{y}(T)) \text{ strongly in } H \times H \text{ as } m_k \to \infty. \tag{2.14}$$

(Indeed, by the Ascoli–Arzela theorem, we can choose $\{z_{m_k}\}$ and $\{y_{m_k}\}$ such that $y_{m_k} \to \overline{y}$ and $z_{m_k} \to \widetilde{z}$ strongly in $C([0,T];H)$.) Now, by the continuity of g in the third variable and by (2.12) and (2.13), we obtain

$$g(x,t,z_{m_k}(x,t))y_{m_k}(x,t) \to g(x,t,\widetilde{z}(x,t))\overline{y}(x,t) \text{ a.e. in } L^2(Q) \text{ as } m_k \to \infty.$$

On the other hand, by $(\mathbf{H_f})$ and by making use of the Hölder inequality and the Sobolev embedding theorem, we have

$$\|g(x,t,z_{m_k})y_{m_k}\|^2_{L^2(Q)} \le C\|y_{m_k}\|^2_Y \|z_{m_k}\|^2_Y \le C.$$

Thus it holds that

$$g(x,t,z_{m_k}(x,t))y_{m_k}(x,t) \to g(x,t,\widetilde{z}(x,t))\overline{y}(x,t) \text{ weakly in } L^2(Q) \text{ as } m_k \to \infty.$$

Then by (2.12), (2.13), (2.14), and the above, we can pass to the limit, as $m_k \to \infty$, in the equation satisfied by y_{m_k} and z_{m_k} to get $\widetilde{y} = \overline{y}$. Hence, we have proved that $y_m \to \widetilde{y}$ strongly in $L^2(Q)$, as $m \to \infty$, and therefore Φ is continuous.

Next we claim that, for any z in the set $\mathcal{A}_\rho$, the function $y \equiv \Phi(z)$ satisfies the following estimate:

$$\|y\|^2_Y \le \widetilde{C}(1+\rho^{2\beta})(\|v\|^2_{L^2(Q)} + (1+\lambda_{\mathcal{K}})|a^{I,\mathcal{K}}|^2_2), \tag{2.15}$$

where $\widetilde{C}$ is a positive constant independent of ρ, $\mathcal{K}$, v, and $a^{I,\mathcal{K}}$.

By (i) of $(\mathbf{H_f})$, we obtain

$$g(x,t,\xi) \ge \mu - \lambda_1 \ \ \forall\, \xi \in \mathbf{R} \text{ and for almost all } (x,t) \in Q. \tag{2.16}$$

Multiplying the first equation of (2.11) by y and making use of (2.16), we get

$$\frac{1}{2}\frac{d}{dt}|y(t)|^2_2 + \mu|y(t)|^2_2 \le |\langle v(t), y(t)\rangle| \ \forall\ t \in [0,T],$$

from which it follows that

$$\frac{d}{dt}|y(t)|^2_2 + \mu|y(t)|^2_2 \le \frac{1}{\mu}|v(t)|^2_2 \ \forall\ t \in [0,T].$$

Integrating the above over $(0,T)$, we get

$$e^{\mu T}|y(T)|^2_2 \le \frac{1}{\mu}e^{\mu T}\int_0^T |v(t)|^2_2 dt + |y(0)|^2_2.$$

Hence, it holds that

$$e^{\mu T}|y^{II,\mathcal{K}}(T)|^2_2 \le \frac{1}{\mu}e^{\mu T}\int_0^T |v(t)|^2_2 dt + |y^{II,\mathcal{K}}(0)|^2_2 + |a^{I,\mathcal{K}}|^2_2.$$

Since $y^{II,\mathcal{K}}(0) = y^{II,\mathcal{K}}(T)$, we have

$$\begin{aligned} |y^{II,\mathcal{K}}(0)|^2_2 &\le (e^{\mu T}-1)^{-1}\left(\frac{1}{\mu}e^{\mu T}\int_0^T |v(t)|^2_2 dt\right) + (e^{\mu T}-1)^{-1}|a^{I,\mathcal{K}}|^2_2 \\ &\le C(\|v\|^2_{L^2(Q)} + |a^{I,\mathcal{K}}|^2_2), \end{aligned}$$

from which it follows that

$$|y(0)|_2^2 \le C(\|v\|_{L^2(Q)}^2 + |a^{I,\mathcal{K}}|_2^2).$$

Then by (2.11) and (2.16) and by making use of the standard energy estimate argument for linear parabolic equations, we can easily have

$$\begin{aligned}\|y\|_{V_2(Q)}^2 \equiv \operatorname{esssup}_{0\le t\le T}|y(t)|_2^2 + \int_0^T \|y(t)\|^2 dt &\le C(\|v\|_{L^2(Q)}^2 + |y(0)|_2^2)\\ &\le C(\|v\|_{L^2(Q)}^2 + |a^{I,\mathcal{K}}|_2^2).\end{aligned} \tag{2.17}$$

On the other hand, we let $q = \frac{n+2}{n-2}\cdot\frac{1}{\beta}$. Then $q > \frac{n+2}{2}$ for $0 < \beta \le \frac{2}{n}$. By $(\mathbf{H_f})$ and by making use of the Hölder inequality and the Sobolev embedding theorem $(Y \hookrightarrow L^{\frac{2(n+2)}{n-2}}(Q))$, we get

$$\begin{aligned}\|(g(x,t,z))^2\|_{L^q(Q)} &\le C\left\{\int_Q (|z|^\beta + 1)^{2q} dxdt\right\}^{\frac{1}{q}} \le C\left\{1 + \left(\int_Q |z|^{\frac{2(n+2)}{n-2}} dxdt\right)^{\frac{1}{q}}\right\}\\ &\le C\left(1 + \|z\|^{2\beta}_{L^{\frac{2(n+2)}{n-2}}(Q)}\right) \le C(1 + \|z\|_Y^{2\beta}) \le C(1+\rho^{2\beta}).\end{aligned} \tag{2.18}$$

By (2.17) and (2.18), and by the estimate in [3] (see p. 50 in [3]), we get

$$\begin{aligned}\int_Q (g(x,t,z))^2 y^2 dxdt &\le C\|(g(x,t,z))^2\|_{L^q(Q)} \cdot \|y\|_{V_2(Q)}^2\\ &\le C(1+\rho^{2\beta})(\|v\|_{L^2(Q)}^2 + |a^{I,\mathcal{K}}|_2^2).\end{aligned} \tag{2.19}$$

Now let $F = v - g(x,t,z)y \in L^2(Q)$. Then $y \in Y$ satisfies the following heat equation:

$$\begin{cases} y_t - \Delta y = F & \text{in } Q,\\ y = 0 & \text{on } \Sigma,\\ y^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}},\ y^{II,\mathcal{K}}(0) = y^{II,\mathcal{K}}(T) & \text{in } \Omega. \end{cases} \tag{2.20}$$

Moreover, by (2.19), it is clear that

$$\int_Q (F(x,t))^2 dxdt \le C(1+\rho^{2\beta})(\|v\|_{L^2(Q)}^2 + |a^{I,\mathcal{K}}|_2^2). \tag{2.21}$$

Multiplying the first equation of (2.20) by $t(-\Delta y(t))$, we get

$$\frac{d}{dt}(t\|y(t)\|^2) + t|\Delta y(t)|_2^2 \le \|y(t)\|^2 + t|F(t)|_2^2.$$

Integrating the above over $(0,T)$, by (2.17) and (2.21), we get

$$T\|y(T)\|^2 \le \int_0^T \|y(t)\|^2 dt + T\int_0^T |F(t)|_2^2 dt \le C(1+\rho^{2\beta})(\|v\|_{L^2(Q)}^2 + |a^{I,\mathcal{K}}|_2^2),$$

from which it follows that

$$\|y^{II,\mathcal{K}}(0)\|^2 = \|y^{II,\mathcal{K}}(T)\|^2 \le C(1+\rho^{2\beta})(\|v\|_{L^2(Q)}^2 + |a^{I,\mathcal{K}}|_2^2).$$

Hence,

$$\|y(0)\|^2 = \|y^{I,\mathcal{K}}(0)\|^2 + \|y^{II,\mathcal{K}}(0)\|^2 \le C(1+\rho^{2\beta})(\|v\|^2_{L^2(Q)} + |a^{I,\mathcal{K}}|_2^2) + \|a^{I,\mathcal{K}}\|^2$$
$$\le C(1+\rho^{2\beta})(\|v\|^2_{L^2(Q)} + (1+\lambda_{\mathcal{K}})|a^{I,\mathcal{K}}|_2^2).$$

Then by (2.20) and (2.21) and by making use of the standard energy estimate argument for the heat equation, we can get (2.15) easily.

Notice that $\beta \le \frac{2}{n} < 1$. Thus for a given $M > 0$, we can take a $\rho > 0$ large enough such that $2\widetilde{C}M^2(1+\rho^{2\beta}) \le \rho^2$, where the constant $\widetilde{C}$ is the one appearing in (2.15). Then we fix an integer $\mathcal{K} \equiv \mathcal{K}(\rho)$ such that (2.11) has a unique solution in Y for each triplet $(v, a^{I,\mathcal{K}}, z) \in L^2(Q) \times H^{I,\mathcal{K}} \times \mathcal{A}_\rho$. Hence, it follows from (2.15) that, for any pair $(v, a^{I,\mathcal{K}}) \in L^2(Q) \times H^{I,\mathcal{K}}$ with $\|v\|^2_{L^2(Q)} + (1+\lambda_{\mathcal{K}})|a^{I,\mathcal{K}}|_2^2 \le 2M^2$, the map $\Phi \equiv \Phi_{\rho,\mathcal{K},v,a^{I,\mathcal{K}}}$ satisfies

$$\|\Phi(z)\|_Y^2 \le 2\widetilde{C}M^2(1+\rho^{2\beta}) \le \rho^2 \text{ for any } z \in \mathcal{A}_\rho, \text{ namely, } \Phi(\mathcal{A}_\rho) \subset \mathcal{A}_\rho.$$

Since $\mathcal{A}_\rho$ is convex and compact in $L^2(Q)$ and Φ is continuous, we conclude that Φ has a fixed point $y \in Y$ which is a solution to (2.10). This completes the proof of Theorem 2.2. □

Next, we shall give some examples to show how to estimate the integers $\mathcal{K}_0$ and $\mathcal{K}$ in Theorems 2.1 and 2.2 in certain particular cases.

Example 2.3. Consider the following equation:

$$\text{(2.22)} \quad \begin{cases} y_t(x,t) - \Delta y(x,t) - e(x,t)y(x,t) = v(x,t) & 0 \le x \le 1, 0 \le t \le 1, \\ y(0,t) = y(1,t) = 0 & 0 \le t \le 1. \end{cases}$$

Here $e \in L^\infty((0,1)\times(0,1))$ and $v \in L^2((0,1)\times(0,1))$. Notice that $\{\sqrt{2}\sin(j\pi x)\}_{j=1}^\infty$ forms an orthonormal basis of $L^2(0,1)$. We write $v(x,t) = \sum_{j=1}^\infty v_j(t)\sin(j\pi x)$ and $y(x,t) = \sum_{j=1}^\infty w_j(t)\sin(j\pi x)$. We assume first that $e(x,t) \equiv a$ for all (x,t) in $[0,1]\times[0,1]$, where a is a real number. Then we have

$$w_j'(t) = -(j\pi)^2 w_j(t) + a w_j(t) + v_j(t), \quad j = 1, 2, \dots.$$

Thus, it holds that, for all $j = 1, 2, \dots,$

$$\text{(2.23)} \qquad w_j(1)\exp\{-a + (j\pi)^2\} - w_j(0) = \int_0^1 v_j(t)\exp\{-a + (j\pi)^2\}dt.$$

Now, for a given natural number N, we choose $a = (N\pi)^2$ and choose $v(x,t)$ such that $\int_0^1 v_N(t)dt \ne 0$; then, by (2.23), we see that (2.22) can never have a solution $y(x,t) = \sum_{j=1}^\infty w_j(t)\sin(j\pi x)$ such that $w_N(0) = w_N(1)$. Namely, (2.22) can never have a $\mathcal{K}$-approximate periodic solution with $\mathcal{K} = N-1$. However, in the case that $a = (N\pi)^2$, for each $a^{I,\mathcal{K}} = \sum_{j=1}^N a_j \sin(j\pi x)$, $v \in L^2((0,1)\times(0,1))$, (2.22) does have a unique solution y satisfying $y^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}}$, $y^{II,\mathcal{K}}(0) = y^{II,\mathcal{K}}(1)$. Namely, (2.22) has a unique $\mathcal{K}$-approximate periodic solution y with $\mathcal{K} = N$, which satisfies $y^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}}$. Moreover, we can easily see from (2.23) that, for each real number a, if we take $\mathcal{K}_0$ such that

$$\text{(2.24)} \qquad (\mathcal{K}_0\pi)^2 \ge a,$$

then it holds that, for each integer $\mathcal{K} \ge \mathcal{K}_0$, $v \in L^2((0,1)\times(0,1))$, and $a^{I,\mathcal{K}} = \sum_{j=1}^N a_j \sin(j\pi x)$, (2.22) has a unique solution y satisfying $y^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}}$, $y^{II,\mathcal{K}}(0) =$

$y^{II,\mathcal{K}}(1)$. Thus, the inequality (2.24) gives a way to calculate the integer $\mathcal{K}_0$ in Theorem 2.1 in this case.

Next, we consider (2.22) where e is a function in the subset $\{e \in L^\infty((0,1)\times(0,1)); \|e\|_{L^\infty((0,1)\times(0,1))} \le M\}$ for a fixed positive number M. In this case, the integer $\mathcal{K}_0$ in Theorem 2.1 can be estimated as follows. Notice first that the constant C in the inequality (2.8) can be obtained from the embedding inequalities (2.6) and (2.7). Now in (2.8) we take $\varepsilon = \frac{1}{5C} < \frac{1}{4C}$ and $C(\varepsilon, M) = \frac{M^2}{\varepsilon} = 5CM^2$. Then the integer $\mathcal{K}_0$ can be calculated from the following inequalities:

$$\begin{aligned} &\exp\{-2((\mathcal{K}_0+1)\pi)^2\} < \frac{1}{4}, \\ &\frac{5C^2M^2}{2((\mathcal{K}_0+1)\pi)^2}(1-\exp\{-2((\mathcal{K}_0+1)\pi)^2\}) < \frac{1}{4}. \end{aligned} \tag{2.25}$$

Finally, we consider the following semilinear equation:

$$\begin{cases} y_t(x,t) - \Delta y(x,t) + f(y(x,t)) = v(x,t) & 0 \le x \le 1, 0 \le t \le 1, \\ y(0,t) = y(1,t) = 0 & 0 \le t \le 1. \end{cases} \tag{2.26}$$

Here the function $f : \mathbf{R} \to \mathbf{R}$ is continuously differentiable and has the following properties: $f(0) = 0$, $f(\xi)\xi \ge 0$ for all $\xi \in \mathbf{R}$, and there exists a positive constant C_0 such that $|f'(\xi)| \le C_0$ for all $\xi \in \mathbf{R}$. Then the function $g(\xi)$ defined by

$$g(\xi) = \begin{cases} \frac{f(\xi)}{\xi}, & \xi \ne 0, \\ \lim_{\xi\to 0} \frac{f(\xi)}{\xi}, & \xi = 0 \end{cases}$$

satisfies $|g(\xi)| \le C_0$ for all $\xi \in \mathbf{R}$. Hence, it holds that, for each $z \in \mathbf{Y}$,

$$|g(z(x,t))| \le C_0, \quad \text{a.e. } (x,t) \in (0,1)\times(0,1). \tag{2.27}$$

Then, by Theorem 2.1, we obtain that there exists a nonnegative integer $\mathcal{K}_0 \equiv \mathcal{K}_0(-\triangle, C_0)$ such that for each $\mathcal{K} \ge \mathcal{K}_0$, $z \in \mathbf{Y}$, $v \in L^2((0,1)\times(0,1))$, and $a^{I,\mathcal{K}} \in H^{I,\mathcal{K}}$, the linearized equation

$$\begin{cases} y_t - \Delta y + g(z)y = v & 0 \le x \le 1, 0 \le t \le 1, \\ y(0,t) = y(1,t) = 0 & 0 \le t \le 1, \\ y^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}},\ y^{II,\mathcal{K}}(0) = y^{II,\mathcal{K}}(T) & 0 \le x \le 1 \end{cases}$$

has a unique solution $y \in \mathbf{Y}$ satisfying

$$\|y\|_Y^2 \le C(C_0, -\triangle)(\|v\|^2_{L^2((0,1)\times(0,1))} + \lambda_{\mathcal{K}}|a^{I,\mathcal{K}}|_2^2). \tag{2.28}$$

Moreover, the aforementioned integer $\mathcal{K}_0$ can be estimated by (2.25) with M being replaced by C_0. Notice that, in this case, both the integer $\mathcal{K}_0$ and the constant $C(C_0, -\Delta)$ in (2.28) are independent of ρ, which is the positive number given in the proof of Theorem 2.2. Then, by (2.27) and (2.28), after carefully checking the proof of Theorem 2.2, we can obtain that, in this case, the integer $\mathcal{K}$ in Theorem 2.2 can be taken as the aforementioned integer $\mathcal{K}_0$. Hence, in this case, the integer $\mathcal{K}$ in Theorem 2.2 can be estimated by (2.25).

3. Pontryagin's maximum principle of optimal controls for the problem $(\mathbf{P}_{\mathcal{K},r})$. In this section, we shall prove Theorem 1.1, namely, the Pontryagin maximum principle for the optimal control problem $(\mathbf{P}_{\mathcal{K},r})$. The basic idea to be used here is to construct a well-posed approximate optimization problem to approach the non-well-posed problem $(\mathbf{P}_{\mathcal{K},r})$. This idea has been used in [10], [12], [13]. However, in the current work, the approximation problem is to ask the infimum of a penalty functional over a suitable subset of the space $Y \times L^2(Q)$. Moreover, this construction works well due to Theorem 2.1, namely, the results of existence and uniqueness of $\mathcal{K}$-approximate periodic solutions for the linear parabolic equations. Throughout this section, we assume that $(\mathbf{A_f})$ holds.

3.1. Formulation of an approximate problem. Let $(y^*, u^*) \in Y \times L^2(Q)$ be optimal for problem $(\mathbf{P}_{\mathcal{K},r})$. By the property $(\mathbf{A_f})$ and by making use of the Sobolev embedding theorem and the Hölder inequality, we see that there exists a positive constant C_0 such that the following holds:

$$\|f'_y(x,t,y)\|_{L^\infty(0,T;L^n(\Omega))} \le C_0(1 + \|y\|_Y^{\frac{2}{n-2}}) \ \forall\ y \in Y. \tag{3.1}$$

Set

$$M = C_0(1 + \|y^*\|_Y^{\frac{2}{n-2}}) + 1. \tag{3.2}$$

Then by Theorem 2.1, there exists an integer $\mathcal{K}_0 \equiv \mathcal{K}_0(M, \Omega, -\Delta, T)$, with $\mathcal{K}_0 > \mathcal{K}$, such that for each function e in the space $L^\infty(0,T;L^n(\Omega))$ satisfying the estimate $\|e\|_{L^\infty(0,T;L^n(\Omega))} \le M$, each function g in the space $L^2(Q)$, and each element $b^{I,\mathcal{K}_0}$ in the space $H^{I,\mathcal{K}_0}$, the following equation has a unique solution y in the space Y:

$$\begin{cases} y_t - \Delta y + ey = g & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y^{I,\mathcal{K}_0}(0) = b^{I,\mathcal{K}_0},\ y^{II,\mathcal{K}_0}(0) = y^{II,\mathcal{K}_0}(T) & \text{in } \Omega. \end{cases} \tag{3.3}$$

Write $H^{II,\mathcal{K}} = H^{II,\mathcal{K},1} \oplus H^{II,\mathcal{K},2}$, where $H^{II,\mathcal{K},1} = span\{X_i\}_{i=\mathcal{K}+1}^{\mathcal{K}_0}$, $H^{II,\mathcal{K},2} = span\{X_i\}_{i=\mathcal{K}_0+1}^{\infty} \equiv H^{II,\mathcal{K}_0}$. Then, for each $y(t) = \sum_{i=1}^{\infty} y_i(t)X_i \in L^2(0,T;H)$ and for each $h = \sum_{i=1}^{\infty} h_i X_i \in H$, we have

$$y(t) = y^{I,\mathcal{K}}(t) + y^{II,\mathcal{K},1}(t) + y^{II,\mathcal{K},2}(t) \equiv \sum_{i=1}^{\mathcal{K}} y_i(t)X_i + \sum_{i=\mathcal{K}+1}^{\mathcal{K}_0} y_i(t)X_i + \sum_{i=\mathcal{K}_0+1}^{\infty} y_i(t)X_i$$

and

$$h = h^{I,\mathcal{K}} + h^{II,\mathcal{K},1} + h^{II,\mathcal{K},2} \equiv \sum_{i=1}^{\mathcal{K}} h_i X_i + \sum_{i=\mathcal{K}+1}^{\mathcal{K}_0} h_i X_i + \sum_{i=\mathcal{K}_0+1}^{\infty} h_i X_i.$$

Now, for each $\varepsilon > 0$, we define a functional $J_\varepsilon : Y \times L^2(Q) \to \mathbf{R}$ by setting

$$\begin{aligned} J_\varepsilon(y,u) &= J(y,u) + \frac{n-2}{2n}\int_Q (y - y^*)^{\frac{2n}{n-2}} dxdt + \frac{1}{2}\int_Q (u - u^*)^2 dxdt \\ &\quad + \frac{1}{2\varepsilon}|y^{II,\mathcal{K},1}(T) - y^{II,\mathcal{K},1}(0)|_2^2 + \frac{1}{2\varepsilon}(d_B(y^{I,\mathcal{K}}(T)) + \varepsilon)^2 \\ &\quad + \frac{1}{2\varepsilon}\int_Q (y_t - \Delta y + f(x,t,y) - \chi_\omega u)^2 dxdt, \end{aligned} \tag{3.4}$$

where $d_B(\cdot)$ denotes the distance function of $\cdot$ to the set $B^{I,\mathcal{K}}(0,r)$ in the space $H^{I,\mathcal{K}}$. One can easily check that the functional J_ε is well-defined. Set

$$\widetilde{Y} = \{y \in Y;\ y^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}},\ y^{II,\mathcal{K},2}(0) = y^{II,\mathcal{K},2}(T)\},$$

and consider the following optimization problem:

$$(\mathbf{P}^{\varepsilon}_{\mathcal{K},r}):\ \ \mathrm{Inf} J_\varepsilon(y,u) \text{ over all } (y,u) \in \widetilde{Y} \times L^2(Q).$$

We shall use the well-posed approximate problem $(\mathbf{P}^{\varepsilon}_{\mathcal{K},r})$ to approach the original optimal control problem $(\mathbf{P}_{\mathcal{K},r})$ and finally derive the Pontryagin maximum principle in a qualified form for the problem $(\mathbf{P}_{\mathcal{K},r})$.

3.2. Properties of the approximate problem $(\mathbf{P}^{\varepsilon}_{\mathcal{K},r})$. First of all, we shall derive a prior estimate for the solutions to the following equation, which will be used later:

$$\left\{\begin{array}{ll} y_t - \Delta y = g & \text{in } Q,\\ y = 0 & \text{on } \Sigma,\\ y^{I,\mathcal{K}}(0) = b^{I,\mathcal{K}},\ y^{II,\mathcal{K},1}(0) = y^{II,\mathcal{K},1}(T) + \xi^{II,\mathcal{K},1} & \text{in } \Omega,\\ y^{II,\mathcal{K},2}(0) = y^{II,\mathcal{K},2}(T) & \text{in } \Omega. \end{array}\right. \tag{3.5}$$

LEMMA 3.1. *Let $g \in L^2(Q)$, $b^{I,\mathcal{K}} \in H^{I,\mathcal{K}}$, $\xi^{II,\mathcal{K},1} \in H^{II,\mathcal{K},1}$, and $y \in Y$ satisfy* (3.5). *Then there exists a positive constant $C \equiv C(\Omega, -\Delta, T)$ depending only on T, Ω, and the operator $(-\Delta)$ such that*

$$\|y\|^2_Y \le C(\|g\|^2_{L^2(Q)} + |b^{I,\mathcal{K}}|^2_2 + |\xi^{II,\mathcal{K},1}|^2_2). \tag{3.6}$$

Proof of Lemma 3.1. We write $y(x,t) = \sum_{i=1}^{\infty} y_i(t)X_i$, $g(x,t) = \sum_{i=1}^{\infty} g_i(t)X_i$, $b^{I,\mathcal{K}} = \sum_{i=1}^{\mathcal{K}} b_i X_i, \xi^{II,\mathcal{K},1} = \sum_{i=\mathcal{K}+1}^{\mathcal{K}_0} \xi_i X_i$, where $b_i,\ \xi_i \in \mathbf{R}$. Then it holds that

$$\frac{d}{dt} y_i(t) + \lambda_i y_i(t) = g_i(t),\ i = 1,2,\ldots,$$

from which we get

$$y_i(t) = e^{-\lambda_i t} y_i(0) + \int_0^t e^{-\lambda_i (t-s)} g_i(s) ds,\ i = 1,2,\ldots.$$

Then by (3.5), we have

$$\left\{\begin{array}{ll} y_i(0) = b_i, & i = 1,2,\ldots,\mathcal{K},\\ y_i(0) = \dfrac{1}{1-e^{-\lambda_i T}} \left\{ \displaystyle\int_0^T e^{-\lambda_i (T-s)} g_i(s) ds + \xi_i \right\}, & i = \mathcal{K}+1, \mathcal{K}+2, \ldots, \mathcal{K}_0,\\ y_i(0) = \dfrac{1}{1-e^{-\lambda_i T}} \displaystyle\int_0^T e^{-\lambda_i (T-s)} g_i(s) ds, & i = \mathcal{K}_0+1, \mathcal{K}_0+2, \ldots. \end{array}\right. \tag{3.7}$$

Notice that $0 < \lambda_1 < \lambda_2 \le \cdots \to +\infty$. A simple computation, together with (3.7), then leads us to

$$\begin{aligned} \|y(\cdot,0)\|^2 = \sum_{i=1}^{\infty} \lambda_i y_i^2(0) &\le C \left(\sum_{i=1}^{\mathcal{K}} b_i^2 + \sum_{i=\mathcal{K}+1}^{\mathcal{K}_0} \xi_i^2 + \sum_{i=1}^{\infty} \int_0^T g_i^2(t) dt \right)\\ &= C(|b^{I,\mathcal{K}}|^2_2 + |\xi^{II,\mathcal{K},1}|^2_2 + \|g\|^2_{L^2(Q)}). \end{aligned}$$

Then, by the standard energy estimate for the heat equation, we get (3.6). This completes the proof of Lemma 3.1. □

Remark 3.1. Lemma 3.1 holds for all integers $\mathcal{K}, \mathcal{K}_0$, with $0 \le \mathcal{K} < \mathcal{K}_0$.

LEMMA 3.2. *Problem* $(\mathbf{P}^{\varepsilon}_{\mathcal{K},r})$ *has at least one solution.*

Proof of Lemma 3.2. Set $d = \mathrm{Inf}\{J_\varepsilon(y,u);\ (y,u) \in \widetilde{Y} \times L^2(Q)\}$, and let $\{(y_m, u_m)\}$ be a sequence in the space $\widetilde{Y} \times L^2(Q)$ such that

$$d \le J_\varepsilon(y_m, u_m) \le d + \frac{1}{m}. \tag{3.8}$$

By (3.4) and (3.8), it follows that the sequence $\{u_m\}$ is bounded in $L^2(Q)$, the sequence $\{(y_m)_t - \Delta y_m + f(x,t,y_m) - \chi_\omega u_m\}$ is bounded in $L^2(Q)$, and the sequence $\{y_m\}$ is bounded in $L^{\frac{2n}{n-2}}(Q)$. From the latter and by making use of the assumption $(\mathbf{A_f})$, we see easily that the sequence $\{f(x,t,y_m)\}$ is bounded in $L^2(Q)$. Applying (3.4) and (3.8) again, we obtain also that the sequence $\{y_m^{II,\mathcal{K},1}(T) - y_m^{II,\mathcal{K},1}(0)\}$ is bounded in H. Set $g_m = (y_m)_t - \Delta y_m$ and $\xi_m^{II,\mathcal{K},1} = y_m^{II,\mathcal{K},1}(0) - y_m^{II,\mathcal{K},1}(T)$. Then the sequences $\{g_m\}$ and $\{\xi_m^{II,\mathcal{K},1}\}$ are bounded in the spaces $L^2(Q)$ and H, respectively. Since $\{y_m\} \subset \widetilde{Y}$, we have $y_m^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}}$ and $y_m^{II,\mathcal{K},2}(0) = y_m^{II,\mathcal{K},2}(T)$. Thus we can apply Lemma 3.1 to get

$$\|y_m\|_Y^2 \le C(\|g_m\|^2_{L^2(Q)} + |a^{I,\mathcal{K}}|_2^2 + |\xi_m^{II,\mathcal{K},1}|_2^2).$$

Hence, $\{y_m\}$ is bounded in Y. Then by the Aubin compactness theorem and the Ascoli–Arzela theorem, we can extract a subsequence of $\{y_m\}$, which, for simplicity of notation, is still denoted by itself such that as $m \to \infty$

$$y_m \to \widetilde{y} \quad \text{weakly in } Y, \tag{3.9}$$

$$y_m \to \widetilde{y} \quad \text{strongly in } L^2(0,T;V) \cap C([0,T];H), \tag{3.10}$$

$$y_m \to \widetilde{y} \quad \text{a.e. in } Q. \tag{3.11}$$

In particular, it follows from (3.10) that

$$(y_m^{II,\mathcal{K},1}(T), y_m^{II,\mathcal{K},1}(0)) \to (\widetilde{y}^{II,\mathcal{K},1}(T), \widetilde{y}^{II,\mathcal{K},1}(0)) \text{ strongly in } H \times H \text{ as } m \to \infty \tag{3.12}$$

and

$$d_B(y_m^{I,\mathcal{K}}(T)) \to d_B(\widetilde{y}^{I,\mathcal{K}}(T)) \ \text{ as } m \to \infty. \tag{3.13}$$

On the other hand, since the sequence $\{u_m\}$ is bounded in the space $L^2(Q)$, we can assume, without loss of generality, that

$$u_m \to \widetilde{u} \text{ weakly in } L^2(Q) \ \text{ as } m \to \infty. \tag{3.14}$$

Since $\{y_m\} \subset \widetilde{Y}$, it follows from (3.10) that $\widetilde{y}^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}}$, $\widetilde{y}^{II,\mathcal{K},2}(0) = \widetilde{y}^{II,\mathcal{K},2}(T)$. Hence, $\widetilde{y} \in \widetilde{Y}$.

Now we claim that

$$f(x,t,y_m) \to f(x,t,\widetilde{y}) \text{ weakly in } L^2(Q) \ \text{ as } m \to \infty. \tag{3.15}$$

Indeed, by (3.11) and $(\mathbf{A_f})$, we obtain

$$f(x,t,y_m) \to f(x,t,\widetilde{y}) \ \text{ a.e. in } Q \text{ as } m \to \infty.$$

On the other hand, the sequence $\{f(x,t,y_m)\}$ is bounded in $L^2(Q)$. Then (3.15) follows immediately. Now, by (3.9), (3.14), and (3.15), we see that as $m \to \infty$

$$(y_m)_t - \Delta y_m + f(x,t,y_m) - \chi_\omega u_m \to \widetilde{y}_t - \Delta\widetilde{y} + f(x,t,\widetilde{y}) - \chi_\omega \widetilde{u} \text{ weakly in } L^2(Q). \tag{3.16}$$

Then by (3.9), from which it follows that $y_m \to \widetilde{y}$ weakly in $L^{\frac{2n}{n-2}}(Q)$ as $m \to \infty$, and by (3.12), (3.13), (3.14), (3.16), and (3.4), we can easily get that

$$\underline{\lim}_{m\to\infty} J_\varepsilon(y_m, u_m) \geq J_\varepsilon(\widetilde{y}, \widetilde{u}) \geq d.$$

This together with (3.8) implies $J_\varepsilon(\widetilde{y}, \widetilde{u}) = d$. Since $(\widetilde{y}, \widetilde{u}) \in \widetilde{Y} \times L^2(Q)$, $(\widetilde{y}, \widetilde{u})$ is a solution to problem $(\mathbf{P}^\varepsilon_{\mathcal{K},r})$. This completes the proof of Lemma 3.2. □

LEMMA 3.3. *Let $(y_\varepsilon, u_\varepsilon)$ be optimal for problem $(\mathbf{P}^\varepsilon_{\mathcal{K},r})$. Then when $\varepsilon \to 0^+$, $u_\varepsilon \to u^*$ strongly in $L^2(Q)$, $y_\varepsilon \to y^*$ strongly in Y.*

Proof of Lemma 3.3. It is clear that $y^* \in \widetilde{Y}$. By the optimality of $(y_\varepsilon, u_\varepsilon)$ to problem $(\mathbf{P}^\varepsilon_{\mathcal{K},r})$, it follows that

$$J_\varepsilon(y_\varepsilon, u_\varepsilon) \leq J_\varepsilon(y^*, u^*) = J(y^*, u^*) + \frac{\varepsilon}{2},$$

which implies

$$\overline{\lim_{\varepsilon\to 0^+}} J_\varepsilon(y_\varepsilon, u_\varepsilon) \leq J(y^*, u^*). \tag{3.17}$$

By (3.4) and (3.17), we see that the sequence $\{y_\varepsilon\}$ is bounded in $L^{\frac{2n}{n-2}}(Q)$, the sequence $\{u_\varepsilon\}$ is bounded in $L^2(Q)$, the sequence $\{y_\varepsilon^{II,\mathcal{K},1}(T) - y_\varepsilon^{II,\mathcal{K},1}(0)\}$ is bounded in H, and the sequence $\{(y_\varepsilon)_t - \Delta y_\varepsilon + f(x,t,y_\varepsilon) - \chi_\omega u_\varepsilon\}$ is bounded in $L^2(Q)$. Then, by the same argument as that in the proof of Lemma 3.2, we can extract a generalized subsequence of $\{\varepsilon\}$, which, for simplicity of notation, is still denoted in the same way such that as $\varepsilon \to 0^+$

$$\begin{array}{ll} u_\varepsilon \to \widetilde{u} & \text{weakly in } L^2(Q), \\ y_\varepsilon \to \widetilde{y} & \text{weakly in } Y, \\ & \text{a.e. in } Q, \\ & \text{strongly in } C([0,T];H) \cap L^2(0,T;V), \end{array} \tag{3.18}$$

$$(y_\varepsilon)_t - \Delta y_\varepsilon + f(x,t,y_\varepsilon) - \chi_\omega u_\varepsilon \to \widetilde{y}_t - \Delta\widetilde{y} + f(x,t,\widetilde{y}) - \chi_\omega \widetilde{u} \text{ weakly in } L^2(Q). \tag{3.19}$$

Since $\{y_\varepsilon\} \subset \widetilde{Y}$, it follows from (3.18) that $\widetilde{y} \in \widetilde{Y}$.

Applying (3.4) and (3.17) again, we obtain

$$d_B(y_\varepsilon^{I,\mathcal{K}}(T)) \to 0 \ \text{ as } \varepsilon \to 0^+, \tag{3.20}$$

$$y_\varepsilon^{II,\mathcal{K},1}(T) - y_\varepsilon^{II,\mathcal{K},1}(0) \to 0 \ \text{ strongly in } H^{II,\mathcal{K},1} \ \text{ as } \varepsilon \to 0^+, \tag{3.21}$$

and

$$(y_\varepsilon)_t - \Delta y_\varepsilon + f(x,t,y_\varepsilon) - \chi_\omega u_\varepsilon \to 0 \ \text{ strongly in } L^2(Q) \ \text{ as } \varepsilon \to 0^+. \tag{3.22}$$

By (3.18) and (3.20), we have $\widetilde{y}^{I,\mathcal{K}}(T) \in B^{I,\mathcal{K}}(0,r)$. By (3.18) and (3.21), we obtain $\widetilde{y}^{II,\mathcal{K},1}(T) = \widetilde{y}^{II,\mathcal{K},1}(0)$. Then by (3.19) and (3.22) and noticing that $\widetilde{y} \in \widetilde{Y}$, we see that $(\widetilde{y}, \widetilde{u})$ satisfies the following:

$$\begin{cases} \widetilde{y}_t - \Delta\widetilde{y} + f(x,t,\widetilde{y}) = \chi_\omega \widetilde{u} & \text{in } Q, \\ \widetilde{y} = 0 & \text{on } \Sigma, \\ \widetilde{y}^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}},\ \widetilde{y}^{II,\mathcal{K}}(0) = \widetilde{y}^{II,\mathcal{K}}(T) & \text{in } \Omega, \\ \widetilde{y}^{I,\mathcal{K}}(T) \in B^{I,\mathcal{K}}(0,r). \end{cases}$$

Hence, $(\widetilde{y},\widetilde{u}) \in Y \times L^2(Q)$ is admissible for problem $(\mathbf{P}_{\mathcal{K},r})$, and therefore

$$J(\widetilde{y},\widetilde{u}) \geq J(y^*,u^*). \tag{3.23}$$

On the other hand, it follows, from (3.4) and (3.18), that

$$\varliminf_{\varepsilon\to 0^+} J_\varepsilon(y_\varepsilon,u_\varepsilon) \geq \varliminf_{\varepsilon\to 0^+} J(y_\varepsilon,u_\varepsilon) \geq J(\widetilde{y},\widetilde{u}),$$

which, together with (3.17), (3.18), (3.23), and (3.4), implies that $(\widetilde{y},\widetilde{u}) = (y^*,u^*)$ and that as $\varepsilon \to 0^+$

$$u_\varepsilon \to u^* \ \text{ strongly in } L^2(Q), \tag{3.24}$$

$$\begin{aligned} y_\varepsilon \to y^* \quad & \text{strongly in } L^{\frac{2n}{n-2}}(Q) \cap C([0,T];H) \cap L^2(0,T;V), \\ & \text{weakly in } Y. \end{aligned} \tag{3.25}$$

In particular,

$$(y_\varepsilon(0),y_\varepsilon(T)) \to (y^*(0),y^*(T)) \ \text{ strongly in } H\times H \ \text{ as } \varepsilon\to 0^+. \tag{3.26}$$

Next, we shall prove that $y_\varepsilon \to y^*$ strongly in Y as $\varepsilon \to 0^+$. To this end, we first claim that

$$f(x,t,y_\varepsilon) \to f(x,t,y^*) \ \text{ strongly in } L^2(Q) \ \text{ as } \varepsilon\to 0^+. \tag{3.27}$$

Notice that

$$|f(x,t,y_\varepsilon) - f(x,t,y^*)| = |b_\varepsilon(x,t)| \cdot |y_\varepsilon - y^*|,$$

where $b_\varepsilon(x,t) = \int_0^1 f_y'(x,t,y^* + \theta(y_\varepsilon - y^*))d\theta$. By assumption $(\mathbf{A_f})$, we have $|b_\varepsilon|^2 \leq C(1+|y^*|+|y_\varepsilon|)^{2(\alpha-1)}$ for some positive constant C independent of ε. Then by the Sobolev embedding theorem and the Hölder inequality, we see that the sequence $\{b_\varepsilon^2\}$ is bounded in $L^{\frac{n}{2}}(Q)$. By (3.25) and by making use of Hölder's inequality, we obtain

$$\int_Q |f(x,t,y_\varepsilon) - f(x,t,y^*)|^2 dxdt \leq \left(\int_Q |b_\varepsilon|^n dxdt\right)^{\frac{2}{n}} \left(\int_Q |y_\varepsilon - y^*|^{\frac{2n}{n-2}} dxdt\right)^{\frac{n-2}{n}} \to 0 \quad \text{as } \varepsilon\to 0^+.$$

Thus we have proved (3.27).

Now we set $g_\varepsilon = (y_\varepsilon)_t - \Delta y_\varepsilon + f(x,t,y_\varepsilon) - \chi_\omega u_\varepsilon$. Then it follows immediately from (3.22) that $g_\varepsilon \to 0$ strongly in $L^2(Q)$ as $\varepsilon \to 0^+$. Set $\varphi_\varepsilon = y_\varepsilon - y^*$ and $F_\varepsilon = \chi_\omega(u_\varepsilon - u^*) + f(x,t,y^*) - f(x,t,y_\varepsilon)$. Then by (3.24) and (3.27), $F_\varepsilon \to 0$ strongly in $L^2(Q)$ as $\varepsilon \to 0^+$. Moreover, it holds that

$$\begin{cases} (\varphi_\varepsilon)_t - \Delta\varphi_\varepsilon = F_\varepsilon & \text{in } Q, \\ \varphi_\varepsilon = 0 & \text{on } \Sigma. \end{cases} \tag{3.28}$$

Since $y_\varepsilon \in \widetilde{Y}$, we have

$$\varphi_\varepsilon^{I,\mathcal{K}}(0) = 0, \ \ \varphi_\varepsilon^{II,\mathcal{K},2}(0) = \varphi_\varepsilon^{II,\mathcal{K},2}(T). \tag{3.29}$$

By (3.21), it follows that

$$\varphi_\varepsilon^{II,\mathcal{K},1}(0) - \varphi_\varepsilon^{II,\mathcal{K},1}(T) \equiv \xi_\varepsilon^{II,\mathcal{K},1} \to 0 \ \text{ strongly in } H \ \text{ as } \varepsilon\to 0^+. \tag{3.30}$$

Then by (3.28), (3.29), and (3.30), we can apply Lemma 3.1 to get

$$\|\varphi_\varepsilon\|_Y^2 \le C(\|F_\varepsilon\|_{L^2(Q)}^2 + |\xi_\varepsilon^{II,\mathcal{K},1}|_2^2) \to 0 \ \text{ as } \varepsilon \to 0^+.$$

This completes the proof of Lemma 3.3. □

LEMMA 3.4. *Let $(y_\varepsilon, u_\varepsilon) \in \widetilde{Y} \times L^2(Q)$ be optimal for problem $(\mathbf{P}^\varepsilon_{\mathcal{K},r})$. Then there exists an $\varepsilon_0 > 0$ such that, for any ε with $0 < \varepsilon \le \varepsilon_0$, there exist a pair $(\mu_\varepsilon, a_\varepsilon^{I,\mathcal{K}}) \in \mathbf{R} \times H^{I,\mathcal{K}}$ and a function $p_\varepsilon \in Y$ satisfying*

$$1 \le \mu_\varepsilon^2 + |a_\varepsilon^{I,\mathcal{K}}|_2^2 \le 2, \tag{3.31}$$

$$\mu_\varepsilon(2u_\varepsilon - u^*) = \chi_\omega p_\varepsilon, \ \ a.e.\ in\ Q, \tag{3.32}$$

$$\begin{cases} (p_\varepsilon)_t + \Delta p_\varepsilon - f_y'(x,t,y_\varepsilon)p_\varepsilon = \mu_\varepsilon\{y_\varepsilon + (y_\varepsilon - y^*)^{\frac{n+2}{n-2}}\} & in\ Q, \\ p_\varepsilon = 0 & on\ \Sigma, \\ p_\varepsilon^{I,\mathcal{K}}(T) = -a_\varepsilon^{I,\mathcal{K}}, \ \ p_\varepsilon^{II,\mathcal{K}}(T) = p_\varepsilon^{II,\mathcal{K}}(0) & in\ \Omega. \end{cases} \tag{3.33}$$

Proof of Lemma 3.4. Write Z for the set $\{z \in Y;\ z^{I,\mathcal{K}}(0) = 0,\ z^{II,\mathcal{K},2}(0) = z^{II,\mathcal{K},2}(T)\}$. Let (z, v) be an arbitrary but fixed pair in the set $Z \times L^2(Q)$. Write $y_{\varepsilon,\lambda} = y_\varepsilon + \lambda z$ and $u_{\varepsilon,\lambda} = u_\varepsilon + \lambda v$, $\lambda > 0$. Then it holds that $y_{\varepsilon,\lambda} \in \widetilde{Y}$ and $u_{\varepsilon,\lambda} \in L^2(Q)$. Moreover, we have

$$y_{\varepsilon,\lambda} \to y_\varepsilon \ \text{ strongly in } Y, \ \ u_{\varepsilon,\lambda} \to u_\varepsilon \ \text{ strongly in } L^2(Q) \ \text{ as } \lambda \to 0^+.$$

By the optimality of $(y_\varepsilon, u_\varepsilon)$ for problem $(\mathbf{P}^\varepsilon_{\mathcal{K},r})$, we see that

$$0 \le \frac{J_\varepsilon(y_{\varepsilon,\lambda}, u_{\varepsilon,\lambda}) - J_\varepsilon(y_\varepsilon, u_\varepsilon)}{\lambda}. \tag{3.34}$$

Notice that, as $\lambda \to 0^+$,

$$\frac{\frac{1}{2\varepsilon}\{(\varepsilon + d_B(y_{\varepsilon,\lambda}^{I,\mathcal{K}}(T)))^2 - (\varepsilon + d_B(y_\varepsilon^{I,\mathcal{K}}(T)))^2\}}{\lambda} \to \frac{d_B(y_\varepsilon^{I,\mathcal{K}}(T)) + \varepsilon}{\varepsilon}\langle a_\varepsilon^{I,\mathcal{K}}, z^{I,\mathcal{K}}(T)\rangle, \tag{3.35}$$

where

$$a_\varepsilon^{I,\mathcal{K}} \in \partial d_B(y_\varepsilon^{I,\mathcal{K}}(T)) = \begin{cases} 0 & \text{if } y_\varepsilon^{I,\mathcal{K}}(T) \in \mathrm{Int}(B^{I,\mathcal{K}}(0,r)), \\ \left\{\dfrac{s y_\varepsilon^{I,\mathcal{K}}(T)}{|y_\varepsilon^{I,\mathcal{K}}(T)|_2}\right\}_{1 \ge s \ge 0} & \text{if } y_\varepsilon^{I,\mathcal{K}}(T) \in \partial B^{I,\mathcal{K}}(0,r), \\ \dfrac{y_\varepsilon^{I,\mathcal{K}}(T)}{|y_\varepsilon^{I,\mathcal{K}}(T)|_2} & \text{if } y_\varepsilon^{I,\mathcal{K}}(T) \notin B^{I,\mathcal{K}}(0,r). \end{cases} \tag{3.36}$$

Here $\partial d_B(\cdot)$ denotes the subdifferential of $d_B(\cdot)$. (See [1] or [2].) Now we claim that the following holds:

$$\frac{f(x,t,y_{\varepsilon,\lambda}) - f(x,t,y_\varepsilon)}{\lambda} \to f_y'(y_\varepsilon)z \ \text{ strongly in } L^2(Q) \ \text{ as } \lambda \to 0^+. \tag{3.37}$$

Indeed, we can write

$$\frac{f(x,t,y_{\varepsilon,\lambda}) - f(x,t,y_\varepsilon)}{\lambda} = b_{\varepsilon,\lambda}(x,t)z,$$

where $b_{\varepsilon,\lambda} = \int_0^1 f_y'(x,t,y_\varepsilon + \theta\lambda z)d\theta$. It is clear that

$$\{(b_{\varepsilon,\lambda} - f_y'(x,t,y_\varepsilon))z\}^2 \to 0 \text{ a.e. in } Q \text{ as } \lambda \to 0^+.$$

By the assumption in $(\mathbf{A_f})$, we have

$$\{(b_{\varepsilon,\lambda} - f_y'(x,t,y_\varepsilon))z\}^2 \le C(1 + (|y_\varepsilon| + |z|)^{2(\alpha-1)})z^2$$

for a certain positive constant C independent of ε and λ. Then by the Sobolev embedding theorem and the Hölder inequality, we see that the function $(1 + (|y_\varepsilon| + |z|)^{2(\alpha-1)})z^2$ is in the space $L^1(Q)$. Then we can use the Lebesgue dominated convergence theorem to get (3.37). Thus, it holds that, as $\lambda \to 0^+$,

$$\begin{aligned} &\frac{1}{2\varepsilon\lambda}\int_Q \{|(y_{\varepsilon,\lambda})_t - \Delta y_{\varepsilon,\lambda} + f(x,t,y_{\varepsilon,\lambda}) - \chi_\omega u_{\varepsilon,\lambda}|^2 \\ &\qquad - |(y_\varepsilon)_t - \Delta y_\varepsilon + f(x,t,y_\varepsilon) - \chi_\omega u_\varepsilon|^2\}dxdt \\ &\to \frac{1}{\varepsilon}\int_Q ((y_\varepsilon)_t - \Delta y_\varepsilon + f(x,t,y_\varepsilon) - \chi_\omega u_\varepsilon)(z_t - \Delta z + f_y'(x,t,y_\varepsilon)z - \chi_\omega v)dxdt. \end{aligned} \tag{3.38}$$

Write μ_ε for the term $\frac{\varepsilon}{d_B(y_\varepsilon^{I,\mathcal{K}}(T))+\varepsilon}$. Then by (3.36), the estimate (3.31) follows immediately.

Set

$$p_\varepsilon = \frac{\mu_\varepsilon}{\varepsilon}\{(y_\varepsilon)_t - \Delta y_\varepsilon + f(x,t,y_\varepsilon) - \chi_\omega u_\varepsilon\} \in L^2(Q).$$

Then by (3.35) and (3.38) and after some simple computation, we can pass to the limit, as $\lambda \to 0^+$, in (3.34) to get

$$\begin{aligned} 0 \le \mu_\varepsilon \Big\{ &\int_Q (2u_\varepsilon - u^*)v dxdt + \int_Q (y_\varepsilon + (y_\varepsilon - y^*)^{\frac{n+2}{n-2}})z dxdt \\ &+ \left\langle \frac{y_\varepsilon^{II,\mathcal{K},1}(T) - y_\varepsilon^{II,\mathcal{K},1}(0)}{\varepsilon}, z^{II,\mathcal{K},1}(T) - z^{II,\mathcal{K},1}(0) \right\rangle \Big\} \\ &+ \langle a_\varepsilon^{I,\mathcal{K}}, z^{I,\mathcal{K}}(T)\rangle \\ &+ \int_Q p_\varepsilon(z_t - \Delta z + f_y'(x,t,y_\varepsilon)z - \chi_\omega v)dxdt. \end{aligned} \tag{3.39}$$

Since the pair (z,v) was arbitrarily taken from the set $Z \times L^2(Q)$, the inequality (3.39) holds for any pair (z,v) in the set $Z \times L^2(Q)$.

By taking $z = 0 \in Z$ in (3.39), we get

$$0 \le \mu_\varepsilon \int_Q (2u_\varepsilon - u^*)v dxdt - \int_Q p_\varepsilon \chi_\omega v dxdt \ \ \forall\, v \in L^2(Q),$$

from which (3.32) follows easily.

By taking $v = 0$ in (3.39), we get

$$\begin{aligned} 0 \le \mu_\varepsilon \Big\{ &\int_Q (y_\varepsilon + (y_\varepsilon - y^*)^{\frac{n+2}{n-2}})z dxdt + \left\langle \frac{y_\varepsilon^{II,\mathcal{K},1}(T) - y_\varepsilon^{II,\mathcal{K},1}(0)}{\varepsilon}, \right. \\ &\left. z^{II,\mathcal{K},1}(T) - z^{II,\mathcal{K},1}(0) \right\rangle \Big\} \\ &+ \langle a_\varepsilon^{I,\mathcal{K}}, z^{I,\mathcal{K}}(T)\rangle + \int_Q p_\varepsilon(z_t - \Delta z + f(x,t,y_\varepsilon)z)dxdt \ \ \forall\, z \in Z. \end{aligned} \tag{3.40}$$

On the other hand, it follows from Lemma 3.3 that $y_\varepsilon \to y^*$ strongly in Y as $\varepsilon \to 0^+$. Thus by (3.1) and (3.2), there exists a positive number ε_0 such that, for each number ε with $0 < \varepsilon \le \varepsilon_0$,

$$\|f_y'(x,t,y_\varepsilon)\|_{L^\infty(0,T;L^n(\Omega))} \le M. \tag{3.41}$$

Then, for each number ε with $0 < \varepsilon \le \varepsilon_0$, the following equation has a unique solution $q_\varepsilon \in Y$ (see (3.3) and notice that $H^{II,\mathcal{K}_0} = H^{II,\mathcal{K},2}$):

$$\begin{cases} (q_\varepsilon)_t + \Delta q_\varepsilon - f_y'(x,t,y_\varepsilon)q_\varepsilon = -\mu_\varepsilon(y_\varepsilon + (y_\varepsilon - y^*)^{\frac{n+2}{n-2}}) & \text{in } Q, \\ q_\varepsilon = 0 & \text{on } \Sigma, \\ q_\varepsilon^{I,\mathcal{K}}(T) = a_\varepsilon^{I,\mathcal{K}},\ q_\varepsilon^{II,\mathcal{K},1}(T) = \mu_\varepsilon(\frac{y_\varepsilon^{II,\mathcal{K},1}(T) - y_\varepsilon^{II,\mathcal{K},1}(0)}{\varepsilon}) & \text{in } \Omega, \\ q_\varepsilon^{II,\mathcal{K},2}(T) = q_\varepsilon^{II,\mathcal{K},2}(0) & \text{in } \Omega. \end{cases} \tag{3.42}$$

Multiplying the first equation of (3.42) by $z \in Z$ and integrating over Q, noticing that, for each z in the set Z, $z^{I,\mathcal{K}}(0) = 0,\ z^{II,\mathcal{K},2}(0) = z^{II,\mathcal{K},2}(T)$, we obtain that, for each ε with $0 < \varepsilon \le \varepsilon_0$,

$$\begin{aligned} &\mu_\varepsilon \int_Q (y_\varepsilon + (y_\varepsilon - y^*)^{\frac{n+2}{n-2}}) z dx dt + \left\langle \mu_\varepsilon \frac{y_\varepsilon^{II,\mathcal{K},1}(T) - y_\varepsilon^{II,\mathcal{K},1}(0)}{\varepsilon}, z^{II,\mathcal{K},1}(T) \right\rangle \\ &\quad + \langle z^{I,\mathcal{K}}(T), a_\varepsilon^{I,\mathcal{K}} \rangle \\ &= \langle z^{II,\mathcal{K},1}(0), q_\varepsilon^{II,\mathcal{K},1}(0) \rangle + \int_Q q_\varepsilon (z_t - \Delta z + f_y'(x,t,y_\varepsilon) z) dx dt, \end{aligned}$$

which together with (3.40) implies that, for each number ε with $0 < \varepsilon \le \varepsilon_0$,

$$\begin{aligned} 0 \le &\int_Q (p_\varepsilon + q_\varepsilon)(z_t - \Delta z + f_y'(x,t,y_\varepsilon) z) dx dt \\ &+ \left\langle z^{II,\mathcal{K},1}(0), q_\varepsilon^{II,\mathcal{K},1}(0) - \mu_\varepsilon \left(\frac{y_\varepsilon^{II,\mathcal{K},1}(T) - y_\varepsilon^{II,\mathcal{K},1}(0)}{\varepsilon} \right) \right\rangle \quad \forall\, z \in Z. \end{aligned} \tag{3.43}$$

However, for each function g in the space $L^2(Q)$ and for each number ε with $0 < \varepsilon \le \varepsilon_0$, the following equation has a unique solution $z_g \in Y$ (see (3.3)):

$$\begin{cases} z_t - \Delta z + f_y'(x,t,y_\varepsilon) z = g & \text{in } Q, \\ z = 0 & \text{on } \Sigma, \\ z^{I,\mathcal{K}}(0) = 0,\ z^{II,\mathcal{K},1}(0) = 0,\ z^{II,\mathcal{K},2}(0) = z^{II,\mathcal{K},2}(T) & \text{in } \Omega. \end{cases}$$

It is clear that this solution z_g is in the set Z. Thus, we can take $z = z_g$ in (3.43) to get $0 \le \int_Q (p_\varepsilon + q_\varepsilon) g dx dt$. Since the function g is arbitrary in the space $L^2(Q)$, it holds that, for each number ε with $0 < \varepsilon \le \varepsilon_0$,

$$p_\varepsilon = -q_\varepsilon \ \text{ over } Q. \tag{3.44}$$

Now the inequality (3.43) is simplified as follows:

$$0 \le \left\langle z^{II,\mathcal{K},1}(0), q_\varepsilon^{II,\mathcal{K},1}(0) - \mu_\varepsilon \left(\frac{y_\varepsilon^{II,\mathcal{K},1}(T) - y_\varepsilon^{II,\mathcal{K},1}(0)}{\varepsilon} \right) \right\rangle \quad \forall\, z \in Z. \tag{3.45}$$

Recall that $Z \equiv \{z \in Y;\ z^{I,\mathcal{K}}(0) = 0,\ z^{II,\mathcal{K},2}(0) = z^{II,\mathcal{K},2}(T)\}$. Thus, we can take $z^{II,\mathcal{K},1}(0)$ to be any element in $H^{II,\mathcal{K},1}$ in (3.45), which implies that, for each number ε with $0 < \varepsilon \le \varepsilon_0$,

$$q_\varepsilon^{II,\mathcal{K},1}(0) = \mu_\varepsilon \left(\frac{y_\varepsilon^{II,\mathcal{K},1}(T) - y_\varepsilon^{II,\mathcal{K},1}(0)}{\varepsilon} \right) = q_\varepsilon^{II,\mathcal{K},1}(T).$$

From this and by (3.42) and (3.44), we can easily obtain (3.33). This completes the proof of Lemma 3.4. □

3.3. The proof of Theorem 1.1. Now we are ready to give the proof of Theorem 1.1.

Proof of Theorem 1.1. By (3.31), we can assume, without loss of generality, that as $\varepsilon \to 0^+$

$$\begin{array}{ll} \mu_\varepsilon \to \mu_0 & \text{in } \mathbf{R}, \\ a_\varepsilon^{I,\mathcal{K}} \to a_0^{I,\mathcal{K}} & \text{strongly in } H^{I,\mathcal{K}}. \end{array} \tag{3.46}$$

We next claim that for any ε with $0 < \varepsilon \le \varepsilon_0$, where ε_0 was given in Lemma 3.4, it holds that

$$\|p_\varepsilon\|_Y \le C. \tag{3.47}$$

Here and in what follows, C stands for a positive constant independent of ε, which may be different in different contexts. It should be mentioned that we cannot directly apply Theorem 2.1 to get (3.47). (See Remark 3.2 following the proof of Theorem 1.1.)

By (3.41), the sequence $\{f_y'(x,t,y_\varepsilon)\}_{0<\varepsilon\le\varepsilon_0}$ is bounded in the space $L^\infty(0,T;L^n(\Omega))$. Then, by the observability inequality for linear parabolic equations based on the Carleman inequality for linear parabolic equations (see [14]), we see that, for any ε with $0 < \varepsilon \le \varepsilon_0$,

$$|p_\varepsilon(0)|_2^2 \le C\left(\|\mu_\varepsilon(y_\varepsilon + (y_\varepsilon - y^*)^{\frac{n+2}{n-2}})\|_{L^2(Q)}^2 + \int_0^T\int_\omega p_\varepsilon^2 dxdt\right).$$

Since the sequence $\{y_\varepsilon\}$ is bounded in the space Y, which is continuously embedded into the space $L^{\frac{2(n+2)}{n-2}}(Q)$, it follows that the sequence $\{(y_\varepsilon - y^*)^{\frac{n+2}{n-2}}\}$ is bounded in the space $L^2(Q)$. Hence, it holds that

$$|p_\varepsilon(0)|_2^2 \le C\left(1 + \int_0^T\int_\omega p_\varepsilon^2 dxdt\right).$$

Notice that the sequence $\{u_\varepsilon\}$ is bounded in the space $L^2(Q)$. Then, by (3.31) and (3.32), we get

$$\int_0^T\int_\omega p_\varepsilon^2 dxdt \le C.$$

Thus it holds that $|p_\varepsilon(0)|_2^2 \le C$ for any ε with $0 < \varepsilon \le \varepsilon_0$, from which we see that, for any ε with $0 < \varepsilon \le \varepsilon_0$, the following holds:

$$|p_\varepsilon(T)|_2^2 = |p_\varepsilon^{I,\mathcal{K}}(T)|_2^2 + |p_\varepsilon^{II,\mathcal{K}}(T)|_2^2 = |a_\varepsilon^{I,\mathcal{K}}|_2^2 + |p_\varepsilon^{II,\mathcal{K}}(0)|_2^2 \le C. \tag{3.48}$$

On the other hand, by the Hölder inequality and the Sobolev embedding theorem, we have that, for any ε with $0 < \varepsilon \le \varepsilon_0$,

$$\begin{aligned} \int_0^t\int_\Omega |f_y'(x,\tau,y_\varepsilon)p_\varepsilon^2| dxd\tau &\le \int_0^t\left\{\left(\int_\Omega |f_y'(x,\tau,y_\varepsilon)|^n dx\right)^{\frac{1}{n}}\left(\int_\Omega p_\varepsilon^{\frac{2n}{n-2}}dx\right)^{\frac{n-2}{2n}}\left(\int_\Omega p_\varepsilon^2 dx\right)^{\frac{1}{2}}\right\} d\tau \\ &\le \|f_y'(x,\tau,y_\varepsilon)\|_{L^\infty(0,T;L^n(\Omega))}\int_0^t \|p_\varepsilon(\tau)\||p_\varepsilon(\tau)|_2 d\tau \\ &\le C\int_0^t \|p_\varepsilon(\tau)\||p_\varepsilon(\tau)|_2 d\tau. \end{aligned} \tag{3.49}$$

Then multiplying the first equation of (3.33) by p_ε, integrating over $\Omega \times (0,t)$, and making use of (3.48), (3.49), and Gronwall's inequality, we get that, for any ε with $0 < \varepsilon \le \varepsilon_0$,

$$\|p_\varepsilon\|^2_{C([0,T];H)} + \|p_\varepsilon\|^2_{L^2(0,T;V)} \le C. \tag{3.50}$$

By (3.50) and applying the Hölder inequality and the Sobolev embedding theorem again, we see that, for any ε with $0 < \varepsilon \le \varepsilon_0$, the following holds:

$$\begin{aligned}\int_0^T \int_\Omega f'_y(x,t,y_\varepsilon)^2 p_\varepsilon^2 dxdt &\le \int_0^T \left\{ \left(\int_\Omega |f'_y(x,t,y_\varepsilon)|^{2\cdot\frac{n}{2}} dx \right)^{\frac{2}{n}} \right. \\ &\quad \left. \times \left(\int_\Omega |p_\varepsilon|^{\frac{2n}{n-2}} dx \right)^{\frac{n-2}{2n}} \left(\int_\Omega |p_\varepsilon|^{\frac{2n}{n-2}} dx \right)^{\frac{n-2}{2n}} \right\} \\ &\le \|f'_y(x,t,y_\varepsilon)\|^2_{L^\infty(0,T;L^n(\Omega))} \int_0^T \|p_\varepsilon(t)\|^2 dt \le C.\end{aligned}$$

Now we can use Lemma 3.1 to get that, for any ε with $0 < \varepsilon \le \varepsilon_0$,

$$\|p_\varepsilon\|^2_Y \le C(\|f'_y(x,t,y_\varepsilon)p_\varepsilon\|^2_{L^2(Q)} + \|\mu_\varepsilon(y_\varepsilon + (y_\varepsilon - y^*)^{\frac{n+2}{n-2}})\|^2_{L^2(Q)} + |a_\varepsilon^{I,\mathcal{K}}|_2^2) \le C.$$

This proves (3.47).

Next, by (3.47) and by making use of the Aubin compactness theorem and the Ascoli–Arzela theorem, we can extract a generalized subsequence of $\{\varepsilon\}_{0<\varepsilon\le\varepsilon_0}$, which, for simplicity of notation, is still denoted in the same way, such that as $\varepsilon \to 0^+$

$$\begin{aligned} &p_\varepsilon \to p \quad && \text{weakly in } Y, \\ &&& \text{a.e. in } Q, \\ &&& \text{strongly in } L^2(0,T;V) \cap C([0,T];H). \end{aligned} \tag{3.51}$$

By (3.51) and the assumption in $(\mathbf{A_f})$, and by making use of Lemma 3.3, one can easily get that

$$f'_y(x,t,y_\varepsilon)p_\varepsilon \to f'_y(x,t,y^*)p \ \text{ weakly in } L^2(Q) \ \text{ as } \varepsilon \to 0^+. \tag{3.52}$$

Then by (3.46), (3.51), and (3.52), we can pass to the limit, as $\varepsilon \to 0^+$, in (3.33) to get

$$\begin{cases} p_t + \Delta p - f'_y(x,t,y^*)p = \mu_0 y^* & \text{in } Q, \\ p = 0 & \text{on } \Sigma, \\ p^{I,\mathcal{K}}(T) = -a_0^{I,\mathcal{K}}, \ \ p^{II,\mathcal{K}}(T) = p^{II,\mathcal{K}}(0) & \text{in } \Omega. \end{cases} \tag{3.53}$$

By Lemma 3.3 and by (3.51), we can pass to the limit, as $\varepsilon \to 0^+$, in (3.32) to get

$$\chi_\omega p = \mu_0 u^* \ \text{ a.e. in } Q. \tag{3.54}$$

Next, we prove that $\mu_0 \neq 0$. Notice that the number μ_0 appears only in the case that $\mathcal{K} > 0$. (In the case $\mathcal{K} = 0$, namely, in the periodic case, $H = H^{II,\mathcal{K}}$ and both $a^{I,\mathcal{K}}$ and $B^{I,\mathcal{K}}(0,r)$ do not appear. Hence μ_0 does not appear.) Now we seek a contradiction, supposing that $\mu_0 = 0$. Then, by (3.31), we get $|a_\varepsilon^{I,\mathcal{K}}|_2^2 \ge \delta > 0$ for some $\delta > 0$ independent of ε. Since $H^{I,\mathcal{K}}$ is a finite dimensional space, $a_0^{I,\mathcal{K}} \neq 0$. On the other hand, it follows from (3.54) that

$$p = 0 \ \text{ a.e. in } \omega \times (0,T).$$

Then, by the unique continuation of solutions to the linear parabolic equations based on the Carleman inequality for the linear parabolic equations (see [14]), we get

$$p = 0 \text{ a.e. in } \Omega \times (0,T),$$

which leads us to a contradiction since $p^{I,\mathcal{K}}(T) = -a_0^{I,\mathcal{K}} \neq 0$. Hence we have proved $\mu_0 \neq 0$. The proof of Theorem 1.1 is completed. □

Remark 3.2. The boundedness of $\{p_\varepsilon\}_{\varepsilon_0 \geq \varepsilon > 0}$ in Y cannot be obtained by just making use of Theorem 2.1. In the problem $(\mathbf{P}_{\mathcal{K},r})$, the nonnegative integer $\mathcal{K}$ is arbitrarily given. However, the integer $\mathcal{K}$ in Theorem 2.1 must be sufficiently large.

Remark 3.3. By a very similar method, we can obtain the Pontryagin maximum principle for optimal controls for the problem $(\mathbf{P}_{\mathcal{K},r})$ even if the cost functional $J(y,u)$ has the following more general form:

$$J(y,u) = \int_0^T [g(t,y) + h(u)]dt,$$

where $g : [0,T] \times H \to \mathbf{R}^+$ is measurable in t and locally Lipschitzian in the second variable, $g(\cdot,0) \in L^\infty(0,T)$, and $h : H \to \overline{\mathbf{R}} \equiv (-\infty, +\infty]$ is a lower semicontinuous and convex functional satisfying $h(u) \geq C_1|u|_2^2 + C_2$. Here $C_1 > 0$ and $C_2 \in \mathbf{R}$. In this case, the only modification for the proof that we need is to use suitable smooth functionals g_ε and h_ε to approach g and h, respectively. (The reader is referred to [1], [13] for the construction of such smooth approximations.)

4. The existence of optimal controls for the problem $(\mathbf{P}_{\mathcal{K},r})$. In this section, we shall prove Theorem 1.2; namely, we prove the existence of optimal solutions for the problem $(\mathbf{P}_{\mathcal{K},r})$.

Proof of Theorem 1.2. Let $M > 0$ be given. By Theorem 2.2, there exists an integer $\mathcal{K} \equiv \mathcal{K}(M) \geq 0$ such that, for each $a^{I,\mathcal{K}} \in H^{I,\mathcal{K}}$ with $(1+\lambda_{\mathcal{K}})|a^{I,\mathcal{K}}|_2^2 \leq M^2$, there exists a control $u \in L^2(Q)$ with $\|u\|_{L^2(Q)}^2 \leq M^2$ such that the equation

$$\begin{cases} y_t - \Delta y + f(x,t,y) = \chi_\omega u & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y^{I,\mathcal{K}}(0) = a^{I,\mathcal{K}},\ y^{II,\mathcal{K}}(0) = y^{II,\mathcal{K}}(T) & \text{in } \Omega \end{cases} \tag{4.1}$$

has a solution $y \in Y$ with the estimate:

$$|y(T)|_2^2 \leq C\|y\|_Y^2 \leq C(M),$$

which implies $|y^{I,\mathcal{K}}(T)|_2^2 \leq C(M)$. Here $C(M)$ denotes a positive constant depending on M.

Now we fix an integer $\mathcal{K}$ as above, an element $a^{I,\mathcal{K}} \in H^{I,\mathcal{K}}$ with $(1+\lambda_{\mathcal{K}})|a^{I,\mathcal{K}}|_2^2 \leq M^2$, and fix a number r such that $r \geq C(M)$. Then problem $(\mathbf{P}_{\mathcal{K},r})$ has a nonempty admissible set; namely, there exists at least one pair $(y,u) \in Y \times L^2(Q)$ (4.1) and $y^{I,\mathcal{K}}(T) \in B^{I,\mathcal{K}}(0,r)$.

Set $d = \text{Inf}\{J(y,u);\ (y,u) \in Y \times L^2(Q)$ satisfying (1.1) and (1.2)$\}$, and let $\{(y_m,u_m)\} \subset Y \times L^2(Q)$ be a minimization sequence of problem $(\mathbf{P}_{\mathcal{K},r})$, namely,

$$d \leq \frac{1}{2}\int_Q (y_m^2 + u_m^2)dxdt \leq d + \frac{1}{m},\ m = 1,2,\ldots, \tag{4.2}$$

and (y_m,u_m) satisfies (4.1) and $y_m^{I,\mathcal{K}}(T) \in B^{I,\mathcal{K}}(0,r)$. By (4.2), we see that the sequence $\{u_m\}$ is bounded in the space $L^2(Q)$.

Multiplying the first equation of (4.1), where (y,u) is replaced by (y_m,u_m), by y_m and then integrating over $\Omega\times(0,T)$, by making use of (i) of $(\mathbf{H_f})$ and the property that $y_m^{II,\mathcal{K}}(0)=y_m^{II,\mathcal{K}}(T)$, we can easily derive

$$|y_m(0)|_2^2\le C(\|u_m\|_{L^2(Q)}^2+|a^{I,\mathcal{K}}|_2^2).$$

Here and in what follows, C denotes a positive constant independent of m, which may be different in different contexts. From the above, using (4.1) with (y,u) being replaced by (y_m,u_m) and (i) of $(\mathbf{H_f})$ again, we obtain

$$\|y_m\|_{V_2(Q)}^2\le C(\|u_m\|_{L^2(Q)}^2+|a^{I,\mathcal{K}}|_2^2). \tag{4.3}$$

Then by (ii) of $(\mathbf{H_f})$ and by making use of the Hölder inequality and the Sobolev embedding theorem $(V_2(Q)\hookrightarrow L^{\frac{2(n+2)}{n}}(Q))$, we get

$$\begin{aligned}\int_Q f(x,t,y_m)^2dxdt&\le C\int_Q|y_m|(1+|y_m|^\beta)^2dxdt\\&\le C\left(\|y_m\|_{L^2(Q)}^2+\|y_m\|_{L^{\frac{2(n+2)}{n}}(Q)}^{\frac{2(n+2)}{n}}\right)\\&\le C\left(\|y_m\|_{L^2(Q)}^2+\|y_m\|_{V_2(Q)}^{\frac{2(n+2)}{n}}\right)\le C.\end{aligned} \tag{4.4}$$

Now by Lemma 3.1, where $\xi^{II,\mathcal{K},1}=0$, we conclude that

$$\|y_m\|_Y^2\le C(\|u_m\|_{L^2(Q)}^2+\|f(x,t,y_m)\|_{L^2(Q)}^2+|a^{I,\mathcal{K}}|_2^2)\le C.$$

Hence, we can extract subsequences of $\{y_m\}$ and $\{u_m\}$, which, for simplicity of notation, are still denoted by themselves, such that as $m\to\infty$

$$\begin{aligned}&y_m\to\widetilde{y}\quad&&\text{weakly in } Y,\\&&&\text{strongly in } L^2(0,T;V)\cap C([0,T];H),\\&&&\text{a.e. in } Q,\end{aligned} \tag{4.5}$$

$$u_m\to\widetilde{u}\ \text{ weakly in } L^2(Q). \tag{4.6}$$

Now by $(\mathbf{H_f})$, we obtain

$$f(x,t,y_m)\to f(x,t,\widetilde{y})\quad\text{a.e. in } Q\ \text{ as } m\to\infty,$$

which together with (4.4) implies

$$f(x,t,y_m)\to f(x,t,\widetilde{y})\ \text{ weakly in } L^2(Q)\ \text{ as } m\to\infty. \tag{4.7}$$

Then by (4.5), (4.6), and (4.7), we can pass to the limit, as $m\to\infty$, in (4.1) with (y,u) being replaced by (y_m,u_m) to get that $(\widetilde{y},\widetilde{u})$ satisfies (4.1). Moreover, it follows from (4.5) that

$$y_m^{I,\mathcal{K}}(T)\to\widetilde{y}^{I,\mathcal{K}}(T)\ \text{ strongly in } H \text{ as } m\to\infty.$$

Hence,

$$|\widetilde{y}^{I,\mathcal{K}}(T)|_2^2=\lim_{m\to\infty}|y_m^{I,\mathcal{K}}(T)|_2^2\le r^2.$$

Therefore $(\widetilde{y},\widetilde{u})$ is admissible for the problem $(\mathbf{P}_{\mathcal{K},r})$.

Finally, by passing to the limit, as $m\to\infty$, in (4.2), we get $d=\frac{1}{2}\int_Q(\widetilde{y}^2+\widetilde{u}^2)dxdt$. Hence, $(\widetilde{y},\widetilde{u})$ is an optimal solution for the problem $(\mathbf{P}_{\mathcal{K},r})$. This completes the proof of Theorem 1.2. □

REFERENCES

[1] V. Barbu, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic, New York, 1993.

[2] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, D. Reidel, Dordrecht, 1986.

[3] Y.-Z. Chen, *Parabolic Partial Differential Equations of Second Order*, Series in Graduate Textbooks in Mathematics, Beijing University Press, Beijing, 2003.

[4] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.

[5] H. Gao and N. H. Pavel, *Optimal control problems for a class of semilinear multisolution elliptic equations*, J. Optim. Theory Appl., 118 (2003), pp. 353–380.

[6] L. Lei, *Identification of parameters through the approximate periodic solutions of a linear parabolic system*, J. Optim. Theory Appl., to appear.

[7] L. Lei, *Approximate periodic solutions of a nonlinear parabolic system and an identification problem*, J. Math. Anal. Appl., 328 (2007), pp. 1396–1416.

[8] X. Li and J. Yong, *Necessary conditions for optimal control of distributed parameter systems*, SIAM J. Control Optim., 29 (1991), pp. 895–908.

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

[10] J. L. Lions, *Some Methods in the Mathematical Analysis of System and Their Control*, Science Press, Beijing, China, Gordon and Breach, New York, 1981.

[11] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Valued Problems and Application*, Springer-Verlag, Berlin, Heidelberg, New York, 1971.

[12] G. Wang, *Optimal controls of 3-dimensional Navier–Stokes equations with state constraints*, SIAM J. Control Optim., 41 (2002), pp. 583–606.

[13] G. Wang and L. Wang, *State-constrained optimal control governed by non-well-posed parabolic differential equations*, SIAM J. Control Optim., 40 (2002), pp. 1517–1539.

[14] G. Wang and L. Wang, *The Carleman inequality and its application to periodic optimal control governed by parabolic system*, J. Optim. Theory Appl., 118 (2003), pp. 429–461.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1779–1801

THE DYNAMIC PROGRAMMING EQUATION FOR THE PROBLEM OF OPTIMAL INVESTMENT UNDER CAPITAL GAINS TAXES*

IMEN BEN TAHAR†, H. METE SONER‡, AND NIZAR TOUZI§

Abstract. This paper considers an extension of the Merton optimal investment problem to the case where the risky asset is subject to transaction costs and capital gains taxes. We derive the dynamic programming equation in the sense of constrained viscosity solutions. We next introduce a family of functions $(V_\varepsilon)_{\varepsilon>0}$, which converges to our value function uniformly on compact subsets, and which is characterized as the unique constrained viscosity solution of an approximation of our dynamic programming equation. In particular, this result justifies the numerical results reported in the accompanying paper [I. Ben Tahar, H. M. Soner, and N. Touzi (2005), *Modeling Continuous-Time Financial Markets with Capital Gains Taxes*, preprint, http://www.cmap.polytechnique.fr/∼touzi/bst06.pdf].

Key words. optimal consumption and investment in continuous time, transaction costs, capital gains taxes, viscosity solutions

AMS subject classifications. 91B28, 49J20, 35D99

DOI. 10.1137/050646044

1. Introduction.

The problem of optimal investment and consumption in financial markets has been introduced by Merton [20, 21]. The explicit solution derived in these papers is widely used among fund managers in practical financial markets. Moreover, this problem became very quickly one of the classical examples of application of the verification theorem in stochastic control theory. Indeed, by direct financial considerations, it is easily seen that the value function of the problem satisfies some homogeneity property, which completely determines its dependence on the wealth state variable. Plugging this information into the corresponding dynamic programming equation (DPE) leads to an ordinary differential equation (ODE) which can be solved explicitly, thus providing a candidate smooth solution to the DPE.

In this paper, we consider the extension of the Merton problem to the case where the risky asset is subject to capital gains taxes. For technical reasons, we also assume that the risky asset is subject to proportional transaction costs. This problem is formulated in the accompanying paper [5]. In contrast with the Merton frictionless model, no explicit solution is available in this context. The main result of [5] is the derivation of an explicit first order expansion of the value function for small tax and interest rate parameters. The numerical results reported in [5] show that the relative error induced by this approximation is of the order of 4%. These numerical results are obtained by comparing the explicit first order expansion to the finite differences approximation of the solution of the corresponding DPE.

The literature on the optimal investment problem under capital gains taxes is not very expanded and is mainly developed in discrete-time binomial models; see

*Received by the editors November 25, 2005; accepted for publication (in revised form) March 13, 2007; published electronically November 14, 2007.
http://www.siam.org/journals/sicon/46-5/64604.html

†CEREMADE, Université Paris Dauphine, Paris, France (bentahar@ceremade.dauphine.fr).

‡Sabanci University, Istanbul, Turkey (msoner@ku.edu.tr). Member of the Turkish Academy of Sciences. The work of this author was partly supported by the Turkish Academy of Sciences and by the Turkish Scientific and Technological Research Institute, TÜBITAK.

§Centre de Mathématiques Appliquées, Ecole Polytechnique Paris, Paris, France (touzi@cmap.polytechnique.fr), and Tanaka Business School, Imperial College London, London, UK (n.touzi@imperial.ac.uk).

[7, 10, 11, 12, 17, 18, 15, 19].

The main purpose of this paper is to justify the approximation of the value function by means of the finite differences scheme applied to the corresponding DPE. Since our optimal control problem is singular, the DPE takes the form of a variational inequality:

$$\min\left\{-\mathcal{L}v,\ \mathbf{g^b}\cdot Dv,\ \mathbf{g^s}\cdot Dv\right\}\ =\ 0 \text{ on } \bar{\mathcal{S}},\qquad v=0 \text{ on } \partial^z\mathcal{S},$$

where $\mathcal{L}$ is a second order differential operator defined in (2.13) $\mathbf{g^b}, \mathbf{g^s}$ are two vector fields defined in (2.15) corresponding to the purchase and sale decisions, $\mathcal{S}$ is the state space, and $\partial^z\mathcal{S}$ is part of the boundary of $\mathcal{S}$. The main difficulty comes from the fact that the vector field $\mathbf{g^s}$ is not locally Lipschitz. Then the standard techniques to prove a uniqueness result for the above partial differential equation (PDE) fail. We then introduce a convenient locally Lipschitz approximation $\mathbf{g^s_\varepsilon}$ of $\mathbf{g^s}$, and we consider the approximating PDE

$$\min\left\{-\mathcal{L}v,\ \mathbf{g^b}\cdot Dv,\ \mathbf{g^s_\varepsilon}\cdot Dv\right\}\ =\ 0 \text{ on } \bar{\mathcal{S}},\qquad v=0 \text{ on } \partial^z\mathcal{S}.$$

The main result of this paper states that the above approximating PDE has a unique continuous viscosity solution V_ε which converges uniformly on compact subsets towards the value function V of our optimal investment problem under capital gains taxes. Applying the general results of Barles and Souganidis [4], we see that this justifies the convergence of the numerical scheme implemented in the accompanying paper [5] towards this unique solution of the approximating PDE.

The paper is organized as follows. Section 2 provides a quick review of the problem of optimal investment under capital gains taxes. The main approximation result is stated in section 3. In section 4, we prove a comparison result for the approximating PDE, which implies the required uniqueness claim. In section 5 we prove the existence of a solution of the approximating PDE by introducing a family of control problems obtained by modifying conveniently our original problem. Finally, section 6 reports the proof of convergence of V_ε towards V uniformly on compact subsets.

Notation. For a domain $\mathbf{D}$ in $\mathbb{R}^n$, we denote by $\mathrm{USC}(\mathbf{D})$ (resp., $\mathrm{LSC}(\mathbf{D})$) the collection of all upper semicontinuous (resp., lower semicontinuous) functions from $\mathbf{D}$ to $\mathbb{R}$. The set of continuous functions from $\mathbf{D}$ to $\mathbb{R}$ is denoted by $\mathrm{C}^0(\mathbf{D}) := \mathrm{USC}(\mathbf{D}) \cap \mathrm{LSC}(\mathbf{D})$. For a parameter $\delta > 0$, we say that a function $f : \mathbf{D} \longrightarrow \mathbb{R}$ has δ-polynomial growth if

$$\sup_{x\in\mathbf{D}} \frac{|f(x)|}{1+|x|^\delta} < \infty.$$

We finally denote $\mathrm{USC}_\delta(\mathbf{D}) := \{f \in \mathrm{USC}(\mathbf{D}) \ :\ f \text{ has } \delta\text{-polynomial growth}\}$. The sets $\mathrm{LSC}_\delta(\mathbf{D})$ and $\mathrm{C}^0_\delta(\mathbf{D})$ are defined similarly.

2. Optimal investment under capital gains taxes.

2.1. Problem formulation. In this section, we quickly review the formulation of the problem of optimal investment under capital gains taxes. We refer the interested reader to the accompanying paper [5] for more details. The financial market consists of a tax-free bank account with constant interest rate $r > 0$ and a risky asset subject to proportional transaction costs and to capital gains taxes. The price process of the risky asset evolves according to the Black–Scholes model:

$$dP_t\ =\ P_t\left(\rho dt + \sigma dW_t\right), \qquad t \geq 0, \tag{2.1}$$

where $\rho > 0$ is a constant instantaneous return of the asset, and $\sigma > 0$ is a constant volatility parameter. The process $W = \{W_t, t \geq 0\}$ is a standard Brownian motion with values in $\mathbb{R}^1$ defined on an underlying complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We shall denote by $\mathbb{F}$ the $\mathbb{P}$-completion of the natural filtration of the Brownian motion.

For technical reasons (see section 4), we assume that the risky asset is also subject to proportional transaction costs defined by the coefficients λ, $\mu \in [0, 1)$, so that the bid and ask prices at time t of the risky asset are given by $(1 - \mu)P_t$ and $(1 + \lambda)P_t$.

A control process is a triple of $\mathbb{F}$-adapted processes $\nu = (C, L, M)$, where

$$C \geq 0 \text{ and } \int_0^T C_t dt < \infty \quad \mathbb{P}\text{-a.s. for all } T > 0\,, \tag{2.2}$$

L and M are nondecreasing right-continuous, $L_{0-} = M_{0-} = 0$, and the jumps of M satisfy

$$\Delta M_t \leq 1 \text{ for } t \geq 0 \quad \mathbb{P}\text{-a.s.} \tag{2.3}$$

Here C_t is the consumption rate at time t, $dL_t \geq 0$ is the amount invested between times t and $t + dt$ to purchase risky assets, and $dM_t \geq 0$ is the *proportion* of risky assets in the portfolio which is sold between times t and $t + dt$. Then, the amount of wealth $Y = \{Y_t, t \geq 0\}$ on the risky asset account is defined by the dynamics

$$dY_t = Y_t \frac{dP_t}{P_t} + dL_t - Y_{t-} dM_t\,, \quad t \geq 0. \tag{2.4}$$

Since $\Delta M_t \leq 1$, the no short-sales constraint $Y \geq 0$ holds. Capital gains are taxed only when the investor sells the risky asset. The amount of capital gains (or losses) is evaluated by comparing the actual price P_t to a tax basis B_t specified by the taxation code. In our framework the tax basis is defined as the weighted average of past purchase prices,

$$B_t := \frac{K_t}{Y_t} P_t \text{ if } Y_t > 0 \text{ and } B_t := P_t \text{ otherwise}, \qquad t \geq 0\,,$$

where

$$dK_t = dL_t - K_{t-} dM_t\,, \qquad t \geq 0. \tag{2.5}$$

The natural initial condition for the process K is zero, as initially there are no prior stocks bought. However, the method of dynamic programming always forces us to consider all possible initial data. Hence we consider the K-equation with general initial data $K_0 = k$. Also a more detailed derivation of this tax basis and its place in actual tax codes is given in subsection 2.2 of the accompanying paper [5].

Finally, we consider a linear taxation rule, with constant tax rate parameter $\alpha \in [0, 1]$, so that the after-tax and after-transaction costs induced by selling the amount $Y_{t-} dM_t$ between times t and $t + dt$ are given by

$$(1 - \mu) Y_{t-} dM_t - \alpha(1 - \mu) \left[Y_{t-} dM_t - \frac{Y_{t-} dM_t}{P_t} B_{t-} \right] = (1 - \mu) \left[(1 - \alpha) Y_{t-} + \alpha K_{t-} \right] dM_t.$$

This justifies the following dynamics for the nonrisky asset component of the wealth process:

$$dX_t = (rX_t - C_t)dt - (1 + \lambda)dL_t + (1 - \mu) \left[(1 - \alpha) Y_{t-} + \alpha K_{t-} \right] dM_t, \quad t \geq 0. \tag{2.6}$$

We denote by $\mathcal{A}$ the set of all control processes and by $S = (X, Y, K)$ the corresponding state process defined by (2.4), (2.5), (2.6). A control process ν is said to be admissible if the no bankruptcy condition

$$Z_t \;:=\; X_t + (1-\mu)\left[(1-\alpha)Y_t + \alpha K_t\right] \;\geq\; 0, \quad t \geq 0, \quad \mathbb{P}\text{-a.s.} \tag{2.7}$$

holds. Here Z_t is the after-tax and after-transaction costs liquidation value of the portfolio at time t. Given an initial condition $S_{0-} = s$, we shall denote by $\mathcal{A}(s)$ the collection of all admissible controls.

The problem of optimal consumption and investment under capital gains taxes is defined by the value function

$$V(s) \;:=\; \sup_{\nu \in \mathcal{A}(s)} \mathbb{E}\left[\int_0^\infty e^{-\beta t} U(C_t)dt\right], \quad \text{where } U(x) \;:=\; \frac{x^p}{p}, \quad x \geq 0, \tag{2.8}$$

and $\beta > 0$, $p \in (0,1)$ are two given constant parameters.

Throughout this paper, we assume that the coefficients of the model satisfy the condition

$$\frac{\beta}{p} - r - \frac{(\delta - r)^2}{2(1-p)\sigma^2} > 0, \tag{2.9}$$

which ensures that the value function of the Merton optimal consumption-investment problem (the case $\lambda = \mu = \alpha = 0$) is finite. In particular, the value function V is finite under condition (2.9).

2.2. The DPE. For an admissible control $\nu \in \mathcal{A}(s)$, the induced state process $S^\nu = (X^\nu, Y^\nu, K^\nu)$ defined by (2.4), (2.5), (2.6) together with some initial data $S_0^\nu = s$ is valued in the state space

$$\bar{\mathcal{S}} := \left\{(x,y,k) \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+ \;:\; z = x + (1-\mu)[(1-\alpha)y + \alpha k] \geq 0\right\}. \tag{2.10}$$

We denote by $\mathcal{S} := \mathrm{int}(\bar{\mathcal{S}})$ the interior of $\bar{\mathcal{S}}$, and we decompose the boundary of this state space into $\partial\mathcal{S} = \partial^y\mathcal{S} \cup \partial^k\mathcal{S} \cup \partial^z\mathcal{S}$, where

$$\partial^y\mathcal{S} = \left\{s \in \mathcal{S} \;:\; y = 0\right\}, \; \partial^k\mathcal{S} = \left\{s \in \mathcal{S} \;:\; k = 0\right\}, \text{ and } \partial^z\mathcal{S} = \left\{s \in \mathcal{S} \;:\; z = 0\right\}.$$

Observe that the value function is not known on the entire boundary of the state space $\mathcal{S}$. It is shown in [5] that the only boundary information is

$$V(s) \;=\; 0 \text{ for all } s \in \partial^z\mathcal{S}. \tag{2.11}$$

The main result of this section states that the value function V defined in (2.8) solves the corresponding DPE

$$F(s, v, Dv, D^2v) \;:=\; \min\left\{-\mathcal{L}v,\; \mathbf{g^b}\cdot Dv,\; \mathbf{g^s}\cdot Dv\right\} \;=\; 0 \text{ on } \bar{\mathcal{S}} \setminus \partial^z\mathcal{S}, \tag{2.12}$$

where $\mathcal{L}$ is the second order differential operator

$$\mathcal{L}\varphi\,(s) := -\beta\varphi\,(s) + rx\varphi_x\,(s) + \rho y\varphi_y\,(s) + \frac{1}{2}\sigma^2y^2\varphi_{yy}\,(s) + \tilde{U}\left(\varphi_x\,(s)\right), \tag{2.13}$$

$\tilde{U}$ is the Fenchel dual defined by

$$\tilde{U}(\xi) \;:=\; \sup_{c>0}\left(U(c) - c\xi\right) \;\text{ for all }\; \xi > 0, \tag{2.14}$$

and $\mathbf{g^b}$, $\mathbf{g^s}$ are the vector fields defined by

$$\mathbf{g^b} := \begin{pmatrix} 1+\lambda \\ -1 \\ -1 \end{pmatrix}, \mathbf{g^s}(s) := \begin{pmatrix} -(1-\mu) \\ \frac{1}{1-\alpha} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{-\alpha}{1-\alpha} \\ 1 \end{pmatrix} \frac{k\,\mathbf{1}_{(y,k)\neq 0}}{(1-\alpha)\,y+\alpha k}. \tag{2.15}$$

The DPE can be written in different forms by taking other vector fields which are parallel to our choices $\mathbf{g^b}, \mathbf{g^s}$. Since our choice for $\mathbf{g^s}$ is discontinuous and this fact is central to many of the technicalities, one may propose to choose a parallel vector field which is continuous. However, in singular stochastic control, if the vector fields appearing in the equation vanish (which is the case here if we choose continuous vector fields), then the first order part of the equation (i.e., the part $\mathbf{g^s} \cdot Dv$ in the above particular case) becomes degenerate. Indeed, this degeneracy is equivalent to the technical difficulties related to the discontinuity of the vector fields. For this reason, it is standard in singular control to choose these vector fields as nondegenerate and close to unit vector fields.

Since we have no knowledge of any a priori regularity of the value function V, we will use the theory of viscosity solutions. This notion allows for a weak formulation of solutions to second order parabolic PDEs and boundary conditions; see [23, 9].

In what follows, we use the following classical notation from viscosity theory. For a locally bounded function $v : \bar{\mathcal{S}} \longrightarrow \mathbb{R}$, we denote the corresponding upper and lower semicontinuous envelopes by

$$v^*(s) := \limsup_{\mathcal{S} \ni s' \to s} v(s') \text{ and } v_*(s) := \liminf_{\mathcal{S} \ni s' \to s} v(s').$$

The notation F_* in the subsequent definition is defined similarly. Observe that $F = F_*$ outside the axis $\{(x,0,0) : x \geq 0\}$.

Definition 2.1. (i) *A locally bounded function v is a constrained viscosity subsolution of* (2.11)–(2.12) *if $v^* \leq 0$ on $\partial^z\mathcal{S}$, and for all $s \in \bar{\mathcal{S}} \setminus \partial^z\mathcal{S}$ and $\varphi \in C^2(\bar{\mathcal{S}})$ with $(v^* - \varphi)(s) = \max_{\bar{\mathcal{S}}}(v^* - \varphi)$ we have $F_*\left(s, v(s), D\varphi(s), D^2\varphi(s)\right) \leq 0$.*

(ii) *A locally bounded function v is a viscosity supersolution of* (2.11)–(2.12) *if $v_* \geq 0$ on $\partial^z\mathcal{S}$, and for all $s \in \mathcal{S}$ and $\varphi \in C^2(\mathcal{S})$ with $(v_* - \varphi)(s) = \min_{\mathcal{S}}(v_* - \varphi)$ we have $F\left(s, v(s), D\varphi(s), D^2\varphi(s)\right) \geq 0$.*

(iii) *A locally bounded function v is a constrained viscosity solution of* (2.11)–(2.12) *if it is a constrained viscosity subsolution and supersolution.*

In the above definition, observe that there is no boundary value assigned to the value function on $\partial^y\mathcal{S} \cup \partial^k\mathcal{S}$. Instead, the subsolution property holds on this boundary. Notice that the supersolution property is satisfied only in the interior of the domain $\mathcal{S}$.

Proposition 2.2. *The value function V is a constrained viscosity solution of* (2.11)–(2.12).

The proof is reported in section 5 for the case $\varepsilon = 0$. In the accompanying paper [5] a numerical scheme based on the finite differences approximation of the DPE (2.11)–(2.12) is implemented. In order for us to justify this algorithm, we need a uniqueness result for this DPE. As it is usually the case for parabolic second order equations, uniqueness follows as a consequence of a comparison result. At this point, a chief difficulty is encountered: the vector field $\mathbf{g^s}$ is not locally Lipschitz on the axis $\{(x,0,0),\ x \geq 0\}$. Because of this problem, the standard techniques for the derivation of a comparison result for the DPE (2.11)–(2.12) fail.

Remark 1. Consider the Lipschitz vector field $\mathbf{G^s} := (-(1-\mu)[(1-\alpha)y+\alpha k], y, k) = [(1-\alpha)y+\alpha k]\mathbf{g^s}$. Then, the supersolutions of (2.11)–(2.12) coincide with those of

$$\min\left\{-\mathcal{L}v,\ \mathbf{g^b}\cdot Dv,\ \mathbf{G^s}\cdot Dv\right\} \geq 0 \text{ on } \bar{\mathcal{S}}\setminus\partial^z\mathcal{S} \text{ and } v=0 \text{ on } \partial^z\mathcal{S}. \tag{2.16}$$

However, these two equations do not have the same set of subsolutions. The reason for this is that the subsolution property must hold also on the boundary $\partial^y\mathcal{S}\cup\partial^k\mathcal{S}$. Since $\mathbf{G^s}(x,0,0)=0$ for every $x\geq 0$, (2.16) provides no information on this axis. Notice, however, that $\lim_{n\to\infty}\mathbf{g^s}(s_n)$ exists for some sequences $s_n\to(x,0,0)$, and might be nonzero, so that (2.12) bears more information on this axis.

This remark justifies that the above mentioned difficulty can be avoided if a priori comparison on the axis $\{(x,0,0)\ :\ x\geq 0\}$ is available.

PROPOSITION 2.3. *Let $\lambda+\mu>0$. Let u be an upper semicontinuous constrained viscosity subsolution of* (2.11)–(2.12) *and v be a lower semicontinuous viscosity supersolution of* (2.11)–(2.12) *with $(u-v)^+\in \mathrm{USC}_p(\bar{\mathcal{S}})$. Assume further that $(u-v)(x,0,0)\leq 0$ for all $x\geq 0$. Then $u\leq v$ on $\bar{\mathcal{S}}$.*

The proof of this comparison result is given at the end of section 4. Unfortunately, this result does not provide uniqueness of a constrained viscosity solution for the DPE (2.11)–(2.12), as we have no a priori comparison of two possible solutions on the axis $\{(x,0,0)\ :\ x\geq 0\}$.

The chief goal of this paper is to obtain an alternative characterization of V by considering a convenient approximating PDE which has a unique solution converging to our value function V. Before turning to this issue, we report the following continuity property from [5] which follows from Proposition 2.3.

PROPOSITION 2.4 (see [5]). *Let $\lambda+\mu>0$. For $s=(x,y,k)\in\bar{\mathcal{S}}$ and $z:=x+(1-\mu)[(1-\alpha)y+\alpha k]$, we have $V(s)=z^p\ \mathcal{V}\left(\frac{y}{z},\frac{k}{z}\right)$, where $\mathcal{V}$ is a Lipschitz-continuous function on $\mathbb{R}^2_+$.*

3. The main results. For every $\varepsilon>0$ and $s=(x,y,k)\in\bar{\mathcal{S}}$, we define

$$f^\varepsilon(s)\ :=\ h\left(\frac{k}{\varepsilon z}\right)^+, \text{ where } z:=x+(1-\mu)\left[(1-\alpha)y+\alpha k\right], \tag{3.1}$$

and h is a nondecreasing $C^2(\mathbb{R}_+)$-function with

$$h=0 \ \text{ on } \ [0,1] \ \text{ and } \ h=1 \ \text{ on } \ [2,\infty).$$

For $\varepsilon=0$, we set $f^0(s)=1$.

We next introduce, for all $\varepsilon\geq 0$, the approximation $\mathbf{g^s_\varepsilon}$ of $\mathbf{g^s}$,

$$\mathbf{g^s_\varepsilon}(s)\ :=\ \mathbf{g^s}\left(x,y,kf^\varepsilon(s)\right) \text{ for } s=(x,y,k)\in\bar{\mathcal{S}}, \tag{3.2}$$

and the corresponding approximation of the DPE (2.11)–(2.12):

$$\min\left\{-\mathcal{L}v,\ \mathbf{g^b}\cdot Dv,\ \mathbf{g^s_\varepsilon}\cdot Dv\right\}=0 \text{ on } \bar{\mathcal{S}}\setminus\partial^z\mathcal{S} \text{ and } v=0 \text{ on } \partial^z\mathcal{S}. \tag{3.3}$$

A constrained viscosity solution of this equation is defined exactly as in Definition 2.1, replacing $\mathbf{g^s}$ by $\mathbf{g^s_\varepsilon}$. For each $\varepsilon>0$ the approximation $\mathbf{g^s_\varepsilon}$ is Lipschitz-continuous on $\bar{\mathcal{S}}\setminus\partial^z\mathcal{S}$, and this property is sufficient to obtain the following comparison result.

THEOREM 3.1. *Let $\lambda+\mu>0$ and $\varepsilon>0$. Let u be an upper semicontinuous constrained viscosity subsolution of* (3.3) *and v be a lower semicontinuous viscosity*

supersolution of (3.3) *with* $(u-v)^+ \in \mathrm{USC}_p(\bar{\mathcal{S}})$. *Assume further that* $u \le v$ *on* $\partial^z\mathcal{S}$. *Then* $u \le v$ *on* $\bar{\mathcal{S}}$.

This result is proved in section 4 and implies, as usual, a uniqueness result for the approximating PDE (3.3) for every $\varepsilon > 0$. We can now state our main DPE characterization of the value function V which justifies the numerical scheme implemented in the accompanying paper [5].

THEOREM 3.2. *For every* $\varepsilon > 0$, *there exists a unique constrained viscosity solution* V_ε *for the nonlinear parabolic PDE* (3.3) *in the class* C^0_p. *Moreover, the family* $(V_\varepsilon)_{\varepsilon>0}$ *is nondecreasing and converges to the value function* V *uniformly on compact subsets of* $\bar{\mathcal{S}}$ *as* $\varepsilon \searrow 0$.

The existence of a solution for the approximating PDE (3.3) is proved in section 5 by conveniently modifying the optimal investment problem under capital gains taxes, and showing that the induced value function V_ε is a constrained viscosity solution of (3.3). Moreover, we will prove in Proposition 6.2 that $0 \le V_\varepsilon \le V$, so that V_ε inherits the p-polynomial growth of V stated in [5]. Together with the comparison result of Theorem (3.1), this shows that V^ε is the unique constrained viscosity solution in C^0_p. The convergence result is proved in section 6.

4. The comparison result. We adapt the standard argument based on the Ishii technique; see Theorem 3.2 and Lemma 3.1 in [9]. The subsequent proof is also inspired from [1]. In comparison to the latter paper, we have the additional difficulty implied by the state constraint $(y,k) \in \mathbb{R}^2_+$. We use the idea of Theorem 7.9 in [9] to account for this avoidance of the continuity assumptions of this theorem. We mention that comparison results for second order PDEs with state constraints have been obtained for specific control problems in [2] and [3] but do not apply to our context. In the subsequent analysis, the key result to avoid the continuity is the observation that

$$\begin{aligned} &\text{for each } s \in \bar{\mathcal{S}} \setminus \partial^z\mathcal{S}, \ \text{ there exists some } \ \zeta_s > 0 \ \text{ such that} \\ &\qquad s - \zeta \mathbf{g^b} \ \in \ \mathcal{S} \ \text{ for every } 0 < \zeta < \zeta_s, \end{aligned} \tag{4.1}$$

together with the following.

LEMMA 4.1. *Let* $v \in \mathrm{LSC}(\bar{\mathcal{S}})$ *be such that* $v(s_0) = \liminf_{\mathcal{S}\ni s\to s_0} v(s)$ *for* $s_0 \in \partial\mathcal{S}$. *Assume that* $\mathbf{g^b}\cdot Dv \ge 0$ *on* $\mathcal{S}$ *in the viscosity sense. Then*

$$\lim_{\ell\searrow 0} v(s - \ell\mathbf{g^b}) \ = \ v(s) \text{ for any } s \in \bar{\mathcal{S}} \setminus \partial^z\mathcal{S}.$$

Proof. Since v is a viscosity supersolution of $\mathbf{g^b}\cdot Dv \ge 0$ on $\mathcal{S}$ and (4.1) holds, we deduce that, for any $s \in \mathcal{S}$, the function $\ell \longmapsto v(s-\ell\mathbf{g^b})$ is well defined and nonincreasing on a neighborhood of 0. In particular, $v(s-\ell\mathbf{g^b}) \le v(s)$ for any $s \in \mathcal{S}$, and $\ell \ge 0$ sufficiently small. For $s_0 \in \partial\mathcal{S}$, it follows from the assumption of the lemma that $v(s_0) = \liminf_{\mathcal{S}\ni s\to s_0} v(s) \ge \liminf_{\mathcal{S}\ni s\to s_0} v(s' - \ell\mathbf{g^b}) \ge v(s_0 - \ell\mathbf{g^b})$. Hence

$$v(s-\ell\mathbf{g^b}) \ \le \ v(s) \text{ for any } s \in \bar{\mathcal{S}} \setminus \partial^z\mathcal{S} \text{ and } \ell \ge 0.$$

This implies that, for any $s \in \bar{\mathcal{S}} \setminus \partial^z\mathcal{S}$,

$$v(s) \ \ge \ \limsup_{\ell\searrow 0} v(s-\ell\mathbf{g^b}) \ \ge \ \liminf_{\ell\searrow 0} v(s-\ell\mathbf{g^b}) \ \ge \ \liminf_{\mathcal{S}\ni s'\to s} v(s') \ \ge \ v(s),$$

completing the proof. □

Another important ingredient of our comparison result is the use of a strict supersolution of the equation

$$\min\{\mathbf{g}^{\mathbf{b}} \cdot Dv,\ \mathbf{g}^{\mathbf{s}}_{\varepsilon} \cdot Dv\} \ = \ 0 \text{ on } \bar{\mathcal{S}} \setminus \partial^z \mathcal{S}. \tag{4.2}$$

This is the only place where the presence of transaction costs is crucial.

LEMMA 4.2. *Let $\lambda + \mu > 0$ and assume that condition (2.9) holds. Then, there exist two positive parameters*

$$0 \ < \ \bar{\eta} \ < \ \frac{\lambda + \mu}{2} \text{ and } \delta \in (p, 1) \text{ with } \frac{\beta}{\delta} - r - \frac{\theta^2}{2(1-\delta)} \ > \ 0$$

such that the function

$$\Phi(s) \ := \ \left(x + (1-\mu)\left[(1-\alpha+\bar{\eta})y + (\alpha+\bar{\eta})k\right]\right)^{\delta} \text{ for } s \ \in \ \bar{\mathcal{S}}$$

is a classical strict supersolution of (4.2).

Proof. We show only that $\mathbf{g}^{\mathbf{s}}_{\varepsilon} \cdot D\Phi > 0$, as the other strict inequalities are easily seen to hold. Setting $\tilde{z} := x + (1-\mu)\left[(1-\alpha+\bar{\eta})y + (\alpha+\bar{\eta})k\right]$, we directly compute that

$$\left(\mathbf{g}^{\mathbf{s}}_{\varepsilon} \cdot D\Phi\right)(s) = \frac{(1-\mu)\bar{\eta}}{1-\alpha}\tilde{z}^{\delta-1}\left[1 + (1-2\alpha)\frac{kf^{\varepsilon}(s)}{(1-\alpha)y + \alpha k f^{\varepsilon}(s)}\right].$$

If $y = k = 0$ or $1 - 2\alpha \geq 0$, the required inequality is trivial. We next assume that $(y,k) \neq 0$ and $1 - 2\alpha < 0$. Then using the fact that $f^{\varepsilon}(s) \leq 1$, it follows that

$$\begin{aligned}\left(\mathbf{g}^{\mathbf{s}}_{\varepsilon} \cdot D\Phi\right)(s) &\geq \frac{(1-\mu)\bar{\eta}}{1-\alpha}\tilde{z}^{\delta-1}\left[1 + (1-2\alpha)\frac{k}{(1-\alpha)y + \alpha k}\right] \\ &= \frac{(1-\mu)\bar{\eta}}{1-\alpha}\tilde{z}^{\delta-1}\frac{(1-\alpha)(y+k)}{(1-\alpha)y + \alpha k} \ > \ 0. \qquad \square\end{aligned}$$

We are now ready for the following proof.

Proof of Theorem 3.1. We start by setting a new notation. We denote by $\tilde{\mathcal{L}}$ the operator

$$\tilde{\mathcal{L}}(s,u,q,Q) \ := \ -\beta u + rxq_1 + \rho y q_2 + \frac{1}{2}\sigma^2 Q_{22}$$

for $s = (x,y,k) \in \bar{\mathcal{S}}$, $u \in \mathbb{R}$, $q = (q_i)_{1 \leq i \leq 3} \in \mathbb{R}^3$, and $Q = (Q_{i,j})_{\substack{1 \leq i \leq 3 \\ 1 \leq j \leq 3}} \in \mathbb{S}(3)$, so that the second order operator $\mathcal{L}$ can be written as

$$\mathcal{L}\varphi(s) \ = \ \tilde{\mathcal{L}}\left(s, \varphi(s), D\varphi(s), D^2\varphi(s)\right) + \tilde{U}(\varphi_x(s)).$$

Let u and v be as in the statement of Theorem 3.1, and let us prove that $u \leq v$ in $\bar{\mathcal{S}}$.

We first observe that we can assume, without loss of generality, that

$$v(s) \ = \ \liminf\{v(s') \ : \ s' \in \mathcal{S} \text{ and } s' \neq s\} \text{ for every } s \in \partial^y \mathcal{S} \cup \partial^k \mathcal{S}. \tag{4.3}$$

Indeed, we may define the function $\underline{v} := v$ on $\mathcal{S} \cup \partial^z \mathcal{S}$ and $\underline{v}(s) := \liminf_{s \neq s' \to s} v(s')$ for $s \in \partial^y \mathcal{S} \cup \partial^k \mathcal{S}$. Then, $\underline{v}$ satisfies the same conditions as v, and if we succeed in proving that $u \leq \underline{v}$, we deduce immediately that $u \leq v$ since the inequality $\underline{v} \leq v$ is trivial.

We now start the proof of the comparison result with the additional condition (4.3). Assume to the contrary that

$$(u-v)(s^*) > 0 \text{ for some } s^* \in \bar{\mathcal{S}}. \tag{4.4}$$

Step 1. Let Φ be the strict supersolution of (4.2) defined in Lemma 4.2, and $\eta > 0$, $\zeta > 0$ be some fixed parameters such that

$$m_0 := (u-v)(s_0) - 2\eta\Phi(s_0) - \zeta|\mathbf{g^b}|^2 = \max_{s\in\bar{\mathcal{S}}} (u - v - 2\eta\Phi) - \zeta|\mathbf{g^b}|^2 > 0 \tag{4.5}$$

by (4.4), where the maximum is attained thanks to the p-polynomial growth condition on $(u-v)^+$ and the fact that $\delta > p$. In particular, it follows from (4.5), together with $\Phi \geq 0$, $u \leq v$ on $\partial^z\mathcal{S}$ and (4.1), that

$$s_0 \in \bar{\mathcal{S}} \setminus \partial^z\mathcal{S} \text{ and } s_0 - \zeta\mathbf{g^b} \in \mathcal{S} \text{ for small } \zeta > 0. \tag{4.6}$$

We next define the mappings on $\bar{\mathcal{S}} \times \bar{\mathcal{S}}$ by

$$\begin{aligned}\Psi_n(s,s') &:= (u-\eta\Phi)(s) - (v+\eta\Phi)(s') - \psi_n(s,s'),\\ \psi_n(s,s') &:= \left|n(s-s') - \zeta\mathbf{g^b}\right|^2 + \zeta|s-s_0|^2 .\end{aligned}$$

Here, $\zeta \in (0,1)$ is some given constant. From the p-polynomial growth condition on $(u-v)^+$ and the fact that $\delta > p$ in the definition of Φ, we see that the upper semicontinuous function Ψ_n attains its maximum at some (s_n, s'_n) in $\bar{\mathcal{S}} \times \bar{\mathcal{S}}$, so that by (4.5),

$$m_n := \Psi_n(s_n, s'_n) = \max_{(s,s')\in\bar{\mathcal{S}}\times\bar{\mathcal{S}}} \Psi_n(s,s') \geq m_0 > 0.$$

By (4.6) and the definition of Ψ_n, we have the inequality $\Psi_n(s_n, s'_n) \geq \Psi_n(s_0, s_0 - \frac{\zeta}{n}\mathbf{g^b})$ which, together with the p-polynomial growth condition on u and v, provides

$$\begin{aligned}\left|n(s_n - s'_n) - \zeta\mathbf{g^b}\right|^2 + \zeta\,|s_n - s_0|^2 &\leq (u-\eta\Phi)(s_n) - (v+\eta\Phi)(s'_n)\\ &\quad -(u-\eta\Phi)(s_0) + (v+\eta\Phi)\left(s_0 - \frac{\zeta}{n}\mathbf{g^b}\right)\\ &\leq \tilde{A}\left(1 + |s_n|^p + |s'_n|^p + \eta|s_n|^\delta + \eta|s'_n|^\delta\right)\end{aligned} \tag{4.7}$$

for some positive constant $\tilde{A}$. We deduce from the last inequality that the sequences $(s_n)_{n\geq 1}$ and $(s'_n)_{n\geq 1}$ are bounded, and we can assume, without loss of generality, that $s_n, s'_n \longrightarrow \hat{s} \in \bar{\mathcal{S}}$ as $n \to \infty$. We now use Lemma 4.1, together with the upper semicontinuity of u and the lower semicontinuity of v, to pass to the limit as $n \to \infty$ in (4.7). This provides

$$\begin{aligned}\limsup_{n\to\infty}\left(\left|n(s_n - s'_n) - \zeta\mathbf{g^b}\right|^2 + \zeta\,|s_n - s_0|^2\right) &\leq (u-\eta\Phi)(\hat{s}) - (v+\eta\Phi)(\hat{s})\\ &\quad - ((u-\eta\Phi)(s_0) - (v-\eta\Phi)(s_0))\\ &\leq 0,\end{aligned}$$

where the last inequality follows from (4.5). Consequently

$$\left|n(s_n - s'_n) - \zeta\mathbf{g^b}\right|^2 \longrightarrow 0 \text{ and } s_n, s'_n \longrightarrow s_0 \text{ as } n \to \infty.$$

In particular, it follows from (4.6) that

$$s'_n \;=\; s_n - \frac{\zeta \mathbf{g^b} + o(1)}{n} \;\in\; \mathcal{S} \text{ and } s_n \;\in\; \bar{\mathcal{S}} \setminus \partial^z \mathcal{S} \text{ for large } n. \tag{4.8}$$

Step 2. For each $n \geq 1$, (s_n, s'_n) is a maximum point of

$$\Psi_n \;:\; (s,s') \longmapsto (u - \eta\Phi)(s) - (v + \eta\Phi)(s') - \psi_n(s,s').$$

Then applying Theorem 3.2 in [9] to the upper semicontinuous functions $u - \eta\Phi$ and to the lower semicontinuous function $v + \eta\Phi$, we deduce that there exist 3×3 symmetric matrices Ξ_n and Υ_n, with $\Xi_n \leq \Upsilon_n$ such that

$$j_n := \big(q_n := D_1\psi(s_n,s'_n) + \eta D\Phi(s_n); Q_n := \Xi_n + \eta D^2\Phi(s_n)\big) \;\in\; \bar{J}^{2,+}_{\bar{\mathcal{S}}\setminus\partial^z\mathcal{S}} u(s_n), \tag{4.9}$$

$$j'_n := \big(q'_n := -D_2\psi(s_n,s'_n) - \eta D\Phi(s'_n); Q'_n := \Upsilon_n - \eta D^2\Phi(s'_n)\big) \;\in\; \bar{J}^{2,-}_{\mathcal{S}} v(s'_n), \tag{4.10}$$

and

$$-\big(2n^2 + ||M_n||\big)\, I \;\leq\; \begin{pmatrix} \Xi_n & 0 \\ 0 & -\Upsilon_n \end{pmatrix} \;\leq\; M_n + \frac{1}{2n^2} M - n^2, \tag{4.11}$$

where

$$M_n \;:=\; D^2\psi(s_n,s'_n) \;=\; 2n^2 \begin{pmatrix} I & -I \\ -I & I \end{pmatrix} + 2\zeta \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix},$$

$$D_1\psi(s,s') = 2n\left(n(s-s') - \zeta\mathbf{g^b}\right) + 2\zeta(s - s_0), \quad -D_2\psi(s,s') = 2n\left(n(s-s') - \zeta\mathbf{g^b}\right).$$

Here the norm of a symmetric matrix M is defined as $\|M\| = \sup\{M\xi\cdot\xi \;:\; |\xi| \leq 1\}$.

By (4.8) and for large $n \geq 1$, the subsolution property of u holds at j_n and the supersolution property of v holds at j'_n, i.e.,

$$\min\Big\{\beta u(s_n) - \tilde{\mathcal{L}}(s_n,q_n,Q_n) - \tilde{U}(q_{n\,1}),\; \mathbf{g^b}\cdot q_n,\; \mathbf{g^s_\varepsilon}(s_n)\cdot q_n\Big\} \leq 0, \tag{4.12}$$

$$\min\Big\{\beta v(s'_n) - \tilde{\mathcal{L}}(s'_n,q'_n,Q'_n) - \tilde{U}(q'_{n\,1}),\; \mathbf{g^b}\cdot q'_n,\; \mathbf{g^s_\varepsilon}(s'_n)\cdot q'_n\Big\} \geq 0. \tag{4.13}$$

Step 3. For each $n \geq 1$,

$$\mathbf{g^b}\cdot q_n - \mathbf{g^b}\cdot q'_n = \eta\mathbf{g^b}\cdot(D\Phi(s_n) + D\Phi(s'_n)) + 2\zeta\mathbf{g^b}\cdot(s_n - s_0).$$

Recall that $s_n, s'_n \longrightarrow s_0 \in \bar{\mathcal{S}} \setminus \partial^z\mathcal{S}$, and $\mathbf{g^b}\cdot\Phi > 0$ on $\bar{\mathcal{S}} \setminus \partial^z\mathcal{S}$; then

$$\lim_{n\to\infty}\big(\mathbf{g^b}\cdot q_n - \mathbf{g^b}\cdot q'_n\big) = 2\eta\mathbf{g^b}\cdot D\Phi(s_0) \;>\; 0. \tag{4.14}$$

We also compute for all $n \geq 1$ that

$$\begin{aligned}\mathbf{g^s_\varepsilon}(s_n)\cdot q_n - \mathbf{g^s_\varepsilon}(s'_n)\cdot q'_n = \eta\left(\mathbf{g^s_\varepsilon}(s_n)\cdot D\Phi(s_n) + \mathbf{g^s_\varepsilon}(s'_n)\cdot D\Phi(s'_n)\right) + 2\zeta\mathbf{g^s_\varepsilon}(s_n)\cdot(s_n - s_0)\\ + \left(\mathbf{g^s_\varepsilon}(s_n) - \mathbf{g^s_\varepsilon}(s'_n)\right)\cdot 2n\left[n(s_n - s'_n) - \zeta\mathbf{g^b}\right].\end{aligned}$$

By the local Lipschitz continuity of the function $\mathbf{g^s_\varepsilon}$ at s_0, there exists some positive constant C_0 such that for large n,

$$\begin{aligned}&|\mathbf{g^s_\varepsilon}(s_n)\cdot q_n - \mathbf{g^s_\varepsilon}(s'_n)\cdot q'_n - \eta\left(\mathbf{g^s_\varepsilon}(s_n)\cdot D\Phi(s_n) + \mathbf{g^s_\varepsilon}(s'_n)\cdot D\Phi(s'_n)\right)|\\ &\quad\leq\; 2\zeta\,|\mathbf{g^s_\varepsilon}(s_n)|\,|s_n - s_0|\,C_0\,|s_n - s'_n|\,2n\left|n(s_n - s'_n) - \zeta\mathbf{g^b}\right|\\ &\quad\leq\; 2\zeta\,|\mathbf{g^s_\varepsilon}(s_n)|\,|s_n - s_0|\,2C_0\left|n(s_n - s'_n) - \zeta\mathbf{g^b}\right|^2 + 2C_0\zeta|\mathbf{g^b}|\left|n(s_n - s'_n) - \zeta\mathbf{g^b}\right|.\end{aligned}$$

Since $s_n \longrightarrow s_0$ and $|n(s_n - s'_n) - \zeta \mathbf{g^b}| \longrightarrow 0$, we get

$$\lim_{n\to\infty} (\mathbf{g}^{\mathbf{s}}_{\varepsilon}(s_n)\cdot q_n - \mathbf{g}^{\mathbf{s}}_{\varepsilon}(s'_n)\cdot q'_n) = 2\eta \mathbf{g}^{\mathbf{s}}_{\varepsilon}\cdot D\Phi(s_0) \ > \ 0. \tag{4.15}$$

We deduce from (4.13), (4.14), and (4.15), together with Lemma 4.2, that for large n,

$$\min\left\{\mathbf{g^b}\cdot q_n,\ \mathbf{g}^{\mathbf{s}}_{\varepsilon}(s_n)\cdot q_n\right\} \geq 2\min\left\{\mathbf{g^b}\cdot D\Phi(s_0),\ \mathbf{g}^{\mathbf{s}}_{\varepsilon}(s_n)\cdot D\Phi(s_0)\right] + o(1) \ > \ 0.$$

Consequently (4.12) implies that for large n,

$$\beta u(s_n) - \tilde{\mathcal{L}}(s_n, q_n, Q_n) - \tilde{U}(q_{n\,1}) \leq 0. \tag{4.16}$$

Step 4. From (4.13) and (4.16), it follows that for large n,

$$\beta u(s_n) - \tilde{\mathcal{L}}(s_n, q_n, Q_n) - \tilde{U}(q_{n\,1}) \leq \ 0 \ \leq \beta v(s'_n) - \tilde{\mathcal{L}}(s'_n, q'_n, Q'_n) - \tilde{U}(q'_{n\,1}).$$

Using the local Lipschitz continuity property of the function $\tilde{U}$, a direct calculation shows that for some positive constant C and for large n,

$$\begin{aligned}
\beta(u(s_n) - v(s'_n)) &\leq \tilde{\mathcal{L}}(s_n, q_n, Q_n) - \tilde{\mathcal{L}}(s'_n, q'_n, Q'_n) + \tilde{U}(q_{n\,1}) - \tilde{U}(q'_{n\,1}) \\
&\leq C\left(|s_n|\zeta|s_n - s_0| + |n(s_n - s'_n) - \zeta\mathbf{g^b}|^2 + |D\Phi(s_n) - D\Phi(s'_n)|\right) \\
&\quad + \frac{\sigma^2}{2}\left(y_n^2(Q_n)_{22} - (y'_n)^2(Q'_n)_{22}\right) \\
&\quad + \eta\left\{\tilde{\mathcal{L}}(s_n, D\Phi(s_n), D^2\Phi(s_n)) + \tilde{\mathcal{L}}(s'_n, D\Phi(s'_n), D^2\Phi(s'_n))\right\}.
\end{aligned}$$

From (4.11), we have that

$$\left(y_n^2(Q_n)_{22} - (y'_n)^2(Q'_n)_{22}\right) \leq 4\zeta y_n(u_n - y'_n) + \frac{\zeta^2}{n^2}y_n.$$

Moreover, the mapping Φ satisfies $\beta\Phi(.) - \tilde{\mathcal{L}}(., D\Phi, D^2\Phi)$ on $\bar{\mathcal{S}} \setminus \partial^z\mathcal{S}$, and hence for some positive constant $\tilde{C}$ and for large n,

$$\begin{aligned}
&\beta[u(s_n) - v(s'_n)] - \eta\Phi(s_n) - \eta\Phi(s'_n) \\
&\quad \leq \tilde{\mathcal{L}}(s_n, q_n, Q_n) - \tilde{\mathcal{L}}(s'_n, q'_n, Q'_n) + \tilde{U}(q_{n\,1}) - \tilde{U}(q'_{n\,1}) \\
&\quad \leq \tilde{C}\left\{\frac{1}{n^2} + |s_n|\zeta|s_n - s_0| + |n(s_n - s'_n) - \zeta\mathbf{g^b}|^2 + |D\Phi(s_n) - D\Phi(s'_n)|\right\},
\end{aligned}$$

where the right-hand side of the inequality goes to zero as $n \to \infty$. This implies

$$\beta[u(s_0) - v(s_0)] - 2\eta\Phi(s_0) = \limsup_{n\to\infty}\ (\beta[u(s_n) - v(s'_n)] - \eta\Phi(s_n) - \eta\Phi(s'_n)) \ \leq \ 0,$$

contradicting (4.5). □

We conclude this section with the following proof.

Proof of Proposition 2.3. We use the same arguments as in the proof of Theorem 3.1, but this time substituting $\mathbf{g^s}$ for $\mathbf{g}^{\mathbf{s}}_{\varepsilon}$. The only difference is the following. The maximizer s_0 in (4.5) is now known to be in $\bar{\mathcal{S}} \setminus (\partial^z\mathcal{S} \cup \{(x, 0, 0) : \ x \geq 0\})$, as it is assumed in the statement of the proposition that $u \leq v$ on $\partial^z\mathcal{S} \cup \{(x, 0, 0) : \ x \geq 0\}$. Then, the sequences $(s_n)_n$ and $(s'_n)_n$, defined in Step 1, are valued in a ball around s_0 which does not intersect the axis $\{(x, 0, 0) : \ x \geq 0\}$. Since $\mathbf{g^s}$ is locally Lipschitz on $\bar{\mathcal{S}} \setminus \{(x, 0, 0) : \ x \geq 0\}$, we just follow along the lines of the previous proof. □

5. An approximating control problem. Let $s = (x, y, k)$ be an initial condition in the state space $\bar{\mathcal{S}}$, and consider a control process $\nu \in \mathcal{A}$, i.e., a triple of $\mathbb{F}$-adapted processes $\nu = (C, L, M)$, with nondecreasing right-continuous processes L, M, $L_{0-} = M_{0-} = 0$ and satisfying conditions (2.2) and (2.3). For every parameter $\varepsilon \geq 0$, we denote by $S^{\varepsilon,s,\nu} = (X^{\varepsilon,s,\nu}, Y^{\varepsilon,s,\nu}, K^{\varepsilon,s,\nu})$ the unique strong solution of

$$dX_t^\varepsilon = (rX_t^\varepsilon - C_t)dt - (1+\lambda)dL_t + (1-\mu)\left[(1-\alpha)Y_{t-}^\varepsilon + \alpha f^\varepsilon(S_{t-}^\varepsilon)K_{t-}^\varepsilon\right] dM_t, \tag{5.1}$$

$$dY_t^\varepsilon = Y_t^\varepsilon\left[\rho dt + \sigma dW_t\right] + dL_t - Y_{t-}^\varepsilon dM_t, \tag{5.2}$$

$$dK_t^\varepsilon = dL_t - f^\varepsilon(S_{t-}^\varepsilon)K_{t-}^\varepsilon dM_t \tag{5.3}$$

with initial condition $S_{0-}^{\varepsilon,s,\nu} = s$. With this definition, observe that the jumps of the state processes $S^{\varepsilon,s,\nu}$ are given by

$$\Delta S_t^{\varepsilon,s,\nu} = -\Delta L_t\, \mathbf{g^b} - \Delta M_t \left[(1-\alpha)Y_{t-}^{\varepsilon,s,\nu} + \alpha f^\varepsilon(S_{t-}^{\varepsilon,s,\nu})K_{t-}^{\varepsilon,s,\nu}\right] \mathbf{g}_\varepsilon^{\mathbf{s}}\left(S_{t-}^{\varepsilon,s,\nu}\right),$$

where the vector fields $\mathbf{g^b}$ and $\mathbf{g}_\varepsilon^{\mathbf{s}}$ are defined as in (2.15) and (3.2).

A control process $\nu = (C, L, M)$ is said to be (s, ε)-admissible if the corresponding state process $S^{\varepsilon,s,\nu}$ is valued in $\bar{\mathcal{S}}$. We shall denote by $\mathcal{A}^\varepsilon(s)$ the collection of all (s, ε)-admissible controls.

For every initial condition $s \in \bar{\mathcal{S}}$, $\varepsilon \geq 0$, and (s, ε)-admissible control $\nu = (C, L, M)$, we introduce the criterion

$$J_T^\varepsilon(s, \nu) := \mathbb{E}\left[\int_0^T e^{-\beta t}U(C_t)dt + e^{-\beta T}U(Z_T^{\varepsilon,s,\nu})\mathbf{1}_{T<\infty}\right], \qquad T \in \mathbb{R}_+ \cup \{\infty\}, \tag{5.4}$$

where U is the power utility function defined in (2.8). The value function V_ε is then defined by

$$V_\varepsilon(s) := \sup_{\nu \in \mathcal{A}^\varepsilon(s)} J_\infty^\varepsilon(s, \nu). \tag{5.5}$$

Remark 2. When $\varepsilon = 0$, the above problem reduces to the optimal investment problem under capital gains taxes reviewed in section 2, in particular $V_0 = V$. For positive ε, the control problem (5.5) can be interpreted as a utility maximization problem with a modified taxation rule. Under this new taxation rule, the tax basis used to evaluate the capital gains is equal to the relative weighted average purchase price as long as the ratio K/Z is larger than 2ε, but it is set to zero when $K/Z < \varepsilon$. Roughly speaking, for $\varepsilon > 0$, the investor pays more taxes than in the original market when the ratio $K/Z < \varepsilon$. Consequently, we expect that V_ε increases towards V as ε goes to zero. This will be proved in Proposition 6.2 below.

The main objective of this section is to prove that the function V_ε is a constrained viscosity solution of the approximating PDE (3.3), thus proving the existence statement in Theorem 3.2. The arguments of this section hold for every $\varepsilon \geq 0$. In particular, the proof of Proposition 2.2 corresponds to the special case $\varepsilon = 0$.

As usual, the key ingredient for deriving the DPE is a dynamic programming principle. We state it here without proof, and we refer the reader to [6, 13, 14].

THEOREM 5.1. *Let $\varepsilon \geq 0$, $s \in \bar{\mathcal{S}}$, and let τ be some $\mathbb{P}$-a.s. finite $\mathbb{F}$-stopping time. Then*

$$V_\varepsilon(s) = \sup_{\nu=(C,L,M)\in\mathcal{A}^\varepsilon(s)} \mathbb{E}\left[\int_0^\tau e^{-\beta t}U(C_t)dt + e^{-\beta\tau}V_\varepsilon\left(S_\tau^{\varepsilon,s,\nu}\right)\right].$$

Before turning to the derivation of the DPE for the problem V_ε, we introduce a notation which will be used frequently in what follows. Let $\varepsilon \geq 0$, $s \in \bar{\mathcal{S}}$, $\nu = (C, L, M) \in \mathcal{A}(s)$, and consider some stopping time τ such that $S^{\varepsilon,s,\nu}_{\tau-} \in \bar{\mathcal{S}}$. Then, it is easy to verify that the strategy $\nu(\tau)$ defined by

$$\nu(\tau)_t \;:=\; \left(\bar{C}, \bar{L}, \bar{M}\right) := \nu_t \mathbf{1}_{[0,\tau[}(t) + (0, L_{\tau-}, M_{\tau-} + (1 - \Delta M_\tau))\, \mathbf{1}_{[\tau,\infty)}(t) \tag{5.6}$$

is in $\mathcal{A}^\varepsilon(s)$, and that

$$\mathbb{E}\left[\int_0^\infty e^{-\beta t} U\left(\bar{C}_t\right) dt\right] = \mathbb{E}\left[\int_0^\tau e^{-\beta t} U\left(\bar{C}_t\right) dt\right]. \tag{5.7}$$

5.1. Supersolution property. In this section, we prove that the value function V_ε is a viscosity supersolution of (3.3) on $\mathcal{S}$ for every $\varepsilon \geq 0$.

Step 1. Fix some $\varepsilon \geq 0$. Recall that $V_\varepsilon \geq 0$ by definition, and in particular $(V_\varepsilon)_*(0) \geq 0$. So it remains to show that, for s_0 in $\mathcal{S}$ and φ in $C^2(\bar{\mathcal{S}})$ such that

$$0 \;=\; \left((V_\varepsilon)_* - \varphi\right)(s_0) = \min_{\mathcal{S}} \left((V_\varepsilon)_* - \varphi\right),$$

the test function φ must satisfy, at the point s_0,

$$\min\left\{-\mathcal{L}\varphi,\ \mathbf{g}^{\mathbf{b}} \cdot D\varphi,\ \mathbf{g}^{\mathbf{s}}_\varepsilon \cdot D\varphi\right\}(s_0) \geq 0.$$

Step 2.1. Let $\eta > 0$ be such that $B(s_0, \eta) \subset \mathcal{S}$, and consider some sequence $(s_n)_{n\geq 1}$ satisfying

$$\begin{aligned} &\text{(i) } B(s_0, \eta) \ni s_n \underset{n\to\infty}{\longrightarrow} s_0, \\ &\text{(ii) } \xi_n \;:=\; V_\varepsilon(s_n) - \varphi(s_n) \;\longrightarrow\; 0 \text{ as } n \to 0. \end{aligned}$$

Fix some (c, ℓ, m) in $(0, \infty)^3$, define the strategy $\nu \in \mathcal{A}$ by

$$\nu_t \;=\; (C_t = c, L_t = \ell\, t, M_t = m\, t),$$

and let $(\tau^n)_{n\geq 0}$ be the stopping times

$$\tau^n \;:=\; \inf\{t \geq 0 \;:\; S^{\varepsilon,s_n,\nu}_t \notin \mathcal{S}\} n \geq 0.$$

Given that for each $n \geq 0$, $s_n \notin \partial^z \mathcal{S}$, and that the strategy ν is continuous, we have

$$\tau^n \;>\; 0 \text{ for all } n \geq 0 \;\text{ and }\; \tau^n \underset{n\to\infty}{\longrightarrow} \tau^0 \quad \mathbb{P}\text{-a.s.} \tag{5.8}$$

Step 2.2. To each $n \geq 1$ we associate the (ε, s_n)-admissible strategy $\nu(\tau^n) = (C^n, L^n, M^n) \in \mathcal{A}^\varepsilon(s_n)$ defined in (5.6). To simplify the notation, we set $S^n := S^{\varepsilon,s_n,\nu(\varepsilon,s_n)}$. For any $\mathbb{P}$-a.s. finite stopping time θ^n, the dynamic programming principle of Theorem 5.1 provides

$$V_\varepsilon(s_n) \geq \mathbb{E}\left[\int_0^{\theta^n \wedge \tau^n/2} e^{-\beta t} U(C^n_t) dt \;+\; e^{-\beta\, \theta^n \wedge \tau^n/2} V_\varepsilon\left(S^n_{\theta^n \wedge \tau^n/2}\right)\right].$$

Notice that $S^n_{\theta^n \wedge \tau^n/2} \in \mathcal{S}$; we then deduce from the inequalities $\varphi \leq (V_\varepsilon)_* \leq V_\varepsilon$ on $\mathcal{S}$ that

$$\xi_n \;+\; \varphi(s_n) \geq \mathbb{E}\left[\int_0^{\theta^n \wedge \tau^n/2} e^{-\beta t} U(C^n_t) dt \;+\; e^{-\beta\theta^n \wedge \tau^n/2} \varphi\left(S^n_{\theta^n \wedge \tau^n/2}\right)\right].$$

By the definition of the strategy $\nu(\tau^n)$, jumps of the process S^n may occur only at the stopping time τ^n, and by definition of the stopping time τ^n, the process $\left\{S_t^n \mathbf{1}_{[0,\tau^n]}(t), t \geq 0\right\}$ is uniformly bounded. Hence, using the Itô formula we get

$$
\begin{aligned}
-\xi_n \leq \mathbb{E}\left[\int_0^{\theta^n \wedge \tau^n/2} e^{-\beta t}\left\{-\mathcal{L}\varphi + \tilde{U}(\varphi_x) - \left(U(C_t^n) - C_t^n \varphi_x\right)\right\}(S_t^n) dt\right] \\
+ \ell\, \mathbb{E}\left[\int_0^{\theta^n \wedge \tau^n/2} e^{-\beta t} \mathbf{g}^{\mathbf{b}} \cdot D\varphi(S_t^n) dt\right] \\
+ m\, \mathbb{E}\left[\int_0^{\theta^n \wedge \tau^n/2} e^{-\beta t}\left[(1-\alpha) Y_t^n + \alpha f^\varepsilon(S_t^n) K_t^n\right]\left(\mathbf{g}_\varepsilon^{\mathbf{s}} \cdot D\varphi\right)(S_t^n) dt\right].
\end{aligned} \tag{5.9}
$$

Step 2.3. Set

$$
\theta_n = \begin{cases} \sqrt{\xi_n} & \text{if } \xi_n > 0, \\ n^{-1} & \text{if } \xi_n = 0. \end{cases}
$$

Since $\theta^n \longrightarrow 0$ and $\tau^n \longrightarrow \tau^0 > 0$ $\mathbb{P}$-a.s. as $n \to \infty$, it follows that for $\mathbb{P}$-a.s., $\theta^n \wedge \tau^n/2 = \theta^n$ for large n. Rewriting (5.9), and taking the limits as $n \to \infty$, we obtain

$$
\begin{aligned}
0 &= \lim_{n\to\infty} -\frac{\xi_n}{\theta_n}, \\
&\leq \liminf_{n\to\infty} \mathbb{E}\left[\frac{1}{\theta_n}\int_0^{\theta^n \wedge \tau^n/2} e^{-\beta t}\left\{-\mathcal{L}\varphi + \tilde{U}(\varphi_x) - \left(U(C_t^n) - C_t^n \varphi_x\right)\right\}(S_t^n) dt\right] \\
&\qquad + \ell\, \mathbb{E}\left[\frac{1}{\theta_n}\int_0^{\theta^n \wedge \tau^n/2} e^{-\beta t} \mathbf{g}^{\mathbf{b}} \cdot D\varphi(S_t^n) dt\right] \\
&\qquad + m\, \mathbb{E}\left[\frac{1}{\theta_n}\int_0^{\theta^n \wedge \tau^n/2} e^{-\beta t}\left[(1-\alpha) Y_t^n + \alpha f^\varepsilon(S_t^n) K_t^n\right]\left(\mathbf{g}_\varepsilon^{\mathbf{s}} \cdot D\varphi\right)(S_t^n) dt\right].
\end{aligned} \tag{5.10}
$$

Since $\varphi \in C^2\left(\bar{\mathcal{S}}\right)$, and the process $\left\{S_t^n \mathbf{1}_{[0,\tau^n/2]}(t), t \geq 0\right\}$ is continuous and uniformly bounded, we get by dominated convergence

$$
\begin{aligned}
&\liminf_{n\to\infty} \mathbb{E}\left[\frac{1}{\theta_n}\int_0^{\theta^n \wedge \tau^n/2} e^{-\beta t}\left\{-\mathcal{L}\varphi + \tilde{U}(\varphi_x) - \left(U(C_t^n) - C_t^n \varphi_x\right)\right\}(S_t^n) dt\right] \\
&\qquad + \ell\, \mathbb{E}\left[\frac{1}{\theta_n}\int_0^{\theta^n \wedge \tau^n/2} e^{-\beta t} \mathbf{g}^{\mathbf{b}} \cdot D\varphi(S_t^n) dt\right] \\
&\qquad + m\, \mathbb{E}\left[\frac{1}{\theta_n}\int_0^{\theta^n \wedge \tau^n/2} e^{-\beta t}\left[(1-\alpha) Y_t^n + \alpha f^\varepsilon(S_t^n) K_t^n\right]\left(\mathbf{g}_\varepsilon^{\mathbf{s}} \cdot D\varphi\right)(S_t^n) dt\right] \\
&\quad = -\mathcal{L}\varphi(s_0) + \tilde{U}(\varphi_x(s_0)) - \left(U(c) - c\,\varphi_x(s_0)\right) \\
&\qquad + \ell\, \mathbf{g}^{\mathbf{b}} \cdot D\varphi(s_0) \;+\; m\; \left[(1-\alpha) y_0 + \alpha f^\varepsilon(s_0) k_0\right] \mathbf{g}_\varepsilon^{\mathbf{s}}(s_0) \cdot D\varphi(s_0).
\end{aligned}
$$

Recall (5.10); then

$$
\begin{aligned}
0 \leq -\mathcal{L}\varphi(s_0) + \tilde{U}(\varphi_x(s_0)) - \left(U(c) - c\,\varphi_x(s_0)\right) \\
+ \ell\, \mathbf{g}^{\mathbf{b}} \cdot D\varphi(s_0) + m\; \left[(1-\alpha) y_0 + \alpha f^\varepsilon(s_0) k_0\right] \mathbf{g}_\varepsilon^{\mathbf{s}}(s_0) \cdot D\varphi(s_0).
\end{aligned} \tag{5.11}
$$

Step 2.4. Observe that $s_0 \in \mathcal{S}$ implies that $[(1-\alpha)y_0 + \alpha f^\varepsilon(s_0)k_0] > 0$. Since $(c,\ell,m) \in (0,\infty)^3$, (5.11) provides

$$0 \le \min\left\{-\mathcal{L}\varphi,\ \mathbf{g^b}\cdot D\varphi,\ \mathbf{g}^{\mathbf{s}}_\varepsilon \cdot D\varphi\right\}(s_0).$$

5.2. Subsolution property. In this section, we prove that the value function V_ε is a constrained viscosity subsolution of (3.3) for every $\varepsilon \ge 0$. In preparation for this proof, we state some intermediate results.

LEMMA 5.2. *Let φ be a mapping in $C^2(\bar{\mathcal{S}})$, and let $s_0 \in \bar{\mathcal{S}}$ such that $\varphi_x(s_0) > 0$. Then there exist $\eta > 0$, $\gamma > 0$, and $c_0 > 0$ such that*

$$\tilde{U}\left(\varphi_x(s)\right) - \left[U(c) - c\varphi_x(s)\right] \ge \gamma\left(c - c_0\right)^+ \text{for all } c \ge 0 \text{ and } s \in B\left(s_0,\eta\right) \cap \bar{\mathcal{S}}.$$

Proof. Since $\varphi_x(s_0) > 0$, we can find some $\eta, \delta > 0$ such that $\varphi_x > \delta$ on $B(s_0,\eta) \cap \bar{\mathcal{S}}$. The mapping $s \longmapsto \mathcal{I}\left(\varphi_x(s)\right) := \left(U'\right)^{-1}\left(\varphi_x(s)\right)$ is then bounded on $B(s_0,\eta) \cap \bar{\mathcal{S}}$, and since U' is a decreasing function, we can find $c_0 > 0$ such that

$$c_0 > \max_{B(s_0,\eta)\cap\bar{\mathcal{S}}} \mathcal{I}\left(\varphi_x\right) \text{ and } \gamma := \min_{B(s_0,\eta)\cap\bar{\mathcal{S}}} \left(\varphi_x - U'\left(c_0\right)\right) > 0.$$

For all $s \in B(s_0,\eta) \cap \bar{\mathcal{S}}$, using the nonnegativity and the convexity of the function $c \in \mathbb{R}_+ \longmapsto \tilde{U}(\varphi_x(s)) - (U(c) - c\,\varphi_x(s))$, we get

$$\begin{aligned}
\tilde{U}(\varphi_x(s)) - (U(c) - c\,\varphi_x(s)) &\ge \tilde{U}(\varphi_x(s)) - (U(c) - c\,\varphi_x(s)) \\
&\quad - \tilde{U}(\varphi_x(s)) + (U(c_0) - c_0\,\varphi_x(s)) \\
&\ge \left(\varphi_x(s) - U'\left(c_0\right)\right)\left(c - c_0\right)^+ \\
&\ge \gamma\ \left(c - c_0\right)^+. \qquad \square
\end{aligned}$$

LEMMA 5.3. *Let $\varphi \in \mathrm{C}^1(\bar{\mathcal{S}})$ and $s_0 \in \bar{\mathcal{S}} \setminus \partial^z\mathcal{S}$. Assume that*

$$\min\left\{\mathbf{g^b}\cdot D\varphi,\ \mathbf{g}^{\mathbf{s}}_\varepsilon \cdot D\varphi\right\}(s_0) > 0.$$

Then, there exist $\eta, \gamma > 0$ such that for $s = (x,y,k) \in B(s_0,\eta) \cap \bar{\mathcal{S}}$ and $s' := s - \ell\mathbf{g^b}\eta - m[(1-\alpha)y + \alpha f^\varepsilon(s)k]\mathbf{g}^{\mathbf{s}}_\varepsilon \in B(s_0,\eta) \cap \bar{\mathcal{S}}$ with $\ell, m \ge 0$,

$$\varphi(s) - \varphi(s') \ge \gamma\ell + \gamma m\ \left[(1-\alpha)y + \alpha f^\varepsilon(s)k\right].$$

Proof. We first observe that $\|\mathbf{g}^{\mathbf{s}}_\varepsilon\|_\infty < \infty$. In view of the definition of $\mathbf{g}^{\mathbf{s}}_\varepsilon$, this follows from

$$0 \le \frac{kf^\varepsilon(s)}{(1-\alpha)y + \alpha k f^\varepsilon(s)} \le \frac{k}{(1-\alpha)y + \alpha k} \le \frac{1}{\alpha},$$

where we used the inequality $f^\varepsilon \le 1$. Set

$$4\gamma := \min\left\{\mathbf{g^b}\cdot D\varphi\,;\, \mathbf{g}^{\mathbf{s}}_\varepsilon \cdot D\varphi\right\}(s_0) > 0.$$

Since $\mathbf{g}^{\mathbf{s}}_\varepsilon$ and $D\varphi$ are continuous on $\bar{\mathcal{S}} \setminus \partial^z\mathcal{S}$, there exists some $\eta > 0$ such that for all $s, s' \in B(s_0,\eta) \cap \bar{\mathcal{S}}$,

$$\begin{aligned}
&\text{(i)} \quad \min\left\{\mathbf{g^b}\cdot D\varphi,\ \mathbf{g}^{\mathbf{s}}_\varepsilon \cdot D\varphi\right\}(s) > 2\gamma, \\
&\text{(ii)} \quad |D\varphi(s) - D\varphi(s')| \le \frac{\gamma}{\|\mathbf{g}^{\mathbf{s}}_\varepsilon\|_\infty}.
\end{aligned}$$

Let s and s' be as in the statement of the lemma. By the mean value theorem, there exists some $s^* \in [s, s'] \subset B(s_0, \eta) \cap \bar{\mathcal{S}}$ such that

$$\begin{aligned}
\varphi(s) - \varphi(s') &= (s - s') \cdot D\varphi(s^*) \\
&= \ell\, \mathbf{g^b} \cdot D\varphi(s^*) + m\left[(1-\alpha)y + \alpha f^\varepsilon(s)k\right] \mathbf{g}^{\mathbf{s}}_\varepsilon(s) \cdot D\varphi(s^*) \\
&= \ell\, \mathbf{g^b} \cdot D\varphi(s^*) + m\left[(1-\alpha)y + \alpha f^\varepsilon(s)k\right] \mathbf{g}^{\mathbf{s}}_\varepsilon(s) \cdot D\varphi(s) \\
&\quad - m\left[(1-\alpha)y + \alpha f^\varepsilon(s)k\right] \mathbf{g}^{\mathbf{s}}_\varepsilon(s) \cdot \left[D\varphi(s) - D\varphi(s^*)\right] \\
&\geq \ell\, \mathbf{g^b} \cdot D\varphi(s^*) + m\left[(1-\alpha)y + \alpha f^\varepsilon(s)k\right] \mathbf{g}^{\mathbf{s}}_\varepsilon(s) \cdot D\varphi(s) \\
&\quad - m\left[(1-\alpha)y + \alpha f^\varepsilon(s)k\right] \|\mathbf{g}^{\mathbf{s}}_\varepsilon\|_\infty |D\varphi(s) - D\varphi(s^*)|. \\
&\geq \ell\, 2\gamma + m\left[(1-\alpha)y + \alpha f^\varepsilon(s)k\right](2\gamma - \gamma) \\
&\geq \gamma\, \ell \,+\, \gamma\, m\left[(1-\alpha)y + \alpha f^\varepsilon(s)k\right]. \qquad \square
\end{aligned}$$

Proof of the subsolution property.

Step 1. For each $\varepsilon \geq 0$, the value function V_ε is bounded from above by V; see Proposition 6.2 below. We also recall from Proposition 4.5 in [5] that for every $s = (x, y, k) \in \bar{\mathcal{S}}$,

$$V(s) \leq V^0(x + (1-\mu)\alpha k, (1-\alpha)y),$$

where the function V^0, defined in [5], is continuous and satisfies

$$V^0(\bar{x}, \bar{y}) \;=\; 0 \text{ for all } (\bar{x}, \bar{y}) \in \mathbb{R}^2 \text{ such that } \bar{x} + (1-\mu)\bar{y} = 0.$$

It then follows that for each $\varepsilon \geq 0$, the lower semicontinuous envelope of V_ε satisfies $(V_\varepsilon)_* \leq 0$ on $\partial^z \mathcal{S}$.

Let $s_0 \in \bar{\mathcal{S}} \setminus \partial^z \mathcal{S}$ and $\varphi \in C^2\left(\bar{\mathcal{S}}\right)$ be such that

$$0 \;=\; (V^*_\varepsilon - \varphi)(s_0) = \max_{\bar{\mathcal{S}}} \left(V^*_\varepsilon - \varphi\right),$$

and assume to the contrary that

$$F_*\left(s_0, \varphi(s_0), D\varphi(s_0), D^2\varphi(s_0)\right) > 0.$$

Observe that the last inequality implies that $\tilde{U}\left(\varphi_x(s_0)\right) < \infty$ and therefore $\varphi_x(s_0) > 0$. Since $\varphi \in C^2(\bar{\mathcal{S}})$, we deduce from Lemmas 5.2 and 5.3 the existence of $\eta, \gamma, c_0 > 0$, with $B(s_0, \eta) \subset \bar{\mathcal{S}} \setminus \partial^z \mathcal{S}$, such that

$$\min\left\{-\mathcal{L}\varphi,\; \mathbf{g^b} \cdot D\varphi,\; \mathbf{g}^{\mathbf{s}}_\varepsilon \cdot D\varphi\right\}(s) \wedge \varphi_x(s) \;>\; 0, \tag{5.12}$$

$$\tilde{U}\left(\varphi_x(s)\right) - \left(U(c) - c\varphi_x(s)\right) \geq \gamma\left(c - c_0\right), \tag{5.13}$$

$$\varphi(s) - \varphi(s') \;\geq\; \gamma l + \gamma m\left[(1-\alpha)y + \alpha f^\varepsilon(s)k\right] \tag{5.14}$$

for all $s \in B(s_0, \eta) \cap \bar{\mathcal{S}}$, and $s' = s - \ell \mathbf{g^b}\eta - m\mathbf{g}^{\mathbf{s}}_\varepsilon \in B(s_0, \eta) \cap \bar{\mathcal{S}}$ for some $\ell, m \geq 0$.

Step 2. Let $\left(s_n = (x_n, y_n, k_n)\right)_{n \geq 1}$ be some sequence such that

$$\begin{aligned}
&\text{(i)} \quad s_n \;\in\; B\left(s_0, \frac{\eta}{2}\right), \\
&\text{(ii)} \quad s_n \underset{n\to\infty}{\longrightarrow} s_0, \\
&\text{(iii)} \quad \xi_n \;:=\; \left|V_\varepsilon(s_n) - V^*_\varepsilon(s_0)\right| \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

For each $n \geq 1$, there exists a strategy $\nu^n = (C^n, L^n, M^n) \in \mathcal{A}^\varepsilon(s_n)$ such that

$$V_\varepsilon(s_n) \leq \xi_n + \mathbb{E}\left[\int_0^\infty e^{-\beta t} U(C_t^n) dt\right].$$

Set $S^n = (X^n, Y^n, K^n) := S^{\varepsilon, s_n, \nu^n}$ for $n \geq 1$, and fix some finite positive time horizon $T > 0$. By the dynamic programming principle of Theorem 5.1,

$$V_\varepsilon(s_n) \leq \xi_n \; + \; \mathbb{E}\left[\int_0^{T\wedge\theta^n} e^{-\beta t} U(C_t^n) dt\right] \; + \; \mathbb{E}\left[e^{-\beta T\wedge\theta^n} V_\varepsilon\left(S^n_{T\wedge\theta^n}\right)\right],$$

where $\theta^n := \inf\{t \geq 0 \, : \, S_t^n \notin B(s_0, \eta)\}$. Since $V_\varepsilon \leq V_\varepsilon^* \leq \varphi$ on $\bar{\mathcal{S}} \setminus \partial^z \mathcal{S}$, and $\xi_n = |V_\varepsilon(s_n) - V_\varepsilon^*(s_0)| = |V_\varepsilon(s_n) - \varphi(s_0)|$, it follows that for all $n \geq 1$,

$$\varphi(s_0) \; - \; \mathbb{E}\left[e^{-\beta T\wedge\theta^n} \varphi\left(S^n_{T\wedge\theta^n}\right)\right] \leq 2\xi_n \; + \; \mathbb{E}\left[\int_0^{T\wedge\theta^n} e^{-\beta t} U(C_t^n) dt\right].$$

Notice that for all $n \geq 1$, the process $\left\{S_t^n \mathbf{1}_{[0, T\wedge\theta^n)}(t), t \geq 0\right\}$ is uniformly bounded; then the Itô formula provides

$$\begin{aligned}
2\,\xi_n \geq\; & \mathbb{E}\left[\int_0^{T\wedge\theta^n} e^{-\beta t}\left[-\mathcal{L}\varphi + \tilde{U}(\varphi_x) - \left(U(C_t^n) - C_t^n \varphi_x\right)\right](S_t^n) dt\right] \\
& + \mathbb{E}\left[\int_0^{T\wedge\theta^n} e^{-\beta t} \mathbf{g^b} \cdot D\varphi(S_t^n) dL_t^{nc}\right] \\
& + \mathbb{E}\left[\int_0^{T\wedge\theta^n} e^{-\beta t}\left[(1-\alpha) Y_t^n + \alpha f^\varepsilon(S_t^n) K_t^n\right]\left(\mathbf{g}_\varepsilon^{\mathbf{s}} \cdot D\varphi\right)(S_t^n) dM_t^{nc}\right] \\
& + \mathbb{E}\left[\sum_{0\leq t<T\wedge\theta^n} e^{-\beta t}\left(\varphi(S_{t-}^n) \; - \; \varphi(S_t^n)\right)\right],
\end{aligned}$$

where L^{nc} and M^{nc} denote the continuous part of L^n and M^n. Recall that φ satisfies (5.12), (5.13), and (5.14); then it follows from the previous inequality that

$$\begin{aligned}
2\,\xi_n \geq\; & \gamma\, e^{-\beta T}\, \mathbb{E}\left[(T\wedge\theta^n) \; + \; L^{nc}_{T\wedge\theta^n} \; + \; \int_0^{T\wedge\theta^n} \left[(1-\alpha) Y_t^n + \alpha f^\varepsilon\left(S_t^n\right) K_t^n\right] dM_t^{nc}\right] \\
& + \gamma e^{-\beta T}\, \mathbb{E}\left[\sum_{0\leq t<T\wedge\theta^n} \Delta L_t^n + \left[(1-\alpha) Y_{t-}^n + \alpha f^\varepsilon\left(S_{t-}^n\right) K_{t-}^n\right] \Delta M_t^n\right] \\
& + e^{-\beta T} \gamma \mathbb{E}\left[\int_0^{T\wedge\theta^n} \left(C_t^n - c_0\right)^+ dt\right], \\
\geq\; & \mathbb{E}[h^n(T\wedge\theta^n)],
\end{aligned}$$

where

$$\begin{aligned}
h^n(T\wedge\theta^n) = \gamma e^{-\beta T}\Bigg\{ & (T\wedge\theta^n) + L^n_{T\wedge\theta^n} + \int_0^{T\wedge\theta^n} \left[(1-\alpha) Y_{t-}^n + \alpha f^\varepsilon\left(S_{t-}^n\right) K_{t-}^n\right] dM_t^n \\
& + \int_0^{T\wedge\theta^n} \left(C_t^n - c_0\right)^+ dt\Bigg\}.
\end{aligned}$$

Step 3. To obtain a contradiction, we show that for a sufficiently small T, there is some constant m_* such that for large $n \geq 1$, $\mathbb{E}\left[h^n(T \wedge \theta^n)\right] \geq m_*$. The following argument is largely inspired from [22].

Step 3.1. We start by providing estimates for $|X^n - x_0|$, $|Y^n - y_0|$, and $|K^n - k_0|$. Fix some $n \geq 1$, and assume that n is sufficiently large so that $\xi_n \leq \eta/2$ holds. Let Λ be the process defined by: $\Lambda_t := (\rho - \frac{\sigma^2}{2})t + \sigma W_t$, and set

$$\Lambda_t^\star := \left|\rho - \frac{\sigma^2}{2}\right| t + \sigma\left(W_t^\star - W_{\star t}\right), \quad \text{where } W_t^\star := \max_{u\in[0,t]} W_u \text{ and } W_{\star t} := \min_{u\in[0,t]} W_u.$$

Since $d\left[Y_t^n e^{-\Lambda_t}\right] = e^{-\Lambda_t} dL_t^n - e^{-\Lambda_t} Y_{t-}^n dM_t^n$, we deduce by a direct calculation that

$$|Y_t^n - y_0| \leq |y_0 - y_n| + y_n|1 - e^{\Lambda_t}| + e^{\Lambda_t^\star} L_t^n + e^{\Lambda_t^\star} \int_0^t Y_{u-}^n dM_u^n. \tag{5.15}$$

The dynamics of the processes K^n and X^n are such that

$$|K_t^n - k_0| \leq |k_0 - k_n| + L_t^n + \int_0^t f^\varepsilon(S_{u-}^n) K_{u-}^n dM_u^n, \tag{5.16}$$

$$|X_t^n - x_0| \leq |x_n - x_0| + |x_n|\left(e^{rt} - 1\right) + e^{rt}\int_0^t e^{-ru} C_u^n du + e^{rt}(1+\lambda)\int_0^t e^{-ru} dL_u^n$$

$$+ e^{rt}\int_0^t e^{-ru}(1-\mu)\left[(1-\alpha)Y_{u-}^n + \alpha f^\varepsilon(S_{u-}^n) K_{u-}^n\right] dM_u^n. \tag{5.17}$$

Step 3.2. We have $|1 - e^{\Lambda_T}| \leq \max\left[e^{\Lambda_T^\star} - 1; 1 - e^{-\Lambda_T^\star}\right]$. Define the set

$$F_T := \left\{\omega \in \Omega \ : \ \max\left[e^{\Lambda_T^\star} - 1; 1 - e^{-\Lambda_T^\star}\right] \leq \min\left[1, \frac{\eta}{4(y_0+1)}\right]\right\}.$$

We claim that it is possible to choose the parameter $T > 0$ such that

$$\mathbb{P}(F_T) \geq \frac{1}{2}, \quad e^{rT} - 1 \leq \frac{\eta}{4(1+|x_0|)}, \quad \text{and} \quad e^{rT} \leq 2. \tag{5.18}$$

Indeed, Doob's maximal Martingale inequalities provide, for $\delta > 0$,

$$\mathbb{P}\{W_T^\star \geq \delta\} \leq \frac{1}{\delta^2}\,\mathbb{E}[W_T^\star]^2 \leq \frac{4}{\delta^2}\,\mathbb{E}[W_T]^2 = \frac{4\,T}{\delta^2}; \text{ similarly } \mathbb{P}\{W_{\star T} \leq \delta\} \leq \frac{4\,T}{\delta^2}.$$

Hence for all $\delta > 0$,

$$\mathbb{P}\{W_T^\star - W_{\star T} \geq \delta\} \leq \mathbb{P}\left\{W_T^\star \geq \frac{\delta}{2}\right\} + \mathbb{P}\left\{W_{\star T} \leq \frac{\delta}{2}\right\} \leq \frac{32\,T}{\delta^2}.$$

We now return to the estimates (5.15), (5.16), (5.17) and recall that $\xi_n \leq \eta/2$. Since T satisfies (5.18), the following inequalities (where A denotes some positive constant depending on (x_0, y_0, k_0)) hold $\mathbb{P}$-a.s. on the set F_T:

$$|X_T^n - x_0| \leq \eta/2 + \eta/4 + 2\int_0^T C_t^n dt + AL_T^n + A\int_0^T G^\varepsilon(S_{t-}^n) dM_t^n, \tag{5.19}$$

$$|Y_T^n - y_0| \leq \eta/2 + \eta/4 + AL_T^n + A\int_0^T G^\varepsilon(S_{t-}^n) dM_t^n, \tag{5.20}$$

$$|K_T^n - k_0| \leq \eta/2 + AL_T^n + A\int_0^T G^\varepsilon(S_{t-}^n) dM_t^n, \tag{5.21}$$

where

$$G^{\varepsilon}(s) \;:=\; (1-\alpha)y + \alpha f^{\varepsilon}(s)k \text{ for } s=(x,y,k)\in \bar{\mathcal{S}}\,.$$

Step 3.3. For ω in F_T, we consider the following cases.

Case 1. $\theta^n(\omega) \geq T$. Then, by the definition of $h^n(T\wedge\theta^n)$, we have $h^n(T\wedge\theta^n) \;\geq\; \gamma\; e^{-\beta T}T$.

Case 2. $\theta^n(\omega) < T$. Recall that S^n is càdlagˇ; then, by the definition of the stopping time θ^n, this happens when $S^n_{\theta^n}(w)\notin B(s_0,\eta]$, i.e.,

$$\max\Big[|X^n_{\theta^n(\omega)}(\omega)-x_0|;\quad |Y^n_{\theta^n(\omega)}(\omega)-y_0|;\quad |K^n_{\theta^n(\omega)}(\omega)-k_0|\Big] \;\geq\; \eta.$$

Subcase 2.1. $|X^n_{\theta^n(\omega)}(\omega)-x_0|\geq\eta$. It follows from (5.19) that at least one of the following inequalities holds:

$$\text{(i)}\quad \int_0^{\theta^n(\omega)} C^n_t dt \;\geq\; \eta/16 \quad\text{or}\quad \text{(ii)}\quad L^n_{\theta^n} + \int_0^{\theta^n(\omega)} G^{\varepsilon}(S^n_{t-})dM^n_t \;\geq\; \frac{\eta}{8\,A}.$$

In inequality (i),

$$\frac{\eta}{16} \leq \int_0^{\theta^n(\omega)} C^n_t dt \;\leq\; c_0 T \;+\; \int_0^{\theta^n(\omega)} (C^n_t - c_0)\,dt.$$

Since it is possible to choose T such that $c_0T \;\leq\; \frac{\eta}{32}$, it follows that

$$\frac{\eta}{16} \leq \frac{\eta}{32} \;+\; \int_0^{\theta^n(\omega)} (C^n_t - c_0)^+\,dt;$$

then $\eta/32 \;\leq\; \int_0^{\theta^n(\omega)} (C^n_t - c_0)^+\,dt$, and it follows that

$$h^n(T\wedge\theta^n) \geq \gamma e^{-\beta T}\;\int_0^{\theta^n(\omega)} (C^n_t - c_0)^+\,dt \;\geq\; \gamma e^{-\beta T}\frac{\eta}{32}.$$

In inequality (ii), it immediately follows that $h^n(T\wedge\theta^n) \;\geq\; \gamma e^{-\beta T}\frac{\eta}{8\,A}$.

Subcase 2.2. $|Y^n_{\theta^n(\omega)}-y_0| \;\geq\; \eta$. Then, it follows from inequality (5.20) that

$$\frac{\eta}{4} \leq A\;\left(L^n_{\theta^n(\omega)} + \int_0^{\theta^n} G^{\varepsilon}(S^n_{t-})dM^n_t\right),$$

and hence, $h^n(T\wedge\theta^n(\omega)) \;\geq\; \gamma e^{-\beta T}\frac{\eta}{4\,A}$.

Subcase 2.3. $|K^n_{\theta^n(\omega)}-k_0| \;\geq\; \eta$. By inequality (5.21) we see that in this case,

$$\frac{\eta}{2} \leq A\;\left(L^n_{\theta^n(\omega)} + \int_0^{\theta^n} G^{\varepsilon}(S^n_{t-})dM^n_t\right),$$

and hence, $h^n(T\wedge\theta^n(\omega)) \;\geq\; \gamma e^{-\beta T}\frac{\eta}{2\,A}$.

From the several cases discussed above, it follows that for $\mathbb{P}$-a.e. ω in F_T,

$$h^n(T\wedge\theta^n(\omega)) \geq m_\star \;:=\; \gamma\min\Big[T,\frac{\eta}{32},\frac{\eta}{8A}\Big],$$

and therefore, for T sufficiently small and large n,

$$\mathbb{E}\left[h^n(T\wedge\theta^n)\right] \geq \mathbb{E}\left[\mathbf{1}_{F_T}h^n(T\wedge\theta^n)\right] \;\geq\; m_\star\mathbb{P}(F_T) \;=\; \frac{m_\star}{2}. \qquad \square$$

Remark 3. Let $\mathcal{A}_0(s)$ be the subset of $\mathcal{A}(s)$ consisting of all controls $\nu=(C,L,M)$ with a Lebesgue absolutely continuous component M. Then, it is clear that the above derivation of the DPE is not altered by this additional restriction. Hence, the value problem of this new control problem coincides with V_ε by the comparison result of Theorem 3.1. The same comment holds if the component L is, or both components L and M are, restricted to be Lebesgue absolutely continuous.

6. The convergence result. We first derive a useful estimate.

LEMMA 6.1. *Let s be in $\bar{\mathcal{S}}$. Then for any $\varepsilon\geq 0$, $\mathcal{A}^\varepsilon(s)\subset\mathcal{A}(s)$, and for all $\nu\in\mathcal{A}(s)$ and $t\geq 0$,*

$$0 \;\leq\; Z_t^{0,s,\nu}-Z_t^{\varepsilon,s,\nu} \;\leq\; 4\varepsilon r\, Z_T^{0,s,\nu\,\star}e^{rt}, \quad \textit{where} \quad Z_t^{0,s,\nu\,\star} := \sup_{u\in[0,t]}|Z_u^{0,s,\nu}|.$$

Proof. Clearly the inclusion $\mathcal{A}^\varepsilon(s)\subset\mathcal{A}(s)$ follows from the inequality $Z^{0,s,\nu}\geq Z^{\varepsilon,s,\nu}$.

Step 1. We first prove that $Z^{\varepsilon,s,\nu}\leq Z^{0,\varepsilon,\nu}$ $\mathbb{P}$-a.s. To see this, we consider a sequence of stopping times $(\tau_n)_{n\geq 0}$ exhausting the jumps of the càdlàg process M, with $\tau_0=0$. The dynamics of the processes $K^{\varepsilon,s,\nu}$ and $K^{0,s,\nu}$ are such that

$$d\left(K^{\varepsilon,s,\nu}-K^{0,s,\nu}\right)_t = -\left(K^{\varepsilon,s,\nu}-K^{0,s,\nu}\right)_{t-}dM_t + \left[1-f^\varepsilon\left(S_{t-}^{\varepsilon,s,\nu}\right)\right]K_{t-}^{\varepsilon,s,\nu}dM_t.$$

Then, for all $n\geq 0$, we have $\mathbb{P}$-a.s. for $t\in[\tau_n,\tau_{n+1})$,

(6.1)
$$\begin{aligned} &K_t^{\varepsilon,s,\nu}-K_t^{0,s,\nu} \\ &\quad= e^{-(M_t^c-M_{\tau_n}^c)}\left(K_{\tau_n}^{\varepsilon,s,\nu}-K_{\tau_n}^{0,s,\nu}+\int_{\tau_n}^t e^{M_u^c-M_{\tau_n}^c}\left[1-f^\varepsilon(S_{u-}^{\varepsilon,s,\nu})\right]K_{u-}^{\varepsilon,s,\nu}dM_u\right).\end{aligned}$$

Since $1-f^\varepsilon\geq 0$, this implies that

(6.2)
$$\begin{aligned} K_t^{\varepsilon,s,\nu}-K_t^{0,s,\nu} &\geq e^{-(M_t^c-M_{\tau_n}^c)}(K_{\tau_n}^{\varepsilon,s,\nu}-K_{\tau_n}^{0,s,\nu}) \\ &= e^{-(M_t^c-M_{\tau_n}^c)}((K_{\tau_n-}^{\varepsilon,s,\nu}-K_{\tau_n-}^{0,s,\nu})(1-\Delta M_{\tau_n}) \\ &\qquad +[1-f^\varepsilon(S_{\tau_n-}^{\varepsilon,s,\nu})]K_{\tau_n-}^{\varepsilon,s,\nu}\Delta M_{\tau_n})\geq 0.\end{aligned}$$

Clearly, $Y^{\varepsilon,s,\nu}=Y^{0,s,\nu}$. Then

$$d\left(Z^{\varepsilon,s,\nu}-Z^{0,s,\nu}\right)_t = r\left(Z^{\varepsilon,s,\nu}-Z^{0,s,\nu}\right)_t dt - r(1-\mu)\alpha\left(K^{\varepsilon,s,\nu}-K^{0,s,\nu}\right)dt.$$

Since $Z_0^{\varepsilon,s,\nu}-Z_0^{0,s,\nu}=0$ and $K^{\varepsilon,s,\nu}\geq K^{0,s,\nu}$, this implies that

$$Z_t^{\varepsilon,s,\nu}-Z_t^{0,s,\nu} = -r(1-\mu)\alpha e^{rt}\int_0^t e^{-ru}\left(K_u^{\varepsilon,s,\nu}-K_u^{0,s,\nu}\right)du \;\leq\; 0. \tag{6.3}$$

Step 2. We next prove the second inequality. Observe that $[1-f_\varepsilon(s)]\,k\leq 2\varepsilon z$ for $s=(x,y,k)\in\bar{\mathcal{S}}$, where $z:=x+(1-\mu)\,[(1-\alpha)y+\alpha k]$. Together with (6.2) and (6.2) this shows that, for all $n\geq 0$ and $t\in[\tau_n,\tau_{n+1})$,

$$K_t^{\varepsilon,s,\nu}-K_t^{0,s,\nu}\leq 2\varepsilon\, e^{-(M_t^c-M_{\tau_n}^c)}\left(Z_{\tau_n-}^{\varepsilon,s,nu}+\int_{\tau_n}^t e^{M_u^c-M_{\tau_n}^c}Z_{u-}^{\varepsilon,s,\nu}dM_u\right).$$

Using the increase of M together with the fact that $Z^{\varepsilon,s,\nu} \le Z^{0,s,\nu}$, as shown in the first step of this proof, this provides

$$K_t^{\varepsilon,s,\nu} - K_t^{0,s,\nu} \le 2\varepsilon Z_t^{0,s,\nu\star}\, e^{-(M_t^c - M_{\tau_n}^c)} \left(1 + \int_{\tau_n}^t e^{M_u^c - M_{\tau_n}^c}\, dM_u\right) \le 4\varepsilon Z_t^{0,s,\nu\star}.$$

The required inequality is obtained by plugging this estimate into (6.3). ∎

PROPOSITION 6.2. *The sequence $(V_\varepsilon)_{\varepsilon>0}$ is nonincreasing and $V_\varepsilon \le V$.*

Proof. The inequality $V_\varepsilon \le V$ follows immediately from the fact that $\mathcal{A}^\varepsilon(s) \subset \mathcal{A}(s)$, as stated in Lemma 6.1. To prove that the sequence $(V_\varepsilon)_{\varepsilon>0}$ is nonincreasing, we shall prove that $\mathcal{A}^{\varepsilon_1}(s) \subset \mathcal{A}^{\varepsilon_2}(s)$ whenever $\varepsilon_1 \ge \varepsilon_2$. To do this, it is sufficient to prove that for any control $\nu = (C, L, M) \in \mathcal{A}_\varepsilon(s)$, the associated process $Z^\varepsilon := X^{\varepsilon,s,\nu} + (1-\mu)\left[(1-\alpha)Y^{\varepsilon,s,\nu} + \alpha K^{\varepsilon,s,\nu}\right]$ is nonincreasing with respect to ε. Recall that

$$f^\varepsilon(s) \;=\; h\left(\frac{k}{\varepsilon z}\right), \quad \text{where } z = x + (1-\mu)[(1-\alpha y + \alpha z],$$

and h is a smooth function. From Remark 3, we may restrict the process M to be absolutely continuous with respect to the Lebesgue measure, i.e., $M_t = \int_0^t m_u du$ for some $\mathbb{F}$-adapted process $\{m_t,\ t \ge 0\}$, as the restriction of the control M to this class produces the same value function V_ε.

Then, by classical results on the regularity of flows of stochastic differential equations (see, e.g., [16]), the processes Z^ε, $Y^\varepsilon := Y^{\varepsilon,s,\nu}$ and $K^\varepsilon := K^{\varepsilon,s,\nu}$ are differentiable in ε, and the processes

$$z_t^\varepsilon \;:=\; e^{-rt}\frac{\partial Z_t^\varepsilon}{\partial \varepsilon}, \quad y_t^\varepsilon \;:=\; \frac{\partial Y_t^\varepsilon}{\partial \varepsilon}, \quad k_t^\varepsilon \;:=\; e^{-rt}\frac{\partial K_t^\varepsilon}{\partial \varepsilon}$$

satisfy $y_t^\varepsilon = 0$ for all $t \ge 0$, $z_0^\varepsilon = k_0^\varepsilon = 0$, and solve the system of ODEs

$$\dot{z}_t^\varepsilon \;=\; -r\alpha k_t \quad \text{and} \quad \dot{k}_t^\varepsilon \;=\; a_t + b_t z_t - c_t k_t,$$

where

$$a_t \;:=\; \frac{(K_t^\varepsilon)^2}{\varepsilon Z_t^\varepsilon}\, h'\left(\frac{K_t^\varepsilon}{\varepsilon Z_t^\varepsilon}\right), \qquad b_t \;:=\; \frac{(K_t^\varepsilon)^2}{\varepsilon (Z_t^\varepsilon)^2}\, h'\left(\frac{K_t^\varepsilon}{\varepsilon Z_t^\varepsilon}\right),$$

and

$$e^{-rt} c_t := r + m_t \left[h\left(\frac{K_t^\varepsilon}{\varepsilon Z_t^\varepsilon}\right) + \frac{K_t^\varepsilon}{\varepsilon Z_t^\varepsilon}\, h'\left(\frac{K_t^\varepsilon}{\varepsilon Z_t^\varepsilon}\right)\right].$$

Differentiating once more with respect to the t-variable, we obtain the following second order differential equation for z^ε:

$$-\ddot{z}_t^\varepsilon \;-\; c_t\, \dot{z}_t^\varepsilon \;-\; r\alpha b_t\, z_t^\varepsilon \;-\; r\alpha a_t \;=\; 0 \quad \text{and} \quad \dot{z}_0^\varepsilon = z_0 = 0. \tag{6.4}$$

We now consider the function

$$\hat{z}_t \;:=\; -\frac{r\alpha}{\varepsilon} \int_0^t \int_0^u \frac{(K_t^\varepsilon)^2}{\varepsilon Z_t^\varepsilon}\, h'\left(\frac{K_t^\varepsilon}{\varepsilon Z_t^\varepsilon}\right) du\, dt \quad \text{for} \quad t \ge 0.$$

Since $\hat{z}_t \le 0$, $\dot{\hat{z}}_t \le 0$, $b_t \ge 0$, and $c \ge 0$, it follows that $\hat{z}_t$ is a supersolution of (6.4). By a standard comparison result, we deduce that $z_t^\varepsilon \le \hat{z}_t$, and therefore $z_t^\varepsilon \le 0$ for all $t \ge 0$. This completes the proof. ∎

Our final result states the convergence of V_ε towards V.

PROPOSITION 6.3. *The sequence $(V_\varepsilon)_{\varepsilon>0}$ is nonincreasing and converges towards V, as $\varepsilon \searrow 0$, uniformly on compact subsets of $\bar{\mathcal{S}}$.*

Proof. Let $(\nu^n = (C^n, L^n, M^n))_{n\geq 1}$ be a maximizing sequence of controls for $V(s)$:

$$V(s) - \frac{1}{n} \;\leq\; \mathbb{E}\left[\int_0^\infty e^{-\beta t} U(C_t^n)dt\right] \quad \text{for all} \quad n \geq 1.$$

By the monotone convergence theorem, we verify that

$$\mathbb{E}\left[\int_0^\infty e^{-\beta t} U(C_t^n)dt\right] = \lim_{T\to\infty} \mathbb{E}\left[\int_0^T e^{-\beta t} U(C_t^n)dt\right].$$

Then $V(s) - \frac{1}{2n} \leq \mathbb{E}[\int_0^{T^n} e^{-\beta t} U(C_t^n)dt]$ for some $T^n > 0$. By Lemma 6.1 we have $Z^{0,s,\nu}_{t\wedge T^n} \geq Z^{\varepsilon,s,\nu}_{t\wedge T^n} \geq Z^{0,s,\nu}_{t\wedge T^n} - 4r\varepsilon Z^{0,s,\nu\star}_{T^n}$ $\mathbb{P}$-a.s. for all $t \geq 0$. Then, the stopping times $\tau(\varepsilon, s, n) := \inf\{t \geq 0 : Z_t^{\varepsilon,s,\nu} \leq 0\}$, $\varepsilon \geq 0$, satisfy

$$\tau(0,s,n)\wedge T^n \geq \tau(\varepsilon,s,n)\wedge T^n \text{ and } \lim_{\varepsilon\to 0} \tau(\varepsilon,s,n)\wedge T^n = \tau(0,s,n)\wedge T^n \;\; \mathbb{P}\text{-a.s.}$$

Hence, by the monotone convergence theorem,

$$\lim_{\varepsilon\to 0} \mathbb{E}\left[\int_0^{\tau(\varepsilon,s,n)\wedge T^n} e^{-\beta t} U(C_t^n)dt\right] = \mathbb{E}\left[\int_0^{\tau(0,s,n)\wedge T^n} e^{-\beta u} U(C_t^n)dt\right].$$

Recall from (5.6) and (5.7) that

$$V_\varepsilon(s) \geq \mathbb{E}\left[\int_0^{\tau(\varepsilon,s,n)\wedge T^n} e^{-\beta t} U(C_t^n)dt\right] \quad \text{and}$$
$$\mathbb{E}\left[\int_0^{\tau(0,s,n)\wedge T^n} e^{-\beta t} U(C_t^n)dt\right] = \mathbb{E}\left[\int_0^{T^n} e^{-\beta u} U(C_u^n)du\right].$$

Then

$$\liminf_{\varepsilon\to 0} V_\varepsilon(s) \geq \mathbb{E}\left[\int_0^{T^n} e^{-\beta t} U(C_t^n)dt\right] \;\geq\; V(s) - \frac{1}{2n}.$$

By arbitrariness of $n \geq 1$, this provides $\liminf_{\varepsilon\to 0} V_\varepsilon(s) \geq V(s)$. Together with Proposition 6.2, this shows that $V_\varepsilon(s) \longrightarrow V(s)$ as $\varepsilon \searrow 0$ for every $s \in \bar{\mathcal{S}}$.

We finally recall from Proposition 2.4 that the limit function V is continuous. Since $(V_\varepsilon)_{\varepsilon>0}$ is a monotonic sequence of continuous functions, it follows from the Dini theorem that the convergence holds uniformly on compact subsets of $\bar{\mathcal{S}}$. □

Acknowledgments. The authors would like to thank the anonymous referees for insightful comments.

REFERENCES

[1] M. Akian, J. L. Menaldi, and A. Sulem (1996), *On an investment-consumption model with transaction costs*, SIAM J. Control Optim., 34, pp. 329–364.

[2] G. Barles and J. Burdeau (1995), *The Dirichlet problem for semilinear second-order degenerate elliptic equations and applications to stochastic exit time control problems*, Comm. Partial Differential Equations, 20, pp. 129–178.

[3] G. Barles and E. Rouy (1998), *A strong comparison result for the Bellman equation arising in stochastic exit time control problems and its applications*, Comm. Partial Differential Equations, 23, pp. 1945–2033.

[4] G. Barles and P. E. Souganidis (1991), *Convergence of approximation schemes for fully nonlinear equations*, Asymptotic Anal., 4, pp. 271–283.

[5] I. Ben Tahar, H. M. Soner, and N. Touzi (2005), *Modeling Continuous-Time Financial Markets with Capital Gains Taxes*, preprint. Available online at http://www.cmap.polytechnique.fr/∼touzi/bst06.pdf.

[6] V. S. Borkar (1989), *Optimal Control of Diffusion Processes*, Pitman Res. Notes Math. Ser. 203, Longman, Harlow, UK; Wiley, New York.

[7] G. M. Constantinides (1983), *Capital market equilibrium with personal taxes*, Econometrica, 51, pp. 611–636.

[8] J. Cox and C. F. Huang (1989), *Optimal consumption and portfolio policies when asset prices follow a diffusion process*, J. Econom. Theory, 49, pp. 33–83.

[9] M. G. Crandall, H. Ishii, and P. L. Lions (1992), *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27, pp. 1–67.

[10] R. M. Dammon, C. S. Spatt, and H. H. Zhang (2001), *Optimal consumption and investment with capital gains taxes*, The Review of Financial Studies, 14, pp. 583–616.

[11] V. DeMiguel and R. Uppal (2005), *Portfolio investment with the exact tax basis via nonlinear programming*, Management Sci., 51, pp. 277–290.

[12] P. Dybvig and H. Koo (1996), *Investment with Taxes*, Working paper, Washington University, Saint Louis, MO.

[13] N. El Karoui (1981), *Les aspects probabilistes du contrôle stochastique*, in Ninth Saint Flour Probability Summer School 1979, Lecture Notes in Math. 876, Springer-Verlag, Berlin, pp. 73–238.

[14] W. H. Fleming and H. M. Soner (1993), *Controlled Markov Processes and Viscosity Solutions*, Appl. Math. 25, Springer-Verlag, New York.

[15] M. Gallmeyer, R. Kaniel, and S. Tompaidis (2006), *Tax management strategies with multiple risky assets*, J. Financial Econom., 80, pp. 243–291.

[16] N. Ikeda and S. Watanabe (1989), *Stochastic Differential Equations and Diffusion Processes*, 2nd ed., North–Holland, Amsterdam.

[17] E. Jouini, P.-F. Koehl, and N. Touzi (1997), *Optimal investment with taxes: An optimal control problem with endogeneous delay*, Nonlinear Anal. Theory Methods Appl., 37, pp. 31–56.

[18] E. Jouini, P.-F. Koehl, and N. Touzi (2000), *Optimal investment with taxes: An existence result*, J. Math. Econom., 33, pp. 373–388.

[19] H. E. Leland (1999), *Optimal Portfolio Management with Transactions Costs and Capital Gains Taxes*, preprint, Institute of Business and Economic Research, University of California at Berkeley, Berkeley, CA. Available online at http://repositories.cdlib.org/iber/finance/RPF-290/

[20] R. C. Merton (1969), *Lifetime portfolio selection under uncertainty: The continuous-time model*, Rev. Econ. Statist., 51, pp. 247–257.

[21] R. C. Merton (1971), *Optimum consumption and portfolio rules in a continuous-time model*, J. Econom. Theory, 3, pp. 373–413.

[22] S. E. Shreve and H. M. Soner (1994), *Optimal investment and consumption with transaction costs*, Ann. Appl. Probab., 4, pp. 609–692.

[23] H. M. Soner (1986), *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24, pp. 552–561.

[24] H.M. Soner (1986), *Optimal control with state-space constraint* II, SIAM J. Control Optim., 24, pp. 1110–1122.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1802–1830

EXACT INTERNAL CONTROLLABILITY FOR THE TWO-DIMENSIONAL MAGNETOHYDRODYNAMIC EQUATIONS*

TEODOR HAVÂRNEANU†, CĂTĂLIN POPA†, AND S. S. SRITHARAN‡

Abstract. In this paper we establish the local exact internal controllability for the two-dimensional magnetohydrodynamic equations. The needed Carleman estimate for the adjoint linearized magnetohydrodynamic equations is also obtained here.

Key words. magnetohydrodynamic equations, controllability, Stokes equations, dynamo equations, observability inequality, Carleman estimates

AMS subject classifications. 35Q35, 76W05, 76D55, 35Q30, 35Q60, 93B05, 93C20, 93B07

DOI. 10.1137/040611884

1. Introduction.

In this paper we study the internal controllability for the magnetohydrodynamic (MHD) equations in two-dimensional bounded domains. We show that the final value (state) of any sufficiently smooth solution (evolution) of the MHD system can be attained by starting from initial states which are "close" enough to the initial state of the target solution and acting in both fluid and magnetic parts of the MHD equations by appropriate locally distributed internal controls. The corresponding three-dimensional result was established by the authors in [7]. The first but less general controllability result for the MHD equations in three-dimensional domains was obtained in [2] (see also [3]). We emphasize here that despite their similarities the two-dimensional and three-dimensional cases are different; they are not particular cases of a more general situation.

As in the three-dimensional case, we shall reduce the local controllability problem for the MHD equations to the global controllability problem for the linearized MHD equations by means of an infinite-dimensional version of the implicit function theorem. This approach is inspired from [9] (see also [5] and [8]), where the related but simpler case of the internal controllability of the Navier–Stokes equations is studied.

To solve the global controllability problem, we approximate it by a family of ad hoc optimal control problems for the same linearized MHD system. The estimates that we need to prove the convergence of the approximation procedure are obtained by using an observability inequality for the adjoint linearized MHD equations. Such an inequality is usually derived from a Carleman inequality for the same equations. For this reason, the main effort here will be directed toward obtaining the required Carleman inequality for the adjoint linearized MHD system. The strategy described here is nothing more than that used in [9] for the case of the Navier–Stokes equations.

Let us mention that two new Carleman inequalities—one for elliptic equations with nonhomogeneous Dirichlet boundary conditions and the other for the Stokes equations, established in [10] and [4], respectively—suggest that our controllability

*Received by the editors July 19, 2004; accepted for publication (in revised form) March 27, 2007; published electronically November 14, 2007.
http://www.siam.org/journals/sicon/46-5/61188.html

†Facultatea de Matematică, Universitatea "Al.I.Cuza," Bdul. Carol I, Nr. 11, 700506 Iaşi, Romania (havi@uaic.ro, cpopa@uaic.ro). The second author's research was supported by ONRIFO under grant N00014-03-1-4017.

‡Department of Mathematics, University of Wyoming, Laramie, WY 82071 (Sri@uwyo.edu). The work of this author was supported by the Army Research Office, Probability and Statistics Program.

result for the MHD system can still be improved. More specifically, it is expected to be able to prove that less regular target solutions for the MHD equations can be attained by local internal action if we use the main result in [10] and adapt the approach in [4] to our situation.

2. Functional framework and main result. Let Ω be a bounded multiconnected open set in $\mathbb{R}^2$ whose boundary $\partial\Omega$ is a finite union of mutually disjoint closed curves of class C^2. Such a set can be made simply connected with a finite number of smooth cuts. This means that there exist p mutually disjoint curves $\Gamma_1, \dots, \Gamma_p$ of class C^2 which are not tangent to $\partial\Omega$ such that $\Omega \setminus (\cup_{i=1}^{p} \Gamma_i)$ is simply connected. Let $T > 0$ be fixed. We set $Q = \Omega \times (0,T)$. We also fix an open subset ω of Ω. The controlled MHD equations (with boundary and initial conditions) we consider are the following:

$$
\begin{aligned}
&\frac{\partial y}{\partial t} - \nu\Delta y + (y\cdot\nabla)y + \nabla p + \nabla\left(\frac{1}{2}\ B^2\right) - (B\cdot\nabla)B \\
&\qquad = f + \chi_\omega u && \text{in } Q,\\
&\frac{\partial B}{\partial t} + \eta\,\widetilde{\operatorname{curl}}(\operatorname{curl} B) + (y\cdot\nabla)B - (B\cdot\nabla)y = P(\chi_\omega v) && \text{in } Q,\\
&\operatorname{div} y = 0,\ \operatorname{div} B = 0 && \text{in } Q,\\
&y = 0,\ B\cdot N = 0,\ \operatorname{curl} B = 0 && \text{on } \Sigma = \partial\Omega\times(0,T),\\
&y(\cdot,0) = y_0,\ B(\cdot,0) = B_0 && \text{in } \Omega.
\end{aligned}
\tag{2.1}
$$

Here $y = (y_1, y_2) : \Omega \times [0,T] \longrightarrow \mathbb{R}^2$ is the velocity vector field, $p : \Omega \times [0,T] \longrightarrow \mathbb{R}$ is the pressure, and $B = (B_1, B_2) : \Omega \times [0,T] \longrightarrow \mathbb{R}^2$ is the magnetic field. System (2.1) is controlled through the vector functions $u = (u_1, u_2) : \Omega \times [0,T] \longrightarrow \mathbb{R}^2$ and $v = (v_1, v_2) : \Omega \times [0,T] \longrightarrow \mathbb{R}^2$. The variables of the functions (fields) y, p, B, u, and v are denoted by $x = (x_1, x_2)$ and t (belonging to Ω and $[0,T]$, respectively). The other symbols in (2.1) denote known (given) quantities (or objects). So, ν and η are the kinematic viscosity and magnetic resistivity, which are supposed to be positive. From now on, for the sake of simplicity and without loss of generality, ν and η will be assumed to be 1. Further, $f = (f_1, f_2) : \Omega \times [0,T] \longrightarrow \mathbb{R}^2$ is the density of the external forces, χ_ω is the characteristic function of ω, P is the Leray projector (put there to "kill" the gradient part of $\chi_\omega v$), and $y_0 : \Omega \longrightarrow \mathbb{R}^2$ and $B_0 : \Omega \longrightarrow \mathbb{R}^2$ are the initial velocity and magnetic fields. The operators curl and $\widetilde{\operatorname{curl}}$ are defined as follows:

$$
\begin{aligned}
&\operatorname{curl} B = \frac{\partial B_2}{\partial x_1} - \frac{\partial B_1}{\partial x_2} && \text{for every vector function } B = (B_1, B_2),\\
&\widetilde{\operatorname{curl}}\, w = \left(\frac{\partial w}{\partial x_2}, -\frac{\partial w}{\partial x_1}\right) && \text{for every scalar function } w.
\end{aligned}
$$

It is well known that $\widetilde{\operatorname{curl}}(\operatorname{curl} B) = -\Delta B + \operatorname{grad}(\operatorname{div} B)$. Finally, N is the unit outer normal to $\partial\Omega$.

When Ω is not simply connected (but it remains multiconnected), to assure the well-posedness of problem (2.1), we have to impose the following additional conditions on B on the cuts Γ_i:

$$
\int_{\Gamma_i} B\cdot N\, d\sigma = 0 \ \text{ in } (0,T), \quad i = 1, \dots, p,
\tag{2.2}
$$

where N is the unit outer normal to Γ_i's. (See Appendix I of [11] for an equivalent form of (2.2).) When Ω is simply connected, conditions (2.2) are no longer necessary.

For the statement of our results and the considerations which follow, some function spaces are required. For each positive integer m and $p > 1$, or $p = +\infty$, we denote (as usual) by $W^{m,p}(\Omega)$ the Sobolev space of functions in $L^p(\Omega)$ whose weak derivatives of order less than or equal to m are also in $L^p(\Omega)$. When $p = 2$, we set $H^1(\Omega) = W^{1,2}(\Omega)$. In a similar way, $H^{2,1}(Q)$ is the space of all functions in $L^2(Q)$ whose first and second order weak derivatives with respect to x_1 and x_2, and first order weak derivative with respect to t, are all in $L^2(Q)$, too. The fractional order Sobolev space $H^{1/2}(\Omega)$ will also be required by a certain argument. Since all the functions involved in equations (2.1) (except p) are actually vector functions (fields), we mostly use some product function spaces: $(L^2(Q))^2$, $(H^1(\Omega))^2$, $(W^{1,\infty}(\Omega))^2$, $(W^{2,p}(\Omega))^2$, $(H^{2,1}(Q))^2$, etc., all endowed with the product norms. The time-dependent function space $L^2(0,T;H^1(\Omega))$ contains all (equivalence classes of) measurable functions from $(0,T)$ to $H^1(\Omega)$ having the square of their $H^1(\Omega)$ norm integrable over $(0,T)$. The spaces $L^\infty(0,T;(W^{1,\infty}(\Omega))^2)$ and $L^\infty(0,T;(W^{2,p}(\Omega))^2)$ are defined similarly. Finally, we need the corresponding Sobolev spaces $W^{1,\infty}(0,T;(W^{1,\infty}(\Omega))^2)$ and $W^{1,\infty}(0,T;(W^{2,p}(\Omega))^2)$ containing the functions in $L^\infty(0,T;(W^{1,\infty}(\Omega))^2)$ and $L^\infty(0,T;(W^{2,p}(\Omega))^2)$ whose first order weak derivatives are in $L^\infty(0,T;(W^{1,\infty}(\Omega))^2)$ and $L^\infty(0,T;(W^{2,p}(\Omega))^2)$, respectively. The norms of all the considered spaces are denoted in the same manner: $|\cdot|_{(L^2(Q))^2}$, $|\cdot|_{(H^1(\Omega))^2}$, $|\cdot|_{(H^{2,1}(Q))^2}$, etc.

The natural functional framework for the MHD equations (2.1) is given by the space H of all weakly divergence-free vector functions in $(L^2(\Omega))^2$ which are tangential to the boundary in a weak sense, endowed with the $(L^2(\Omega))^2$ norm. Our considerations here require the following two $(H^1(\Omega))^2$ versions (subspaces) of H:

$$V_1 = \{y \in (H^1(\Omega))^2 : \operatorname{div} y = 0 \ \text{ in } \Omega \ \text{ and } \ y = 0 \ \text{ on } \partial\Omega\}$$

and

$$V_2 = \{B \in (H^1(\Omega))^2 : \operatorname{div} B = 0 \ \text{ in } \Omega \ \text{ and } \ B \cdot N = 0 \ \text{ on } \partial\Omega\}$$

if Ω is simply connected, or

$$V_2 = \{B \in (H^1(\Omega))^2 : \operatorname{div} B = 0 \ \text{ in } \Omega,\ B \cdot N = 0 \ \text{ on } \partial\Omega,$$
$$\text{and } \int_{\Gamma_i} B \cdot N \, d\sigma = 0,\ i = 1, \dots, p\}$$

if Ω is not simply connected (but it is multiconnected).

A specific feature of the MHD system (which distinguishes it from the Navier–Stokes equations) is the fact that the left-hand side of the second equation in (2.1) is in H (as a function of x) if (y, B, p) is a *strong* solution of (2.1), that is, if $y, B \in (H^{2,1}(Q))^2$, $y \in V_1$, $B \in V_2$ (as functions of x), and $\operatorname{curl} B = 0$ on Σ. Indeed, let $\phi \in H^1(\Omega)$, arbitrary. Using a Green-type formula, we have

$$\int_\Omega \widetilde{\operatorname{curl}}(\operatorname{curl} B) \cdot \nabla\phi \, dx$$
$$= \int_\Omega \operatorname{curl} B \operatorname{curl} \nabla\phi \, dx - \int_{\partial\Omega} \operatorname{curl} B (\nabla\phi \cdot \mathtt{T}) d\sigma = 0,$$

because $\operatorname{curl}\nabla\phi = 0$ in Ω and $\operatorname{curl} B = 0$ on $\partial\Omega$. Here $\mathtt{T}$ represents the unit tangent vector $(-N_2, N_1)$ to $\partial\Omega$, where $(N_1, N_2) = N$. Integrating by parts, we also have

$$\begin{aligned}\int_\Omega ((y\cdot\nabla)B - (B\cdot\nabla)y)\cdot\nabla\phi\,dx \\ = -\sum_{i,j=1}^{2}\int_\Omega \frac{\partial y_i}{\partial x_j}\,\frac{\partial B_j}{\partial x_i}\,\phi\,dx - \int_\Omega y\cdot\nabla(\operatorname{div} B)dx + \int_{\partial\Omega}\phi((y\cdot\nabla)B)\cdot N\,d\sigma \\ + \sum_{i,j=1}^{2}\int_\Omega \frac{\partial B_i}{\partial x_j}\,\frac{\partial y_j}{\partial x_i}\,\phi\,dx + \int_\Omega B\cdot\nabla(\operatorname{div} y)dx - \sum_{j=1}^{2}\int_{\partial\Omega}\phi N_j B\cdot\nabla y_j\,d\sigma = 0,\end{aligned}$$

because $y \in V_1$ and $B \in V_2$. (As $y = 0$ on $\partial\Omega$, the vectors $\nabla y_1(x), \nabla y_2(x)$, and $N(x)$ have the same directions for those $x \in \partial\Omega$ at which $\nabla y_1(x)$ and $\nabla y_2(x)$ are different from the vector 0.) Since the left-hand side of (2.1) is in H, the right-hand side must be in H, too. This is the reason of the presence of the Leray projector P before the control parameter $\chi_\omega v$. So, the gradient-type term which is generally associated with an equation like $\operatorname{div} B = 0$ is zero in the second equation in (2.1). For more information about the mathematical setting and analysis of the MHD equations one can consult the bibliography in [2].

Let us fix $\widetilde{y}, \widetilde{p}$, and $\widetilde{B}$ that satisfy both MHD equations and boundary conditions in (2.1):

$$\begin{aligned}&\frac{\partial\widetilde{y}}{\partial t} - \Delta\widetilde{y} + (\widetilde{y}\cdot\nabla)\widetilde{y} + \nabla\widetilde{p} + \nabla\left(\frac{1}{2}\,\widetilde{B}^2\right) - (\widetilde{B}\cdot\nabla)\widetilde{B} = f && \text{in } Q,\\ &\frac{\partial\widetilde{B}}{\partial t} + \widetilde{\operatorname{curl}}(\operatorname{curl}\widetilde{B}) + (\widetilde{y}\cdot\nabla)\widetilde{B} - (\widetilde{B}\cdot\nabla)\widetilde{y} = 0 && \text{in } Q,\\ &\operatorname{div}\widetilde{y} = 0,\ \operatorname{div}\widetilde{B} = 0 && \text{in } Q,\\ &\widetilde{y} = 0,\ \widetilde{B}\cdot N = 0,\ \operatorname{curl}\widetilde{B} = 0 && \text{on } \Sigma.\end{aligned} \tag{2.3}$$

Now we can state the controllability result we have described before.

THEOREM 2.1. *Let Ω be an open, bounded, and multiconnected subset of $\mathbb{R}^2$ whose boundary $\partial\Omega$ is a finite union of mutually disjoint closed curves of class C^2, and let ω be an open subset of Ω. Let $f \in (L^2(Q))^2$ and let $\widetilde{y}, \widetilde{B} \in W^{1,\infty}(0,T;(W^{2,p}(\Omega))^2)$, with $p > 2$, satisfy equations* (2.3). *Then there is $\eta > 0$ such that for any $(y_0, B_0) \in V_1 \times V_2$ which satisfy*

$$|y_0 - \widetilde{y}(\cdot,0)|_{(H^1(\Omega))^2} + |B_0 - \widetilde{B}(\cdot,0)|_{(H^1(\Omega))^2} \le \eta$$

there exist $(u, v) \in (L^2(Q))^4$ and $(y, B, p) \in (H^{2,1}(Q))^4 \times L^2(0,T;H^1(\Omega))$ that satisfy (2.1), (2.2), *and*

$$y(x,T) = \widetilde{y}(x,T),\ B(x,T) = \widetilde{B}(x,T) \quad \textit{a.e. } x \in \Omega.$$

The proof of Theorem 2.1 follows the same lines as the proof of the corresponding three-dimensional result (Theorem 2.1 in [7]). For this reason, we only outline it (see [7] and [2] for details) but emphasize those points where some differences (caused by the different boundary conditions for B and the lower dimension) appear.

3. Carleman inequality for the adjoint linearized MHD equations. Let $\widetilde{y}$ and $\widetilde{B}$ satisfy (2.3). The adjoint linearized (around $\widetilde{y}$ and $\widetilde{B}$) MHD equations are the following:

$$
\begin{aligned}
&\frac{\partial z}{\partial t} + \Delta z + (\widetilde{y}\cdot\nabla)z - z\cdot(\nabla\widetilde{y}) \\
&\qquad - (\widetilde{B}\cdot\nabla)C - C\cdot(\nabla\widetilde{B}) + \nabla q = h && \text{in } Q,\\
&\frac{\partial C}{\partial t} - \widetilde{\operatorname{curl}}(\operatorname{curl} C) + P((\widetilde{y}\cdot\nabla)C + C\cdot(\nabla\widetilde{y}) \\
&\qquad - (\widetilde{B}\cdot\nabla)z + z\cdot(\nabla\widetilde{B})) = H && \text{in } Q,\\
&\operatorname{div} z = 0,\ \operatorname{div} C = 0 && \text{in } Q,\\
&z = 0,\ C\cdot N = 0,\ \operatorname{curl} C = 0 && \text{on } \Sigma.
\end{aligned}
\tag{3.1}
$$

Here $h : \Omega\times[0,T] \longrightarrow \mathbb{R}^2$ and $H : \Omega\times[0,T] \longrightarrow \mathbb{R}^2$ are two given vector functions, and $z\cdot(\nabla\widetilde{y})$ is the vector field of components $z\cdot\partial\widetilde{y}/\partial x_i$, $i = 1, 2$; $C\cdot(\nabla\widetilde{B})$, $C\cdot(\nabla\widetilde{y})$, and $z\cdot(\nabla\widetilde{B})$ are defined in the same way.

The presence of the Leray projector P in (3.1) comes from the fact that the sum $(\widetilde{y}\cdot\nabla)C + C\cdot(\nabla\widetilde{y}) - (\widetilde{B}\cdot\nabla)z + z\cdot(\nabla\widetilde{B})$ is not generally in the space H (defined in the preceding section) but the other terms in the second equation in (3.1) are there (see Theorem 3.1). The Leray projector can be replaced by the gradient of a pseudopressure r in the following way: For any solution (z, q, C) of (3.1) there exists some function $r : \Omega\times[0,T] \longrightarrow \mathbb{R}$ such that

$$
\begin{aligned}
&(\widetilde{y}\cdot\nabla)C + C\cdot(\nabla\widetilde{y}) - (\widetilde{B}\cdot\nabla)z + z\cdot(\nabla\widetilde{B}) \\
&\qquad = P((\widetilde{y}\cdot\nabla)C + C\cdot(\nabla\widetilde{y}) - (\widetilde{B}\cdot\nabla)z + z\cdot(\nabla\widetilde{B})) + \nabla r.
\end{aligned}
\tag{3.2}
$$

To express the Carleman inequality for equations (3.1), some suitable weight functions are needed. Let us fix an open subset ω_0 of ω such that $\omega_0 \subset\subset \omega$. Since Ω is bounded and connected, there exist functions $\psi \in C^2(\overline{\Omega})$ such that

$$
\psi > 0 \ \text{ in } \Omega, \quad \psi = 0 \ \text{ on } \partial\Omega, \quad \text{and} \quad |\nabla\psi| > 0 \ \text{ in } \overline{\Omega}\setminus\omega_0.
$$

Let us fix such a function ψ, too. We set

$$
\varphi(x,t) = \frac{e^{\lambda\psi(x)}}{(t(T-t))^8} \quad \text{and} \quad \alpha(x,t) = \frac{e^{\lambda\psi(x)} - e^{2\lambda|\psi|_{C(\overline{\Omega})}}}{(t(T-t))^8}
$$

for $\lambda > 0$. We denote by $\widehat{\varphi}$ and $\widehat{\alpha}$ the values taken by φ and α on the boundary $\partial\Omega$ (where $\psi = 0$):

$$
\widehat{\varphi}(t) = \frac{1}{(t(T-t))^8} \quad \text{and} \quad \widehat{\alpha}(t) = \frac{1 - e^{2\lambda|\psi|_{C(\overline{\Omega})}}}{(t(T-t))^8}\,\cdot
$$

For establishing the Carleman inequality two variants of φ and α are also needed:

$$
\overline{\varphi}(x,t) = \frac{e^{-\lambda\psi(x)}}{(t(T-t))^8} \quad \text{and} \quad \overline{\alpha}(x,t) = \frac{e^{-\lambda\psi(x)} - e^{2\lambda|\psi|_{C(\overline{\Omega})}}}{(t(T-t))^8}\,\cdot
$$

We set $Q_\omega = \omega \times (0,T)$ and $Q_{\omega_0} = \omega_0 \times (0,T)$. Now we are prepared to present the Carleman inequality for equations (3.1).

THEOREM 3.1. *Let Ω be an open, bounded, and multiconnected subset of $\mathbb{R}^2$ whose boundary $\partial\Omega$ is a finite union of mutually disjoint closed curves of class C^2, and let ω, ω_0, and ω_1 be open subsets of Ω such that $\omega_0 \subset\subset \omega_1 \subset\subset \omega$. We set $Q_{\omega_1} = \omega_1 \times (0,T)$. Let $\widetilde{y}, \widetilde{B} \in L^\infty(0,T;(W^{2,p}(\Omega))^2)$ with $p > 2$. Then there exists $\lambda_0 > 0$ such that for any $\lambda > \lambda_0$ one can find $s_0(\lambda) > 0$ and $c(\lambda) > 0$ that for $s > s_0(\lambda)$ the following inequality holds:*

$$
\begin{aligned}
&\int_Q e^{2s\alpha} \left(\frac{1}{s\varphi} \left(\left| \frac{\partial z}{\partial t} \right|^2 + \left| \frac{\partial C}{\partial t} \right|^2 + \sum_{i,j=1}^{2} \left(\left| \frac{\partial^2 z}{\partial x_i \partial x_j} \right|^2 + \left| \frac{\partial^2 C}{\partial x_i \partial x_j} \right|^2 \right) \right) \right. \\
&\qquad \left. + s\varphi(|\nabla z|^2 + |\nabla C|^2) + s^3\varphi^3(|z|^2 + |C|^2) \right) dx\,dt \\
&\le c(\lambda) \left(\int_{Q_\omega} e^{2s\alpha} s^3 \varphi^3 (|z|^2 + |C|^2) dx\,dt + \int_{Q_{\omega_1}} e^{2s\alpha} s^{\frac{11}{4}} \widehat{\varphi}^{\frac{11}{4}} (q^2 + r^2) dx\,dt \right. \\
&\qquad \left. + \int Q(e^{2s\alpha} + e^{2s\widehat{\alpha}} s^{\frac{3}{4}} \widehat{\varphi}^{\frac{3}{4}})(|h|^2 + |H|^2) dx\,dt \right)
\end{aligned} \tag{3.3}
$$

for all $h, H \in (L^2(Q))^2$ which satisfy $\operatorname{div} h = \operatorname{div} H = 0$ in Q and $H \cdot N = 0$ on Σ, and all corresponding solutions $(z, C, q) \in (H^{2,1}(Q))^4 \times L^2(0,T;H^1(\Omega))$ of system (3.1) and $r \in L^2(0,T;H^1(\Omega))$ which satisfy (3.2).

To make the proof of Theorem 3.1 easier to follow for the reader, we divide it into several intermediate statements, which will be taken under consideration by turns and then put together at the end. First we present two Carleman-type inequalities for the adjoint Stokes equations with the null Dirichlet boundary condition, that is,

$$
\begin{aligned}
&\frac{\partial z}{\partial t} + \Delta z + \nabla q = g && \text{in } Q, \\
&\operatorname{div} z = 0 && \text{in } Q, \\
&z = 0 && \text{on } \Sigma,
\end{aligned} \tag{3.4}
$$

and for the adjoint equations of certain dynamo-type equations with solutions having null curl at the boundary,

$$
\begin{aligned}
&\frac{\partial C}{\partial t} - \widetilde{\operatorname{curl}}(\operatorname{curl} C) = PG && \text{in } Q, \\
&\operatorname{div} C = 0 && \text{in } Q, \\
&C \cdot N = 0,\ \operatorname{curl} C = 0 && \text{on } \Sigma.
\end{aligned} \tag{3.5}
$$

Here $g : \Omega \times [0,T] \longrightarrow \mathbb{R}^2$ and $G : \Omega \times [0,T] \longrightarrow \mathbb{R}^2$ are two given vector functions.

THEOREM 3.2. *Let Ω, ω, ω_0, and ω_1 be open subsets of $\mathbb{R}^2$ as in the statement of Theorem 3.1. Then there exists $\lambda_0 > 0$ such that for any $\lambda > \lambda_0$ one can find*

$s_0(\lambda) > 0$ and $c(\lambda) > 0$ such that for $s > s_0(\lambda)$ we have

$$\int_Q e^{2s\alpha}\left(\frac{1}{s\varphi}\left(\left|\frac{\partial z}{\partial t}\right|^2+\sum_{i,j=1}^{2}\left|\frac{\partial^2 z}{\partial x_i\partial x_j}\right|^2\right)+s\varphi|\nabla z|^2+s^3\varphi^3|z|^2\right)dx\,dt$$

$$\le c(\lambda)\Bigg(\int_{Q_\omega} e^{2s\alpha}s^3\varphi^3|z|^2dx\,dt+\int_{Q_{\omega_1}} e^{2s\alpha}s^{\frac{11}{4}}\widehat{\varphi}^{\frac{11}{4}}q^2dx\,dt \tag{3.6}$$

$$+\int_Q(e^{2s\alpha}+e^{2s\widehat{\alpha}}s^{\frac{3}{4}}\widehat{\varphi}^{\frac{3}{4}})|g|^2dx\,dt+\int_Q e^{2s\alpha}s^{\frac{1}{2}}\widehat{\varphi}^{\frac{1}{2}}|\operatorname{div} g|^2dx\,dt\Bigg)$$

for all $g \in L^2(0,T;(H^1(\Omega))^2)$ and all corresponding solutions $(z,q) \in (H^{2,1}(Q))^2 \times L^2(0,T;H^1(\Omega))$ of system (3.4).

The analogous result for equations (3.5) is as follows.

THEOREM 3.3. *Let Ω, ω, ω_0, and ω_1 be open subsets of $\mathbb{R}^2$ as in the statement of Theorem* 3.1. *Then there exists $\lambda_0 > 0$ such that for any $\lambda > \lambda_0$ one can find $s_0(\lambda) > 0$ and $c(\lambda) > 0$ such that for $s > s_0(\lambda)$ we have*

$$\int_Q e^{2s\alpha}\left(\frac{1}{s\varphi}\left(\left|\frac{\partial C}{\partial t}\right|^2+\sum_{i,j=1}^{2}\left|\frac{\partial^2 C}{\partial x_i\partial x_j}\right|^2\right)+s\varphi|\nabla C|^2+s^3\varphi^3|C|^2\right)dx\,dt$$

$$\le c(\lambda)\Bigg(\int_{Q_\omega} e^{2s\alpha}s^3\varphi^3|C|^2dx\,dt+\int_{Q_{\omega_1}} e^{2s\alpha}s^{\frac{11}{4}}\widehat{\varphi}^{\frac{11}{4}}r^2dx\,dt \tag{3.7}$$

$$+\int_Q(e^{2s\alpha}+e^{2s\widehat{\alpha}}s^{\frac{3}{4}}\widehat{\varphi}^{\frac{3}{4}})|G|^2dx\,dt+\int_Q e^{2s\alpha}s^{\frac{1}{2}}\widehat{\varphi}^{\frac{1}{2}}|\operatorname{div} G|^2dx\,dt\Bigg)$$

for all $G \in L^2(0,T;(H^1(\Omega))^2)$ and all corresponding solutions $C \in (H^{2,1}(Q))^2$ of system (3.5) *and $r \in L^2(0,T;H^1(\Omega))$ which satisfy*

$$G = PG + \nabla r. \tag{3.8}$$

To prove Theorems 3.2 and 3.3, we shall couple two kinds of estimates. First we establish Carleman inequalities for (3.4) and (3.5), viewed as parabolic systems in the unknowns z and C, respectively (so, ∇q and ∇r are passed in the right-hand side near g and G). Since q and r satisfy Poisson equations (obtained by applying the divergence operator to both sides of (3.4) and (3.5)), the needed estimates for ∇q and ∇r are derived by using an adequate Carleman inequality for elliptic equations with nonhomogeneous Dirichlet boundary conditions (obtained by Imanuvilov in [9]). So, let us begin by presenting the Carleman estimates for (3.4) and (3.5), viewed as parabolic systems in z and C.

LEMMA 3.1. *Let Ω be an open, bounded, and connected subset of $\mathbb{R}^2$ having the boundary $\partial\Omega$ of class C^2, and let ω be an open subset of Ω. Then there exists $\lambda_0 > 0$ such that for any $\lambda > \lambda_0$ one can find $s_0(\lambda) > 0$ and $c(\lambda) > 0$ such that for $s > s_0(\lambda)$ we have*

$$\int_Q e^{2s\alpha}\left(\frac{1}{s\varphi}\left(\left|\frac{\partial z}{\partial t}\right|^2+\sum_{i,j=1}^{2}\left|\frac{\partial^2 z}{\partial x_i\partial x_j}\right|^2\right)+s\varphi|\nabla z|^2+s^3\varphi^3|z|^2\right)dx\,dt \tag{3.9}$$

$$\le c(\lambda)\left(\int_{Q_\omega} e^{2s\alpha}s^3\varphi^3|z|^2dx\,dt+\int_Q e^{2s\alpha}|\nabla q|^2dx\,dt+\int_Q e^{2s\alpha}|g|^2dx\,dt\right)$$

for all $g \in (L^2(Q))^2$ *and all corresponding solutions* $(z, q) \in (H^{2,1}(Q))^2 \times L^2(0, T; H^1(\Omega))$ *of system* (3.4).

Estimate (3.9) was essentially obtained by Imanuvilov in [8]. (We refer the reader to Lemma 2.2 in [9] or Lemma 3.1 in [6], too.)

The analogous estimate for equations (3.5) is contained in the following statement.

LEMMA 3.2. *Let* Ω *and* ω *be open subsets of* $\mathbb{R}^2$ *as in the statement of Lemma* 3.1. *Then there exists* $\lambda_0 > 0$ *such that for any* $\lambda > \lambda_0$ *one can find* $s_0(\lambda) > 0$ *and* $c(\lambda) > 0$ *such that for* $s > s_0(\lambda)$ *we have*

$$\begin{aligned} &\int_Q e^{2s\alpha} \left(\frac{1}{s\varphi} \left(\left| \frac{\partial C}{\partial t} \right|^2 + \sum_{i,j=1}^{2} \left| \frac{\partial^2 C}{\partial x_i \partial x_j} \right|^2 \right) + s\varphi |\nabla C|^2 + s^3 \varphi^3 |C|^2 \right) dx\, dt \\ &\qquad \leq c(\lambda) \left(\int_{Q_\omega} e^{2s\alpha} s^3 \varphi^3 |C|^2 dx\, dt + \int_Q e^{2s\alpha} |\nabla r|^2 dx\, dt + \int_Q e^{2s\alpha} |G|^2 dx\, dt \right) \end{aligned} \tag{3.10}$$

for all $G \in (L^2(Q))^2$ *and all corresponding solutions* $C \in (H^{2,1}(Q))^2$ *of system* (3.5) *and* $r \in L^2(0, T; H^1(\Omega))$ *which satisfy* (3.8).

Proof. Inequality (3.10) can be established in almost the same way as its three-dimensional analogue. (See inequality (3.58) in [2] together with its proof.) However, because of the different (though similar) boundary conditions here, and for the reader's convenience, we shall sketch the proof, emphasizing the points where some differences appear.

Let us first describe in a few words the idea behind the proof. The part of the solutions of equations (3.5) outside Q_ω can be removed in the right-hand side of the estimates by varying (increasing) two parameters s and λ introduced into equations by performing a suitable change of unknown function. So, let us set $D = e^{s\alpha} C$. Passing to D in (3.5), we obtain the following system:

$$\begin{aligned} &\frac{\partial D}{\partial t} + \Delta D + s^2 \lambda^2 \varphi^2 |\nabla \psi|^2 D - 2s\lambda\varphi (\nabla\psi \cdot \nabla) D \\ &\qquad - s\lambda^2 \varphi |\nabla\psi|^2 D - s\lambda\varphi \Delta\psi D - s \frac{\partial \alpha}{\partial t} D = e^{s\alpha} PG \quad \text{in } Q, \\ &\operatorname{div} D = s\lambda\varphi (\nabla\psi \cdot D) \quad \text{in } Q, \\ &D \cdot N = 0, \ \operatorname{curl} D = s\lambda\varphi |\nabla\psi| D \cdot \mathtt{T} \quad \text{on } \Sigma, \\ &D(\cdot, 0) = D(\cdot, T) = 0 \quad \text{in } \Omega, \end{aligned} \tag{3.11}$$

where $\mathtt{T}$ is the unit tangent vector $(-N_2, N_1)$ to $\partial\Omega$.

If we set

$$P(x,t)D = -\Delta D - s^2\lambda^2\varphi^2 |\nabla\psi|^2 D - s\lambda^2\varphi |\nabla\psi|^2 D + s\lambda\varphi\Delta\psi D + s \frac{\partial\alpha}{\partial t} D,$$

$$R(x,t)D = -2s\lambda\varphi (\nabla\psi \cdot \nabla) D - 2s\lambda^2\varphi |\nabla\psi|^2 D,$$

we may write equations (3.11) as

$$\frac{\partial D}{\partial t} + R(x,t)D - P(x,t)D = e^{s\alpha} PG \ \text{ in } Q. \tag{3.12}$$

Multiplying (3.12) by itself, integrating over Q, and neglecting some nonnegative terms, we obtain the inequality

$$I \leq J + \frac{1}{2} \int_Q e^{2s\alpha} |PG|^2 dx\, dt, \tag{3.13}$$

where

$$I = -\int_Q P(x,t)D \cdot R(x,t)D \, dx \, dt, \quad J = \int_Q \frac{\partial D}{\partial t} \cdot P(x,t)D \, dx \, dt.$$

Multiplication of the five terms of $P(x,t)D$ by the two terms of $R(x,t)D$ in (3.13) produces ten terms (integrals): $I = \sum_{i=1}^{10} I_i$. Eight of them (I_3 through I_{10}) can be estimated identically as the corresponding terms in the three-dimensional case. We refer the reader to [2] (see also [6]) for the expression of those estimates and for other details. So, let us examine the other two terms (which contain ΔD).

Integrating by parts twice and using the fact that $N = (-1/|\nabla\psi|)\nabla\psi$, we obtain

$$\begin{aligned} I_1 &= -2s\lambda \int_Q \varphi(\nabla\psi \cdot \nabla)D \cdot \Delta D \, dx \, dt \\ &\geq -s\lambda^2 \int_Q \varphi |\nabla\psi|^2 |\nabla D|^2 dx \, dt - cs\lambda \int_Q \varphi |\nabla D|^2 dx \, dt \\ &\quad + 2s\lambda \int_\Sigma \varphi |\nabla\psi| \sum_{i=1}^{2} \left(\frac{\partial D_i}{\partial N} \right)^2 d\sigma \, dt - s\lambda \int_\Sigma \varphi |\nabla\psi| \, |\nabla D|^2 d\sigma \, dt, \end{aligned} \tag{3.14}$$

where c (here and throughout this proof) denotes a positive constant depending on ψ only.

Using Green's formula, we have (after some calculation)

$$\begin{aligned} I_2 &= -2s\lambda^2 \int_Q \varphi |\nabla\psi|^2 D \cdot \Delta D \, dx \, dt \\ &\geq \frac{3}{2} \, s\lambda^2 \int_Q \varphi |\nabla\psi|^2 |\nabla D|^2 dx \, dt - cs\lambda^4 \int_Q \varphi |D|^2 dx \, dt \\ &\quad - 2s\lambda^2 \int_\Sigma \varphi |\nabla\psi|^2 \sum_{i,j=1}^{2} D_i \frac{\partial D_i}{\partial x_j} N_j \, d\sigma \, dt \quad \text{for } \lambda \geq 1. \end{aligned} \tag{3.15}$$

(See [2] or [6] for more details.) Let us now estimate the surface integral in (3.15). Using first the second boundary condition in (3.11) and then observing that $D \cdot \nabla(D \cdot \nabla\psi) = 0$ on Σ (thanks to the first boundary condition there), we obtain

$$\begin{aligned} -\sum_{i,j=1}^{2} D_i \frac{\partial D_i}{\partial x_j} N_j &= (\operatorname{curl} D) D \cdot \mathtt{T} - \sum_{i,j=1}^{2} D_i \frac{\partial D_j}{\partial x_i} N_j \\ &= s\lambda\varphi |\nabla\psi| (D \cdot \mathtt{T})^2 - |\nabla\psi|^{-1} \sum_{i,j=1}^{2} D_i D_j \frac{\partial^2 \psi}{\partial x_i \partial x_j} \quad \text{on } \Sigma. \end{aligned} \tag{3.16}$$

Thus, we have

$$\begin{aligned} &-2s\lambda^2 \int_\Sigma \varphi |\nabla\psi|^2 \sum_{i,j=1}^{2} D_i \frac{\partial D_i}{\partial x_j} N_j \, d\sigma \, dt \\ &\geq 2s^2\lambda^3 \int_\Sigma \varphi^2 |\nabla\psi|^3 |D|^2 d\sigma \, dt - cs\lambda^2 \sum_{i=1}^{2} \int_0^T \widehat{\varphi}(t) \left(\int_{\partial\Omega} D_i^2 d\sigma \right) dt. \end{aligned} \tag{3.17}$$

To estimate the integral over $\partial\Omega$ in (3.17), we apply the trace theorem and an interpolation inequality:

$$
\begin{aligned}
s\lambda^2 \int_{\partial\Omega} D_i^2 d\sigma &\le c_1 s\lambda^2 |D_i|^2_{H^{\frac{1}{2}}(\Omega)} \le c_2 s\lambda^2 |D_i|_{L^2(\Omega)} |D_i|_{H^1(\Omega)} \\
&\le c_3 \left(s^{\frac{3}{2}} \lambda^3 |D_i|^2_{L^2(\Omega)} + s^{\frac{1}{2}} \lambda |D_i|^2_{L^2(\Omega)} + s^{\frac{1}{2}} \lambda |\nabla D_i|^2_{L^2(\Omega)} \right).
\end{aligned} \tag{3.18}
$$

Now, inserting first (3.18) into (3.17) and then (3.17) into (3.15), we obtain

$$
\begin{aligned}
I_2 \ge{}& \frac{3}{2}\, s\lambda^2 \int_Q \varphi |\nabla\psi|^2 |\nabla D|^2 dx\, dt \\
&- c \left(s\lambda^4 \int_Q \varphi |D|^2 dx\, dt + s^{\frac{3}{2}} \lambda^3 \int_Q \varphi |D|^2 dx\, dt + s^{\frac{1}{2}} \lambda \int_Q \varphi |\nabla D|^2 dx\, dt \right) \\
&+ 2s^2\lambda^3 \int_\Sigma \varphi^2 |\nabla\psi|^3 |D|^2 d\sigma\, dt \quad \text{for } \lambda \ge 1 \text{ and } s \ge 1.
\end{aligned} \tag{3.19}
$$

We next estimate J. Using successively two Green-type formulas, the initial, final, and boundary conditions in (3.11), and, finally, performing an integration by parts with respect to t, we have

$$
\begin{aligned}
&-\int_Q \Delta D \cdot \frac{\partial D}{\partial t}\, dx\, dt \\
&\quad = \int_Q \widetilde{\operatorname{curl}}(\operatorname{curl} D) \cdot \frac{\partial D}{\partial t}\, dx\, dt - \int_Q \nabla(\operatorname{div} D) \cdot \frac{\partial D}{\partial t}\, dx\, dt \\
&\quad = \frac{1}{2} \int_Q \frac{\partial}{\partial t} (\operatorname{curl} D)^2 dx\, dt - \int_\Sigma \operatorname{curl} D\ \frac{\partial D}{\partial t} \cdot \mathtt{T}\, d\sigma\, dt \\
&\qquad + \frac{1}{2} \int_Q \frac{\partial}{\partial t} (\operatorname{div} D)^2 dx\, dt - \int_\Sigma \operatorname{div} D\ \frac{\partial}{\partial t} (D \cdot N) d\sigma\, dt \\
&\quad = s\lambda \int_\Sigma \varphi |\nabla\psi| (D \cdot \mathtt{T})\ \frac{\partial}{\partial t}\ (D \cdot \mathtt{T}) d\sigma\, dt = -\frac{1}{2}\ s\lambda \int_\Sigma \widehat{\varphi}' |\nabla\psi|\, |D|^2 d\sigma\, dt.
\end{aligned}
$$

Putting this last form of $-\int_Q \Delta D \cdot (\partial D/\partial t) dx\, dt$ into the expression of J, we obtain

$$
J \le -\frac{1}{2}\ s\lambda \int_\Sigma \widehat{\varphi}' |\nabla\psi|\, |D|^2 d\sigma\, dt + c(\lambda) s^2 \int_Q \varphi^{\frac{17}{8}} |D|^2 dx\, dt \quad \text{for } \lambda \ge 1,\ s \ge 1, \tag{3.20}
$$

where $c(\lambda)$ is a positive parameter depending on ψ, T, and λ (see [2] or [6] for more details).

To be able to eliminate the surface integrals in (3.14), (3.19), and (3.20), as well as those in the estimates for $I_3, \dots, I_{10}$, we need to repeat all the previous considerations for $\overline{D} = e^{s\overline{\alpha}} C$. Changing C by $\overline{D}$ in (3.5), we obtain (3.11) where, in all places, λ is replaced by $-\lambda$. So, we have

$$
\frac{\partial \overline{D}}{\partial t} + \overline{R}(x,t)\overline{D} - \overline{P}(x,t)\overline{D} = e^{s\overline{\alpha}} PG \ \text{ in } Q,
$$

where the operators $\overline{P}$ and $\overline{Q}$ are defined as P and Q but with $-\lambda$ instead of λ; consequently, $\overline{\varphi}$ and $\overline{\alpha}$ now replace φ and α there. In the same way as before, we have

$$
\overline{I} \le \overline{J} + \frac{1}{2} \int_Q e^{2s\overline{\alpha}} |PG|^2 dx\, dt, \tag{3.21}
$$

where

$$\overline{I} = -\int_Q \overline{P}(x,t)\overline{D}\cdot\overline{R}(x,t)\overline{D}\,dx\,dt, \qquad \overline{J} = \int_Q \frac{\partial \overline{D}}{\partial t}\cdot\overline{P}(x,t)\overline{D}\,dx\,dt.$$

We write $\overline{I}$ as a sum of ten terms, too: $\overline{I} = \sum_{i=1}^{10} \overline{I}_i$. Each of these terms can be estimated in the same way as its corresponding I_i. For $\overline{I}_1$ and $\overline{I}_2$ we obtain the inequalities

$$\begin{aligned}\overline{I}_1 \geq & -s\lambda^2\int_Q \overline{\varphi}|\nabla\psi|^2|\nabla\overline{D}|^2dx\,dt - cs\lambda\int_Q \overline{\varphi}|\nabla\overline{D}|^2dx\,dt \\ & - 2s\lambda\int_\Sigma \varphi|\nabla\psi|\sum_{i=1}^{2}\left(\frac{\partial\overline{D}_i}{\partial N}\right)^2 d\sigma\,dt - s\lambda\int_\Sigma \varphi|\nabla\psi|\,|\nabla\overline{D}|^2 d\sigma\,dt,\end{aligned} \tag{3.22}$$

$$\begin{aligned}\overline{I}_2 \geq & \frac{3}{2}\,s\lambda^2\int_Q \overline{\varphi}|\nabla\psi|^2|\nabla\overline{D}|^2dx\,dt \\ & - c\left(s\lambda^4\int_Q \overline{\varphi}|\overline{D}|^2dx\,dt + s^{\frac{3}{2}}\lambda^3\int_Q \overline{\varphi}|\overline{D}|^2dx\,dt + s^{\frac{1}{2}}\lambda\int_Q \overline{\varphi}|\nabla\overline{D}|^2dx\,dt\right) \\ & - 2s^2\lambda^3\int_\Sigma \varphi^2|\nabla\psi|^2|D|^2d\sigma\,dt\end{aligned} \tag{3.23}$$

for all $\lambda \geq 1$ and $s \geq 1$. Notice that we have taken into account the fact that $\overline{\varphi} = \varphi$ and $\overline{D} = D$ on Σ (because $\psi = 0$ on $\partial\Omega$). In the same way as the estimate for J was derived, we obtain

$$\overline{J} \leq \frac{1}{2}\,s\lambda\int_\Sigma \widehat{\varphi}'|\nabla\psi|\,|D|^2d\sigma\,dt + c(\lambda)s^2\int_Q \overline{\varphi}^{\frac{17}{8}}|\overline{D}|^2dx\,dt \quad \text{for } \lambda \geq 1,\ s \geq 1. \tag{3.24}$$

Now we add inequalities (3.13) and (3.21). Because of the different signs before the corresponding surface integrals in $I_2, \overline{I}_2, I_3, \overline{I}_3, I_5, \overline{I}_5, I_9$, and $\overline{I}_9$ (see [2] or [6] for the definition of the other quantities), all those surface integrals are canceled. (The different signs come from the odd exponents of the powers of λ.) It remains to see how the four surface integrals in $I_1 + \overline{I}_1$ can be removed. By careful calculations involving both (3.16) and (3.18), we obtain

$$\begin{aligned}& 2s\lambda\int_\Sigma \varphi|\nabla\psi|\sum_{i=1}^{2}\left(\frac{\partial D_i}{\partial N}\right)^2 d\sigma\,dt - s\lambda\int_\Sigma \varphi|\nabla\psi|\,|\nabla D|^2d\sigma\,dt \\ & \quad - 2s\lambda\int_\Sigma \varphi|\nabla\psi|\sum_{i=1}^{2}\left(\frac{\partial\overline{D}_i}{\partial N}\right)^2 d\sigma\,dt + s\lambda\int_\Sigma \varphi|\nabla\psi|\,|\nabla\overline{D}|^2d\sigma\,dt \\ & = -4s^2\lambda^2\int_\Sigma \varphi^2|\nabla\psi|\sum_{i,j=1}^{2}\frac{\partial^2\psi}{\partial x_i\partial x_j}\,D_iD_j\,d\sigma\,dt \\ & \geq -c\left(s^3\lambda^3\int_Q \varphi^3|D|^2dx\,dt + s\lambda\int_Q \varphi|\nabla D|^2dx\,dt\right).\end{aligned} \tag{3.25}$$

(We refer the reader to [2] or [6] for all the details.) So, adding (3.13) and (3.21) and using the estimates (3.14), (3.19), (3.20), and (3.22) through (3.25), as well as

the estimates for $I_3, \dots, I_{10}$ and $\overline{I}_3, \dots, \overline{I}_{10}$ (which can be taken from [2] or [6]), we obtain

$$\begin{aligned} & s\lambda^2 \int_Q \varphi |\nabla \psi|^2 |\nabla D|^2 dx\, dt + s^3\lambda^4 \int_Q \varphi^3 |\nabla \psi|^4 |D|^2 dx\, dt \\ & \quad \le cs\lambda \int_Q \varphi |\nabla D|^2 dx\, dt + cs^3\lambda^3 \int_Q \varphi^3 |D|^2 dx\, dt \\ & \quad + c(\lambda) s^2 \int_Q \varphi^3 |D|^2 dx\, dt + \int_Q e^{2s\alpha} |PG|^2 dx\, dt \\ & \qquad\qquad \text{for } \lambda \ge 1,\ s \ge 1. \end{aligned} \tag{3.26}$$

The key point is now coming. Since, by the third property of ψ, $|\nabla \psi| \ge \rho$ in $\overline{\Omega} \setminus \omega_0$ for some $\rho > 0$, from (3.26), we have

$$\begin{aligned} & \rho^2 s\lambda^2 \int_{Q \setminus Q_{\omega_0}} \varphi |\nabla D|^2 dx\, dt + \rho^4 s^3\lambda^4 \int_{Q \setminus Q_{\omega_0}} |D|^2 dx\, dt \\ & \quad \le cs\lambda \int_Q \varphi |\nabla D|^2 dx\, dt + cs^3\lambda^3 \int_Q \varphi^3 |D|^2 dx\, dt \\ & \quad + c(\lambda) s^2 \int_Q \varphi^3 |D|^2 dx\, dt + \int_Q e^{2s\alpha} |PG|^2 dx\, dt \\ & \qquad\qquad \text{for } \lambda \ge 1,\ s \ge 1. \end{aligned} \tag{3.27}$$

As the powers of s and λ before integrals of $\varphi|\nabla D|^2$ and $\varphi^3|D|^2$ are greater in the left-hand side of (3.27) than in the right-hand side, we can remove the part of $\varphi|\nabla D|^2$ and $\varphi^3|D|^2$ outside Q_{ω_0} in the right-hand side by simply increasing the parameters λ and s. So, taking first $\lambda > \lambda_0 = (c+1)\max(\rho^{-2}, \rho^{-4})$ and then $s > s_0(\lambda) = c(\lambda)\lambda^{-3}(\rho^4\lambda - c - 1)^{-1}$, where c and $c(\lambda)$ are those in (3.27), we obtain

$$\begin{aligned} & s\lambda \int_Q \varphi |\nabla D|^2 dx\, dt + s^3\lambda^3 \int_Q \varphi^3 |D|^2 dx\, dt \\ & \quad \le c(\lambda) \left(s \int_{Q_{\omega_0}} \varphi |\nabla D|^2 dx\, dt + s^3 \int_{Q_{\omega_0}} \varphi^3 |D|^2 dx\, dt + \int_Q e^{2s\alpha} |PG|^2 dx\, dt \right) \\ & \qquad\qquad \text{for } \lambda > \lambda_0,\ s > s_0(\lambda). \end{aligned}$$

Coming back to C ($D = e^{s\alpha} C$), we can rewrite the above inequality as

$$\begin{aligned} & s\lambda \int_Q e^{2s\alpha} \varphi |\nabla C|^2 dx\, dt + s^3\lambda^3 \int_Q e^{2s\alpha} \varphi^3 |C|^2 dx\, dt \\ & \quad \le c(\lambda) \left(s \int_{Q_{\omega_0}} e^{2s\alpha} \varphi |\nabla C|^2 dx\, dt + s^3 \int_{Q_{\omega_0}} e^{2s\alpha} \varphi^3 |C|^2 dx\, dt + \int_Q e^{2s\alpha} |PG|^2 dx\, dt \right) \\ & \qquad\qquad \text{for } \lambda > \lambda_0 \text{ and } s > s_0(\lambda), \end{aligned} \tag{3.28}$$

where λ_0 is possibly greater than that before.

The integral of $e^{2s\alpha}\varphi|\nabla C|^2$ in the right-hand side of (3.28) can be eliminated as

in [2]. So, we have

$$
\begin{aligned}
&s\lambda\int_Q e^{2s\alpha}\varphi|\nabla C|^2dx\,dt + s^3\lambda^3\int_Q e^{2s\alpha}\varphi^3|C|^2dx\,dt\\
&\qquad\le c(\lambda)\left(s^3\int_{Q_{\omega_1}} e^{2s\alpha}\varphi^3|C|^2dx\,dt + \int_Q e^{2s\alpha}|PG|^2dx\,dt\right)\\
&\qquad\qquad\text{for } \lambda>\lambda_0 \text{ and } s>s_0(\lambda),
\end{aligned}
\tag{3.29}
$$

where ω_1 is an open subset of ω such that $\omega_0 \subset\subset \omega_1 \subset\subset \omega$.

It remains to show how estimates such as (3.29) can also be obtained for the first and second order derivatives of C with respect to time and space variables, respectively. To be able to estimate the weighted L^2 norm of $\partial C/\partial t$ in (3.10), we scalarly multiply (3.5) by $e^{2s\alpha}\varphi^{-1}\partial C/\partial t$. Thus, the desired estimate is reduced to that of the integral of $e^{2s\alpha}\varphi^{-1}\partial C/\partial t\cdot\widetilde{\operatorname{curl}}(\operatorname{curl} C)$ over Q. Using a version of Green's formula together with the boundary conditions in (3.5) and then integrating by parts with respect to t, we obtain

$$
\begin{aligned}
&-\int_Q e^{2s\alpha}\varphi^{-1}\ \frac{\partial C}{\partial t}\cdot\widetilde{\operatorname{curl}}(\operatorname{curl} C)dx\,dt\\
&\quad=-\int_Q \operatorname{curl}\left(e^{2s\alpha}\varphi^{-1}\ \frac{\partial C}{\partial t}\right)\operatorname{curl} C\,dx\,dt+\int_\Sigma e^{2s\alpha}\varphi^{-1}\operatorname{curl} C\ \frac{\partial C}{\partial t}\cdot \mathtt{T}\,d\sigma\,dt\\
&\quad=\frac{1}{2}\int_Q e^{2s\alpha}\left(2s\varphi^{-1}\ \frac{\partial\alpha}{\partial t}-\varphi^{-2}\ \frac{\partial\varphi}{\partial t}\right)|\operatorname{curl} C|^2dx\,dt\\
&\qquad+\lambda\int_Q e^{2s\alpha}(2s-\varphi^{-1})\ \operatorname{curl} C\ \widetilde{\operatorname{curl}}\,\psi\cdot\frac{\partial C}{\partial t}\ dx\,dt.
\end{aligned}
$$

Since $|\operatorname{curl} C|^2\le 2|\nabla C|^2$, the last relation leads to the inequality

$$
\int_Q e^{2s\alpha}(s\varphi)^{-1}\left|\frac{\partial C}{\partial t}\right|^2dx\,dt\le c(\lambda)\left(s\int_Q e^{2s\alpha}\varphi|\nabla C|^2dx\,dt+\int_Q e^{2s\alpha}|PG|^2dx\,dt\right)
$$

$$
\text{for } s\ge 1.
$$

This together with (3.29) yields

$$
\begin{aligned}
&\int_Q e^{2s\alpha}(s\varphi)^{-1}\left|\frac{\partial C}{\partial t}\right|^2dx\,dt\\
&\qquad\le c(\lambda)\left(s^3\int_{Q_{\omega_1}} e^{2s\alpha}\varphi^3|C|^2dx\,dt+\int_Q e^{2s\alpha}|PG|^2dx\,dt\right)\\
&\qquad\qquad\text{for } \lambda>\lambda_0 \text{ and } s>s_0(\lambda).
\end{aligned}
\tag{3.30}
$$

Now multiplying (3.5) by $e^{2s\alpha}\varphi^{-1}\Delta C$ and integrating over Q, after some arrangements involving (3.30), we obtain

$$
\begin{aligned}
&\int_Q e^{2s\alpha}(s\varphi)^{-1}|\Delta C|^2dx\,dt\\
&\qquad\le c(\lambda)\left(s^3\int_{Q_{\omega_1}} e^{2s\alpha}\varphi^3|C|^2dx\,dt+\int_Q e^{2s\alpha}|PG|^2dx\,dt\right)\\
&\qquad\qquad\text{for } \lambda>\lambda_0 \text{ and } s>s_0(\lambda).
\end{aligned}
\tag{3.31}
$$

An estimate similar to (3.31) (whose derivation is based on it) can be obtained for $\sum_{i,j=1}^{2} \left|\partial^2 C/\partial x_i \partial x_j\right|^2$, too. (We refer the reader to [2] for more details.) Taking this last estimate together with (3.30) and (3.31), we obtain (3.10), which finishes the proof. □

As we have already mentioned, q in (3.4) and r satisfying (3.8) are solutions to some Poisson equations. Indeed, applying the divergence operator to both sides of (3.4) and (3.5), we see that q and r satisfy

$$\Delta q = \operatorname{div} g \quad \text{and} \quad \Delta r = \operatorname{div} G \quad \text{in } Q. \tag{3.32}$$

So, to estimate the involved weighted L^2 norms of ∇q and ∇r, we may use Imanuvilov's Carleman inequality for second order uniformly elliptic equations with nonhomogeneous Dirichlet boundary conditions (obtained in [9]). Applying this inequality for the first Poisson equation in (3.32) twice and using the energy a priori estimate for the Stokes equations, too, we can establish the following estimate for q.

LEMMA 3.3. *Let Ω, ω, ω_0, and ω_1 be open subsets of $\mathbb{R}^2$ as in the statement of Theorem* 3.1. *Then there exists $\lambda_0 > 0$ such that for any $\lambda > \lambda_0$ one can find $s_0(\lambda) > 0$ and $c(\lambda) > 0$ such that for $s > s_0(\lambda)$ we have*

$$\begin{aligned} &\int_Q e^{2s\alpha} |\nabla q|^2 dx\, dt \\ &\quad \le c(\lambda) \bigg(\int_Q e^{2s\widehat{\alpha}} s^{\frac{11}{4}} \widehat{\varphi}^3 |z|^2 dx\, dt + \int_{Q_{\omega_1}} e^{2s\alpha} s^{\frac{11}{4}} \widehat{\varphi}^{\frac{11}{4}} q^2 dx\, dt \\ &\qquad + \int_Q e^{2s\widehat{\alpha}} s^{\frac{3}{4}} \widehat{\varphi}^{\frac{3}{4}} |g|^2 dx\, dt + \int_Q e^{2s\alpha} s^{\frac{1}{2}} \widehat{\varphi}^{\frac{1}{2}} (\operatorname{div} g)^2 dx\, dt \bigg) \end{aligned} \tag{3.33}$$

for all $g \in L^2(0,T;(H^1(\Omega))^2)$ and all corresponding solutions $(z,q) \in (H^{2,1}(Q))^2 \times L^2(0,T;H^1(\Omega))$ of system (3.4).

The proof of the above lemma is identical to that of its three-dimensional version (see [6] or [7]). In a completely similar way (we may follow the proof of Lemma 3.3 in [6] step by step, because $-\widetilde{\operatorname{curl}}(\operatorname{curl} C) = \Delta C$ when $\operatorname{div} C = 0$) we can obtain the corresponding estimate for ∇r. (See [7] for the three-dimensional variant.)

LEMMA 3.4. *Let Ω, ω, ω_0, and ω_1 be open subsets of $\mathbb{R}^2$ as in the statement of Theorem* 3.1. *Then there exists $\lambda_0 > 0$ such that for any $\lambda > \lambda_0$ one can find $s_0(\lambda) > 0$ and $c(\lambda) > 0$ such that for $s > s_0(\lambda)$ we have*

$$\begin{aligned} &\int_Q e^{2s\alpha} |\nabla r|^2 dx\, dt \\ &\quad \le c(\lambda) \bigg(\int_Q e^{2s\widehat{\alpha}} s^{\frac{11}{4}} \widehat{\varphi}^3 |C|^2 dx\, dt + \int_{Q_{\omega_1}} e^{2s\alpha} s^{\frac{11}{4}} \widehat{\varphi}^{\frac{11}{4}} r^2 dx\, dt \\ &\qquad + \int_Q e^{2s\widehat{\alpha}} s^{\frac{3}{4}} \widehat{\varphi}^{\frac{3}{4}} |G|^2 dx\, dt + \int_Q e^{2s\alpha} s^{\frac{1}{2}} \widehat{\varphi}^{\frac{1}{2}} (\operatorname{div} G)^2 dx\, dt \bigg) \end{aligned} \tag{3.34}$$

for all $G \in L^2(0,T;(H^1(\Omega))^2)$ and all corresponding solutions $C \in (H^{2,1}(Q))^2$ of system (3.5) *and $r \in L^2(0,T;H^1(\Omega))$ which satisfy* (3.8).

The proofs of Theorems 3.2 and 3.3 come to an end by simply coupling (3.9) with (3.33) and (3.10) with (3.34).

Proof of Theorem 3.1. To finish the proof, we have to take inequalities (3.6) and (3.7) together with g and G given by

$$g = -(\widetilde{y}\cdot\nabla)z + z\cdot(\nabla\widetilde{y}) + (\widetilde{B}\cdot\nabla)C + C\cdot(\nabla\widetilde{B}) + h,$$

$$G = -(\widetilde{y}\cdot\nabla)C - C\cdot(\nabla\widetilde{y}) + (\widetilde{B}\cdot\nabla)z - z\cdot(\nabla\widetilde{B}) + H.$$

We have

$$\operatorname{div} g = -\sum_{i,j=1}^{2}\frac{\partial\widetilde{y}_j}{\partial x_i}\,\frac{\partial z_i}{\partial x_j} + \nabla z\cdot\nabla\widetilde{y} + z\cdot\Delta\widetilde{y} + \sum_{i,j=1}^{2}\frac{\partial\widetilde{B}_j}{\partial x_i}\,\frac{\partial C_i}{\partial x_j} + \nabla C\cdot\nabla\widetilde{B} + C\cdot\Delta\widetilde{B},$$

$$\operatorname{div} G = -\sum_{i,j=1}^{2}\frac{\partial\widetilde{y}_j}{\partial x_i}\,\frac{\partial C_i}{\partial x_j} - \nabla C\cdot\nabla\widetilde{y} - C\cdot\Delta\widetilde{y} + \sum_{i,j=1}^{2}\frac{\partial\widetilde{B}_j}{\partial x_i}\,\frac{\partial z_i}{\partial x_j} - \nabla z\cdot\nabla\widetilde{B} - z\cdot\Delta\widetilde{B}.$$

(Here one has taken into account the fact that z, C, h, and H are divergence-free.) As $\widetilde{y}, \widetilde{B} \in L^\infty(0,T;(W^{2,p}(\Omega))^2)$ with $p > 2$, by the Sobolev imbedding theorem we also have $\widetilde{y}, \widetilde{B} \in L^\infty(0,T;(W^{1,\infty}(\Omega))^2)$, so a.e. in Q we can obtain

$$|g| \le c(|\nabla z| + |\nabla C| + |z| + |C|) + |h|, \tag{3.35}$$

$$|G| \le c(|\nabla z| + |\nabla C| + |z| + |C|) + |H|, \tag{3.36}$$

$$|\operatorname{div} g| \le c(|\nabla z| + |\nabla C|) + |\Delta\widetilde{y}|\,|z| + |\Delta\widetilde{B}|\,|C|, \tag{3.37}$$

$$|\operatorname{div} G| \le c(|\nabla z| + |\nabla C|) + |\Delta\widetilde{y}|\,|C| + |\Delta\widetilde{B}|\,|z|. \tag{3.38}$$

By (3.35) and (3.36), we have

$$\begin{aligned}
&\int_Q \left(e^{2s\alpha} + e^{2s\widehat{\alpha}} s^{\frac{3}{4}}\widehat{\varphi}^{\frac{3}{4}}\right)|g|^2 dx\,dt \\
&\quad\le c\int_Q \left(e^{2s\alpha} + e^{2s\widehat{\alpha}} s^{\frac{3}{4}}\widehat{\varphi}^{\frac{3}{4}}\right)(|\nabla z|^2 + |\nabla C|^2 + |z|^2 + |C|^2)dx\,dt \\
&\qquad + \int_Q \left(e^{2s\alpha} + e^{2s\widehat{\alpha}} s^{\frac{3}{4}}\widehat{\varphi}^{\frac{3}{4}}\right)|h|^2 dx\,dt,
\end{aligned} \tag{3.39}$$

$$\begin{aligned}
&\int_Q \left(e^{2s\alpha} + e^{2s\widehat{\alpha}} s^{\frac{3}{4}}\widehat{\varphi}^{\frac{3}{4}}\right)|G|^2 dx\,dt \\
&\quad\le c\int_Q \left(e^{2s\alpha} + e^{2s\widehat{\alpha}} s^{\frac{3}{4}}\widehat{\varphi}^{\frac{3}{4}}\right)(|\nabla z|^2 + |\nabla C|^2 + |z|^2 + |C|^2)dx\,dt \\
&\qquad + \int_Q \left(e^{2s\alpha} + e^{2s\widehat{\alpha}} s^{\frac{3}{4}}\widehat{\varphi}^{\frac{3}{4}}\right)|H|^2 dx\,dt.
\end{aligned} \tag{3.40}$$

By (3.37), we can write

$$\begin{aligned}
\int_Q e^{2s\alpha} s^{\frac{1}{2}}\widehat{\varphi}^{\frac{1}{2}}|\operatorname{div} g|^2 dx\,dt &\le c\int_Q e^{2s\alpha} s^{\frac{1}{2}}\widehat{\varphi}^{\frac{1}{2}}(|\nabla z|^2 + |\nabla C|^2)dx\,dt \\
&\quad + c\int_Q e^{2s\alpha} s^{\frac{1}{2}}\widehat{\varphi}^{\frac{1}{2}}(|\Delta\widetilde{y}|^2|z|^2 + |\Delta\widetilde{B}|^2|C|^2)dx\,dt.
\end{aligned} \tag{3.41}$$

Using Hölder's inequality, the fact that $\widetilde{y}, \widetilde{B} \in L^\infty(0,T;(W^{2,p}(\Omega))^2)$ with $p>2$, and the continuity of the inclusion $(H^1(\Omega))^2 \subset (L^{2p/(p-2)}(\Omega))^2$, we derive

$$\begin{aligned}
&\int_Q e^{2s\alpha}\widehat{\varphi}^{\frac{1}{2}}|\Delta\widetilde{y}|^2|z|^2dx\,dt \\
&\quad\le \int_0^T \widehat{\varphi}^{\frac{1}{2}}|\Delta\widetilde{y}|^2_{(L^p(\Omega))^2}|e^{s\alpha}z|^2_{(L^{\frac{2p}{p-2}}(\Omega))^2}dt \\
&\quad\le c_1\int_0^T \widehat{\varphi}^{\frac{1}{2}}|e^{s\alpha}z|^2_{(H^1(\Omega))^2}dt \\
&\quad\le c_2(\lambda)\int_Q e^{2s\alpha}\widehat{\varphi}^{\frac{1}{2}}(|\nabla z|^2+s^2\varphi^2|z|^2)dx\,dt
\end{aligned} \tag{3.42}$$

and

$$\begin{aligned}
&\int_Q e^{2s\alpha}\widehat{\varphi}^{\frac{1}{2}}|\Delta\widetilde{B}|^2|C|^2dx\,dt \\
&\quad\le c(\lambda)\int_Q e^{2s\alpha}\widehat{\varphi}^{\frac{1}{2}}(|\nabla C|^2+s^2\varphi^2|C|^2)dx\,dt.
\end{aligned} \tag{3.43}$$

Putting inequalities (3.41) through (3.43) together, we obtain

$$\begin{aligned}
&\int_Q e^{2s\alpha}s^{\frac{1}{2}}\widehat{\varphi}^{\frac{1}{2}}|\operatorname{div} g|^2dx\,dt \\
&\quad\le c(\lambda)\int_Q e^{2s\alpha}\left(s^{\frac{1}{2}}\varphi^{\frac{1}{2}}\left(|\nabla z|^2+|\nabla C|^2\right)+s^{\frac{5}{2}}\varphi^{\frac{5}{2}}\left(|z|^2+|C|^2\right)\right)dx\,dt.
\end{aligned} \tag{3.44}$$

In a similar way, using (3.38), we have

$$\begin{aligned}
&\int_Q e^{2s\alpha}s^{\frac{1}{2}}\widehat{\varphi}^{\frac{1}{2}}|\operatorname{div} G|^2dx\,dt \\
&\quad\le c(\lambda)\int_Q e^{2s\alpha}\left(s^{\frac{1}{2}}\varphi^{\frac{1}{2}}\left(|\nabla z|^2+|\nabla C|^2\right)+s^{\frac{5}{2}}\varphi^{\frac{5}{2}}\left(|z|^2+|C|^2\right)\right)dx\,dt.
\end{aligned} \tag{3.45}$$

Now, we add inequalities (3.6) and (3.7), and take estimates (3.39), (3.40), (3.44), and (3.45) into account. Thus, we obtain inequality (3.3) by simply taking s sufficiently large. This finishes the proof of Theorem 3.1. ∎

4. Observability inequalities for the adjoint linearized MHD equations. Using the Carleman inequality (3.3), one can estimate weighted L^2 norms of solutions (z,C) of equations (3.1) taken over the entire Q by weighted L^2 norms of their values taken over Q_ω only. Such estimates are called observability inequalities. We shall express the needed weights by means of the following version of function α (which is no longer $+\infty$ at $t=0$):

$$\beta(x,t)=\frac{e^{\lambda\psi(x)}-e^{2\lambda|\psi|_{C(\overline{\Omega})}}}{(\theta(t)(T-t))^8},$$

where $\lambda>0$ and θ is an increasing C^∞ function such that $\theta(0)>0$ and $\theta(t)=t$ for $t\in[T/2,T]$. The restriction of β on $\partial\Omega$ is denoted by $\widehat{\beta}$:

$$\widehat{\beta}(t)=\frac{1-e^{2\lambda|\psi|_{C(\overline{\Omega})}}}{(\theta(t)(T-t))^8}\,.$$

The first observability inequality we need can be expressed as follows.

THEOREM 4.1. *Let Ω be an open subset of $\mathbb{R}^2$ as in the statement of Theorem 3.1 and let ω be an open subset of Ω. Let $\widetilde{y}, \widetilde{B} \in W^{1,\infty}(0,T;(W^{2,p}(\Omega))^2)$ with $p>2$. Then there exists $\lambda_0>0$ such that for any $\lambda>\lambda_0$ one can find $s_0(\lambda)>0$ and $\delta_0(\lambda)\in(1/2,1)$ such that for $s>s_0(\lambda)$ and $1/2<\delta<\delta_0(\lambda)$ there is some $c(\lambda,s,\delta)>0$ such that the following inequality holds:*

$$\begin{aligned}(4.1)\qquad &\int_Q e^{2s\widehat{\beta}}(T-t)^8\left(|z|^2+|C|^2\right)dx\,dt\\ &\qquad\le c(\lambda,s,\delta)\left(\int_{Q_\omega} e^{2s\delta\widehat{\beta}}\left(|z|^2+|C|^2\right)dx\,dt+\int_Q e^{2s\delta\widehat{\beta}}\left(|h|^2+|H|^2\right)dx\,dt\right)\end{aligned}$$

for all $h, H\in(L^2(Q))^2$ which satisfy $\operatorname{div} h=\operatorname{div} H=0$ in Q and $H\cdot N=0$ on Σ, and all corresponding solutions $(z,C,q)\in(H^{2,1}(Q))^4\times L^2(0,T;H^1(\Omega))$ of system (3.1).

Proof. The complete proof is quite long and can be found in [7]. (See also [6] for a related situation.) We emphasize that there is no significant difference between the two-dimensional and three-dimensional cases. However, for the reader's convenience, we shall provide an outline of the proof, referring the reader to [7] (or [6]) for details.

The main objective of the proof is to remove q and r in (3.3). To be able to do it, we are forced to pass from z, q, h, C, r, and H to their primitives with respect to t. The new functions satisfy parabolic equations similar to (3.1) (which enables us to apply Carleman inequality (3.3) to them, too) but also certain stationary Stokes equations, which taken under consideration will lead us to the elimination of q and r.

So, for $t\in[0,T]$, we consider the auxiliary functions

$$\overline{z}(x,t)=\int_{\frac{T}{2}}^t z(x,\tau)d\tau,\qquad \overline{q}(x,t)=\int_{\frac{T}{2}}^t q(x,\tau)d\tau,\quad \overline{h}(x,t)=\int_{\frac{T}{2}}^t h(x,\tau)d\tau,$$

$$\overline{C}(x,t)=\int_{\frac{T}{2}}^t C(x,\tau)d\tau,\quad \overline{r}(x,t)=\int_{\frac{T}{2}}^t r(x,\tau)d\tau,\quad \overline{H}(x,t)=\int_{\frac{T}{2}}^t H(x,\tau)d\tau.$$

Integrating equations (3.1) from $T/2$ to t and taking (3.2) into account, one easily sees that $\overline{z},\overline{q},\overline{h}$ and $\overline{c},\overline{r},\overline{H}$ satisfy

$$\begin{aligned}&\frac{\partial\overline{z}}{\partial t}+\Delta\overline{z}+(\widetilde{y}\cdot\nabla)\overline{z}-\overline{z}\cdot(\nabla\widetilde{y})-(\widetilde{B}\cdot\nabla)\overline{C}-\overline{C}\cdot(\nabla\widetilde{B})+\nabla\overline{q}\\ &\quad=\overline{h}+\int_{\frac{T}{2}}^t\left(\frac{\partial\widetilde{y}}{\partial\tau}\cdot\nabla\right)\overline{z}\,d\tau-\int_{\frac{T}{2}}^t\overline{z}\cdot\left(\nabla\frac{\partial\widetilde{y}}{\partial\tau}\right)d\tau\\ &\qquad-\int_{\frac{T}{2}}^t\left(\frac{\partial\widetilde{B}}{\partial\tau}\cdot\nabla\right)\overline{C}d\tau-\int_{\frac{T}{2}}^t\overline{C}\cdot\left(\nabla\frac{\partial\widetilde{B}}{\partial\tau}\right)d\tau+z\left(\cdot,\frac{T}{2}\right)\ \text{in } Q,\end{aligned}$$

$$\begin{aligned}(4.2)\qquad &\frac{\partial\overline{C}}{\partial t}-\widetilde{\operatorname{curl}}(\operatorname{curl}\overline{C})+(\widetilde{y}\cdot\nabla)\overline{C}+\overline{C}\cdot(\nabla\widetilde{y})-(\widetilde{B}\cdot\nabla)\overline{z}+\overline{z}\cdot(\nabla\widetilde{B})+\nabla\overline{r}\\ &\quad=\overline{H}+\int_{\frac{T}{2}}^t\left(\frac{\partial\widetilde{y}}{\partial\tau}\cdot\nabla\right)\overline{C}\,d\tau+\int_{\frac{T}{2}}^t\overline{C}\cdot\left(\nabla\frac{\partial\widetilde{y}}{\partial\tau}\right)d\tau\\ &\qquad-\int_{\frac{T}{2}}^t\left(\frac{\partial\widetilde{B}}{\partial\tau}\cdot\nabla\right)\overline{z}\,d\tau+\int_{\frac{T}{2}}^t\overline{z}\cdot\left(\nabla\frac{\partial\widetilde{B}}{\partial\tau}\right)d\tau+C\left(\cdot,\frac{T}{2}\right)\ \text{in } Q,\\ &\operatorname{div}\overline{z}=0,\quad \operatorname{div}\overline{C}=0\qquad\text{in } Q,\\ &\overline{z}=0,\quad \overline{C}\cdot N=0,\quad \operatorname{curl}\overline{C}=0\qquad\text{on }\Sigma.\end{aligned}$$

(We have also used the fact that $z = \partial\overline{z}/\partial t$ and $C = \partial\overline{C}/\partial t$.)

Since equations (4.2) are similar to equations (3.1) (taken together with (3.2)), we can argue as in the proof of Theorem 3.1 to derive a Carleman inequality like (3.3) for the solutions of (4.2). Indeed, it is enough to treat the eight integral terms in (4.2) in the same way in which the corresponding terms in g and G have been treated in the proof of Theorem 3.1 to obtain estimates as (3.39), (3.40), (3.44), and (3.45) for them. Thus we can assert that there exists $\lambda_0 > 0$ such that for any $\lambda > \lambda_0$ one can find $s_0(\lambda) > 0$ and $c(\lambda) > 0$ such that for $s > s_0(\lambda)$ one has

$$\begin{aligned}
&\int_Q e^{2s\alpha}\left(\frac{1}{s\varphi}\left(\left|\frac{\partial\overline{z}}{\partial t}\right|^2+\left|\frac{\partial\overline{C}}{\partial t}\right|^2+\sum_{i,j=1}^{2}\left(\left|\frac{\partial^2\overline{z}}{\partial x_i\partial x_j}\right|^2+\left|\frac{\partial^2\overline{C}}{\partial x_i\partial x_j}\right|^2\right)\right)\right.\\
&\qquad\left.+\ s\varphi(|\nabla\overline{z}|^2+|\nabla\overline{C}|^2)+s^3\varphi^3(|\overline{z}|^2+|\overline{C}|^2)\right)dxdt \qquad (4.3)\\
&\quad\le c(\lambda)\left(\int_{Q_\omega} e^{2s\alpha}s^3\varphi^3(|\overline{z}|^2+|\overline{C}|^2)dxdt+\int_{Q_{\omega_1}} e^{2s\alpha}s^{\frac{11}{4}}\widehat{\varphi}^{\frac{11}{4}}(\overline{q}^2+\overline{r}^2)\,dxdt\right.\\
&\qquad\left.+\int_Q\left(e^{2s\alpha}+e^{2s\widehat{\alpha}}s^{\frac{3}{4}}\widehat{\varphi}^{\frac{3}{4}}\right)\left(|\overline{h}|^2+|\overline{H}|^2+\left|z\left(\cdot,\frac{T}{2}\right)\right|^2+\left|C\left(\cdot,\frac{T}{2}\right)\right|^2\right)dxdt\right).
\end{aligned}$$

The next step is to eliminate the local terms containing $\overline{q}$ and $\overline{r}$ in the right-hand side of (4.3). To do so, we shall try to estimate those terms by local expressions of $\overline{z}, \overline{C}, z, C, \overline{h}$, and $\overline{H}$.

Clearly we may assume that $\overline{q}$ and $\overline{r}$ satisfy

$$\int_{\omega_1}\overline{q}(x,t)dx = 0 \quad\text{and}\quad \int_{\omega_1}\overline{r}(x,t)dx = 0 \quad \text{for all } t\in[0,T]$$

(because, otherwise, we can pass from q and r to $q - (\operatorname{meas}\omega_1)^{-1}\int_{\omega_1} q\,dx$ and $r - (\operatorname{meas}\omega_1)^{-1}\int_{\omega_1} r\,dx$). So we may apply Proposition 1.2 in [11] and obtain

$$\begin{aligned}
&\int_0^T\int_{\omega_1} e^{2s\alpha}s^{\frac{11}{4}}\widehat{\varphi}^{\frac{11}{4}}(\overline{q}^2+\overline{r}^2)dxdt\\
&\qquad\le c_1(\delta)\int_0^T\int_{\omega_1} e^{2s\delta\widehat{\alpha}}(\overline{q}^2+\overline{r}^2)dxdt\\
&\qquad\le c_2(\delta)\int_0^T e^{2s\delta\widehat{\alpha}}\left(|\nabla\overline{q}|^2_{((H_0^1(\omega_1))^2)'}+|\nabla\overline{r}|^2_{((H_0^1(\omega_1))^2)'}\right)dt \qquad (4.4)\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{for } \frac{1}{2}<\delta<\delta_0(\lambda)<1,
\end{aligned}$$

where

$$\delta_0(\lambda) = \frac{e^{2\lambda|\psi|_{C(\overline{\Omega})}}-e^{\lambda|\psi|_{C(\overline{\Omega})}}}{e^{2\lambda|\psi|_{C(\overline{\Omega})}}-1}\,.$$

(Obviously $\delta_0(\lambda) < 1$ for any $\lambda > 0$, and taking $\lambda > \log 2/|\psi|_{C(\overline{\Omega})}$, we have $\delta_0(\lambda) > 1/2$ as well.)

To estimate the $((H_0^1(\omega_1))^2)'$ norms of $\nabla\overline{q}$ and $\nabla\overline{r}$ in (4.4), we shall use equations (4.2) again, but this time regarded as an elliptic system. In fact we separately consider two elliptic systems, one for $\overline{z},\overline{q},\overline{h}$ and another one for $\overline{C},\overline{r},\overline{H}$, namely,

$$\begin{aligned}
\Delta\overline{z}+\nabla\overline{q} &= \overline{h}-(\widetilde{y}\cdot\nabla)\overline{z}+\overline{z}\cdot(\nabla\widetilde{y})+(\widetilde{B}\cdot\nabla)\overline{C}+\overline{C}\cdot(\nabla\widetilde{B})\\
&\quad+\int_{\frac{T}{2}}^{t}\left(\frac{\partial\widetilde{y}}{\partial\tau}\cdot\nabla\right)\overline{z}\,d\tau-\int_{\frac{T}{2}}^{t}\overline{z}\cdot\left(\nabla\frac{\partial\widetilde{y}}{\partial\tau}\right)d\tau\\
&\quad-\int_{\frac{T}{2}}^{t}\left(\frac{\partial\widetilde{B}}{\partial\tau}\cdot\nabla\right)\overline{C}\,d\tau-\int_{\frac{T}{2}}^{t}\overline{C}\cdot\left(\nabla\frac{\partial\widetilde{B}}{\partial\tau}\right)d\tau\\
&\quad-z+z\left(\cdot,\frac{T}{2}\right) \qquad \text{in } Q,\\
\operatorname{div}\overline{z} &= 0 \qquad \text{in } Q
\end{aligned}\tag{4.5}$$

and

$$\begin{aligned}
\Delta\overline{C}+\nabla\overline{r} &= \overline{H}-(\widetilde{y}\cdot\nabla)\overline{C}-\overline{C}\cdot(\nabla\widetilde{y})+(\widetilde{B}\cdot\nabla)\overline{z}-\overline{z}\cdot(\nabla\widetilde{B})\\
&\quad+\int_{\frac{T}{2}}^{t}\left(\frac{\partial\widetilde{y}}{\partial\tau}\cdot\nabla\right)\overline{C}\,d\tau+\int_{\frac{T}{2}}^{t}\overline{C}\cdot\left(\nabla\frac{\partial\widetilde{y}}{\partial\tau}\right)d\tau\\
&\quad-\int_{\frac{T}{2}}^{t}\left(\frac{\partial\widetilde{B}}{\partial\tau}\cdot\nabla\right)\overline{z}\,d\tau+\int_{\frac{T}{2}}^{t}\overline{z}\cdot\left(\nabla\frac{\partial\widetilde{B}}{\partial\tau}\right)d\tau\\
&\quad-C+C\left(\cdot,\frac{T}{2}\right) \qquad \text{in } Q,\\
\operatorname{div}\overline{C} &= 0 \qquad \text{in } Q.
\end{aligned}\tag{4.6}$$

(We recall that $-\widetilde{\operatorname{curl}}(\operatorname{curl}\overline{C})=\Delta\overline{C}$ when $\operatorname{div}\overline{C}=0$.)

To be able to estimate the $((H_0^1(\omega_1))^2)'$ norm of $\nabla\overline{q}$, we need to split the solutions $(\overline{z},\overline{q})$ of (4.5), considered in ω, into two components. Let $(\overline{z}_1,\overline{q}_1)$ be the solution of the following steady state Stokes equations with homogeneous Dirichlet boundary condition:

$$\begin{aligned}
\Delta\overline{z}_1+\nabla\overline{q}_1 &= \overline{h}-(\widetilde{y}\cdot\nabla)\overline{z}+\overline{z}\cdot(\nabla\widetilde{y})+(\widetilde{B}\cdot\nabla)\overline{C}+\overline{C}\cdot(\nabla\widetilde{B})\\
&\quad+\int_{\frac{T}{2}}^{t}\left(\frac{\partial\widetilde{y}}{\partial\tau}\cdot\nabla\right)\overline{z}\,d\tau-\int_{\frac{T}{2}}^{t}\overline{z}\cdot\left(\nabla\frac{\partial\widetilde{y}}{\partial\tau}\right)d\tau\\
&\quad-\int_{\frac{T}{2}}^{t}\left(\frac{\partial\widetilde{B}}{\partial\tau}\cdot\nabla\right)\overline{C}\,d\tau-\int_{\frac{T}{2}}^{t}\overline{C}\cdot\left(\nabla\frac{\partial\widetilde{B}}{\partial\tau}\right)d\tau\\
&\quad-z+z\left(\cdot,\frac{T}{2}\right) \qquad \text{in } \omega,\\
\operatorname{div}\overline{z}_1 &= 0 \qquad \text{in } \omega,\\
\overline{z}_1 &= 0 \qquad \text{on } \partial\omega.
\end{aligned}\tag{4.7}$$

Subtracting (4.5) and (4.7), one sees that $\overline{z}_2=\overline{z}-\overline{z}_1$ and $\overline{q}_2=\overline{q}-\overline{q}_1$ satisfy the

homogeneous Stokes equations,

$$\begin{aligned} \Delta \overline{z}_2 + \nabla \overline{q}_2 = 0 \qquad & \text{in } \omega, \\ \operatorname{div} \overline{z}_2 = 0 \qquad & \text{in } \omega. \end{aligned} \tag{4.8}$$

From (4.7), using the well-known estimate for the weak solution of the steady state Stokes equations with null boundary conditions as well, it follows that

$$\begin{aligned} &|\nabla \overline{q}_1|_{((H_0^1(\omega_1))^2)'} \\ &\quad \le c\Bigg(|\overline{h}|_{(L^2(\omega))^2} + |z|_{(L^2(\omega))^2} + |-(\widetilde{y}\cdot\nabla)\overline{z} + \overline{z}\cdot(\nabla\widetilde{y}) + (\widetilde{B}\cdot\nabla)\overline{C} + \overline{C}\cdot(\nabla\widetilde{B})|_{((H_0^1(\omega))^2)'} \\ &\qquad + \left| \int_{\frac{T}{2}}^t \left| \left(\frac{\partial \widetilde{y}}{\partial \tau}\cdot\nabla\right)\overline{z} - \overline{z}\cdot\left(\nabla\frac{\partial \widetilde{y}}{\partial \tau}\right) - \left(\frac{\partial \widetilde{B}}{\partial \tau}\cdot\nabla\right)\overline{C} - \overline{C}\cdot\left(\nabla\frac{\partial \widetilde{B}}{\partial \tau}\right) \right|_{((H_0^1(\omega))^2)'} d\tau \right| \\ &\qquad + \left| z\left(\cdot, \frac{T}{2}\right) \right|_{(L^2(\omega))^2} \Bigg). \end{aligned}$$

Estimating the $((H_0^1(\omega))^2)'$ norms of all the products in the right-hand side of the above inequality in the usual way (recalling that $\widetilde{y}, \widetilde{B} \in W^{1,\infty}(0,T;(W^{2,p}(\Omega))^2)$ with $p > 2$), we can write

$$\begin{aligned} &|\nabla \overline{q}_1|_{((H_0^1(\omega_1))^2)'} \\ &\quad \le c\Bigg(|\overline{h}|_{(L^2(\omega))^2} + |\overline{z}|_{(L^2(\omega))^2} + |\overline{C}|_{(L^2(\omega))^2} \\ &\qquad + \left| \int_{\frac{T}{2}}^t \left(|\overline{z}|_{(L^2(\omega))^2} + |\overline{C}|_{(L^2(\omega))^2} \right) d\tau \right| \\ &\qquad + |z|_{(L^2(\omega))^2} + \left| z\left(\cdot, \frac{T}{2}\right) \right|_{(L^2(\omega))^2} \Bigg). \end{aligned} \tag{4.9}$$

To estimate the $((H_0^1(\omega_1))^2)'$ norm of $\nabla \overline{q}_2$, we first remark that $\overline{q}_2$ is harmonic in ω (because, by (4.8), $\Delta \overline{q}_2 = -\Delta \operatorname{div} z_2 = 0$ in ω). So, applying the Laplace operator to (4.8), we obtain

$$\Delta^2 \overline{z}_2 = 0 \text{ in } \omega.$$

A standard interior estimate for the solutions of homogeneous elliptic equations then gives

$$|\overline{z}_2|_{(H^2(\omega_1))^2} \le c|\overline{z}_2|_{(L^2(\omega))^2} \le c\left(|\overline{z}|_{(L^2(\omega))^2} + |\overline{z}_1|_{(L^2(\omega))^2} \right). \tag{4.10}$$

Now using the estimate for the weak solution of (4.7) (interpreted as a stationary Stokes system) once again, in the same way as before, we obtain

$$\begin{aligned} &|\overline{z}_1|_{(L^2(\omega))^2} \\ &\quad \le c\Bigg(|\overline{h}|_{(L^2(\omega))^2} + |\overline{z}|_{(L^2(\omega))^2} + |\overline{C}|_{(L^2(\omega))^2} \\ &\qquad + \left| \int_{\frac{T}{2}}^t \left(|\overline{z}|_{(L^2(\omega))^2} + |\overline{C}|_{(L^2(\omega))^2} \right) d\tau \right| \\ &\qquad + |z|_{(L^2(\omega))^2} + \left| z\left(\cdot, \frac{T}{2}\right) \right|_{(L^2(\omega))^2} \Bigg). \end{aligned} \tag{4.11}$$

From (4.8) it follows that

$$|\nabla \overline{q}_2|_{((H_0^1(\omega_1))^2)'} \le |\overline{z}_2|_{(H^1(\omega_1))^2}.$$

This last inequality taken together with (4.10) and (4.11) yields

$$\begin{aligned} &|\nabla \overline{q}_2|_{((H_0^1(\omega_1))^2)'} \\ &\quad \le c\bigg(|\overline{h}|_{(L^2(\omega))^2} + |\overline{z}|_{(L^2(\omega))^2} + |\overline{C}|_{(L^2(\omega))^2} \\ &\quad\quad + \left| \int_{\frac{T}{2}}^t \left(|\overline{z}|_{(L^2(\omega))^2} + |\overline{C}|_{(L^2(\omega))^2} \right) d\tau \right| \\ &\quad\quad + |z|_{(L^2(\omega))^2} + \left| z\left(\cdot, \frac{T}{2}\right)\right|_{(L^2(\omega))^2} \bigg). \end{aligned} \tag{4.12}$$

Since $\overline{q} = \overline{q}_1 + \overline{q}_2$, inequalities (4.9) and (4.12) give

$$\begin{aligned} &|\nabla \overline{q}|_{((H_0^1(\omega_1))^2)'} \\ &\quad \le c\bigg(|\overline{h}|_{(L^2(\omega))^2} + |\overline{z}|_{(L^2(\omega))^2} + |\overline{C}|_{(L^2(\omega))^2} \\ &\quad\quad + \left| \int_{\frac{T}{2}}^t \left(|\overline{z}|_{(L^2(\omega))^2} + |\overline{C}|_{(L^2(\omega))^2} \right) d\tau \right| \\ &\quad\quad + |z|_{(L^2(\omega))^2} + \left| z\left(\cdot, \frac{T}{2}\right)\right|_{(L^2(\omega))^2} \bigg). \end{aligned} \tag{4.13}$$

Treating equations (4.6) in the same way as equations (4.5) we obtain

$$\begin{aligned} &|\nabla \overline{r}|_{((H_0^1(\omega_1))^2)'} \\ &\quad \le c\bigg(|\overline{H}|_{(L^2(\omega))^2} + |\overline{z}|_{(L^2(\omega))^2} + |\overline{C}|_{(L^2(\omega))^2} \\ &\quad\quad + \left| \int_{\frac{T}{2}}^t \left(|\overline{z}|_{(L^2(\omega))^2} + |\overline{C}|_{(L^2(\omega))^2} \right) d\tau \right| \\ &\quad\quad + |C|_{(L^2(\omega))^2} + \left| C\left(\cdot, \frac{T}{2}\right)\right|_{(L^2(\omega))^2} \bigg). \end{aligned} \tag{4.14}$$

Now we take inequalities (4.4), (4.13), and (4.14) together to obtain

$$\begin{aligned} &\int_{Q_{\omega_1}} e^{2s\alpha} s^{\frac{11}{4}} \widehat{\varphi}^{\frac{11}{4}} (\overline{q}^2 + \overline{r}^2) dxdt \\ &\quad \le c(\lambda, s, \delta) \bigg(\int_{Q_\omega} e^{2s\delta\widehat{\alpha}} (|z|^2 + |C|^2) dxdt + \int_{Q_\omega} e^{2s\delta\widehat{\alpha}} (|h|^2 + |H|^2) dxdt \\ &\quad\quad + \left| z\left(\cdot, \frac{T}{2}\right)\right|^2_{(L^2(\omega))^2} + \left| C\left(\cdot, \frac{T}{2}\right)\right|^2_{(L^2(\omega))^2} \bigg) \qquad \text{for } \frac{1}{2} < \delta < \delta_0(\lambda). \end{aligned} \tag{4.15}$$

Next inserting (4.15) into the Carleman-type inequality (4.3), we obtain

$$\begin{aligned}
&\int_Q e^{2s\alpha}\frac{1}{\varphi}(|z|^2+|C|^2)dxdt \\
&\qquad \le c(\lambda,s,\delta)\bigg(\int_{Q_\omega} e^{2s\delta\widehat{\alpha}}(|z|^2+|C|^2)dxdt+\int_Q e^{2s\delta\widehat{\alpha}}(|h|^2+|H|^2)dxdt \\
&\qquad\quad +\left|z\left(\cdot,\frac{T}{2}\right)\right|^2_{(L^2(\Omega))^2}+\left|C\left(\cdot,\frac{T}{2}\right)\right|^2_{(L^2(\Omega))^2}\bigg) \\
&\qquad\qquad \text{for } \lambda>\lambda_0,\ s>s_0(\lambda),\ \text{and } \frac{1}{2}<\delta<\delta_0(\lambda).
\end{aligned} \tag{4.16}$$

In a standard way (see [9], [6], or [7] for details) we can convert inequality (4.16) into a similar one in which the new weight functions $e^{2s\widehat{\beta}}, (T-t)^8$, and $e^{2s\delta\widehat{\beta}}$ replace $e^{2s\alpha}$, $1/\varphi$, and $e^{2s\delta\widehat{\alpha}}$, respectively, namely,

$$\begin{aligned}
&\int_Q e^{2s\widehat{\beta}}(T-t)^8(|z|^2+|C|^2)dxdt \\
&\qquad \le c(\lambda,s,\delta)\bigg(\int_{Q_\omega} e^{2s\delta\widehat{\beta}}(|z|^2+|C|^2)dxdt+\int_Q e^{2s\delta\widehat{\beta}}(|h|^2+|H|^2)dxdt \\
&\qquad\quad +\left|z\left(\cdot,\frac{T}{2}\right)\right|^2_{(L^2(\Omega))^2}+\left|C\left(\cdot,\frac{T}{2}\right)\right|^2_{(L^2(\Omega))^2}\bigg) \\
&\qquad\qquad \text{for } \lambda>\lambda_0,\ s>s_0(\lambda),\ \text{and } \frac{1}{2}<\delta<\delta_0(\lambda).
\end{aligned} \tag{4.17}$$

Arguing by contradiction one can eliminate the $(L^2(\Omega))^2$ norms of $z(\cdot,T/2)$ and $C(\cdot,T/2)$ in (4.17) and show that (4.1) is true for suitable constants $c(\lambda,s,\delta)$. (We again refer the reader to [9], [6], or [7] for details.) So the proof is complete. □

From the observability inequality (4.1), one can derive the next one, which is in fact an L^2 estimate of z and C taken on Ω at the moment $t=0$ by their values taken on ω but at all the moments $t\in[0,T]$. (We refer the reader to [7] or [6] for the proof.)

THEOREM 4.2. *Under the hypotheses of Theorem* 4.1 *there exists* $\lambda_0>0$ *such that for any* $\lambda>\lambda_0$ *one can find* $s_0(\lambda)>0$ *and* $\delta_0(\lambda)\in(1/2,1)$ *such that for* $s>s_0(\lambda)$ *and* $1/2<\delta<\delta_0(\lambda)$ *there is some* $c(\lambda,s,\delta)>0$ *such that the following inequality holds:*

$$\begin{aligned}
&\int_\Omega \left(|z(x,0)|^2+|C(x,0)|^2\right)dx \\
&\qquad \le c(\lambda,s,\delta)\left(\int_{Q_\omega} e^{2s\delta\widehat{\beta}}\left(|z|^2+|C|^2\right)dx\,dt+\int_Q e^{2s\delta\widehat{\beta}}\left(|h|^2+|H|^2\right)dx\,dt\right)
\end{aligned} \tag{4.18}$$

for all $h,H\in(L^2(Q))^2$ *which satisfy* $\operatorname{div} h=\operatorname{div} H=0$ *in* Q *and* $H\cdot N=0$ *on* Σ, *and all corresponding solutions* $(z,C,q)\in(H^{2,1}(Q))^4\times L^2(0,T;H^1(\Omega))$ *of system* (3.1).

Thus, we have the appropriate tools to approach the controllability problem we are dealing with.

Let us also remark that, as we have already mentioned in the introduction, the results contained in Theorems 4.1 and 4.2 could be substantially improved if we should use the more general Carleman inequality for weak solutions of elliptic equations

with $H^{-1}(\Omega)$ right-hand side in [10] instead of the Carleman inequality for strong solutions of elliptic equations with $L^2(\Omega)$ right-hand side in [9]. Indeed, it seems that applying the new Carleman inequality but keeping the approach here (that is, the use of the primitives with respect to t of the solutions of (3.1)), one could show that the observability inequalities (4.1) and (4.18) (with slightly modified $\widehat{\beta}$) are still valid for $\widetilde{y}$ and $\widetilde{B}$ in $(L^\infty(Q))^2$ with $\partial\widetilde{y}/\partial t$ and $\partial\widetilde{B}/\partial t$ in $L^2(0,T;(L^\infty(\Omega))^2\cap(H^\gamma(\Omega))^2)$, where γ may be any positive exponent. But we can expect even more than this: If we should use the Carleman inequality in [10] but follow the approach in [4], then we could obtain observability inequalities like (4.1) and (4.18) (with modified weight functions) for $\widetilde{y}$ and $\widetilde{B}$ in $(L^\infty(Q))^2$ with $\partial\widetilde{y}/\partial t$ and $\partial\widetilde{B}/\partial t$ in $L^2(0,T;(L^\sigma(\Omega))^2)$ only, where σ is larger than 1. (See [4] for the analogous result for the Navier–Stokes equations.)

5. Global exact null controllability for the linearized MHD equations. The linearization of the controlled MHD equations (2.1) with $\nu=\eta=1$ around $(\widetilde{y},\widetilde{B})$ is the following:

$$
\begin{aligned}
&\frac{\partial y}{\partial t}-\Delta y+(\widetilde{y}\cdot\nabla)y+(y\cdot\nabla)\widetilde{y}-(\widetilde{B}\cdot\nabla)B-(B\cdot\nabla)\widetilde{B}\\
&\qquad+\nabla(\widetilde{B}\cdot B)+\nabla p=f+\chi_\omega u && \text{in } Q,\\
&\frac{\partial B}{\partial t}+\widetilde{\operatorname{curl}}(\operatorname{curl}B)+(\widetilde{y}\cdot\nabla)B+(y\cdot\nabla)\widetilde{B}\\
&\qquad-(\widetilde{B}\cdot\nabla)y-(B\cdot\nabla)\widetilde{y}=F+P(\chi_\omega v) && \text{in } Q,\\
&\operatorname{div}y=0,\ \operatorname{div}B=0 && \text{in } Q,\\
&y=0,\ B\cdot N=0,\ \operatorname{curl}B=0 && \text{on } \Sigma,\\
&y(\cdot,0)=y_0,\ B(\cdot,0)=B_0 && \text{in } \Omega,
\end{aligned}
\tag{5.1}
$$

where $(\widetilde{y},\widetilde{B})$ (together with $\widetilde{p}$) is a solution of (2.3). We also associate condition (2.2) to (5.1).

It is known that if $\widetilde{y},\widetilde{B}\in W^{1,\infty}(0,T;(W^{1,\infty}(\Omega))^2)$, $f,F,u,v\in(L^2(Q))^2$ with $\operatorname{div}F=0$ in Q and $F\cdot N=0$ on Σ, and $(y_0,B_0)\in V_1\times V_2$, then the boundary initial-value problem (5.1) has a unique solution $(y,B,p)\in(H^{2,1}(Q))^4\times L^2(0,T;H^1(\Omega))$ (p is unique up to a constant). In addition, the solution satisfies the following estimate:

$$
\begin{aligned}
&|y|_{(H^{2,1}(Q))^2}+|B|_{(H^{2,1}(Q))^2}+|\nabla p|_{(L^2(Q))^2}\\
&\qquad\le c\left(|y_0|_{(H^1(\Omega))^2}+|B_0|_{(H^1(\Omega))^2}+|f|_{(L^2(Q))^2}+|F|_{(L^2(Q))^2}\right.\\
&\qquad\quad\left.+|u|_{(L^2(Q))^2}+|v|_{(L^2(Q))^2}\right).
\end{aligned}
\tag{5.2}
$$

Two weighted L^2 spaces are needed to formulate the global controllability result for the linear equations (5.1). The space $L^2(Q,(T-t)^{-8}e^{-2s\widehat{\beta}})$ consists of all (equivalence classes of) measurable functions $f:Q\longrightarrow\mathbb{R}$ with $(T-t)^{-4}e^{-s\widehat{\beta}}f\in L^2(Q)$, that is,

$$
\int_Q\frac{1}{(T-t)^8}\ e^{-2s\widehat{\beta}}|f|^2dx\,dt<\infty.
$$

The space $L^2(Q,e^{-2s\delta\widehat{\beta}})$ is defined similarly.

THEOREM 5.1. *Let Ω and ω be as in the statement of Theorem 2.1 and let $\widetilde{y}, \widetilde{B} \in W^{1,\infty}(0,T;(W^{2,p}(\Omega))^2)$ with $p > 2$. Then there are $\lambda > 0$, $s > 0$, and $\delta \in (1/2,1)$ such that for any $f, F \in (L^2(Q,(T-t)^{-8}e^{-2s\widehat{\beta}}))^2$ with $\operatorname{div} F = 0$ in Q and $F \cdot N = 0$ on Σ, $(y_0, B_0) \in V_1 \times V_2$, and $\delta' \in (1/2,\delta)$, there exists $(u,v,y,B,p) \in (L^2(Q))^4 \times (H^{2,1}(Q))^4 \times L^2(0,T;H^1(\Omega))$ which satisfies (5.1), (2.2) and the final conditions*

$$y(x,T) = 0 \quad \text{and} \quad B(x,T) = 0 \quad \text{a.e.} \quad x \in \Omega$$

and which has the following decay at $t = T$:

$$u, v, y, B \in (L^2(Q, e^{-2s\delta\widehat{\beta}}))^2,$$

$$e^{-s\delta'\widehat{\beta}}y, e^{-s\delta'\widehat{\beta}}B \in (H^{2,1}(Q))^2.$$

Proof. We need an approximation of the function $\widehat{\beta}$ which should take a finite value at $t = T$, too. For instance, for $\varepsilon > 0$ we could define $\widehat{\beta}_\varepsilon$ as

$$\widehat{\beta}_\varepsilon(t) = \frac{1 - e^{2\lambda|\psi|_{C(\overline{\Omega})}}}{(\theta(t)(T - t + \varepsilon))^8}\,.$$

Let us now fix $\lambda > 0$, $s > 0$, and $\delta \in (1/2, \delta_0(\lambda))$ such that inequalities (4.1) and (4.2) hold. Then for $\varepsilon > 0$ we consider the corresponding auxiliary optimal control problem:

Minimize

$$(P_\varepsilon) \qquad \begin{aligned} &\frac{1}{2}\int_Q e^{-2s\delta\widehat{\beta}}\left(|u|^2 + |v|^2\right)dx\,dt + \frac{1}{2}\int_Q e^{-2s\delta\widehat{\beta}_\varepsilon}\left(|y|^2 + |B|^2\right)dx\,dt \\ &\quad + \frac{1}{2\varepsilon}\int_\Omega \left(|y(x,T)|^2 + |B(x,T)|^2\right)dx \end{aligned}$$

over all $(u,v) \in (L^2(Q))^4$, where (y,B) satisfies (5.1) and (2.2) (together with some p).

It is known that problem (P_ε) has a unique solution $(u_\varepsilon, v_\varepsilon, y_\varepsilon, B_\varepsilon, p_\varepsilon)$ for any $\varepsilon > 0$.

The idea of the proof is to regard the limit of $(u_\varepsilon, v_\varepsilon, y_\varepsilon, B_\varepsilon, p_\varepsilon)$ when $\varepsilon \longrightarrow 0$ as a possible solution of the controllability problem for system (5.1). To prove the convergence of $(u_\varepsilon, v_\varepsilon, y_\varepsilon, B_\varepsilon, p_\varepsilon)$ we need to obtain L^2 estimates for u_ε and v_ε. To achieve this, we shall combine Pontryagin's maximum principle, applied to problems (P_ε), with the observability inequalities (4.1) and (4.2), applied to the adjoint system.

Let $(u_\varepsilon, v_\varepsilon, y_\varepsilon, B_\varepsilon, p_\varepsilon)$ be the solution of problem (P_ε). Pontryagin's maximum principle asserts that there exists a dual process $(z_\varepsilon, C_\varepsilon, q_\varepsilon)$ which, together with $(u_\varepsilon, v_\varepsilon, y_\varepsilon, B_\varepsilon, p_\varepsilon)$, satisfies the adjoint equations

$$(5.3) \qquad \begin{aligned} &\frac{\partial z_\varepsilon}{\partial t} + \Delta z_\varepsilon + (\widetilde{y}\cdot\nabla)z_\varepsilon - z_\varepsilon\cdot(\nabla\widetilde{y}) \\ &\qquad - (\widetilde{B}\cdot\nabla)C_\varepsilon - C_\varepsilon\cdot(\nabla\widetilde{B}) + \nabla q_\varepsilon = e^{-2s\delta\widehat{\beta}_\varepsilon}y_\varepsilon \quad \text{in } Q, \\ &\frac{\partial C_\varepsilon}{\partial t} - \widetilde{\operatorname{curl}}(\operatorname{curl} C_\varepsilon) + P((\widetilde{y}\cdot\nabla)C_\varepsilon + C_\varepsilon\cdot(\nabla\widetilde{y}) \\ &\qquad - (\widetilde{B}\cdot\nabla)z_\varepsilon + z_\varepsilon\cdot(\nabla\widetilde{B})) = e^{-2s\delta\widehat{\beta}_\varepsilon}B_\varepsilon \quad \text{in } Q, \\ &\operatorname{div} z_\varepsilon = 0,\ \operatorname{div} C_\varepsilon = 0 \quad \text{in } Q, \\ &z_\varepsilon = 0,\ C_\varepsilon\cdot N = 0,\ \operatorname{curl} C_\varepsilon = 0 \quad \text{on } \Sigma, \\ &z_\varepsilon(\cdot,T) = -\frac{1}{\varepsilon}\,y_\varepsilon(\cdot,T),\ C_\varepsilon(\cdot,T) = -\frac{1}{\varepsilon}\,B_\varepsilon(\cdot,T) \quad \text{in } \Omega \end{aligned}$$

and the following maximum conditions:

$$u_\varepsilon = \chi_\omega e^{2s\delta\widehat{\beta}} z_\varepsilon, \quad v_\varepsilon = \chi_\omega e^{2s\delta\widehat{\beta}} C_\varepsilon \quad \text{a.e. in } Q. \tag{5.4}$$

We notice that system (5.3) has the form (3.1) with $h = e^{-2s\delta\widehat{\beta}_\varepsilon} y_\varepsilon$ and $H = e^{-2s\delta\widehat{\beta}_\varepsilon} B_\varepsilon$.

Let us integrate the derivative

$$\frac{d}{dt}\int_\Omega (y_\varepsilon \cdot z_\varepsilon + B_\varepsilon \cdot C_\varepsilon)dx$$

from 0 to T and then use (5.1), (5.3), and (5.4). We obtain

$$\begin{aligned}
&\int_{Q_\omega} e^{2s\delta\widehat{\beta}} \left(|z_\varepsilon|^2 + |C_\varepsilon|^2\right) dx\,dt + \int_Q e^{-2s\delta\widehat{\beta}_\varepsilon} \left(|y_\varepsilon|^2 + |B_\varepsilon|^2\right) dx\,dt \\
&\qquad + \frac{1}{\varepsilon}\int_\Omega \left(|y_\varepsilon(x,T)|^2 + |B_\varepsilon(x,T)|^2\right) dx \\
&\qquad = -\int_Q (f\cdot z_\varepsilon + F\cdot C_\varepsilon)dx\,dt - \int_\Omega (y_0(x)\cdot z_\varepsilon(x,0) + B_0(x)\cdot C_\varepsilon(x,0))dx \\
&\qquad \le \left(\int_Q e^{-2s\widehat{\beta}} \frac{1}{(T-t)^8} \left(|f|^2 + |F|^2\right) dx\,dt\right)^{\frac{1}{2}} \\
&\qquad\quad \times \left(\int_Q e^{2s\widehat{\beta}} (T-t)^8 (|z_\varepsilon|^2 + |C_\varepsilon|^2) dx\,dt\right)^{\frac{1}{2}} \\
&\qquad\quad + \left(|y_0|^2_{(L^2(\Omega))^2} + |B_0|^2_{(L^2(\Omega))^2}\right)^{\frac{1}{2}} \left(\int_\Omega (|z_\varepsilon(x,0)|^2 + |C_\varepsilon(x,0)|^2) dx\right)^{\frac{1}{2}}.
\end{aligned} \tag{5.5}$$

Taking inequality (5.5) together with the observability inequalities (4.1) and (4.18) applied to system (5.3), we have

$$\begin{aligned}
&\int_{Q_\omega} e^{2s\delta\widehat{\beta}} \left(|z_\varepsilon|^2 + |C_\varepsilon|^2\right) dx\,dt + \int_Q e^{-2s\delta\widehat{\beta}_\varepsilon} \left(|y_\varepsilon|^2 + |B_\varepsilon|^2\right) dx\,dt \\
&\qquad \le c\left(|y_0|^2_{(L^2(\Omega))^2} + |B_0|^2_{(L^2(\Omega))^2} + \int_Q e^{-2s\widehat{\beta}} \frac{1}{(T-t)^8} \left(|f|^2 + |F|^2\right) dx\,dt\right),
\end{aligned} \tag{5.6}$$

where c is a positive parameter depending on λ, s, and δ. Passing from z_ε and C_ε to u_ε and v_ε in (5.6) by using (5.4), we obtain

$$\begin{aligned}
&\int_{Q_\omega} e^{-2s\delta\widehat{\beta}} \left(|u_\varepsilon|^2 + |v_\varepsilon|^2\right) dx\,dt + \int_Q e^{-2s\delta\widehat{\beta}_\varepsilon} \left(|y_\varepsilon|^2 + |B_\varepsilon|^2\right) dx\,dt \\
&\qquad \le c\left(|y_0|^2_{(L^2(\Omega))^2} + |B_0|^2_{(L^2(\Omega))^2} + \int_Q e^{-2s\widehat{\beta}} \frac{1}{(T-t)^8} \left(|f|^2 + |F|^2\right) dx\,dt\right),
\end{aligned} \tag{5.7}$$

where the positive constant c is independent of ε. Moreover, now putting (5.5), (4.1), (4.2), and (5.6) together, we have

$$\int_\Omega (|y_\varepsilon(x,T)|^2 + |B_\varepsilon(x,T)|^2) dx \le c\varepsilon \tag{5.8}$$

for some positive constant c independent of ε.

By virtue of (5.7) and (5.2), there exists $(u, v, y, B, p) \in (L^2(Q))^4 \times (H^{2,1}(Q))^4 \times L^2(0,T;H^1(\Omega))$ such that, on a subsequence of $\{\varepsilon\}$, as $\varepsilon \longrightarrow 0$,

$$\begin{aligned} &u_\varepsilon \longrightarrow u, \quad v_\varepsilon \longrightarrow v && \text{weakly in } (L^2(Q))^2,\\ &y_\varepsilon \longrightarrow y, \quad B_\varepsilon \longrightarrow B && \text{weakly in } (H^{2,1}(Q))^2,\\ &p_\varepsilon \longrightarrow p && \text{weakly in } L^2(0,T;H^1(\Omega)). \end{aligned}$$

Thus, letting $\varepsilon \longrightarrow 0$ in (5.1), where u, v, y, B, and p are replaced by u_ε, v_ε, y_ε, B_ε, and p_ε, we see that u, v, y, B, and p satisfy (5.1), too. In addition, by (5.8), we have

$$y(x,T) = 0 \ \text{ and } \ B(x,T) = 0 \ \text{ a.e. } \ x \in \Omega.$$

Finally, the indicated decay of u, v, y, and B at $t = T$ is obtained as in [6]. So, the proof of Theorem 5.1 is complete. □

6. Proof of Theorem 2.1. We shall reduce the controllability of the solution $(\widetilde{y}, \widetilde{B})$ of the MHD equations (2.3) to the controllability of the null solution of a certain version of (2.3) by subtracting (2.1) and (2.3). The differences $y - \widetilde{y}$, $B - \widetilde{B}$, and $p - \widetilde{p}$, also denoted by y, B, and p, satisfy

$$\begin{aligned} &\frac{\partial y}{\partial t} - \Delta y + (y\cdot\nabla)y + (\widetilde{y}\cdot\nabla)y + (y\cdot\nabla)\widetilde{y}\\ &\qquad - (B\cdot\nabla)B - (\widetilde{B}\cdot\nabla)B - (B\cdot\nabla)\widetilde{B}\\ &\qquad + \nabla\left(\frac{1}{2}\ B^2\right) + \nabla(\widetilde{B}\cdot B) + \nabla p = \chi_\omega u && \text{in } Q,\\ &\frac{\partial B}{\partial t} + \widetilde{\operatorname{curl}}(\operatorname{curl} B) + (y\cdot\nabla)B + (\widetilde{y}\cdot\nabla)B + (y\cdot\nabla)\widetilde{B}\\ &\qquad - (B\cdot\nabla)y - (\widetilde{B}\cdot\nabla)y - (B\cdot\nabla)\widetilde{y} = P(\chi_\omega v) && \text{in } Q,\\ &\operatorname{div} y = 0, \ \operatorname{div} B = 0 && \text{in } Q,\\ &y = 0, \ B\cdot N = 0, \ \operatorname{curl} B = 0 && \text{on } \Sigma,\\ &y(\cdot,0) = y_0, \ B(\cdot,0) = B_0 && \text{in } \Omega. \end{aligned} \tag{6.1}$$

The differences $y_0 - \widetilde{y}(\cdot,0)$ and $B_0 - \widetilde{B}(\cdot,0)$ have been denoted here by y_0 and B_0, too. In this way, the original controllability problem can be replaced by that of finding (u, v, y, B, p), which satisfies (6.1) but also $y(\cdot,T) = 0$ and $B(\cdot,T) = 0$ a.e. in Ω.

The null controllability problem for (6.1) can be reformulated as an invertibility property for a certain nonlinear map. Let us define this map in what follows after introducing the needed function spaces.

We take $\lambda > 0$, $s > 0$, and $\delta \in (1/2, 1)$ as in the statement of Theorem 5.1. Let $1/2 < \delta' < \delta$. We denote by $X(Q)$ the space of all $(u, v, y, B, p) \in (L^2(Q))^4 \times (H^{2,1}(Q))^4 \times L^2(0,T;H^1(\Omega))$ which satisfy $(e^{-s\delta'\widehat{\beta}}y, e^{-s\delta'\widehat{\beta}}B) \in (H^{2,1}(Q))^4$, $\partial y/\partial t - \Delta y + (\widetilde{y}\cdot\nabla)y + (y\cdot\nabla)\widetilde{y} - (\widetilde{B}\cdot\nabla)B + (B\cdot\nabla)\widetilde{B} + \nabla(\widetilde{B}\cdot B) + \nabla p - \chi_\omega u \in (L^2(Q,(T-t)^{-8}e^{-2s\widehat{\beta}}))^2$, $\partial B/\partial t + \widetilde{\operatorname{curl}}\,(\operatorname{curl} B) + (\widetilde{y}\cdot\nabla)B + (y\cdot\nabla)\widetilde{B} - (\widetilde{B}\cdot\nabla)y - (B\cdot\nabla)\widetilde{y} - P(\chi_\omega v) \in (L^2(Q,(T-t)^{-8}e^{-2s\widehat{\beta}}))^2$, and

$$\begin{aligned} &\operatorname{div} y = 0, \ \operatorname{div} B = 0 && \text{in } Q,\\ &y = 0, \ B\cdot N = 0, \ \operatorname{curl} B = 0 && \text{on } \Sigma,\\ &y(\cdot,T) = 0, \ B(\cdot,T) = 0 && \text{a.e. in } \Omega. \end{aligned}$$

The space $X(Q)$ becomes a Banach space if we endow it with the norm

$$
\begin{aligned}
|(u,v,y,B,p)|_{X(Q)} = \Bigg(& |u|^2_{(L^2(Q))^2} + |v|^2_{(L^2(Q))^2} \\
& + |e^{-s\delta'\widehat{\beta}} y|^2_{(H^{2,1}(Q))^2} + |e^{-s\delta'\widehat{\beta}} B|^2_{(H^{2,1}(Q))^2} + |p|^2_{L^2(0,T;H^1(\Omega))} \\
& + \int_Q e^{-2s\widehat{\beta}} \frac{1}{(T-t)^8} \Bigg(\left| \frac{\partial y}{\partial t} - \Delta y + (\widetilde{y}\cdot\nabla) y + (y\cdot\nabla)\widetilde{y} \right. \\
& \left. - (\widetilde{B}\cdot\nabla) B - (B\cdot\nabla)\widetilde{B} + \nabla(\widetilde{B}\cdot B) + \nabla p - \chi_\omega u \right|^2 \\
& + \left| \frac{\partial B}{\partial t} + \widetilde{\operatorname{curl}}\,(\operatorname{curl} B) + (\widetilde{y}\cdot\nabla) B + (y\cdot\nabla)\widetilde{B} \right. \\
& \left. - (\widetilde{B}\cdot\nabla) y - (B\cdot\nabla)\widetilde{y} - P(\chi_\omega v) \right|^2 \Bigg) \, dx\, dt \Bigg)^{\frac{1}{2}}.
\end{aligned}
$$

We denote by $Y(Q)$ the space of all $(f,F,y_0,B_0) \in (L^2(Q,(T-t)^{-8}e^{-2s\widehat{\beta}}))^4 \times V_1 \times V_2$ such that $\operatorname{div} F = 0$ in Q and $F\cdot N = 0$ on Σ, which becomes a Banach space, too, if we endow it with the product norm

$$
\begin{aligned}
& |(f,F,y_0,B_0)|_{Y(Q)} \\
& \quad = \left(\int_Q e^{-2s\widehat{\beta}} \frac{1}{(T-t)^8} \left(|f|^2 + |F|^2 \right) dx\, dt + |y_0|^2_{(H^1(\Omega))^2} + |B_0|^2_{(H^1(\Omega))^2} \right)^{1/2}.
\end{aligned}
$$

Now the nonlinear map $\mathcal{A} : X(Q) \longrightarrow Y(Q)$ is defined as follows: for $(u,v,y,B,p) \in X(Q)$,

$$
\begin{aligned}
& \mathcal{A}(u,v,y,B,p) \\
& \quad = \Bigg(\frac{\partial y}{\partial t} - \Delta y + (y\cdot\nabla) y + (\widetilde{y}\cdot\nabla) y + (y\cdot\nabla)\widetilde{y} \\
& \qquad - (B\cdot\nabla) B - (\widetilde{B}\cdot\nabla) B - (B\cdot\nabla)\widetilde{B} \\
& \qquad + \nabla\left(\frac{1}{2} B^2 \right) + \nabla(\widetilde{B}\cdot B) + \nabla p - \chi_\omega u, \\
& \qquad \frac{\partial B}{\partial t} + \widetilde{\operatorname{curl}}\,(\operatorname{curl} B) + (y\cdot\nabla) B + (\widetilde{y}\cdot\nabla) B + (y\cdot\nabla)\widetilde{B} \\
& \qquad - (B\cdot\nabla) y - (\widetilde{B}\cdot\nabla) y - (B\cdot\nabla)\widetilde{y} - P(\chi_\omega v), \\
& \qquad\quad y(\cdot,0), B(\cdot,0) \Bigg).
\end{aligned}
$$

One can show (see [7] or [6]) that if $\delta' > 1/2$, then $\mathcal{A}(u,v,y,B,p) \in Y(Q)$ for $(u,v,y,B,p) \in X(Q)$.

It is easy to calculate the differential of $\mathcal{A}$:

$$\begin{aligned}
&((d\mathcal{A})(\overline{u},\overline{v},\overline{y},\overline{B},\overline{p}))(u,v,y,B,p)\\
&\quad=\Bigg(\frac{\partial y}{\partial t}-\Delta y+(\overline{y}\cdot\nabla)y+(y\cdot\nabla)\overline{y}+(\widetilde{y}\cdot\nabla)y+(y\cdot\nabla)\widetilde{y}\\
&\qquad-(\overline{B}\cdot\nabla)B-(B\cdot\nabla)\overline{B}-(\widetilde{B}\cdot\nabla)B-(B\cdot\nabla)\widetilde{B}\\
&\qquad+\nabla(\overline{B}\cdot B)+\nabla(\widetilde{B}\cdot B)+\nabla p-\chi_\omega u,\\
&\qquad\frac{\partial B}{\partial t}+\widetilde{\operatorname{curl}}\,(\operatorname{curl}B)+(\overline{y}\cdot\nabla)B+(y\cdot\nabla)\overline{B}+(\widetilde{y}\cdot\nabla)B+(y\cdot\nabla)\widetilde{B}\\
&\qquad-(\overline{B}\cdot\nabla)y-(B\cdot\nabla)\overline{y}-(\widetilde{B}\cdot\nabla)y-(B\cdot\nabla)\widetilde{y}-P(\chi_\omega v),\\
&\qquad\quad y(\cdot,0),B(\cdot,0)\Bigg).
\end{aligned}$$

This differential is continuous (see [7]).

Moreover, $\mathcal{A}(0,0,0,0,0)=(0,0,0,0)$. So, we have to show that there exists $\eta>0$ such that for any $(f,F,y_0,B_0)\in Y(Q)$ satisfying

$$|(f,F,y_0,B_0)|_{Y(Q)}<\eta$$

one can find $(u,v,y,B,p)\in X(Q)$ such that

$$\mathcal{A}(u,v,y,B,p)=(f,F,y_0,B_0).$$

(In fact, here it suffices to take $f=F=0$.) According to an infinite-dimensional version of the implicit function theorem (see [1, p. 101]), a sufficient condition assuring such a local invertibility property for $\mathcal{A}$ around $(0,0,0,0,0)$ is that $(d\mathcal{A})(0,0,0,0,0):X(Q)\longrightarrow Y(Q)$ should be an epimorphism. But this expresses nothing else than the global null controllability property for system (5.1) stated by Theorem 5.1. Thus, the proof of Theorem 2.1 is finished.

REFERENCES

[1] V. M. Alekseev, V. M. Tikhomirov, and S. V. Fomin, *Optimal Control*, Contemp. Soviet Math., Consultant Bureau, New York, 1987.

[2] V. Barbu, T. Havârneanu, C. Popa, and S. S. Sritharan, *Exact controllability for the magnetohydrodynamic equations*, Commun. Pure Appl. Math., 56 (2003), pp. 732–783.

[3] V. Barbu, T. Havârneanu, C. Popa, and S. S. Sritharan, *Local exact controllability for the magnetohydrodynamic equations revisited*, Adv. Differential Equations, 10 (2005), pp. 481–504.

[4] E. Fernández-Cara, S. Guerrero, O. Yu. Imanuvilov, and J.-P. Puel, *Local exact controllability of the Navier-Stokes system*, J. Math. Pures Appl. (9), 83 (2004), pp. 1501–1542.

[5] A. V. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul National University, Research Institute of Mathematics, Global Analysis Research Center, Seoul, 1996.

[6] T. Havârneanu, C. Popa, and S. S. Sritharan, *Exact controllability for the three-dimensional Navier-Stokes equations with the Navier slip boundary conditions*, Indiana Univ. Math. J., 54 (2005), pp. 1303–1350.

[7] T. Havârneanu, C. Popa, and S. S. Sritharan, *Exact internal controllability for the magnetohydrodynamic equations in multi-connected domains*, Adv. Differential Equations, 11 (2006), pp. 893–929.

[8] O. Yu. Imanuvilov, *Boundary controllability of parabolic equations*, Sb. Math., 186 (1995), pp. 879–900.

[9] O. Yu. Imanuvilov, *Remarks on exact controllability for the Navier-Stokes equations*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 39–72.
[10] O. Yu. Imanuvilov and J.-P. Puel, *Global Carleman estimates for weak elliptic nonhomogeneous Dirichlet problem*, Int. Math. Res. Not., 16 (2003), pp. 883–913.
[11] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, 3rd ed., Stud. Math. Appl. 2, North-Holland, Amsterdam, 1984.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1831–1848

LQG BALANCING FOR CONTINUOUS-TIME INFINITE-DIMENSIONAL SYSTEMS*

MARK R. OPMEER†

Abstract. In this paper we study the existence of linear quadratic Gaussian (LQG)-balanced realizations for continuous-time infinite-dimensional systems. LQG-balanced realizations are those for which the optimal cost operator for the system and its dual system are equal (and diagonal). The class of systems we consider is that of distributional resolvent linear systems which includes well-posed linear systems as a subclass. We prove the existence of LQG-balanced realizations under a finite cost condition for both the system and its dual system. We also show that an LQG-balanced realization of a well-posed transfer function is well-posed. We further show that approximately controllable and observable LQG-balanced realizations are unique up to a unitary state-space transformation. Finally, we show that the spectrum of the product of the optimal cost operator of a system and its dual system is independent of the particular realization. Our method of proof shows the connections with coprime factorizations, Lyapunov-balanced realizations, and discrete-time systems. The main reason for studying LQG-balanced realizations is that truncated LQG-balanced realizations provide a good approximation of the original system. We show that, under certain conditions, this is also true in the infinite-dimensional case by proving an error bound in the gap-metric.

Key words. balanced realization, infinite-dimensional system, LQG-balanced realization, coprime factorization, Riccati equation

AMS subject classifications. 47A48, 47N70, 93B28, 93C55

DOI. 10.1137/050638229

1. Introduction.

Simple models are normally preferred over complex ones in control systems design. Sometimes it is obvious how to construct a simple model for a physical system, but sometimes it is not obvious what the characteristics essential to the controller design of a physical system are. One way of obtaining a simple model in the latter case is to first obtain a sophisticated model that takes every aspect of possible interest into account and then perform model reduction on this sophisticated model. A simple model reduction procedure was introduced by Moore [9] and is now a textbook subject (see, e.g., Zhou and Doyle [24, Chapter 7]). The method proposed by Moore consists of truncating a balanced realization. A balanced realization (also called Lyapunov- or internally balanced) is a realization for which the controllability and observability gramians are equal and diagonal. Lyapunov-balanced realizations are popular because they are relatively easy to compute and there exists an error bound in the H-infinity norm on the basis of which one can show that compensators based on the reduced order model have a certain performance when applied to the full order system. The Lyapunov-balanced realization method is applicable only to stable systems. Alternatively for unstable systems one can use truncations of a linear quadratic Gaussian (LQG)-balanced realization, which for rational transfer functions always exists. An LQG-balanced realization is a realization for which the optimal cost operator for the system and its dual system (with respect to the standard quadratic cost functional) are equal and diagonal. This method was proposed by Verriest [20], [21] and further developed by Jonckheere and Silverman [7].

*Received by the editors August 16, 2005; accepted for publication (in revised form) April 2, 2007; published electronically November 14, 2007.
http://www.siam.org/journals/sicon/46-5/63822.html
†Department of Mathematics, University of California Davis, One Shields Avenue, Davis, CA 95616-8633 (opmeer@math.ucdavis.edu).

For an alternative treatment see Mustafa and Glover [10], and for the discrete-time case see Hoffmann, Prätzel-Wolters, and Zerz [6]. The computation of an LQG-balanced realization can also be performed reasonably efficiently, and there exists an error bound in the gap-metric which provides advantages similar to those of the H-infinity error bound for the truncated Lyapunov-balanced realization.

In the case that the system is infinite-dimensional, the model/controller approximation becomes essential. One would like to use the methods of balanced truncation and LQG-balanced truncation in this case, too.

The existence of Lyapunov-balanced and LQG-balanced realizations for irrational transfer functions is nontrivial. A necessary and sufficient condition for the existence of Lyapunov-balanced realizations in discrete time was given by Young [23], [22] (see [19, section 9.5] for the continuous-time case). A necessary and sufficient condition for the existence of LQG-balanced realizations for discrete-time systems was given in [16]. The first of the main results of the present article shows the analogous result for the continuous-time case.

As in the finite-dimensional case it is essential for controller design to have convergence in the H-infinity norm (for Lyapunov-balanced realizations) or the gap-metric (for LQG-balanced realizations); see [2]. Additional assumptions need to be made to ensure this. Under appropriate additional assumptions, a priori error bounds ensuring such convergence were given in [5] for continuous-time Lyapunov-balanced realizations and in [1] for discrete-time Lyapunov-balanced realizations. The second of the main aims of the present article is to provide a priori error bounds in the gap-metric for LQG-balanced realizations in both discrete and continuous time.

The class of continuous-time systems we consider is very general: it includes virtually all causal time-invariant linear systems studied in the literature. Details on this class of systems are given in section 4.

The proofs of our results are based on the discrete-time case [16] supplemented by recent results on coprime factorizations [3] and on the Cayley transform and the linear quadratic regulator (LQR) problem [15], [12].

2. LQG-balanced realizations: The finite-dimensional case. In this section we review some of the results on finite-dimensional LQG-balanced realizations. We consider systems of the form

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \quad y(t) = Cx(t) + Du(t), \tag{1}$$

where A, B, C, D are matrices of compatible dimensions. We consider the linear quadratic regulator (LQR) problem for the cost functional

$$J(x_0, u) := \int_0^\infty \|u(t)\|^2 + \|y(t)\|^2 \, dt,$$

where y is given in terms of x_0 and u by (1). The LQR problem consists of finding for a given x_0 that u for which $J(x_0, u)$ is minimal. As is well known, this problem has a unique solution when the system is minimal: the optimal input u^{opt} is given by the state feedback $u^{\mathrm{opt}}(t) = -(I + D^*D)^{-1}(D^*C + B^*Q)x(t)$, where Q is the unique nonnegative solution of the Riccati equation

$$A^*Q + QA + C^*C = (C^*D + QB)(I + D^*D)^{-1}(D^*C + B^*Q),$$

and the optimal cost is given by $J(x_0, u^{\mathrm{opt}}) = \langle x_0, Qx_0 \rangle$. By duality the "optimal filter cost" is given by $\langle x_0, Px_0 \rangle$, where P is the unique nonnegative solution of the

Riccati equation

$$PA^* - AP + BB^* = (BD^* + PC^*)(I + DD^*)^{-1}(DB^* + CP).$$

The quantity $\langle x_0, Px_0\rangle$ can be interpreted as a measure of the difficulty of reconstructing the initial state x_0 from noisy measurements. The eigenvalues of the product PQ are similarity invariants; their square roots are called the LQG-characteristic values of the system. These invariants can be interpreted as a measure of how important the subspace generated by the eigenvector is for the compensator design. This can be seen from the LQG-balanced realization. An LQG-balanced realization is a realization such that $P = Q = \Lambda$, where Λ is the diagonal matrix containing the LQG-characteristic values. Let λ_i be the square root of an eigenvalue of PQ with eigenvector x_i of length one. Then, in the LQG-balanced realization, the optimal cost with initial condition x_i is λ_i and the difficulty of reconstructing this initial state from noisy measurements is also λ_i. The idea behind LQG-balanced truncation is to restrict the system to the subspace generated by the eigenvectors corresponding to the largest eigenvalues. Since this subspace is most important for compensator design, the system obtained by LQG-balanced truncation seems to be a reasonable approximation. As mentioned in the introduction there is also an error bound which justifies the above heuristics. Let δ_g denote the gap-metric (see Zhou and Doyle [24, Chapter 7]), Σ the original (n-dimensional) system, and Σ_k the k-dimensional LQG-balanced truncation. Then

$$\delta_g(\Sigma, \Sigma_k) \le 2 \sum_{i=k+1}^{n} \frac{\lambda_i}{\sqrt{1+\lambda_i^2}};$$

see Mustafa and Glover [10, section 8.4.5].

3. Discrete-time systems. In this section we review the results in [16] on discrete-time infinite-dimensional LQG-balanced realizations and give some extensions. The discrete-time case is a key ingredient for the proof in continuous time.

Let $\mathcal{U}$, $\mathcal{X}$, $\mathcal{Y}$ be separable Hilbert spaces and

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathcal{L}\left(\begin{bmatrix} \mathcal{X} \\ \mathcal{U} \end{bmatrix}, \begin{bmatrix} \mathcal{X} \\ \mathcal{Y} \end{bmatrix}\right).$$

Such a block operator will be called a *discrete-time system*. We will also denote such a block operator using the notation $[A, B; C, D]$ (we denote a block row of operators by $[X, Y]$ and a block column by $[X; Y]$). We denote the set of nonnegative integers by $\mathbb{Z}^+$. For a given initial state $x_0 \in \mathcal{X}$ and input $u : \mathbb{Z}^+ \to \mathcal{U}$ define the state $x : \mathbb{Z}^+ \to \mathcal{X}$ and output $y : \mathbb{Z}^+ \to \mathcal{Y}$ by

$$x_{n+1} = Ax_n + Bu_n, \quad x_0 = x^0, \quad y_n = Cx_n + Du_n. \tag{2}$$

A sequence $h : \mathbb{Z}^+ \to \mathcal{H}$ is called Z-transformable if the power series

$$\sum_{i=0}^{\infty} h_i z^i$$

has a positive radius of convergence. The Z-transform of a Z-transformable sequence h is defined to be the sum of this series and is denoted by $\hat{h}$. For operators A, B, C, D as above define the *transfer function* $G : \mathbb{D}_r \to \mathcal{L}(\mathcal{U}, \mathcal{Y})$ by

$$G(z) = D + \sum_{i=0}^{\infty} CA^i B z^{i+1},$$

where $\mathbb{D}_r$ is defined to be the largest disc centered at the origin for which the above sum converges (note that it definitely converges on the disc centered at the origin with radius $1/r(A)$, where $r(A)$ is the spectral radius of the operator A). If the input sequence u is Z-transformable, then the output sequence y is also Z-transformable, and if $x^0 = 0$, then the Z-transform of the output is given by

$$\hat{y}(z) = G(z)\hat{u}(z)$$

on some neighborhood of the origin. The function $D + Cz(I - zA)^{-1}B$ is called the *characteristic function* of the discrete-time system. Note that the transfer function and the characteristic function are equal on some neighborhood of the origin but may not be identically equal. A discrete-time system is called a *realization* of the function G if $G(z) = D + Cz(I - zA)^{-1}B$ on some neighborhood of the origin. Any $\mathcal{L}(\mathcal{U}, \mathcal{Y})$-valued function that is holomorphic at the origin can be realized as the transfer function of some discrete-time system. This discrete-time system is far from unique.

3.1. Lyapunov-balanced realizations in discrete time. Although we are studying LQG-balanced realizations, we do this by relating them to Lyapunov-balanced realizations. In this subsection we review some results on Lyapunov-balanced realizations that are needed in what follows. To define what we exactly mean by a Lyapunov-balanced realization we first have to define the input and output maps and the gramians of a discrete-time system.

The input map of a discrete-time system is defined for finitely nonzero $u : \mathbb{Z}^- \to \mathcal{U}$ by (here $\mathbb{Z}^-$ is the set of negative integers)

$$\mathcal{B}u := \sum_{i=0}^{\infty} A^i B u_{-i-1}.$$

A discrete-time system is called *approximately controllable* if the range of $\mathcal{B}$ is dense in $\mathcal{X}$, and it is called *input stable* if $\mathcal{B}$ extends to a bounded operator from $l^2(\mathbb{Z}^-; \mathcal{U})$ to $\mathcal{X}$. For an input stable discrete-time system we define the *controllability gramian* $L_B \in \mathcal{L}(\mathcal{X})$ by $L_B := \mathcal{B}\mathcal{B}^*$.

The output map of a discrete-time system is defined for $x \in \mathcal{X}$ by

$$(\mathcal{C}x)_k := CA^k x, \quad k \in \mathbb{Z}^+.$$

A discrete-time system is called *approximately observable* if $\mathcal{C}$ is one-to-one, and it is called *output stable* if $\mathcal{C}$ is a bounded operator from $\mathcal{X}$ to $l^2(\mathbb{Z}^+; \mathcal{Y})$. For an output stable discrete-time system we define the *observability gramian* $L_C \in \mathcal{L}(\mathcal{X})$ by $L_C := \mathcal{C}^*\mathcal{C}$. A discrete-time system is called *minimal* if it is both approximately controllable and approximately observable.

The Hankel operator $\mathcal{H}$ of a discrete-time system is defined for finitely nonzero $u : \mathbb{Z}^- \to \mathcal{U}$ by

$$(\mathcal{H}u)_k = \sum_{i=0}^{\infty} CA^i B u_{k-i-1}, \quad k \in \mathbb{Z}^+.$$

Note that $\mathcal{H} = \mathcal{C}\mathcal{B}$ and that $\mathcal{H}$ depends only on the transfer function of the system.

DEFINITION 3.1. *A discrete-time system is called* Lyapunov-balanced *if it is input and output stable and $L_B = L_C$, and it is called* compact Lyapunov-balanced *if, in addition, $L_B = L_C$ is compact.*

Young [23], [22] proved that every holomorphic uniformly bounded function on the unit disc (i.e., every element of $\mathbf{H}^\infty(\mathbb{D}, \mathcal{L}(\mathcal{U}, \mathcal{Y}))$) has a minimal Lyapunov-balanced realization. He also noted that if the Hankel operator that has this given function as symbol is compact then there exists a minimal compact Lyapunov-balanced realization. A simplification of the proof of Young can be found in Peller [17, section 11.2] and an alternative proof can be found in Staffans [19, section 9.5]. Young [23], [22] has also shown that minimal Lyapunov-balanced realizations are unique up to a unitary transformation in the state space. Let $[A, B; C, D]$ be a compact Lyapunov-balanced realization and denote by $P : \mathcal{X} \to \mathcal{X}$ the projection onto the subspace spanned by the eigenvectors of $L_B = L_C$ corresponding to the largest n eigenvalues (eigenvalues are counted with multiplicity and it is assumed here that the $n+1$st eigenvalue is different from the nth eigenvalue). Then $[PAP, PB; CP, D]$ is called the n-dimensional *truncated Lyapunov-balanced realization*. Note that the n-dimensional truncated Lyapunov-balanced realization may not be defined for every $n \in \mathbb{Z}^+$ due to repeated eigenvalues. Since eigenvalues of compact operators have finite multiplicity it is defined for infinitely many values of n. When we mention n-dimensional truncated Lyapunov-balanced realizations in what follows we will implictely assume that n is such that this notion is well defined. A truncated Lyapunov-balanced realization is not unique, since the Lyapunov-balanced realization is not. However, since two minimal Lyapunov-balanced realizations of the same transfer function are unitarily equivalent, so are all n-dimensional truncated balanced realizations. Consequently the transfer function of an n-dimensional truncated balanced realization G^n is well defined. The whole idea of Lyapunov-balanced realizations is that G^n is a good approximation of G. That this is indeed the case under certain conditions was proven by Bonnet [1]. The result of Bonnet is the discrete-time version of the continuous-time result in [5]. We summarize the results of Young and Bonnet in the following theorem. We note that the singular values of a compact operator T are the square roots of the eigenvalues of T^*T and that an operator is called *nuclear* if it is compact and its singular values form a summable sequence. The *Hankel singular values* of a system are the singular values of its Hankel operator.

THEOREM 3.2. 1. *Every function in* $\mathbf{H}^\infty(\mathbb{D}, \mathcal{L}(\mathcal{U}, \mathcal{Y}))$ *has a minimal Lyapunov-balanced realization. If the Hankel operator of the function is compact, then it has a minimal compact Lyapunov-balanced realization.*

2. *If, in addition, the Hankel operator of the function is nuclear and the input and output spaces are finite-dimensional, then*

$$\|G - G^n\|_\infty \leq 2 \sum_{i=n+1}^{\infty} \sigma_i,$$

where G^n *is the transfer function of a truncated compact Lyapunov-balanced realization of* G *and the* σ_i *are the Hankel singular values.*

Part 2 of the above theorem was proven in [1], following the continuous-time version in [5], only for the case where the Hankel singular values are distinct. As indicated in [5] the generalization to the case of possibly repeating Hankel singular values is not difficult, except notationally. Details may be found in [14, Chapter 10].

3.2. LQG-balanced realizations in discrete time. In this subsection we summarize the results obtained in [16] on LQG-balanced realizations in discrete time. We also extend these by obtaining an error bound on truncated compact LQG-balanced realizations.

To exactly define the concept of an LQG-balanced discrete-time system we first consider the LQR problem. This problem is as follows: for given $x^0 \in \mathcal{X}$ find an input u that minimizes

$$J(x^0, u) := \sum_{n=0}^{\infty} \|u_n\|^2 + \|y_n\|^2, \tag{3}$$

where y is given in terms of x^0 and u by (2). We introduce the following concept: a discrete-time system satisfies the *finite cost condition* if for every $x^0 \in \mathcal{X}$ there exists a $u \in l^2(\mathbb{Z}^+;\mathcal{U})$ such that the corresponding output $y \in l^2(\mathbb{Z}^+;\mathcal{Y})$. It is well known (see, e.g., [4]) that if the finite cost condition is satisfied, then for every $x^0 \in \mathcal{X}$ there exists a unique $u^{\mathrm{opt}} \in l^2(\mathbb{Z}^+;\mathcal{U})$ that minimizes the cost function (3) and there exists a bounded nonnegative operator Q such that the minimal cost is given by $\langle Qx^0, x^0\rangle$. This operator Q is called the *optimal cost operator*. Similarly, if for the *dual system*

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^* = \begin{bmatrix} A^* & C^* \\ B^* & D^* \end{bmatrix}$$

the finite cost condition is satisfied, then there exists an optimal cost operator P for this dual system. This operator P is called the *dual optimal cost operator* of the original system.

Definition 3.3. *A discrete-time system is called* LQG-balanced *if it and its dual both satisfy the finite cost condition and $P = Q$, and it is called* compact LQG-balanced *if, in addition, $P = Q$ is compact.*

Below we state not only the main results obtained in [16], but also some main steps in the proof. These intermediate results are also necessary to obtain the continuous-time analogues.

The optimal cost operator satisfies the following Riccati equation:

$$A^*QA - Q + C^*C = (C^*D + A^*QB)(I + D^*D + B^*QB)^{-1}(B^*QA + D^*C).$$

The optimal input u^{opt} can be given by a state feedback. To explain this we consider the concept of an admissible state feedback pair.

Definition 3.4. *An admissible state feedback pair for a discrete-time system is a pair $[K, F] \in \mathcal{L}([\mathcal{X},\mathcal{U}],\mathcal{U})$ such that $I - F$ is boundedly invertible. The closed-loop system is given by*

$$A^{\mathrm{cl}} := A + B(I - F)^{-1}K, \quad B^{\mathrm{cl}} := B(I - F)^{-1},$$

$$C^{\mathrm{cl}} := \begin{bmatrix} (I - F)^{-1}K \\ C + D(I - F)^{-1}K \end{bmatrix}, \quad D^{\mathrm{cl}} := \begin{bmatrix} (I - F)^{-1} \\ D(I - F)^{-1} \end{bmatrix}.$$

This closed-loop system is obtained by adding the equation $u_n = Kx_n + Fu_n + r_n$ to (2), considering $[u; y]$ as the new output and r as the new input, and solving. The state feedback pair

$$K := -(I + D^*D + B^*QB)^{-1/2}(D^*C + B^*QA), \quad F := I - (I + D^*D + B^*QB)^{-1/2}$$

is admissible and with zero input and initial condition x^0 the output of the closed-loop system is exactly $[u^{\mathrm{opt}}; y^{\mathrm{opt}}]$, the optimal input and output for the system $[A, B; C, D]$. The closed-loop system with this specific admissible state feedback pair will be called the *optimal closed-loop system* corresponding to the system $[A, B; C, D]$.

We give some properties of the optimal closed-loop system that were proven in [16] and some that follow from results in [3]. To formulate these we first recall the concept of a (normalized) right coprime factor.

DEFINITION 3.5. *A function* $[M;N] \in \mathbf{H}^\infty(\mathbb{D}, \mathcal{L}(\mathcal{U}, [\mathcal{U};\mathcal{Y}]))$ *is called a right* factor *of a function* G *if* M *is invertible on some neighborhood of the origin and* $G = NM^{-1}$ *on this region.* M *and* N *as above are called right* coprime *if there exists* $[\tilde{X}, \tilde{Y}] \in \mathbf{H}^\infty(\mathbb{D}, \mathcal{L}([\mathcal{U},\mathcal{Y}],\mathcal{U}))$ *such that* $\tilde{X}M - \tilde{Y}N = I$ *on the unit disc.* $[M;N]$ *as above is called* normalized *if* $M^*M + N^*N = I$ *almost everywhere on the unit circle.*

We note that normalized right coprime factors are unique up to a unitary transformation in $\mathcal{L}(\mathcal{U})$. The following theorem relates factorizations to the optimal closed-loop system.

THEOREM 3.6. *If the system* $[A, B; C, D]$ *satisfies the finite cost condition, then the transfer function of its optimal closed-loop system is a normalized right factor of its transfer function. If, in addition, the dual system also satisfies the finite cost condition, then this factor is right coprime.*

Proof. The first part of the proof follows from Corollary 5.8 of [16]. The second part follows from Lemma 6.7 of [16] and the discrete-time version of [3, Corollary 7.2]. □

Given a realization $[\check{A}, \check{B}; [\check{C}_1; \check{C}_2], [\check{D}_1; \check{D}_2]]$ of a factor $[M;N]$, we can obtain a realization

$$A := \check{A} - \check{B}\check{D}_1^{-1}\check{C}_1, \quad B := \check{B}\check{D}_1^{-1}, \quad C := \check{C}_2 - \check{D}_2\check{D}_1^{-1}\check{C}_1, \quad D := \check{D}_2\check{D}_1^{-1} \tag{4}$$

of NM^{-1}. This follows from [16, Lemma 5.7]. The next result shows how one can obtain an LQG-balanced realization from a Lyapunov-balanced realization of a normalized right coprime factor.

THEOREM 3.7. *Suppose that* G *has a normalized right coprime factor* $[M;N]$. *Let* $[\check{A}, \check{B}; [\check{C}_1; \check{C}_2], [\check{D}_1; \check{D}_2]]$ *be a minimal Lyapunov-balanced realization of this normalized right coprime factor. Define* $[A, B; C, D]$ *by* (4). *Then* $[A, B; C, D]$ *and its dual both satisfy the finite cost condition; its optimal cost operator is* L, *and its dual optimal cost operator is* $L(I - L^2)^{-1}$, *where* L *is the (controllability and observability) gramian of the Lyapunov-balanced realization.*

Proof. That $[M;N]$ has a minimal Lyapunov-balanced realization follows from Theorem 3.2. The rest follows from the first lines of the proof of [16, Theorem 8.2]. □

We note that $I - L$ has a bounded inverse since the Hankel singular values of the optimal closed-loop system are all strictly smaller than one. This last fact follows from the coprimeness of the factorization as in [3, Corollary 7.2].

From the system $[A, B; C, D]$ in Theorem 3.7 we obtain the LQG-balanced realization $[SAS^{-1}, SB; CS^{-1}, D]$, where $S := (I - L^2)^{-1/4}$. The following theorem summarizes some properties of LQG-balanced realizations.

THEOREM 3.8. 1. *Let* $[A_i, B_i; C_i, D_i]$ *with* $i = 1, 2$ *be discrete-time systems that satisfy the finite cost condition and whose duals also satisfy the finite cost condition. If these two systems have the same transfer function, then the nonzero elements of* $\sigma(P_1Q_1)$ *equal the nonzero elements of* $\sigma(P_2Q_2)$.

2. *If* $[A, B; C, D]$ *and its dual both satisfy the finite cost condition, then its transfer function has an LQG-balanced realization.*

3. *If* $[A_i, B_i; C_i, D_i]$ *with* $i = 1, 2$ *are two minimal LQG-balanced realizations of the same transfer function, then there exists a unitary* $U \in \mathcal{L}(\mathcal{X})$ *such that* $[A_1, B_1; C_1, D_1] = [UA_2U^{-1}, UB_2; C_2U^{-1}, D_2]$.

Proof. 1. This is [16, Lemma 7.2] up to an additional assumption that was made there. There it was assumed that the systems were approximately observable. The reason for this was that in [16, Lemma 6.9] this assumption was needed. It was shown in [3, Lemma 4.9] how this assumption can be removed from [16, Lemma 6.9] and this implies that it can also be removed from [16, Lemma 7.2].

2. This is [16, Theorem 8.2].

3. This is [16, Lemma 8.3]. □

The square roots of the nonzero elements of $\sigma(PQ)$ are called the *LQG-characteristic values*. According to part 1 of Theorem 3.8 they do not depend on the realization but only on the transfer function.

We now introduce the metric in which LQG-balanced approximations converge under suitable assumptions. Assume both G_1 and G_2 have normalized right coprime factors $[M_1; N_1]$ and $[M_2; N_2]$, respectively. Define for $i = 1, 2$ the set $Z_i \subset H^2(\mathbb{D}, [U; Y])$ by $Z_i = \{(M_i v; N_i v) : v \in H^2(\mathbb{D}; U)\}$, and let P_i be the orthogonal projection from $H^2(\mathbb{D}, [U; Y])$ onto Z_i. Further define

$$\delta_g(G_1, G_2) = \|P_1 - P_2\|.$$

Note that this does not depend on the particular normalized right coprime factors chosen. The function δ_g is called the *gap-metric*. More information on the gap-metric can be found in [24, Chapter 17] and for nonrational functions in [25]. What is important for us is that

$$\delta_g(G_1, G_2) \leq \left\| \begin{bmatrix} M_1 \\ N_1 \end{bmatrix} - \begin{bmatrix} M_2 \\ N_2 \end{bmatrix} \right\|_\infty. \tag{5}$$

We define truncated LQG-balanced realizations similarly to truncated Lyapunov-balanced realizations. The following new result provides an a priori error bound in the gap-metric for truncated compact LQG-balanced realizations.

THEOREM 3.9. *Suppose a discrete-time system satisfies the following assumptions:*

- *the finite cost condition is satisfied,*
- *the finite cost condition for the dual system is satisfied,*
- *the product PQ of the optimal cost operator and the dual optimal cost operator is nuclear, and*
- *the input and output spaces are finite-dimensional.*

Then the transfer function G has a compact LQG-balanced realization and

$$\delta_g(G, G^n) \leq 2 \sum_{i=n+1}^{\infty} \frac{\mu_i}{\sqrt{1 + \mu_i^2}}, \tag{6}$$

where G^n is the transfer function of an n-dimensional truncated LQG-balanced realization of G.

Proof. From Theorem 3.6 it follows that the transfer function of the given system has a normalized right coprime factor. We show that the Hankel operator of this normalized right coprime factor is nuclear. It follows from [16, Lemmas 5.1 and 6.9] that $L_B L_C = (I + PQ)^{-1} PQ$, where L_B is the controllability gramian of the optimal closed-loop system and L_C is its observability gramian. Since the product PQ is assumed compact, this shows that the product $L_B L_C$ is compact. The eigenvalues are related by

$$\mu_i = \frac{\sigma_i}{\sqrt{1 - \sigma_i^2}}, \quad \sigma_i = \frac{\mu_i}{\sqrt{1 + \mu_i^2}}, \tag{7}$$

where the μ_i are the square roots of the eigenvalues of PQ and the σ_i are the square roots of the eigenvalues of $L_B L_C$. This shows that the square roots of the eigenvalues of $L_B L_C$ are summable. Denote the Hankel operator of the normalized right coprime factor by Γ. As in [16, Lemma 7.2] it follows that the spectrum of $\Gamma^*\Gamma$ equals the spectrum of $L_B L_C$ and the point spectrum of $\Gamma^*\Gamma$ equals the point spectrum of $L_B L_C$ (both with the possible exception of zero). This shows that $\Gamma^*\Gamma$ has only point spectrum (with the possible exception of zero) and that the square roots of the eigenvalues are summable. This shows that the Hankel operator is nuclear.

Denote the normalized eigenvectors of the gramian L of the Lyapunov-balanced realization of the normalized coprime factor by e_n. Since for the optimal control operators of the LQG-balanced realization we have $P^{\mathrm{bal}} = Q^{\mathrm{bal}} = L(I - L^2)^{-1/2}$ from Theorem 3.7 we see that this LQG-balanced realization is actually compact LQG-balanced, and the corresponding orthonormal basis is $\{e_n\}$. We note that the projections associated to Lyapunov-balanced truncation and to LQG-balanced truncation are equal, since the orthonormal bases (including order) are identical. We conclude that the system obtained by applying (4) to the truncated Lyapunov-balanced realization is the truncated LQG-balanced realization. From Theorem 3.2, (5), and (7) we now obtain the estimate (6). □

4. Resolvent linear systems. In this section we recall the concept of a distributional resolvent linear system introduced in [12].

A finite-dimensional linear system is usually described by specifying four matrices A, B, C, D and defining for a given initial state x_0 and an input function $u \in L^2_{\mathrm{loc}}(0,\infty;\mathbb{C}^m)$ the state $x \in C(0,\infty;\mathbb{C}^n)$ and the output $y \in L^2_{\mathrm{loc}}(0,\infty;\mathbb{C}^p)$ as the unique solutions of

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \quad y(t) = Cx(t) + Du(t). \tag{8}$$

As is well known, these unique solutions are given explicitly by

$$\begin{aligned} x(t) &= \mathrm{e}^{At}x_0 + \int_0^t \mathrm{e}^{A(t-s)}Bu(s)\,ds, \\ y(t) &= C\mathrm{e}^{At}x_0 + \int_0^t C\mathrm{e}^{A(t-s)}Bu(s)\,ds + Du(t). \end{aligned} \tag{9}$$

If we Laplace-transform (8) and solve for x and y, we obtain

$$\begin{aligned} \hat{x}(s) &= (sI - A)^{-1}x_0 + (sI - A)^{-1}B\hat{u}(s), \\ \hat{y}(s) &= C(sI - A)^{-1}x_0 + \big(C(sI - A)^{-1}B + D\big)\,\hat{u}(s). \end{aligned} \tag{10}$$

Our approach to infinite-dimensional systems will be to generalize situation (10) rather than situation (8) or (9).

We first study the generalizations of the matrix-valued functions $(sI - A)^{-1}$, $(sI - A)^{-1}B$, $C(sI - A)^{-1}$, and $C(sI - A)^{-1}B + D$.

DEFINITION 4.1. *A* resolvent linear system *on a triple of Hilbert spaces* $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ *consists of a nonempty connected open subset* Λ *of the complex plane and four operator valued functions* $\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}$ *satisfying*

$\mathfrak{a} : \Lambda \to \mathcal{L}(\mathcal{X})$ *satisfies*

$$\mathfrak{a}(\beta) - \mathfrak{a}(\alpha) = (\alpha - \beta)\mathfrak{a}(\beta)\mathfrak{a}(\alpha) \quad \textit{for all } \alpha, \beta \in \Lambda; \tag{11}$$

$\mathfrak{b} : \Lambda \to \mathcal{L}(\mathcal{U}, \mathcal{X})$ *satisfies*

$$\mathfrak{b}(\beta) - \mathfrak{b}(\alpha) = (\alpha - \beta)\mathfrak{a}(\beta)\mathfrak{b}(\alpha) \quad \textit{for all } \alpha, \beta \in \Lambda; \tag{12}$$

$\mathfrak{c} : \Lambda \to \mathcal{L}(\mathcal{X}, \mathcal{Y})$ *satisfies*

$$\mathfrak{c}(\beta) - \mathfrak{c}(\alpha) = (\alpha - \beta)\mathfrak{c}(\alpha)\mathfrak{a}(\beta) \quad \textit{for all } \alpha, \beta \in \Lambda; \tag{13}$$

$\mathfrak{d} : \Lambda \to \mathcal{L}(\mathcal{U}, \mathcal{Y})$ *satisfies*

$$\mathfrak{d}(\beta) - \mathfrak{d}(\alpha) = (\alpha - \beta)\mathfrak{c}(\beta)\mathfrak{b}(\alpha) \quad \textit{for all } \alpha, \beta \in \Lambda. \tag{14}$$

The function $\mathfrak{a}$ *is called the* pseudoresolvent*,* $\mathfrak{b}$ *is the* incoming wave function*,* $\mathfrak{c}$ *is the* outgoing wave function*, and* $\mathfrak{d}$ *is the* characteristic function *of the resolvent linear system.*

The motivation for introducing this class of systems is the following connection with discrete-time systems.

DEFINITION 4.2. *Let* $\alpha > 0$*. The Cayley transform with parameter* α *of a resolvent linear system with* $\alpha \in \Lambda$ *is the discrete-time system*

$$A_d := -I + 2\alpha\ \mathfrak{a}(\alpha), \quad B_d := \sqrt{2\alpha}\ \mathfrak{b}(\alpha), \tag{15}$$

$$C_d := \sqrt{2\alpha}\ \mathfrak{c}(\alpha), \qquad D_d := \mathfrak{d}(\alpha). \tag{16}$$

Remark 4.3. The Cayley transform with parameter α gives a one-to-one correspondence between the set of resolvent linear systems with $\alpha \in \Lambda$ and the set of discrete-time systems.

The following relation between the characteristic function of a resolvent linear system and that of its Cayley transform is easily proven: $G(s) = G_d(z)$, where $z := (\alpha - s)/(\alpha + s)$.

We define two subclasses of resolvent linear systems for which one can make sense of the dynamical system (10).

DEFINITION 4.4. *A* distributional resolvent linear system *is a resolvent linear system with the additional property that there exist constants* $\alpha, \beta > 0$ *and a polynomial* p *such that*

$$\Lambda_E := \{s \in \mathbb{C} : \operatorname{Re} s \geq \beta, \quad |\operatorname{Im} s| \leq \mathrm{e}^{\alpha \operatorname{Re} s}\} \subset \Lambda \tag{17}$$

and

$$\|\mathfrak{a}(s)\| \leq p(|s|) \quad \textit{for all } s \in \Lambda_E. \tag{18}$$

A region Λ_E as above is called an exponential region (see Figure 1 for a sketch of the boundary of such a region).

It is easily seen using the functional equations from Definition 4.1 that the functions $\mathfrak{b}$, $\mathfrak{c}$, and $\mathfrak{d}$ of a distributional resolvent linear system are also bounded in norm by a polynomial on the exponential region Λ_E.

DEFINITION 4.5. *A distributional resolvent linear system is called* exponentially bounded *if there exist a constant* $\gamma > 0$ *and a polynomial* p *such that*

$$\Lambda_H := \{s \in \mathbb{C} : \operatorname{Re} s \geq \gamma\} \subset \Lambda \tag{19}$$

and

$$\|\mathfrak{a}(s)\| \leq p(|s|) \quad \textit{for all } s \in \Lambda_H. \tag{20}$$

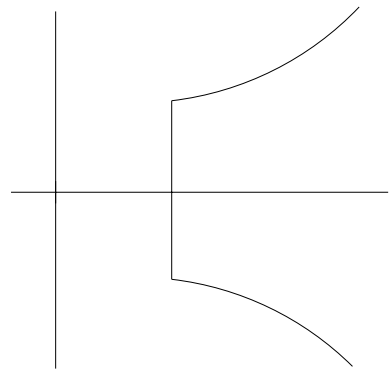


FIG. 1. *A typical example of the boundary of an exponential region.*

Note that the difference between Definitions 4.4 and 4.5 is in the region considered.

Remark 4.6. The term "exponentially bounded" stems from time-domain properties of this subclass. In [12] exponentially bounded distributional resolvent linear systems were called integrated resolvent linear systems. In view of the time-domain results in [13] the term exponentially bounded distributional resolvent linear system, however, seems to be more appropriate.

Remark 4.7. In what follows we will need the following well-known characterization of Laplace transformable Banach space valued distributions by Schwartz. The image of the Schwartz–Laplace transformable Banach space valued distributions is exactly the set of polynomially bounded analytic functions defined on some right half-plane. For details see [18]. A generalization of this Laplace transform was given by Kunstmann. He defined the Laplace transform in such a way that the image of the set of Laplace transformable distributions is exactly the set of functions that are analytic and polynomially bounded on an exponential region (see Kunstmann [8]).

DEFINITION 4.8. *The state x and output y of a distributional resolvent linear system corresponding to the initial state $x_0 \in \mathcal{X}$ and the input u (a $\mathcal{U}$-valued Kunstmann–Laplace transformable distribution) are defined through their Kunstmann–Laplace transforms as*

$$\hat{x}(s) := \mathfrak{a}(s)x_0 + \mathfrak{b}(s)\hat{u}(s), \quad \hat{y}(s) := \mathfrak{c}(s)x_0 + \mathfrak{d}(s)\hat{u}(s). \tag{21}$$

For the case of exponentially bounded distributional resolvent linear systems, if we restrict u to be Schwartz–Laplace transformable, then x and y are Schwartz–Laplace transformable.

For a distributional resolvent linear system we define the set of *stable input-output pairs*

$$\mathcal{V}(x_0) := \left\{ \begin{bmatrix} u \\ y \end{bmatrix} \in \begin{bmatrix} L^2(\mathbb{R}^+;\mathcal{U}) \\ L^2(\mathbb{R}^+;\mathcal{Y}) \end{bmatrix} : y \text{ satisfies (21)} \right\}.$$

DEFINITION 4.9. *We say that a distributional resolvent linear system satisfies the* finite cost condition *if for every $x_0 \in \mathcal{X}$ the set $\mathcal{V}(x_0)$ is nonempty.*

For $\alpha > 0$ the mapping $\mathcal{H}_d : \mathbf{H}^2(\mathbb{C}_0^+;\mathcal{H}) \to \mathbf{H}^2(\mathbb{D};\mathcal{H})$, where $\mathcal{H}$ is a Hilbert space, is unitary. Here $\mathcal{H}_d$ is defined by

$$(\mathcal{H}_d g)(z) = \frac{\sqrt{2\alpha}}{1+z} \; g\left(\alpha \, \frac{1-z}{1+z}\right), \tag{22}$$

with its inverse given by

$$(\mathcal{H}_d^{-1} f)(s) = \frac{\sqrt{2\alpha}}{\alpha + s} \; f\left(\frac{\alpha - s}{\alpha + s}\right). \tag{23}$$

$\mathbb{C}_0^+$ is the right half-plane and $\mathbf{H}^2$ is a Hardy space. The following theorem shows that, for a suitably chosen parameter α, there is a one-to-one relationship between the stable input-output pairs of a distributional resolvent linear system and those of its Cayley transform.

THEOREM 4.10. *Let $(\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d})$ be a distributional resolvent linear system with $\alpha \in \Lambda_E$, where $\alpha > 0$. Let $[A_d, B_d; C_d, D_d]$ be its Cayley transform with parameter α. Then $(u; y) \in \mathcal{V}(x_0)$ if and only if $(\mathcal{H}_d u; \mathcal{H}_d y) \in \mathcal{V}_d(x_0)$.*

The following is [12, Lemma 9].

LEMMA 4.11. *For a distributional resolvent linear system on a triple of Hilbert spaces for which the finite cost condition is satisfied there exists a nonnegative operator $Q \in \mathcal{L}(\mathcal{X})$ such that the optimal cost for the cost function*

$$\int_0^\infty \|u(t)\|^2 + \|y(t)\|^2 \, dt \tag{24}$$

is given by $\langle Q x_0, x_0 \rangle$. This Q satisfies the Riccati equation

$$\begin{aligned} &-\mathfrak{a}(\alpha)^* Q - Q\mathfrak{a}(\alpha) + 2\alpha \mathfrak{a}(\alpha)^* Q \mathfrak{a}(\alpha) + \mathfrak{c}(\alpha)^* \mathfrak{c}(\alpha) \\ &= (\mathfrak{c}(\alpha)^* \mathfrak{d}(\alpha) - Q\mathfrak{b}(\alpha) + 2\alpha \mathfrak{a}(\alpha)^* Q \mathfrak{b}(\alpha)) \\ &\quad (I + \mathfrak{d}(\alpha)^* \mathfrak{d}(\alpha) + 2\alpha \mathfrak{b}(\alpha)^* Q \mathfrak{b}(\alpha))^{-1} \\ &\quad (\mathfrak{d}(\alpha)^* \mathfrak{c}(\alpha) - \mathfrak{b}(\alpha)^* Q + 2\alpha \mathfrak{b}(\alpha)^* Q \mathfrak{a}(\alpha)) \end{aligned}$$

for all $\alpha \in \Lambda_E$.

The operator Q mentioned above is called the *optimal cost operator* of the distributional resolvent linear system. We now study admissible state feedbacks.

DEFINITION 4.12. *An* admissible state feedback pair *for a* distributional resolvent linear system *is a pair $[\mathfrak{k}, \mathfrak{f}] : \Lambda_E \to \mathcal{L}(\mathcal{X} \times \mathcal{U}, \mathcal{U})$ that satisfies*

$$\begin{aligned} \mathfrak{k}(\beta) - \mathfrak{k}(\alpha) &= (\alpha - \beta)\mathfrak{k}(\alpha)\mathfrak{a}(\beta), \\ \mathfrak{f}(\beta) - \mathfrak{f}(\alpha) &= (\alpha - \beta)\mathfrak{k}(\beta)\mathfrak{b}(\alpha), \end{aligned}$$

and such that $(I - \mathfrak{f}(s))^{-1}$ exists and is polynomially bounded on some exponential region.

The closed-loop system of a distributional resolvent linear system with an admissible state feedback pair is the distributional resolvent linear system

$$\begin{aligned} \mathfrak{a}^{\mathrm{cl}} &:= \mathfrak{a} + \mathfrak{b}(I - \mathfrak{f})^{-1}\mathfrak{k}, \quad \mathfrak{b}^{\mathrm{cl}} := \mathfrak{b}(I - \mathfrak{f})^{-1}, \\ \mathfrak{c}^{\mathrm{cl}} &:= \begin{bmatrix} (I - \mathfrak{f})^{-1}\mathfrak{k} \\ \mathfrak{c} + \mathfrak{d}(I - \mathfrak{f})^{-1}\mathfrak{k} \end{bmatrix}, \quad \mathfrak{d}^{\mathrm{cl}} := \begin{bmatrix} (I - \mathfrak{f})^{-1} \\ \mathfrak{d}(I - \mathfrak{f})^{-1} \end{bmatrix}. \end{aligned}$$

It can be easily checked that this is indeed a distributional resolvent linear system. The exponential region on which this closed-loop system is defined is the largest exponential region contained in the intersection of the exponential region on which

the original system was defined and the exponential region on which $(I - \mathfrak{f})^{-1}$ exists and is polynomially bounded.

The following is [12, Lemma 8].

LEMMA 4.13. *For a distributional resolvent linear system on a triple of Hilbert spaces for which the finite cost condition is satisfied there exists an admissible state feedback pair such that the optimal control* u^{opt} *for the cost function* (24) *is given by* $\hat{u}^{\mathrm{opt}}(s) = (I - \mathfrak{f}(s))^{-1}\mathfrak{k}(s)x_0$ *for* $s \in \Lambda_E$.

Remark 4.14. A proper proof of Lemma 4.13 is given in [12, Lemma 8]. We do want to mention the main idea of the proof. First, one Cayley-transforms the system with a suitable parameter α. One then defines $\mathfrak{k}(\alpha)$ and $\mathfrak{f}(\alpha)$ in terms of the optimal admissible state feedback pair $[K, F]$ of the Cayley-transformed system. The functions $\mathfrak{k}$ and $\mathfrak{f}$ are then extended to Λ_E using the functional equations from Definition 4.12. In the specific case of the optimal state feedback it is simple to prove that $(I - \mathfrak{f})^{-1}$ exists and is polynomially bounded on an exponential region: its Cayley transform equals the denominator M_d of the normalized right factor mentioned in Theorem 3.6. So the Cayley transform of $(I - \mathfrak{f})^{-1}$ is in H^∞ of the unit disc, from which it follows that $(I - \mathfrak{f})^{-1}$ is in H^∞ of the right half-plane.

DEFINITION 4.15. *An* admissible state feedback pair *for an* exponentially bounded distributional resolvent linear system *is a pair* $[\mathfrak{k}, \mathfrak{f}] : \Lambda_H \to \mathcal{L}(\mathcal{X} \times \mathcal{U}, \mathcal{U})$ *that satisfies*

$$\mathfrak{k}(\beta) - \mathfrak{k}(\alpha) = (\alpha - \beta)\mathfrak{k}(\alpha)\mathfrak{a}(\beta),$$
$$\mathfrak{f}(\beta) - \mathfrak{f}(\alpha) = (\alpha - \beta)\mathfrak{k}(\beta)\mathfrak{b}(\alpha),$$

and such that $(I - \mathfrak{f}(s))^{-1}$ *exists and is polynomially bounded on some right half-plane.*

The closed-loop system of an exponentially bounded distributional resolvent linear system and an admissible state feedback pair in the sense of Definition 4.15 is easily seen to be an exponentially bounded distributional resolvent linear system. Theorem 4.13 holds for exponentially bounded distributional resolvent linear systems with admissible state feedback operator now understood in the stronger sense of Definition 4.15.

DEFINITION 4.16. *The* dual *of a resolvent linear system* $\mathfrak{a}$, $\mathfrak{b}$, $\mathfrak{c}$, $\mathfrak{d}$ *is the resolvent linear system*

$$\mathfrak{a}^d(s) := \mathfrak{a}(\bar{s})^*, \quad \mathfrak{b}^d(s) := \mathfrak{c}(\bar{s})^*, \quad \mathfrak{c}^d(s) := \mathfrak{b}(\bar{s})^*, \quad \mathfrak{d}^d(s) := \mathfrak{d}(\bar{s})^*.$$

Note that the dual of a distributional resolvent linear system is a distributional resolvent linear system and that the dual of an exponentially bounded distributional resolvent linear system is an exponentially bounded distributional resolvent linear system.

The concept of approximate observability has a natural generalization to distributional resolvent linear systems.

DEFINITION 4.17. *A distributional resolvent linear system is said to be* approximately observable *if for zero input the output is zero if and only if the initial state is zero.*

Note that a distributional resolvent linear system is approximately observable if and only if $\mathfrak{c}(s)x_0 = 0$ for all $s \in \Lambda_E$ implies $x_0 = 0$.

DEFINITION 4.18. *A distributional resolvent linear system is said to be* approximately controllable *if its dual system is approximately observable. It is called* minimal *if it is both approximately controllable and approximately observable.*

It is easily seen that a distributional resolvent linear system is approximately controllable (observable) if and only if its Cayley transform with a parameter $\alpha \in \Lambda_E$ is.

We denote by $\mathbf{H}^\infty(\mathbb{C}_0^+, \mathcal{E})$ the Hardy space of uniformly bounded $\mathcal{E}$-valued analytic functions defined on the right half-plane, where $\mathcal{E}$ is a Banach space.

DEFINITION 4.19. *Let Λ_E be an exponential region. A function $G : \Lambda_E \to \mathcal{L}(\mathcal{U}, \mathcal{Y})$, is said to have a* right factorization *if there exist $N \in \mathbf{H}^\infty(\mathbb{C}_0^+, \mathcal{L}(\mathcal{U}, \mathcal{Y}))$ and $M \in \mathbf{H}^\infty(\mathbb{C}_0^+, \mathcal{L}(\mathcal{U}))$ such that $M(s)^{-1}$ exists for all $s \in \Lambda_E$ and $G = NM^{-1}$ on Λ_E.*

This factorization is called normalized *if $[M; N]$ is inner, i.e., if for almost all $\omega \in \mathbb{R}$ we have*

$$M_b(i\omega)^* M_b(i\omega) + N_b(i\omega)^* N_b(i\omega) = I,$$

where M_b and N_b are the boundary functions of M and N, respectively.

This factorization is called right coprime *if there exist $\tilde{X} \in \mathbf{H}^\infty(\mathbb{C}_0^+, \mathcal{L}(\mathcal{U}))$, $\tilde{Y} \in \mathbf{H}^\infty(\mathbb{C}_0^+, \mathcal{L}(\mathcal{Y}, \mathcal{U}))$ such that*

$$\tilde{X}M - \tilde{Y}N = I \quad \text{on } \mathbb{C}_0^+. \tag{25}$$

Using the Cayley transform, we obtain from Theorem 3.6 the following theorem.

THEOREM 4.20. *If a distributional resolvent linear system satisfies the finite cost condition, then its characteristic function has a normalized right factor. If, in addition, the dual finite cost condition is satisfied, then this factor is right coprime.*

The above theorem is a slight generalization of [3, Theorem 8.9]. The proof is almost identical; one simply replaces the reciprocal transform used there by the Cayley transform with a suitable parameter (i.e., positive and in Λ_E). The relation between characteristic functions mentioned in Remark 4.3 is of course essential.

5. Well-posed linear systems. We now show how the well-known class of well-posed linear systems fits into our framework.

DEFINITION 5.1. *A resolvent linear system is called* well-posed *if*

1. *the pseudoresolvent is the resolvent of the generator of a strongly continuous semigroup T;*
2. *for every $x \in \mathcal{X}$ the function $\mathfrak{c}x$ restricts to a function in $\mathbf{H}^2(\mathbb{C}_\omega^+; \mathcal{Y})$, where ω is some real number strictly larger than the growth bound of T;*
3. *for every $x \in \mathcal{X}$ the function $\mathfrak{b}^d x$ restricts to a function in $\mathbf{H}^2(\mathbb{C}_\omega^+; \mathcal{U})$, where ω is some real number strictly larger than the growth bound of T; and*
4. *$\mathfrak{d}$ restricts to a function in $\mathbf{H}^\infty(\mathbb{C}_\omega^+; \mathcal{L}(\mathcal{U}, \mathcal{Y}))$, where ω is some real number strictly larger than the growth bound of T.*

The above definition is equivalent to the usual time-domain definition.

DEFINITION 5.2. *An* admissible state feedback pair *for a* well-posed linear system *is a pair $[\mathfrak{k}, \mathfrak{f}] : \mathbb{C}_\omega^+ \to \mathcal{L}(\mathcal{X} \times \mathcal{U}, \mathcal{U})$ that satisfies*

$$\mathfrak{k}(\beta) - \mathfrak{k}(\alpha) = (\alpha - \beta)\mathfrak{k}(\alpha)\mathfrak{a}(\beta),$$
$$\mathfrak{f}(\beta) - \mathfrak{f}(\alpha) = (\alpha - \beta)\mathfrak{k}(\beta)\mathfrak{b}(\alpha),$$

and such that for every $x \in \mathcal{X}$ the function $\mathfrak{k}x$ restricts to a function in $\mathbf{H}^2(\mathbb{C}_\omega^+; \mathcal{U})$, the function $\mathfrak{f}$ restricts to a function in $\mathbf{H}^\infty(\mathbb{C}_\omega^+; \mathcal{L}(\mathcal{U}))$, and $(I - \mathfrak{f}(s))^{-1}$ exists and is uniformly bounded on some right half-plane.

The above definition is equivalent to the time-domain definition in [19]. In [19] it is shown that the closed-loop system of a well-posed linear system with an admissible state feedback in the sense of Definition 5.2 is a well-posed linear system.

6. LQG-balanced realizations. In this section we prove the continuous-time analogues of the discrete-time results of section 3.2.

DEFINITION 6.1. *For a distributional resolvent linear system that satisfies both the finite cost condition and the dual finite cost condition the nonzero elements of the set* $\sqrt{\sigma(PQ)}$*, where* Q *is the optimal cost operator of the system and* P *is the optimal cost operator of the dual system, are called* LQG-characteristic values.

The following theorem shows that the LQG-characteristic values depend only on the characteristic function.

THEOREM 6.2. *Two distributional resolvent linear systems that both satisfy both the finite cost condition and the dual finite cost condition and whose characteristic functions are equal on an exponential region have the same set of LQG-characteristic values.*

Proof. Cayley-transform both distributional resolvent linear systems with a parameter which is in the exponential region of both. The transfer functions of the Cayley-transformed systems then agree in some neighborhood of zero. It follows from Theorem 3.8 that the LQG-characteristic values of these Cayley-transformed systems are equal. The LQG-characteristic values of a distributional resolvent linear system and its Cayley transform are equal, since the optimal cost operators are equal (which follows from Theorem 4.10). Hence it follows that the LQG-characteristic values of the two distributional resolvent linear systems are equal. □

DEFINITION 6.3. *A distributional resolvent linear system is called* LQG-balanced *if it and its dual both satisfy the finite cost condition and if the optimal cost operators of the system and that of its dual are equal. It is called* compact LQG-balanced *if, in addition, this operator is compact.*

The following theorem gives a necessary and sufficient condition for the existence of LQG-balanced realizations.

THEOREM 6.4. *An* $\mathcal{L}(\mathcal{U},\mathcal{Y})$*-valued holomorphic function, defined and polynomially bounded on an exponential region, has a normalized right coprime factor if and only if it has an LQG-balanced realization.*

Proof. Assume the given function $\mathfrak{d}$ has a normalized right coprime factor $[M;N]$. It follows from [11] or [19, section 9.5] that $[M;N]$ has a minimal well-posed Lyapunov-balanced realization $\mathfrak{a}_L$, $\mathfrak{b}_L$, $[\mathfrak{c}_L^1;\mathfrak{c}_L^2]$, $[M;N]$. Consider the well-posed linear system $\mathfrak{a}_L$, $\mathfrak{b}_L$, $\mathfrak{c}_L^2$, N and the feedback pair $[\mathfrak{k},\mathfrak{f}] := [-\mathfrak{c}_L^1, I-M]$. This feedback pair is admissible for the given system: the algebraic relations easily follow from the fact that the Lyapunov-balanced system is a resolvent linear system (even a well-posed linear system). Since $(I-\mathfrak{f}(s))^{-1} = M^{-1}$ it remains to show that M^{-1} is polynomially bounded on some exponential region. This follows from the equation $M^{-1} = X - Y\mathfrak{d}$ on Λ_E, which follows from the Bezout equation (25). The closed-loop system of the above system with the given feedback pair is $\mathfrak{a}_L - \mathfrak{b}_L M^{-1}\mathfrak{c}_L^1$, $\mathfrak{b}_L M^{-1}$, $[-M^{-1}\mathfrak{c}_L^1; \mathfrak{c}_L^2 - NM^{-1}\mathfrak{c}_L^1]$, $[M^{-1}; NM^{-1}]$. It follows that this is a distributional resolvent linear system. We drop one of the components and obtain the following distributional resolvent linear system:

$$\mathfrak{a}_s := \mathfrak{a}_L - \mathfrak{b}_L M^{-1}\mathfrak{c}_L^1, \quad \mathfrak{b}_s := \mathfrak{b}_L M^{-1}, \quad \mathfrak{c}_s := \mathfrak{c}_L^2 - NM^{-1}\mathfrak{c}_L^1, \quad \mathfrak{d}_s := NM^{-1}. \tag{26}$$

Now choose $\alpha > 0$ in the intersection of the exponential regions of all the systems considered above and Cayley-transform these systems with this parameter. It is obvious from the constructions and Theorem 3.7 that the system (26) has L as its optimal cost operator and $L(I-L^2)^{-1}$ as its dual optimal cost operator, where L is the gramian of the Lyapunov-balanced realization. Define $S := (I-L^2)^{-1/4}$, and define $\mathfrak{a}_l := S\mathfrak{a}_s S^{-1}$, $\mathfrak{b}_l := S\mathfrak{b}$, $\mathfrak{c}_l := \mathfrak{c}S^{-1}$, $\mathfrak{d}_l = \mathfrak{d}_s$. We conclude that $\mathfrak{a}_l$, $\mathfrak{b}_l$, $\mathfrak{c}_l$, $\mathfrak{d}_l$ is

LQG-balanced. Since $\mathfrak{d}_l = NM^{-1} = \mathfrak{d}$ this distributional resolvent linear system is an LQG-balanced realization of $\mathfrak{d}$.

The converse trivially follows from Theorem 4.20. ∎

Theorem 6.4 can be rephrased in terms of realizations as follows. Here Theorem 4.20 is used.

COROLLARY 6.5. *For a distributional resolvent linear system that satisfies both the finite cost condition and the dual finite cost condition there exists an LQG-balanced distributional resolvent linear system such that the characteristic functions of these two systems are equal on some exponential region.*

The following two corollaries show that in the special cases of exponentially bounded distributional resolvent linear systems and well-posed linear systems the LQG-balanced realization belongs to the same class.

COROLLARY 6.6. *For an exponentially bounded distributional resolvent linear system that satisfies both the finite cost condition and the dual finite cost condition there exists an LQG-balanced exponentially bounded distributional resolvent linear system such that the characteristic functions of these two systems are equal on some right half-plane.*

Proof. This follows from the proof of Theorem 6.4 noting that the Bezout equation now shows that the feedback is admissible in the sense of Definition 4.15. ∎

COROLLARY 6.7. *For a well-posed linear system that satisfies both the finite cost condition and the dual finite cost condition there exists an LQG-balanced well-posed linear system such that the characteristic functions of these two systems are equal on some right half-plane.*

Proof. This follows from the proof of Theorem 6.4 noting that the Bezout equation now shows that the feedback is admissible in the sense of Definition 5.2. ∎

Let $\mathfrak{a}$, $\mathfrak{b}$, $\mathfrak{c}$, $\mathfrak{d}$ be an LQG-balanced distributional resolvent linear system, and let $U \in \mathcal{L}(\mathcal{X})$ be unitary. Then obviously $U\mathfrak{a}U^*$, $U\mathfrak{b}$, $\mathfrak{c}U^*$, $\mathfrak{d}$ is also an LQG-balanced distributional resolvent linear system. The next theorem shows that these are all LQG-balanced distributional resolvent linear systems with characteristic function $\mathfrak{d}$ if we assume a minimality assumption on the state space.

THEOREM 6.8. *If two distributional resolvent linear systems whose characteristic functions agree on some exponential region are both LQG-balanced, approximately controllable, and approximately observable, then there exists a unitary state-space transformation between them.*

Proof. Choose a parameter that is in the exponential region of both systems, and Cayley-transform both systems with this parameter. The resulting systems are LQG-balanced, approximately controllable, and approximately observable and have the same transfer function. It follows from part 3 of Theorem 3.8 that these discrete-time systems are unitarily equivalent. From this it follows that the distributional resolvent linear systems are unitarily equivalent. ∎

The gap-metric in continuous time is defined in exactly the same way as was done in discrete time in section 3.2, but with the unit disc $\mathbb{D}$ replaced by the right half-plane $\mathbb{C}_0^+$. It is easily seen that the distance between two systems equals the distance between their Cayley transforms (taken with the same parameter, obviously).

Using the Cayley transform, the following theorem follows immediately from Theorem 3.9.

THEOREM 6.9. *Suppose a distributional resolvent linear system satisfies the following assumptions:*

- *the finite cost condition is satisfied,*

- *the finite cost condition for the dual system is satisfied,*
- *the product PQ of the optimal cost operator and the dual optimal cost operator is nuclear, and*
- *the input and output spaces are finite-dimensional.*

Then there exists a compact LQG-balanced distributional resolvent linear system whose characteristic function equals the characteristic function of the original system on some exponential region and

$$\delta_g(\mathfrak{d}, G^n) \leq 2 \sum_{i=n+1}^{\infty} \frac{\mu_i}{\sqrt{1+\mu_i^2}}, \tag{27}$$

where G^n is the transfer function of an n-dimensional truncated LQG-balanced realization.

7. Conclusions. In this article we have obtained existence and uniqueness results for LQG-balanced realizations for continuous-time infinite-dimensional systems. We also obtained a priori error bounds in the gap-metric for both the continuous-time and the discrete-time cases.

REFERENCES

[1] C. Bonnet, *Convergence and convergence rate of the balanced realization truncations for infinite-dimensional discrete-time systems*, Systems Control Lett., 20 (1993), pp. 353–359.

[2] R. F. Curtain, *Model reduction for control design for distributed parameter systems*, in Research Directions in Distributed Parameter Systems (Raleigh, NC, 2000), Frontiers Appl. Math. 27, SIAM, Philadelphia, 2003, pp. 95–121.

[3] R. F. Curtain and M. R. Opmeer, *Normalized doubly coprime factorizations for infinite-dimensional linear systems*, Math. Control Signals Systems, 18 (2006), pp. 1–31.

[4] R. F. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.

[5] K. Glover, R. F. Curtain, and J. R. Partington, *Realisation and approximation of linear infinite-dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863–898.

[6] J. Hoffmann, D. Prätzel-Wolters, and E. Zerz, *A balanced canonical form for discrete-time minimal systems using characteristic maps*, Linear Algebra Appl., 277 (1998), pp. 63–81.

[7] E. A. Jonckheere and L. M. Silverman, *A new set of invariants for linear systems—application to reduced order compensator design*, IEEE Trans. Automat. Control, 28 (1983), pp. 953–964.

[8] P. C. Kunstmann, *Laplace transform theory for logarithmic regions*, in Evolution Equations and Their Applications in Physical and Life Sciences (Bad Herrenalb, 1998), Lecture Notes in Pure and Appl. Math. 215, Dekker, New York, 2001, pp. 125–138.

[9] B. C. Moore, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–32.

[10] D. Mustafa and K. Glover, *Minimum Entropy H_∞ Control*, Lecture Notes in Control and Inform. Sci. 146, Springer-Verlag, Berlin, 1990.

[11] R. Ober and S. Montgomery-Smith, *Bilinear transformation of infinite-dimensional state-space systems and balanced realizations of nonrational transfer functions*, SIAM J. Control Optim., 28 (1990), pp. 438–465.

[12] M. R. Opmeer, *Infinite-dimensional linear systems: A distributional approach*, Proc. London Math. Soc., 91 (2005), pp. 738–760.

[13] M. R. Opmeer, *Distribution semigroups and control systems*, J. Evol. Equ., 6 (2006), pp. 145–159.

[14] M. R. Opmeer, *Model Reduction for Controller Design for Infinite-Dimensional Systems*, Ph.D. thesis, University of Groningen, Groningen, The Netherlands 2006; available at http://irs.ub.rug.nl/ppn/296880116.

[15] M. R. Opmeer and R. F. Curtain, *New Riccati equations for well-posed linear systems*, Systems Control Lett., 52 (2004), pp. 339–347.

[16] M. R. Opmeer and R. F. Curtain, *Linear quadratic Gaussian balancing for discrete-time infinite-dimensional linear systems*, SIAM J. Control Optim., 43 (2004), pp. 1196–1221.
[17] V. V. Peller, *Hankel Operators and Their Applications*, Springer Monogr. Math., Springer-Verlag, New York, 2003.
[18] L. Schwartz, *Théorie des distributions*, Publications de l'Institut de Mathématique de l'Université de Strasbourg, No. IX-X, Nouvelle édition, entièrement corrigée, refondue et augmentée, Hermann, Paris, 1966.
[19] O. Staffans, *Well-Posed Linear Systems*, Encyclopedia Math. Appl. 103, Cambridge University Press, Cambridge, UK, 2005.
[20] E. I. Verriest, *Low sensitivity design and optimal order reduction for the LQG-problem*, in Proceedings of the 24th Midwest Symposium on Circuits and Systems, Albuquerque, NM, 1981, pp. 365–369.
[21] E. I. Verriest, *Suboptimal lqg-design via balanced realizations*, in Proceedings of the 20th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1981, pp. 686–687.
[22] N. J. Young, *Balanced, normal, and intermediate realizations of nonrational transfer functions*, IMA J. Math. Control Inform., 3 (1986), pp. 43–58.
[23] N. J. Young, *Balanced realizations in infinite dimensions*, in Operator Theory and Systems (Amsterdam, 1985), Oper. Theory Adv. Appl. 19, Birkhäuser, Basel, 1986, pp. 449–471.
[24] K. Zhou and J. C. Doyle, *Essentials of Robust Control*, Prentice–Hall, Upper Saddle River, NJ, 1998.
[25] S. Q. Zhu, *Graph topology and gap topology for unstable systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 848–855.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1849–1881

STABILITY OF DISCONTINUOUS DIFFUSION COEFFICIENTS AND INITIAL CONDITIONS IN AN INVERSE PROBLEM FOR THE HEAT EQUATION*

ASSIA BENABDALLAH†, PATRICIA GAITAN‡, AND JÉRÔME LE ROUSSEAU†

Abstract. We consider the heat equation with a discontinuous diffusion coefficient and give uniqueness and stability results for both the diffusion coefficient and the initial condition from a measurement of the solution on an arbitrary part of the boundary and at some arbitrary positive time. The key ingredient is the derivation of a Carleman-type estimate. The diffusion coefficient is assumed to be discontinuous across an interface with a monotonicity condition.

Key words. parabolic equations, Carleman estimates, inverse problem, stability estimate, discontinuous coefficients

AMS subject classifications. 35K05, 35R30

DOI. 10.1137/050640047

1. Introduction.

This article is devoted to the question of the identification of a diffusion coefficient, c, for a heat transfer problem in a bounded domain, with the main particularity that c is discontinuous. Such regularity can be encountered in the case of embedded materials.

Let $\Omega \subset \mathbb{R}^n$ be a bounded connected open set. The set $\overline{\Omega}$ is assumed to be a $\mathscr{C}^2$ submanifold with boundary in $\mathbb{R}^n$ (see, e.g., [12, Definition 1.2.1.2]). We set $\Gamma = \partial\Omega$. Let Ω_0 and Ω_1 be two nonempty open subsets of Ω such that

$$\Omega_0 \Subset \Omega \quad \text{and} \quad \Omega_1 = \Omega \setminus \overline{\Omega}_0.$$

We denote by $S = \overline{\Omega}_0 \cap \overline{\Omega}_1$ the interface, which is assumed to be $\mathscr{C}^2$. We shall use the notation $\Omega' = \Omega_0 \cup \Omega_1$. It should be emphasized here that the position of the interface itself is not assumed to be known.

Let $T > 0$. We consider the following transmission problem for the heat equation:

$$\begin{cases} \partial_t y - \nabla \cdot (c \nabla y) = 0 & \text{in } (0,T) \times \Omega', \\ y(t,x) = h(t,x) & \text{on } (0,T) \times \Gamma, \\ \text{transmission conditions (TC1)} & \text{on } (0,T) \times S, \\ y(0,x) = y_0(x), & \text{in } \Omega, \end{cases} \tag{1.1}$$

with

$$y_{|[0,T]\times S_0} = y_{|[0,T]\times S_1}, \quad c_0 \partial_n y_{|[0,T]\times S_0} = c_1 \partial_n y_{|[0,T]\times S_1}, \tag{TC1}$$

where $f_{|S_i}$ is the trace on S of $f_{|\Omega_i}$ and where

$$c = \begin{cases} c_0 & \text{in } \Omega_0, \\ c_1 & \text{in } \Omega_1, \end{cases} \qquad \widetilde{c} = \begin{cases} \widetilde{c}_0 & \text{in } \Omega_0, \\ \widetilde{c}_1 & \text{in } \Omega_1. \end{cases}$$

*Received by the editors September 10, 2005; accepted for publication (in revised form) March 27, 2007; published electronically November 16, 2007.
http://www.siam.org/journals/sicon/46-5/64004.html

†Université Aix-Marseille I, Laboratoire d'Analyse Topologie Probabilités (LATP), CNRS UMR 6632, Marseille, France (assia@cmi.univ-mrs.fr, jlerous@cmi.univ-mrs.fr).

‡Université Aix-Marseille II, Laboratoire d'Analyse Topologie Probabilités (LATP), CNRS UMR 6632, Marseille, France (gaitan@cmi.univ-mrs.fr).

The boundary condition $h(t,x)$ shall be kept fixed. The diffusion coefficient c shall be kept independent of time, t. If we change the diffusion coefficient c into $\widetilde{c}$, we let $\widetilde{y}$ be the solution of (1.1) associated with $\widetilde{c}$ and $\widetilde{y}_0$ for the initial condition. In the case studied here, the interface remains unchanged when changing coefficients. Its position is, however, not known.

We assume that we can measure both the normal flux $\partial_n \partial_t y$ on $\gamma \subset \partial\Omega$ on the time interval (t_0, T) for some $t_0 \in (0,T)$ and Δy in Ω at time $T' \in (t_0, T)$. In the case of piecewise-*constant* diffusion coefficients, i.e., $c_{|\Omega_i}$, $i = 0, 1$, is constant, our main results are (i) the injectivity of the map

$$L^\infty(\Omega) \times L^2(\Omega) \to L^2((t_0,T)\times\gamma) \times L^2(\Omega),$$
$$(c, y_0) \mapsto (\partial_n \partial_t y, \Delta y(T'))$$

(uniqueness); (ii) the stability for the diffusion coefficient, c (Theorem 3.9): there exists $C > 0$ such that

$$|c - \widetilde{c}|^2_{L^\infty(\Omega)} \le C|\partial_n(\partial_t y - \partial_t \widetilde{y})|^2_{L^2((0,T)\times\gamma)} + C|\Delta y(T',.) - \Delta\widetilde{y}(T',.)|^2_{L^2(\Omega')};$$

and (iii) the stability for the initial condition, y_0 (Theorem 5.5): there exists $C > 0$ such that

$$|y_0 - \widetilde{y}_0|_{L^2(\Omega)} \le C/\left|\ln\left(|(y-\widetilde{y})(T')|_{H^2(\Omega')} + |\partial_n(\partial_t y - \partial_t\widetilde{y})|^2_{L^2((0,T)\times\gamma)}\right)\right|$$

for $|(y-\widetilde{y})(T')|_{H^2(\Omega')} + |\partial_n(\partial_t y - \partial_t\widetilde{y})|^2_{L^2((0,T)\times\gamma)}$ sufficiently small. At the end of section 3 we shall observe that we may require only that the difference of the coefficients, $c - \widetilde{c}$, be piecewise-constant.

The key ingredient to these stability results is a *global* Carleman estimate for the operator $\partial_t - \nabla\cdot(c\nabla(.))$ and the open set Ω.

The use of Carleman estimates to achieve uniqueness and stability results in inverse problems is now well established. Some authors make use of *local* Carleman inequalities and deduce uniqueness and Hölder estimates (see [19], [18], and references cited therein). Others make use of *global* Carleman inequalities and deduce Lipschitz stability results (and hence uniqueness results). We shall follow this second approach. To our knowledge, this method was first used in [16] and then by others, e.g., [2]. We refer the reader to [13, Chapter 8], [14, section 28.2–3] for discussion of local Carleman estimates, and [17] for the parabolic case. For global estimates we refer the reader to [11] and [10].

Stability results for parabolic equations are recent, to our knowledge (see [19], [9]). Apart from [16] there are few results on Lipschitz stability, even for linear cases.

One of the main difficulties in the present problem is to deal with discontinuous diffusion coefficients. Controllability for such parabolic equations has been studied by [8]. A null-controllability property is proved via an observability inequality for the adjoint system, which is deduced in [8] from a *global* Carleman estimate, yet assuming a *monotonicity* on the coefficients c in connection to the observation location: roughly speaking, the observation zone has to be located in the region where the diffusion coefficient is the smallest. Here, to achieve a stability result we have to derive a Carleman estimate for the difference of the two solutions, y, $\widetilde{y}$. This difference is the solution of a nonhomogeneous parabolic equation (with discontinuous coefficient); because of the discontinuity of the diffusion coefficients it does not satisfy the appropriate transmission conditions (TC1), on the interfaces S, defined above. For this reason, under

the same monotonicity assumption as in [8], we derive a peculiar Carleman estimate which includes additional *interface terms* (see Theorem 2.2).

To obtain a stability result, one has to "manage" the dependence of some constants with respect to (w.r.t.) the parameters, s and λ, that appear in the weight functions used in the Carleman estimate (see (2.4) in section 2). The interface terms require some careful treatment. In particular, a stationary-phase argument is used to obtain a sufficiently sharp asymptotic estimate of these terms for s and λ large. Usually, stability estimates are obtained by letting the parameter s become large. Here we also make use of the second parameter λ (see section 4).

As we are concerned with parabolic equations, we have to assume that the observation of the solution occurs at some positive time, $T' > 0$. The suppression of this assumption remains an open problem, and it appears in all articles deriving Lipschitz stability estimates from global Carleman inequalities (see the discussion in the introduction of [16]). However, at the end of section 3 we show that if the position of the interface S is known, we can localize in space the observation at time T'.

The article is organized as follows. In section 2 we derive a Carleman estimate adapted to our problem. In section 3 we prove a stability result for the piecewise-constant diffusion coefficient c when one of the solutions, say $\widetilde{y}$, is in a particular class of solutions. In section 4 we prove that this class is nonempty. As mentioned above we also slightly relax the piecewise-constant condition by imposing solely that the difference of the coefficients be piecewise-constant. In section 5 we prove a stability result for the initial condition under some additional assumptions, particularly on the initial condition itself. The appendix provides some basic regularity properties for the solutions to parabolic equation with nonsmooth coefficients and provides a technical lemma.

We now give some notation and important assumptions. We denote by n the outward unit normal to Ω_1 on S and also the outward unit normal to Ω on Γ. Let S_0 (resp., S_1) be the side of the interface S corresponding to the positive (resp., negative) direction of the normal n.

Note that we do not assume that either Ω_0 or Ω_1 is a connected open set. We shall, however, assume that they are formed with a finite number of connected open sets, say $\Omega_{0,1}, \ldots, \Omega_{0,p_0}$ and $\Omega_{1,1}, \ldots, \Omega_{1,p_1}$, $p_0, p_1 \in \mathbb{N}$. We shall then denote by S_{ij} the interface (possibly empty) between $\Omega_{0,i}$ and $\Omega_{1,j}$.

We make the following assumption.

ASSUMPTION 1.1. *The diffusion coefficient satisfies* $c_i = c_{|\Omega_i} \in \mathscr{C}^1(\overline{\Omega}_i)$, $i = 0, 1$, *and is independent of time* t.

ASSUMPTION 1.2. $c_{0|S} \geq c_{1|S}$ *and* $0 < c_{min} \leq c(x) \leq c_{max}$, $x \in \Omega'$.

Remark 1.3. Assumption 1.1 will be significantly strengthened in section 3 to obtain a stability result. Yet, for some of the results such as the Carleman estimate proved in section 2 and the regularity properties proved in section 4, which can be of some use elsewhere, Assumption 1.1 is sufficient.

We let γ be a subset of the boundary Γ satisfying the following.

ASSUMPTION 1.4. *The interior of* γ *is nonempty w.r.t. the topology on* Γ *induced by the Euclidean topology on* $\mathbb{R}^n$. *Each component of* Ω_1 *contains part of the interior of* γ *in its boundary.*

Examples of situations in which Assumption 1.4 is satisfied are given in Figure 1.

To obtain a Carleman estimate we introduce a geometric assumption, following [8].

ASSUMPTION 1.5. Geometric condition (GC). *We assume that there exist two*

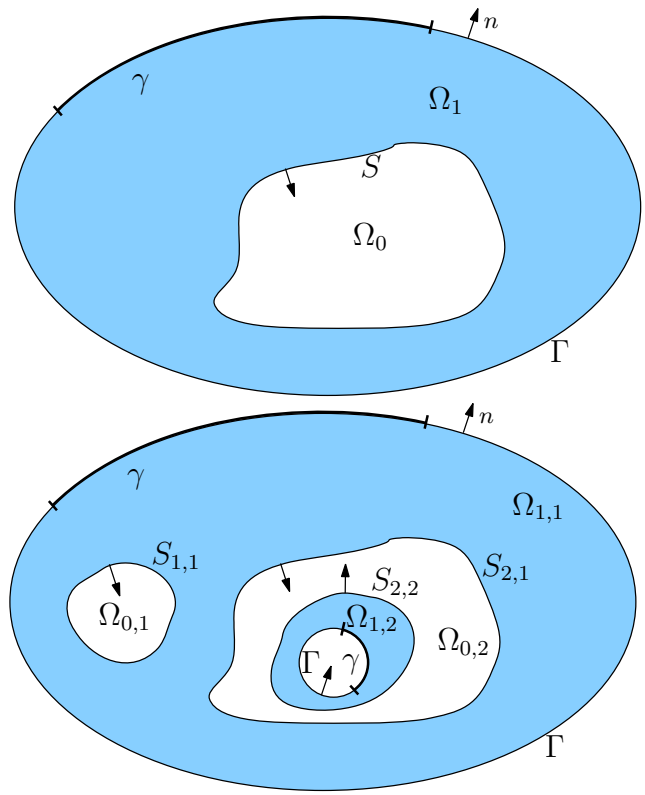


FIG. 1. *Geometric situations in which Assumption* 1.4 *and the geometric condition (GC) are satisfied. Shaded is* Ω_1. *Arrows represent the normal unit vector* n.

disjoint open subsets, $\mathcal{O}^{(1)}$, $\mathcal{O}^{(2)} \Subset \Omega_0$, *and two vector fields,* $\zeta^{(i)} \in \mathscr{C}^1(\overline{\Omega}_0, \mathbb{R}^2)$, $i = 1, 2$, *such that*

$$\zeta^{(i)}(x) \cdot n(x) > 0 \ \ \forall x \in S, \quad i = 1, 2,$$

$$\zeta^{(i)}(x) \cdot n(x) > 0 \ \ \forall x \in \partial\mathcal{O}^{(i)}, \ \ i = 1, 2,$$

$$\zeta^{(i)}(x) \neq 0 \ \ \forall x \in \Omega_0 \setminus \mathcal{O}^{(i)}, \ \ i = 1, 2$$

(n *is the outward unit normal to* Ω_1 *on* S *and the inward unit normal to* $\mathcal{O}^{(i)}$ *on* $\partial\mathcal{O}^{(i)}$, $i = 1, 2$*). Let* $x^{(i)}$ *be the integral curves of* $\zeta^{(i)}$, *i.e.,*

$$\begin{cases} \frac{dx^{(i)}(t)}{dt} = \zeta^{(i)}(x^{(i)}(t)), & t > 0, \\ x^{(i)}(0) = x_0, & x_0 \in S. \end{cases}$$

We also assume that there exists $\mathcal{T} > 0$ *such that for all* $x_0 \in S$ *there exists* $t^{(i)}(x_0) < \mathcal{T}$ *satisfying*

$$x^{(i)}(t) \in \Omega_0 \setminus \mathcal{O}^{(i)} \quad \text{for } 0 < t < t^{(i)}(x_0), \ x_0 \in S, \ i = 1, 2,$$

$$x^{(i)}(t^{(i)}(x_0)) \in \partial\mathcal{O}^{(i)} \quad \textit{for } x_0 \in S,\ i = 1, 2.$$

Note that in Assumption 1.5, there is no restriction to having Ω_0 composed of p_0 components. The examples given in Figure 1 satisfy Assumption 1.5.

We denote by $W^{m,p}(\Omega)$, $m \in \mathbb{N}$, $1 \le p \le \infty$, the usual Sobolev space defined by

$$W^{m,p} = \{u \in L^p(\Omega);\ \partial^\alpha u \in L^p(\Omega) \text{ for } |\alpha| \le m\},$$

where $\alpha = (\alpha_1, \dots, \alpha_n)$ is a multi-index and differentiation is to be understood in the weak sense. As usual we write $H^m(\Omega) = W^{m,2}(\Omega)$. For the definition of $W^{r,p}$ for $r \in \mathbb{R} \setminus \mathbb{N}$ we refer, for instance, to [1].

2. A Carleman estimate. We prove here a Carleman-type estimate with a boundary term on γ in the right-hand side (r.h.s.) of the estimate. For this purpose we shall first introduce a particular type of weight functions, which are constructed using the following lemma.

LEMMA 2.1. *Assume that there exist two disjoint open subsets, $\mathcal{O}^{(1)}$, $\mathcal{O}^{(2)} \Subset \Omega_0$, satisfying Assumption* 1.5. *Let γ be a subset of $\Gamma = \partial\Omega$ satisfying Assumption* 1.4, *and $B^{(i)}$ and $\widetilde{B}^{(i)}$, $i = 1, 2$, be open balls such that $B^{(1)} \Subset \widetilde{B}^{(1)} \Subset \mathcal{O}^{(1)}$ and $B^{(2)} \Subset \widetilde{B}^{(2)} \Subset \mathcal{O}^{(2)}$. Then there exist two functions $\widetilde{\beta}^{(1)}$ and $\widetilde{\beta}^{(2)}$ such that*

$$\widetilde{\beta}^{(1)}(x) = \begin{cases} \widetilde{\beta}_0^{(1)} & \textit{in } \Omega_0, \\ \widetilde{\beta}_1 & \textit{in } \overline{\Omega}_1, \end{cases} \qquad \widetilde{\beta}^{(2)}(x) = \begin{cases} \widetilde{\beta}_0^{(2)} & \textit{in } \Omega_0, \\ \widetilde{\beta}_1 & \textit{in } \overline{\Omega}_1, \end{cases}$$

and the functions $\widetilde{\beta}_0^{(1)}$, $\widetilde{\beta}_0^{(2)}$, and $\widetilde{\beta}_1$ satisfy the following properties: $\widetilde{\beta}_1 \in \mathscr{C}^2(\overline{\Omega}_1)$, $\widetilde{\beta}_1 > 0$ in Ω_1, and

$$\widetilde{\beta}_1 = 0 \quad \textit{on } \Gamma \setminus \gamma, \quad \partial_n \widetilde{\beta}_1 < 0 \quad \textit{on } \Gamma \setminus \gamma,$$

$$\widetilde{\beta}_1 = 2 \quad \textit{on } S, \quad \partial_n \widetilde{\beta}_1 < 0 \quad \textit{on } S,$$

and

$$|\nabla \widetilde{\beta}_1| > 0 \quad \textit{in } \overline{\Omega}_1;$$

for $i = 1, 2$, $\widetilde{\beta}_0^{(i)} \in \mathscr{C}^2(\overline{\Omega}_0)$, $\widetilde{\beta}_0^{(i)} > 0$ in Ω_0,

$$\widetilde{\beta}_0^{(i)} = \widetilde{\beta}_1 = 2 \quad \textit{on } S, \quad i = 1, 2,$$

$$c_0 \partial_n \widetilde{\beta}_0^{(i)} = c_1 \partial_n \widetilde{\beta}_1 \quad \textit{on } S, \quad i = 1, 2,$$

$$\widetilde{\beta}_0^{(1)} \ge 2\widetilde{\beta}_0^{(2)} \quad \textit{in } \widetilde{B}^{(2)}, \tag{2.1}$$

$$\widetilde{\beta}_0^{(2)} \ge 2\widetilde{\beta}_0^{(1)} \quad \textit{in } \widetilde{B}^{(1)}, \tag{2.2}$$

and

$$|\nabla \widetilde{\beta}_0^{(i)}| > 0 \quad \textit{in } \overline{\Omega}_0 \setminus B^{(i)}, \quad i = 1, 2. \tag{2.3}$$

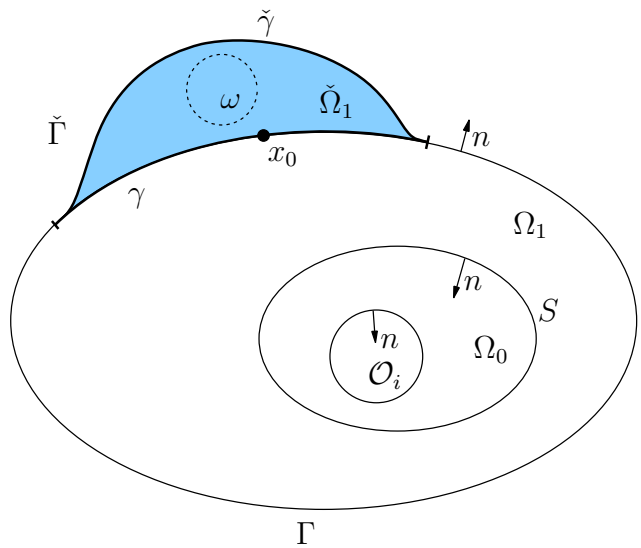


FIG. 2. *Geometrical situation for the proof Lemma* 2.1. *The shaded area represents* $\check{\Omega}_1 \setminus \Omega_1 = \check{\Omega} \setminus \Omega$.

Proof. For the construction of $\widetilde{\beta}_0^{(i)}$, $i = 1, 2$, supported in the connected components of Ω_0, we refer to [8, Lemma 3.2]. We briefly show how the function $\widetilde{\beta}_1$ is constructed. In actuality, the procedure described here has to be performed in each connected component of Ω_1, which is possible since each component contains part of the interior of γ in its boundary by Assumption 1.4.

Let x_0 be in the interior of γ. We can enlarge the open set Ω_1 locally around x_0 while preserving the $\mathscr{C}^2$ regularity of the boundary. Such a procedure is performed in a neighborhood U of x_0 such that $U \cap \Gamma \subset \gamma$. (This can be done by locally straightening out the boundary γ, as $\overline{\Omega}$ is assumed to be a $\mathscr{C}^2$ submanifold with boundary in $\mathbb{R}^n$ [12, Definition 1.2.1.2].) This enlarging procedure affects only γ and leaves $\Gamma \setminus \gamma$ untouched. We call the new boundary $\check{\Gamma}$. We denote by $\check{\Omega}_1$ the extension of Ω_1 and $\check{\Omega}$ that of Ω ($\Omega_1 \subset \check{\Omega}_1$, $\Omega \subset \check{\Omega}$, and $\check{\Gamma} = \partial\check{\Omega}$). Let ω be an open subset such that $\omega \Subset \check{\Omega}_1 \setminus \Omega_1$. The geometry we describe here is illustrated in Figure 2. Following [8, 11], there exists $\mu \in \mathscr{C}^2(\overline{\check{\Omega}})$ that satisfies

$$\mu = 0, \quad \partial_n \mu < 0 \quad \text{on } \check{\Gamma},$$

$$\mu = 2, \quad c_0 \partial_n \widetilde{\beta}_0^{(i)} = c_1 \partial_n \mu < 0 \quad \text{on } S,$$

$$|\nabla \mu| > 0 \quad \text{in } \check{\Omega}_1 \setminus \omega.$$

The function $\widetilde{\beta}_1 := \mu_{|\overline{\Omega}_1}$ satisfies the required properties. ∎

Choosing two functions $\widetilde{\beta}^{(i)}$, $i = 1, 2$, as in the previous lemma, we introduce $\beta^{(i)} = \widetilde{\beta}^{(i)} + K$ with $K = m\|\widetilde{\beta}^{(1)}\|_\infty = m\|\widetilde{\beta}^{(2)}\|_\infty$ and $m > 1$. For $\lambda > 0$ and $t \in (t_0, T)$, we define the following weight functions:

$$\varphi^{(i)}(t,x) = \frac{e^{\lambda\beta^{(i)}(x)}}{(t-t_0)(T-t)}, \qquad \eta^{(i)}(t,x) = \frac{e^{\lambda\overline{\beta}} - e^{\lambda\beta^{(i)}(x)}}{(t-t_0)(T-t)}, \quad i = 1, 2, \tag{2.4}$$

with $\overline{\beta} = 2m\|\widetilde{\beta}^{(i)}\|_\infty$, $i = 1, 2$ (see [8], [10]). We let $t_0 \in (0, T)$, and we set $Q = (t_0, T) \times \Omega$, $Q' = (t_0, T) \times \Omega'$ and recall that $\Omega' = \Omega_0 \cup \Omega_1$.

Let $g \in H^1([t_0, T], H^{\frac{1}{2}}(S))$. We introduce transmission conditions (TC2) on the interval $[t_0, T]$,

$$\text{(TC2)} \qquad \begin{aligned} &q_{|[t_0,T]\times S_0} = q_{|[t_0,T]\times S_1}, \\ &c_0 \partial_n q_{|[t_0,T]\times S_0} = c_1 \partial_n q_{|[t_0,T]\times S_1} + g(t,x), \end{aligned}$$

for a function q which is H^2 in each open set Ω_i, $i = 0, 1$.

We introduce

$$\aleph_g = \Big\{ q \in H^1(t_0, T, H_0^1(\Omega));\ q_{|_{(t_0,T)\times\Omega_i}} \in L^2(t_0, T, H^2(\Omega_i)),\ i = 0, 1, \\ \text{and } q \text{ satisfies (TC2) a.e. w.r.t. } t \Big\}.$$

THEOREM 2.2. *Let γ be a subset of the boundary Γ of an open set Ω of $\mathbb{R}^n$ that satisfies Assumption* 1.5*, and let γ satisfy Assumption* 1.4*. Let c satisfy Assumptions* 1.1 *and* 1.2*. Assume further that $c_{0|_S} - c_{1|_S} \geq \Delta > 0$. Let $g \in H^1(t_0, T, H^{\frac{1}{2}}(S))$. There exists $\lambda_1 = \lambda_1(\Omega, \gamma, \mathcal{O}^{(1)}, \mathcal{O}^{(2)}, c_{min}, c_{max}, \Delta) > 0$, $s_1 = s_1(\lambda_1) > 0$, and a positive constant $C = C(\Omega, \gamma, \mathcal{O}^{(1)}, \mathcal{O}^{(2)}, c_{min}, c_{max}, \Delta)$ so that the following estimate holds:*

$$\begin{aligned} (2.5) \qquad & |M_1^{(1)}(e^{-s\eta^{(1)}} q)|^2_{L^2(Q')} + |M_1^{(2)}(e^{-s\eta^{(2)}} q)|^2_{L^2(Q')} \\ &+ |M_2^{(1)}(e^{-s\eta^{(1)}} q)|^2_{L^2(Q')} + |M_2^{(2)}(e^{-s\eta^{(2)}} q)|^2_{L^2(Q')} \\ &+ s\lambda^2 \iint_Q (e^{-2s\eta^{(1)}} \varphi^{(1)} + e^{-2s\eta^{(2)}} \varphi^{(2)})\ |\nabla q|^2\ dx\ dt \\ &+ s^3\lambda^4 \iint_Q (e^{-2s\eta^{(1)}} {\varphi^{(1)}}^3 + e^{-2s\eta^{(2)}} {\varphi^{(2)}}^3)\ |q|^2\ dx\ dt \\ &\leq C \Bigg[s\lambda \int_{t_0}^T \int_\gamma (e^{-2s\eta^{(1)}} \varphi^{(1)} + e^{-2s\eta^{(2)}} \varphi^{(2)})\ |\partial_n q|^2\ d\sigma\ dt \\ &\quad + \iint_{Q'} (e^{-2s\eta^{(1)}} + e^{-2s\eta^{(2)}})\ |\partial_t q \pm \nabla \cdot (c\nabla q)|^2\ dx\ dt \\ &\quad + s\lambda \int_{t_0}^T \int_S (e^{-2s\eta^{(1)}} \varphi^{(1)} + e^{-2s\eta^{(2)}} \varphi^{(2)})\ |g|^2\ d\sigma\ dt \\ &\quad + \int_{t_0}^T \int_S (e^{-2s\eta^{(1)}} {\varphi^{(1)}}^4 + e^{-2s\eta^{(2)}} {\varphi^{(2)}}^4)\ |g|^2\ d\sigma\ dt \\ &\quad + s^{-2} \int_{t_0}^T \int_S (e^{-2s\eta^{(1)}} + e^{-2s\eta^{(2)}}) |\partial_t g|^2 d\sigma\ dt \Bigg] \end{aligned}$$

for $s \geq s_1$, $\lambda \geq \lambda_1$, and all $q \in \aleph_g$, with M_1 and M_2 to be defined in (2.9)–(2.10).

We recall that $\Omega' = \Omega_0 \cup \Omega_1$ and $Q' = \Omega' \times (t_0, T)$. For a function ρ with a trace on the interface S, from both sides, defined in some sense, we shall denote by ρ_i the trace of $\rho_{|_{\Omega_i}}$ on S, $i = 0, 1$, when there is no ambiguity; in the case $\rho_0 = \rho_1$ we shall simply write ρ. We shall use the notation $[\rho]_S = \rho_0 - \rho_1$ for the jump of ρ across the interface S. We shall adopt Einstein's summation convention for repeated indices.

Proof. We consider $s > 0$ and $q \in \aleph_g$. Let us set $f = \partial_t q + \nabla \cdot (c\nabla q)$ (we treat the case of the operator $\partial_t + \nabla \cdot (c\nabla)$ in the proof; the other case can be treated similarly).

Then $f \in L^2(Q')$ (because of transmission conditions (TC2), observe that q is not in the domain of the operator $\nabla \cdot (c\nabla)$).

In the first part of the proof we shall write η, φ, M_1, etc., in place of $\eta^{(i)}$, $\varphi^{(i)}$, $M_1^{(i)}$, etc., $i = 1, 2$, and treat the two cases at once. We set $\psi = e^{-s\eta}q$. We observe that $\psi(t_0) = \psi(T) = 0$ and since q satisfies transmission conditions (TC2) (and $q(t,.)_{|\Omega_i} \in H^2(\Omega_i)$ a.e. w.r.t. t), we have (a.e. w.r.t. t)

$$c_0 \partial_n \psi_{0|_S}(t,.) = c_1 \partial_n \psi_{1|_S}(t,.) + g_s(t,.) \quad \text{on } S, \tag{2.6}$$

$$\nabla_\tau \psi_{0|_S}(t,.) = \nabla_\tau \psi_{1|_S}(t,.) \quad \text{on } S, \tag{2.7}$$

$$\psi_{0|_S}(t,.) = \psi_{1|_S}(t,.) \quad \text{on } S, \tag{2.8}$$

where $g_s = e^{-s\eta}g$ and ∇_τ denotes the component of the gradient that is tangential to S.

The function ψ satisfies in each Ω_i, $i = 0, 1$,

$$M_1\psi + M_2\psi = f_s$$

with

$$M_1\psi = \nabla \cdot (c\nabla\psi) + s^2\lambda^2\varphi^2|\nabla\beta|^2 c\psi + s(\partial_t\eta)\psi, \tag{2.9}$$

$$M_2\psi = \partial_t\psi - 2s\lambda\varphi c\nabla\beta \cdot \nabla\psi - 2s\lambda^2\varphi c|\nabla\beta|^2\psi, \tag{2.10}$$

$$f_s = e^{-s\eta}f + s\lambda\varphi\nabla \cdot (c\nabla\beta)\psi - s\lambda^2\varphi c|\nabla\beta|^2\psi. \tag{2.11}$$

We have

$$|M_1\psi|^2_{L^2(Q')} + |M_2\psi|^2_{L^2(Q')} + 2(M_1\psi, M_2\psi)_{L^2(Q')} = |f_s|^2_{L^2(Q')}.$$

With the same notation as in [8, Theorem 3.3], we write $(M_1\psi, M_2\psi)_{L^2(Q')}$ as a sum of nine terms I_{ij}, $1 \le i, j \le 3$, where I_{ij} is the inner product of the ith term in the expression of $M_1\psi$ and the jth term in the expression of $M_2\psi$.

As compared to the proof of the Carleman estimate in [8, Theorem 3.3] we only need to adjust the computation of I_{11}, I_{12}, and I_{13} to the present case. In fact the other terms do not involve transmission conditions (2.6) in their computation and thus remain unchanged from the terms obtained in [8].

The term I_{11} follows as

$$I_{11} = \iint_{Q'} \nabla \cdot (c\nabla\psi)\, \partial_t\psi \, dxdt = -\iint_{Q'} c\nabla\psi \cdot \partial_t(\nabla\psi)\, dxdt + \sum_{i=0,1} (-1)^{i+1} \int_{t_0}^{T}\int_S c_i \partial_n\psi_i \, \partial_t\psi \, d\sigma dt,$$

by integration by parts; the surface integral on Γ vanishes since $\partial_t\psi = 0$ there. Noting that $\nabla\psi \cdot \partial_t(\nabla\psi) = \frac{1}{2}\partial_t(|\nabla\psi|^2)$, the first term vanishes since ψ, and thus $\nabla\psi$, vanish at $t = t_0$ and $t = T$ and c is independent of t. For the remaining surface terms we use (2.6), which yields

$$I_{11} = -\int_{t_0}^{T}\int_S g_s \, \partial_t\psi \, d\sigma dt = \int_{t_0}^{T}\int_S \partial_t(g_s)\, \psi \, d\sigma dt,$$

since $g_s \in H^1(t_0, T, H^{\frac{1}{2}}(S))$.

The term I_{12} is given by

$$I_{12} = -2s\lambda \iint_{Q'} \varphi \nabla \cdot (c\nabla\psi) c\nabla\beta \cdot \nabla\psi \, dxdt = 2s\lambda \iint_{Q'} c\nabla\psi \cdot \nabla(\varphi c \nabla\beta \cdot \nabla\psi) \, dxdt$$
$$- 2s\lambda \int_{t_0}^{T}\int_{\Gamma} \varphi c^2 (\nabla\beta\cdot\nabla\psi)(\partial_n\psi) \, d\sigma dt + 2s\lambda \sum_{i=0,1} (-1)^i \int_{t_0}^{T}\int_{S} \varphi c_i^2 (\nabla\beta_i\cdot\nabla\psi_i)(\partial_n\psi_i) \, d\sigma dt.$$

This integration by parts is justified since $\psi(t,.)$ is in H^2 in each Ω_i, $i = 0, 1$. Denoting by I'_{12} the remaining volume integral, we obtain

$$I'_{12} = 2s\lambda^2 \iint_{Q'} \varphi c^2 |\nabla\psi \cdot \nabla\beta|^2 \, dxdt + 2s\lambda \iint_{Q'} \varphi c \, \partial_{x_i}(c\partial_{x_j}\beta)\partial_{x_i}\psi\partial_{x_j}\psi \, dxdt$$
$$+ s\lambda \iint_{Q'} \varphi c^2 \partial_{x_j}\beta\partial_{x_j}|\nabla\psi|^2 \, dxdt.$$

We further compute the last volume integral, denoted by I''_{12}. Observe that $|\nabla\psi|^2_{|\Omega_i}$ is in $W^{1,1}(\Omega_i)$ since $\psi_{|\Omega_i}(t,.) \in H^2(\Omega_i)$, $i = 0, 1$. This allows us to further integrate by parts, since $(c^2\varphi\partial_{x_j}\beta)_{|\Omega_i} \in \mathscr{C}^1(\Omega_i)$, $i = 0, 1$, and yields

$$I''_{12} = -s\lambda \iint_{Q'} \partial_{x_j}(\varphi c^2 \partial_{x_j}\beta)|\nabla\psi|^2 \, dxdt + s\lambda \int_{t_0}^{T}\int_{\Gamma} \varphi c^2 \partial_n\beta |\nabla\psi|^2 \, d\sigma dt$$
$$+ s\lambda \sum_{i=0,1} (-1)^{i+1} \int_{t_0}^{T}\int_{S} \varphi c_i^2 \partial_n\beta_i |\nabla\psi_i|^2 \, d\sigma dt.$$

The remaining volume integral can be further expanded into

$$I'''_{12} = -s\lambda \iint_{Q'} \varphi\partial_{x_j}(c^2\partial_{x_j}\beta)|\nabla\psi|^2 \, dxdt - s\lambda^2 \iint_{Q'} \varphi c^2 |\nabla\beta|^2 |\nabla\psi|^2 \, dxdt.$$

Collecting the surface integrals in a term denoted by J_{12}, we find

$$I_{12} = -s\lambda^2 \iint_{Q'} \varphi c^2 |\nabla\beta|^2 |\nabla\psi|^2 \, dxdt + 2s\lambda^2 \iint_{Q'} c^2\varphi |\nabla\psi \cdot \nabla\beta|^2 \, dxdt + X_1 + J_{12},$$

where

$$X_1 = 2s\lambda \iint_{Q'} \varphi c \, \partial_{x_i}(c\partial_{x_j}\beta)\partial_{x_i}\psi\partial_{x_j}\psi \, dxdt - s\lambda \iint_{Q'} \varphi\partial_{x_j}(c^2\partial_{x_j}\beta)|\nabla\psi|^2 \, dxdt.$$

We now observe that since β is constant on S we have

$$(\nabla\beta \cdot \nabla\psi_i)_{|S} = (\partial_n\beta\partial_n\psi_i)_{|S}, \quad i = 0, 1.$$

Writing $|\nabla\psi|^2 = |\nabla_\tau\psi|^2 + |\partial_n\psi|^2$, we find

$$J_{12} = s\lambda \sum_{i=0,1} (-1)^i \int_{t_0}^{T}\int_{S} \varphi c_i^2 \partial_n\beta_i |\partial_n\psi_i|^2 \, d\sigma dt$$
$$- s\lambda \sum_{i=0,1} (-1)^i \int_{t_0}^{T}\int_{S} \varphi c_i^2 \partial_n\beta_i |\nabla_\tau\psi_i|^2 \, d\sigma dt - s\lambda \int_{t_0}^{T}\int_{\Gamma} \varphi c^2 \partial_n\beta |\partial_n\psi|^2 \, d\sigma dt,$$

where we have used that $\psi_{|\Sigma}$ is constant. Recall that $\nabla_\tau \psi_0 = \nabla_\tau \psi_1$ and that $c_0 \partial_n \beta_0 = c_1 \partial_n \beta_1$ on S. From transmission conditions (TC2) we have

$$|c_0 \partial_n \psi_0|^2 = |c_1 \partial_n \psi_1|^2 + |g_s|^2 + 2(c_1(\partial_n \psi_1) g_s) \quad \text{on } S.$$

We thus obtain

$$\begin{aligned} J_{12} = s\lambda \int_{t_0}^{T}\!\!\int_S \varphi [\partial_n \beta]_S |c_1 \partial_n \psi_1|^2 \, d\sigma dt + s\lambda \int_{t_0}^{T}\!\!\int_S \varphi \partial_n \beta_0 |g_s|^2 \, d\sigma dt \\ - s\lambda \int_{t_0}^{T}\!\!\int_S \varphi [c]_S (c \partial_n \beta) |\nabla_\tau \psi|^2 \, d\sigma dt - s\lambda \int_{t_0}^{T}\!\!\int_\Gamma \varphi c^2 \partial_n \beta |\partial_n \psi|^2 \, d\sigma dt + Y_1, \end{aligned}$$

with

$$Y_1 = 2s\lambda \int_{t_0}^{T}\!\!\int_S \varphi c_1 \partial_n \psi_1 \, \partial_n \beta_0 \, g_s \, d\sigma dt. \tag{2.12}$$

We thus have

$$\begin{aligned} I_{12} = -s\lambda^2 \iint_{Q'} \varphi c^2 |\nabla \beta|^2 |\nabla \psi|^2 \, dxdt + 2s\lambda^2 \iint_{Q'} c^2 \varphi |\nabla \psi \cdot \nabla \beta|^2 \, dxdt \\ + s\lambda \int_{t_0}^{T}\!\!\int_S \varphi [\partial_n \beta]_S |c_1 \partial_n \psi_1|^2 \, d\sigma dt + s\lambda \int_{t_0}^{T}\!\!\int_S \varphi \partial_n \beta_0 |g_s|^2 \, d\sigma dt \\ - s\lambda \int_{t_0}^{T}\!\!\int_S \varphi [c]_S (c \partial_n \beta) |\nabla_\tau \psi|^2 \, d\sigma dt - s\lambda \int_{t_0}^{T}\!\!\int_\Gamma \varphi c^2 \partial_n \beta |\partial_n \psi|^2 \, d\sigma dt \\ + X_1 + Y_1. \end{aligned}$$

The term I_{13} is given by

$$\begin{aligned} I_{13} = -2s\lambda^2 \iint_{Q'} \varphi \nabla \cdot (c \nabla \psi) c |\nabla \beta|^2 \psi \, dxdt = 2s\lambda^2 \iint_{Q'} c \nabla \psi \cdot \nabla (\varphi c |\nabla \beta|^2 \psi) \, dxdt \\ + 2s\lambda^2 \sum_{i=0,1} (-1)^i \int_{t_0}^{T}\!\!\int_S \varphi (c_i \partial_n \psi_i) c_i |\nabla \beta_i|^2 \psi \, d\sigma dt, \end{aligned}$$

where we have used that $\psi_{|\Gamma} = 0$. Expanding the integrand in the volume integral and using (TC2) in the surface term, we obtain

$$I_{13} = 2s\lambda^2 \iint_{Q'} \varphi c^2 |\nabla \beta|^2 |\nabla \psi|^2 \, dxdt + X_2 + Y_2,$$

where

$$\begin{aligned} X_2 = 2s\lambda^2 \iint_{Q'} \varphi c \nabla \psi \cdot \nabla (c |\nabla \beta|^2) \psi \, dxdt + 2s\lambda^3 \iint_{Q'} \varphi c^2 \nabla \psi \cdot \nabla \beta |\nabla \beta|^2 \psi \, dxdt \\ + 2s\lambda^2 \int_{t_0}^{T}\!\!\int_S \varphi (c \partial_n \beta) [\partial_n \beta]_S (c_1 \partial_n \psi_1) \psi \, d\sigma dt, \end{aligned}$$

since $\nabla_\tau \beta_{|S} = 0$ and

$$Y_2 = 2s\lambda^2 \int_{t_0}^{T}\!\!\int_S \varphi c_0 (\partial_n \beta_0)^2 g_s \psi \, d\sigma dt.$$

Following the proof of Theorem 3.3 in [8], we find

$$I_{21} = \frac{1}{2}s^2\lambda^2 \iint_{Q'} \varphi^2 c|\nabla\beta|^2\partial_t|\psi|^2 \, dxdt = -\frac{1}{2}s^2\lambda^2 \iint_{Q'} \partial_t(\varphi^2)c|\nabla\beta|^2|\psi|^2 \, dxdt$$

and

$$\begin{aligned} I_{22} &= -s^3\lambda^3 \iint_{Q'} \varphi^3 c^2|\nabla\beta|^2\nabla\beta\cdot\nabla(|\psi|^2) \, dxdt \\ &= 3s^3\lambda^4 \iint_{Q'} \varphi^3 c^2|\nabla\beta|^4|\psi|^2 \, dxdt + s^3\lambda^3 \int_{t_0}^{T}\int_S \varphi^3|c\partial_n\beta|^2[\partial_n\beta]_S|\psi|^2 \, d\sigma dt + X_3, \end{aligned}$$

with X_3 given by

$$X_3 = s^3\lambda^3 \iint_{Q'} \varphi^3\nabla\cdot(c^2|\nabla\beta|^2\nabla\beta)|\psi|^2 \, dxdt.$$

The terms I_{23}, I_{31} are given by

$$I_{23} = -2s^3\lambda^4 \iint_{Q'} \varphi^3 c^2|\nabla\beta|^4|\psi|^2 \, dxdt,$$

$$I_{31} = \frac{1}{2}s \iint_{Q'} \partial_t\eta\partial_t(|\psi|^2) \, dxdt = -\frac{1}{2}s \iint_{Q'} \partial_t^2\eta|\psi|^2 \, dxdt.$$

The term I_{32} is given by

$$\begin{aligned} I_{32} = -s^2\lambda \iint_{Q'} \varphi(\partial_t\eta)c\nabla\beta\cdot\nabla(|\psi|^2) \, dxdt &= s^2\lambda^2 \iint_{Q'} \varphi(\partial_t\eta)c|\nabla\beta|^2|\psi|^2 \, dxdt \\ &\quad + s^2\lambda \iint_{Q'} \varphi\nabla\cdot((\partial_t\eta)c\nabla\beta)|\psi|^2 \, dxdt, \end{aligned}$$

since $\psi_{|S_0} = \psi_{|S_1}$. Finally, the term I_{33} is given by

$$I_{33} = -2s^2\lambda^2 \iint_{Q'} \varphi c(\partial_t\eta)|\nabla\beta|^2|\psi|^2 \, dxdt.$$

Collecting the terms I_{ij} just computed, we obtain

$$\begin{aligned} &|M_1\psi|^2_{L^2(Q')} + |M_2\psi|^2_{L^2(Q')} + 4s\lambda^2 \iint_{Q'} c^2\varphi|\nabla\psi\cdot\nabla\beta|^2 \, dxdt \\ &\quad + 2s\lambda^2 \iint_{Q'} \varphi c^2|\nabla\beta|^2|\nabla\psi|^2 \, dxdt + 2s^3\lambda^4 \iint_{Q'} \varphi^3 c^2|\nabla\beta|^4|\psi|^2 \, dxdt \\ &\quad + 2s\lambda \int_{t_0}^{T}\int_S \varphi[\partial_n\beta]_S|c_1\partial_n\psi_1|^2 \, d\sigma dt - 2s\lambda \int_{t_0}^{T}\int_\Gamma \varphi c^2\partial_n\beta|\partial_n\psi|^2 \, d\sigma dt \\ &\quad - 2s\lambda \int_{t_0}^{T}\int_S \varphi[c]_S(c\partial_n\beta)|\nabla_\tau\psi|^2 \, d\sigma dt + 2s^3\lambda^3 \int_{t_0}^{T}\int_S \varphi^3|c\partial_n\beta|^2[\partial_n\beta]_S|\psi|^2 \, d\sigma dt \\ &\quad + 2s\lambda \int_{t_0}^{T}\int_S \varphi\partial_n\beta_0|g_s|^2 \, d\sigma dt \\ &\quad = |f_s|^2_{L^2(Q')} - 2(I_{11} + X_1 + Y_1 + X_2 + Y_2 + I_{21} + X_3 + I_{31} + I_{32} + I_{33}). \end{aligned} \tag{2.13}$$

We now consider the surface terms I_{11}, Y_1, Y_2 involving the function g_s and write

$$|I_{11}| = \left| \int_{t_0}^{T}\int_{S} \partial_t(g_s)\, \psi \, d\sigma dt \right| \leq Cs^{-2} \int_{t_0}^{T}\int_{S} |\partial_t g_s|^2 \, d\sigma dt + Cs^2 \int_{t_0}^{T}\int_{S} |\psi|^2 \, d\sigma dt. \tag{2.14}$$

In the proof of Lemma 2.1 we are free to choose β such that $\partial_n\beta_1/c_0 \leq -1$. Since we assume $c_0 - c_1 \geq \Delta$ we obtain

$$[\partial_n\beta]_S = \partial_n\beta_0 - \partial_n\beta_1 = \frac{\partial_n\beta_1}{c_0}(c_1 - c_0) \geq \Delta > 0. \tag{2.15}$$

The second term in (2.14) can thus be absorbed by the term

$$2s^3\lambda^3 \int_{t_0}^{T}\int_{S} \varphi^3 |c\partial_n\beta|^2 [\partial_n\beta]_S |\psi|^2 \, d\sigma dt$$

in (2.13) for s sufficiently large.

The term Y_1 in (2.12) can be estimated by

$$\begin{aligned} |Y_1| &= \left| 2s\lambda \int_{t_0}^{T}\int_{S} \varphi c_1 \partial_n\psi_1 \, \partial_n\beta_0 \, g_s \, d\sigma dt \right| \\ &\leq C_\varepsilon s\lambda \int_{t_0}^{T}\int_{S} \varphi |g_s|^2 \, d\sigma dt + \varepsilon s\lambda \int_{t_0}^{T}\int_{S} \varphi |c_1\partial_n\psi_1|^2 (\partial_n\beta_0)^2 \, d\sigma dt, \quad \varepsilon > 0. \end{aligned} \tag{2.16}$$

For ε sufficiently small, the second surface term in (2.16) can be "absorbed" by the term

$$2s\lambda \int_{t_0}^{T}\int_{S} \varphi [\partial_n\beta]_S |c_1\partial_n\psi_1|^2 \, d\sigma dt$$

in (2.13) by (2.15).

The term Y_2 can be estimated by

$$\begin{aligned} |Y_2| = \left| 2s\lambda^2 \int_{t_0}^{T}\int_{S} \varphi c_0 (\partial_n\beta_0)^2 g_s \psi \, d\sigma dt \right| &\leq Cs\lambda \int_{t_0}^{T}\int_{S} \varphi |g_s|^2 \, d\sigma dt \\ &\quad + Cs\lambda^3 \int_{t_0}^{T}\int_{S} \varphi c_0^2 (\partial_n\beta_0)^4 |\psi|^2 \, d\sigma dt. \end{aligned}$$

Observing that $\varphi \leq CT^4\varphi^3$, the second surface term can be absorbed by the term

$$2s^3\lambda^3 \int_{t_0}^{T}\int_{S} \varphi^3 |c\partial_n\beta|^2 [\partial_n\beta]_S |\psi|^2 \, d\sigma dt$$

in (2.13) for s sufficiently large by (2.15). The two previous "absorption processes" are the points in the proof where the hypothesis $c_0 - c_1 \geq \Delta > 0$ is needed.

Note also that

$$\begin{aligned} s^{-2}|\partial_t g_s|^2 &\leq Cs^{-2}e^{-2s\eta}|\partial_t g|^2 + C(\partial_t\eta)^2 e^{-2s\eta}|g|^2 \\ &\leq Cs^{-2}e^{-2s\eta}|\partial_t g|^2 + CT^2\varphi^4 e^{-2s\eta}|g|^2, \end{aligned}$$

where we have used that $|\partial_t \eta| \leq CT\varphi^2$ [8, equation (90)] (which makes use of the particular choices made above for K and $\overline{\beta}$, which implies that $\overline{\beta} \leq 2\beta$).

Applying the technique presented in the proof of Theorem 3.3 in [8], the previous observations yield the following Carleman estimate (we use the notation $\eta^{(i)}$ instead of η):

(2.17)

$$
\begin{aligned}
&|M_1^{(i)}(e^{-s\eta^{(i)}}q)|^2_{L^2(Q')} + |M_2^{(i)}(e^{-s\eta^{(i)}}q)|^2_{L^2(Q')} + s\lambda^2 \iint_Q e^{-2s\eta^{(i)}}\ \varphi^{(i)}|\nabla q|^2\ dx\ dt \\
&\qquad + s^3\lambda^4 \iint_Q e^{-2s\eta^{(i)}}\ {\varphi^{(i)}}^3|q|^2\ dx\ dt \\
&\leq C\left[s\lambda \int_{t_0}^T\!\!\int_\gamma e^{-2s\eta^{(i)}}\ \varphi^{(i)}|\partial_n q|^2\ d\sigma\ dt + s^3\lambda^4 \int_{t_0}^T\!\!\int_{\widetilde{B}^{(i)}} e^{-2s\eta^{(i)}}{\varphi^{(i)}}^3|q|^2 dx dt\right. \\
&\qquad + \iint_{Q'} e^{-2s\eta^{(i)}}\ |\partial_t q - \nabla\cdot(c\nabla q)|^2\ dx\ dt + s^{-2}\int_{t_0}^T\!\!\int_S e^{-2s\eta^{(i)}}|\partial_t g|^2 d\sigma\ dt \\
&\qquad\qquad \left. + \int_{t_0}^T\!\!\int_S e^{-2s\eta^{(i)}}\ {\varphi^{(i)}}^4|g|^2\ d\sigma\ dt + s\lambda\int_{t_0}^T\!\!\int_S e^{-2s\eta^{(i)}}\ \varphi^{(i)}|g|^2\ d\sigma\ dt\right],
\end{aligned}
$$

for $i = 1, 2$, and for $\lambda \geq \lambda_0(\Omega, \gamma, \mathcal{O}^{(1)}, \mathcal{O}^{(2)}, c_{min}, c_{max}, \Delta)$ and $s \geq s_0(\lambda_0)$ (the sets $\widetilde{B}^{(i)}$ were introduced in Lemma 2.1). Note that the condition $[c]_s \geq 0$ is needed to obtain the previous estimate.

Adding (2.17) for $i = 1, 2$, we deduce (2.5) with the same argumentation as in the proof of Theorem 3.4 in [8]. The terms integrated over $(t_0, T) \times \widetilde{B}^{(i)}$ are absorbed by other terms using properties (2.1)–(2.3) of $\widetilde{\beta}^{(i)}$, $i = 1, 2$. □

Remark 2.3. The Carleman estimate that was just derived is peculiar because of the presence of terms integrated on the interface S. In particular, two terms involve the function g with different powers for the parameters s and λ and for the weight functions $\varphi^{(i)}$, $i = 1, 2$. This Carleman estimate is the key ingredient in the subsequent analysis. The interface terms will require some special treatment. The two parameters s and λ will also have an important role to play in the next section.

Remark 2.4. In the case $g = 0$, the previous Carleman estimate simplifies. By inspection of the proof of Theorem 2.2, observe that in the case $g = 0$, the condition $c_{0|_S} - c_{1|_S} \geq 0$ is sufficient to obtain the Carleman estimate [8]. The case of $c_{0|_S} - c_{1|_S} < 0$ remains open in the case of a dimension greater than or equal to 2. (In the one-dimensional case a Carleman estimate for the heat operator, $\partial_t \pm \partial_x(c\partial_x)$, in arbitrary situations, can be found in [3, 4].)

3. Uniqueness and stability estimate for the diffusion coefficients. In this section we establish a uniqueness result for the discontinuous diffusion coefficient c as well as a stability inequality. This inequality estimates the discrepancy in the coefficients c and $\widetilde{c}$ of two materials (with the same geometry) with an upper bound given by some Sobolev norms of the difference between the solutions y and $\widetilde{y}$ to

$$
\begin{cases}
\partial_t \widetilde{y} - \nabla\cdot(\widetilde{c}\nabla\widetilde{y}) = 0 & \text{in } (0,T)\times\Omega, \\
\widetilde{y}(t,x) = h(t,x) & \text{on } (0,T)\times\Gamma, \\
\text{transmission conditions } (\widetilde{\text{TC1}}) & \text{on } [0,T]\times S, \\
\widetilde{y}(0) = \widetilde{y}_0, &
\end{cases}
\tag{3.1}
$$

with

$$(\widetilde{\text{TC1}}) \qquad \widetilde{y}_{|[0,T]\times S_0} = \widetilde{y}_{|[0,T]\times S_1}, \quad \widetilde{c}_0\partial_n\widetilde{y}_{|[0,T]\times S_0} = \widetilde{c}_1\partial_n\widetilde{y}_{|[0,T]\times S_1},$$

and

$$\begin{cases} \partial_t y - \nabla\cdot(c\nabla y) = 0 & \text{in } (0,T)\times\Omega, \\ y(t,x) = h(t,x) & \text{on } (0,T)\times\Gamma, \\ \text{transmission conditions (TC1)} & \text{on } [0,T]\times S, \\ y(0) = y_0. \end{cases} \tag{3.2}$$

The Carleman estimate proved in the previous section will be the key ingredient in the proof of such a stability estimate.

We introduce

$$\xi = c - \widetilde{c} = \begin{cases} \xi_0 = c_0 - \widetilde{c}_0 & \text{in } \Omega_0, \\ \xi_1 = c_1 - \widetilde{c}_1 & \text{in } \Omega_1. \end{cases}$$

We set $u = y - \widetilde{y}$ and $v = \partial_t u$. Then v is solution to the following problem:

$$\begin{cases} \partial_t v - \nabla\cdot(c\nabla v) = \nabla\cdot(\xi\nabla\partial_t\widetilde{y}) & \text{in } (0,T)\times\Omega', \\ v = 0 & \text{on } (0,T)\times\Gamma, \\ \text{transmission conditions (TC2)} & \text{on } [0,T]\times S, \end{cases} \tag{3.3}$$

with

$$\text{(TC2)} \qquad \begin{cases} v_{|_{[0,T]\times S_0}} = v_{|_{[0,T]\times S_1}}, \\ c_0\partial_n v_{|_{[0,T]\times S_0}} = c_1\partial_n v_{|_{[0,T]\times S_1}} + g(t,x), \end{cases}$$

where

$$g(t,x) = \xi_1\partial_n\partial_t\widetilde{y}_{|_{[0,T]\times S_1}} - \xi_0\partial_n\partial_t\widetilde{y}_{|_{[0,T]\times S_0}} = \alpha\partial_n\partial_t\widetilde{y}_{|_{[0,T]\times S_0}},$$

with $\alpha = \xi_{1|_{S_1}}\frac{\widetilde{c}_{0|_{S_0}}}{\widetilde{c}_{1|_{S_1}}} - \xi_{0|_{S_0}}$.

Let $T' = \frac{1}{2}(T+t_0)$. We make the following assumption.

ASSUMPTION 3.1. *The solutions $\widetilde{y}$ and y belong to $H^2(t_0,T,H^1(\Omega))$ and are such that $y_{|\Omega_i} \in H^1(t_0,T,H^2(\Omega_i))$, $\widetilde{y}_{|\Omega_i} \in H^2(t_0,T,H^2(\Omega_i))$, $i=0,1$. Furthermore, $\widetilde{y}$ satisfies the following:*

1. *Let $r>0$. The solution $\widetilde{y}$ is such that $|\Delta\widetilde{y}(T')| \geq r > 0$ in Ω'.*
2. *$\widetilde{y}_{|\Omega_i}$ is in a bounded domain of $W^{2,\infty}(t_0,T,H^2(\Omega_i))$, $i=0,1$: there exists $M>0$ such that*

$$|\widetilde{y}_{|\Omega_i}(t,.)|^2_{H^2(\Omega_i)} + |\partial_t\widetilde{y}_{|\Omega_i}(t,.)|^2_{H^2(\Omega_i)} + |\partial_t^2\widetilde{y}_{|\Omega_i}(t,.)|^2_{H^2(\Omega_i)} \leq M, \quad i=0,1,$$

   a.e. for $t\in(t_0,T)$.
3. *$\Delta\partial_t\widetilde{y}_{|\Omega_i}$ is in a bounded domain of $L^2(t_0,T,L^\infty(\Omega_i))$, $i=0,1$: there exists $K>0$ such that*

$$\int_{t_0}^T |\Delta\partial_t\widetilde{y}_{|\Omega_i}(t,.)|^2_{L^\infty(\Omega_i)}dt \leq K^2, \qquad i=0,1.$$

In section 4 we shall show that for *any* initial conditions y_0, $\widetilde{y}_0$ in $L^2(\Omega)$ we can achieve the properties listed in Assumption 3.1 by using some particular boundary conditions $h(t,x)$.

From Assumption 3.1, the functions $\widetilde{y}$ and v are such that $\widetilde{y}_{|\Omega_i}, v_{|\Omega_i} \in H^2(\Omega_i)$, $i=0,1$. Then $g \in H^1(t_0,T,H^{\frac{1}{2}}(S))$. The second equality in transmission condition (TC2) thus takes place in the space $H^{\frac{1}{2}}(S)$. Observe that $v=\partial_t(y-\widetilde{y}) \in \aleph_g$ from the above assumption. We can thus apply Carleman estimate (2.5) to v.

We shall use the notation of the proof of Theorem 2.2. We set $\psi^{(i)}=e^{-s\eta^{(i)}}v$, $i=1,2$. With the operator $M_2^{(i)}$ defined in (2.10) we introduce, following [2],

$$I^{(i)}=\int_{t_0}^{T'}\!\!\int_{\Omega'} M_2^{(i)}\psi^{(i)}\ {\varphi^{(i)}}^{\frac{3}{2}}\psi^{(i)}\ dxdt, \quad i=1,2, \quad \text{and} \quad I=\frac{1}{2}(I^{(1)}+I^{(2)}).$$

Note the additional ${\varphi^{(i)}}^{\frac{3}{2}}$ factor as compared to [2]. This will be of importance below.

We have the following estimates.

LEMMA 3.2. *Let $\lambda \geq \lambda_1$ and $s \geq s_1$; then*

$$\begin{aligned}|I| \leq Cs^{-3/2}\lambda^{-2}\Bigg[& s\lambda\int_{t_0}^{T}\!\!\int_{\gamma}(e^{-2s\eta^{(1)}}\varphi^{(1)}+e^{-2s\eta^{(2)}}\varphi^{(2)})\ |\partial_n v|^2\ d\sigma\ dt \\ & + \iint_{Q'}(e^{-2s\eta^{(1)}}+e^{-2s\eta^{(2)}})\ |\partial_t v-\nabla\cdot(c\nabla v)|^2\ dx\ dt \\ & + s\lambda\int_{t_0}^{T}\!\!\int_{S}(e^{-2s\eta^{(1)}}\varphi^{(1)}+e^{-2s\eta^{(2)}}\varphi^{(2)})\ |g|^2\ d\sigma\ dt \\ & + \int_{t_0}^{T}\!\!\int_{S}(e^{-2s\eta^{(1)}}{\varphi^{(1)}}^4+e^{-2s\eta^{(2)}}{\varphi^{(2)}}^4)\ |g|^2\ d\sigma\ dt \\ & + s^{-2}\int_{t_0}^{T}\!\!\int_{S}(e^{-2s\eta^{(1)}}+e^{-2s\eta^{(2)}})|\partial_t g|^2 d\sigma\ dt\Bigg].\end{aligned}$$

Proof. Observe that

$$|I^{(i)}| \leq \frac{1}{2}s^{-3/2}\lambda^{-2}\left(|M_2^{(i)}\psi^{(i)}|^2_{L^2(Q')}+s^3\lambda^4\iint_Q {\varphi^{(i)}}^3 e^{-2s\eta^{(i)}}|v|^2dxdt\right), \qquad i=1,2.$$

(We recall that $Q'=(t_0,T)\times\Omega'$.) Thus

$$\begin{aligned}|I| \leq \frac{1}{2}s^{-3/2}\lambda^{-2}\Bigg(&|M_2^{(1)}\psi^{(1)}|^2_{L^2(Q')}+|M_2^{(2)}\psi^{(2)}|^2_{L^2(Q')} \\ &+s^3\lambda^4\iint_Q(e^{-2s\eta^{(1)}}{\varphi^{(1)}}^3+e^{-2s\eta^{(2)}}{\varphi^{(2)}}^3)\ |v|^2\ dx\ dt\Bigg),\end{aligned}$$

which yields the result from Carleman estimate (2.5). ∎

LEMMA 3.3. *Let $\lambda \geq \lambda_1$ and $s \geq s_1$; then*

$$\begin{aligned}
&\int_{\Omega'} \left(e^{-2s\eta^{(1)}} {\varphi^{(1)}}^{\frac{3}{2}} + e^{-2s\eta^{(2)}} {\varphi^{(2)}}^{\frac{3}{2}} \right) (T',x)\ |v(T',.)|^2\ dx \\
&\quad \leq Cs^{-3/2}\lambda^{-2} \left[s\lambda \int_{t_0}^{T}\!\!\int_{\gamma} (e^{-2s\eta^{(1)}} \varphi^{(1)} + e^{-2s\eta^{(2)}} \varphi^{(2)})\ |\partial_n v|^2\ d\sigma\ dt \right. \\
&\qquad + \iint_{Q'} (e^{-2s\eta^{(1)}} + e^{-2s\eta^{(2)}})\ |\nabla \cdot (\xi \nabla \partial_t \widetilde{y})|^2\ dx\ dt \\
&\qquad + s\lambda \int_{t_0}^{T}\!\!\int_{S} (e^{-2s\eta^{(1)}} \varphi^{(1)} + e^{-2s\eta^{(2)}} \varphi^{(2)})\ |g|^2\ d\sigma\ dt \\
&\qquad + \int_{t_0}^{T}\!\!\int_{S} (e^{-2s\eta^{(1)}} {\varphi^{(1)}}^4 + e^{-2s\eta^{(2)}} {\varphi^{(2)}}^4)\ |g|^2\ d\sigma\ dt \\
&\qquad \left. + s^{-2} \int_{t_0}^{T}\!\!\int_{S} (e^{-2s\eta^{(1)}} + e^{-2s\eta^{(2)}}) |\partial_t g|^2 d\sigma\ dt \right].
\end{aligned}$$

Proof. We evaluate integral $I^{(i)}$, $i = 1, 2$, using (2.10),

$$\begin{aligned}
I^{(i)} &= \int_{t_0}^{T'}\!\!\int_{\Omega'} \left(\partial_t \psi^{(i)} - 2s\lambda \varphi^{(i)} c \nabla \beta^{(i)} \cdot \nabla \psi^{(i)} - 2s\lambda^2 \varphi^{(i)} c |\nabla \beta^{(i)}|^2 \psi^{(i)} \right) {\varphi^{(i)}}^{\frac{3}{2}} \psi^{(i)}\, dxdt \\
&= \frac{1}{2} \int_{t_0}^{T'}\!\!\int_{\Omega'} {\varphi^{(i)}}^{\frac{3}{2}} \partial_t |\psi^{(i)}|^2\ dxdt - s\lambda \int_{t_0}^{T'}\!\!\int_{\Omega'} {\varphi^{(i)}}^{\frac{5}{2}} c \nabla \beta^{(i)} \cdot \nabla |\psi^{(i)}|^2\ dxdt \\
&\qquad - 2s\lambda^2 \int_{t_0}^{T'}\!\!\int_{\Omega'} {\varphi^{(i)}}^{\frac{5}{2}} c |\nabla \beta^{(i)}|^2 |\psi^{(i)}|^2 dxdt \\
&= \frac{1}{2} \int_{t_0}^{T'}\!\!\int_{\Omega'} {\varphi^{(i)}}^{\frac{3}{2}} \partial_t |\psi^{(i)}|^2\ dxdt + s\lambda \int_{t_0}^{T'}\!\!\int_{\Omega'} \nabla \cdot ({\varphi^{(i)}}^{\frac{5}{2}} c \nabla \beta^{(i)}) |\psi^{(i)}|^2\ dxdt \\
&\qquad - 2s\lambda^2 \int_{t_0}^{T'}\!\!\int_{\Omega'} {\varphi^{(i)}}^{\frac{5}{2}} c |\nabla \beta^{(i)}|^2 |\psi^{(i)}|^2 dxdt,
\end{aligned}$$

by integration by parts, without any remaining integral over $(t_0, T') \times S$ by condition transmission (2.8). With an integration by parts w.r.t. t in the first integral, we then obtain

$$\begin{aligned}
&\frac{1}{2} \int_{\Omega'} {\varphi^{(i)}}^{\frac{3}{2}} |\psi^{(i)}(T',.)|^2\ dx = I^{(i)} + \frac{1}{2} s\lambda^2 \int_{t_0}^{T'}\!\!\int_{\Omega'} {\varphi^{(i)}}^{\frac{5}{2}} c |\nabla \beta^{(i)}|^2 |\psi^{(i)}|^2 dxdt \\
&\quad - s\lambda \int_{t_0}^{T'}\!\!\int_{\Omega'} {\varphi^{(i)}}^{\frac{5}{2}} \nabla \cdot (c \nabla \beta^{(i)}) |\psi^{(i)}|^2\ dxdt + \frac{3}{4} \int_{t_0}^{T'}\!\!\int_{\Omega'} (\partial_t \varphi^{(i)}) {\varphi^{(i)}}^{\frac{1}{2}} |\psi^{(i)}|^2\ dxdt, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, 2,
\end{aligned} \tag{3.4}$$

since ${\varphi^{(i)}}^{\frac{3}{2}} \psi^{(i)}(t_0, .) = 0$. Adding (3.4) for $i = 1, 2$, we obtain

$$\begin{aligned}
&\int_{\Omega'} \left(e^{-2s\eta^{(1)}} {\varphi^{(1)}}^{\frac{3}{2}} + e^{-2s\eta^{(2)}} {\varphi^{(2)}}^{\frac{3}{2}} \right) (T',x)\ |v(T',.)|^2\ dx \\
&\leq 4|I| + C(s\lambda^2 + s\lambda + 1) \int_{t_0}^{T'}\!\!\int_{\Omega'} \left(e^{-2s\eta^{(1)}(t,x)} {\varphi^{(1)}}^{\frac{5}{2}} + e^{-2s\eta^{(2)}(t,x)} {\varphi^{(2)}}^{\frac{5}{2}} \right) |v|^2 dxdt,
\end{aligned} \tag{3.5}$$

observing that $|\partial_t \varphi^{(i)}| \le CT{\varphi^{(i)}}^2$, $i = 1, 2$. We use Carleman estimate (2.5) to obtain an upper bound for the last term in (3.5), which yields the result by Lemma 3.2. ∎

We shall now assume the following.

ASSUMPTION 3.4. *The diffusion coefficients c and $\widetilde{c}$ are piecewise-constant, in the sense that $c_{|\Omega_i}$ (resp., $\widetilde{c}_{|\Omega_i}$) are constant in each connected component of Ω_i, $i = 0, 1$. We define*

$$c_{0,j} = c_{|\Omega_{0,j}},\ j = 1, \dots, p_0, \quad c_{1,j} = c_{|\Omega_{1,j}},\ j = 1, \dots, p_1,$$

with similar notation for $\widetilde{c}$ and ξ.

In this case observe that, in Ω',

$$\begin{aligned} v(T', x) = c\Delta u(T', x) + \xi\Delta\widetilde{y}(T', x) &= \sum_{j=1}^{p_0} c_{0,j}\Delta u(T', x)\chi_{\Omega_{0,j}} + \sum_{j=1}^{p_1} c_{1,j}\Delta u(T', x)\chi_{\Omega_{1,j}} \\ &\quad + \sum_{j=1}^{p_0} \xi_{0,j}\Delta\widetilde{y}(T', x)\chi_{\Omega_{0,j}} + \sum_{j=1}^{p_1} \xi_{1,j}\Delta\widetilde{y}(T', x)\chi_{\Omega_{1,j}}, \end{aligned}$$

from the equation satisfied by u expressed at time T' and the definition of v above.

From Lemma 3.3, we obtain

$$\begin{aligned} &\int_{\Omega'} \left(e^{-2s\eta^{(1)}} {\varphi^{(1)}}^{\frac{3}{2}} + e^{-2s\eta^{(2)}} {\varphi^{(2)}}^{\frac{3}{2}} \right) (T', x) |\xi\Delta\widetilde{y}(T', x)|^2 dx \\ &\quad \le C \int_{\Omega'} \left(e^{-2s\eta^{(1)}} {\varphi^{(1)}}^{\frac{3}{2}} + e^{-2s\eta^{(2)}} {\varphi^{(2)}}^{\frac{3}{2}} \right) (T', x) |c\Delta u(T', x)|^2 dx \\ &\qquad + Cs^{-3/2}\lambda^{-2} \Bigg[s\lambda \int_{t_0}^{T}\int_{\gamma} (e^{-2s\eta^{(1)}}\varphi^{(1)} + e^{-2s\eta^{(2)}}\varphi^{(2)})\ |\partial_n v|^2\ d\sigma\ dt \\ &\qquad\qquad + \iint_{Q'} (e^{-2s\eta^{(1)}} + e^{-2s\eta^{(2)}})\ |\nabla\cdot(\xi\nabla\partial_t\widetilde{y})|^2\ dx\ dt \\ &\qquad\qquad + s\lambda \int_{t_0}^{T}\int_{S} (e^{-2s\eta^{(1)}}\varphi^{(1)} + e^{-2s\eta^{(2)}}\varphi^{(2)})\ |g|^2\ d\sigma\ dt \\ &\qquad\qquad + \int_{t_0}^{T}\int_{S} (e^{-2s\eta^{(1)}}{\varphi^{(1)}}^4 + e^{-2s\eta^{(2)}}{\varphi^{(2)}}^4)\ |g|^2\ d\sigma\ dt \\ &\qquad\qquad + s^{-2} \int_{t_0}^{T}\int_{S} (e^{-2s\eta^{(1)}} + e^{-2s\eta^{(2)}}) |\partial_t g|^2 d\sigma\ dt \Bigg]. \end{aligned} \tag{3.6}$$

From Assumption 3.1 we find that

$$|\Delta\partial_t\widetilde{y}(t, x)|^2 \le k^2(t)|\Delta\widetilde{y}(T', x)|^2 \quad \text{in each } (t_0, T)\times\Omega_i,\ i = 0, 1,$$

for

$$k(t) = \frac{1}{r}\sup_{i=0,1} |\Delta\partial_t\widetilde{y}_{|\Omega_i}(t, .)|_{L^\infty(\Omega_i)}.$$

From Assumption 3.1, $k \in L^2(t_0, T)$ and $|k|_{L^2(t_0,T)} \le K' = \frac{1}{r}K$. These observations yield the following estimation of the third term in the r.h.s. of (3.6):

$$\begin{aligned} &\int_{t_0}^{T}\int_{\Omega_i} (e^{-2s\eta^{(1)}} + e^{-2s\eta^{(2)}})\ |\xi\Delta\partial_t\widetilde{y}|^2\ dx\ dt \\ &\quad \le K'^2 \sum_{i=0,1}\sum_{j=1}^{p_i} |\xi_{i,j}|^2 \int_{\Omega_{i,j}} \left(e^{-2s\eta^{(1)}(T',x)} + e^{-2s\eta^{(2)}(T',x)} \right) |\Delta\widetilde{y}(T', x)|^2 dx, \end{aligned} \tag{3.7}$$

where we have used that

$$e^{-2s\eta^{(i)}(t,x)} \leq e^{-2s\eta^{(i)}(T',x)}, \quad x \in \Omega,\ t \in (t_0,T),\ i=1,2.$$

Observing that $0 < C \leq \varphi^{(i)}$, $i=1,2$, since $\beta^{(i)} \geq 0$ and $\frac{1}{(T-t)(t-t_0)} \geq C > 0$, we obtain

$$\int_{t_0}^{T}\int_{\Omega_i} (e^{-2s\eta^{(1)}} + e^{-2s\eta^{(2)}})\ |\xi\Delta\partial_t\widetilde{y}|^2\ dx\ dt$$
$$\leq K'^2 \sum_{i=0,1}\sum_{j=1}^{p_i} |\xi_{i,j}|^2 \int_{\Omega_{i,j}} \left(e^{-2s\eta^{(1)}}\varphi^{(1)\,\frac{3}{2}} + e^{-2s\eta^{(2)}(T',x)}\varphi^{(2)\,\frac{3}{2}}\right)(T',x)|\Delta\widetilde{y}(T',x)|^2 dx.$$

We now treat the interface terms that appear in the r.h.s. of the Carleman estimate. Recall that

$$g(t,x) = \alpha\partial_n\partial_t\widetilde{y}_{|[0,T]\times S_0}, \quad \alpha = \xi_1\frac{\widetilde{c}_0}{\widetilde{c}_1} - \xi_0.$$

Note that $\eta(t,.)$ is constant on S. We denote this constant by $\eta(t,S)$. More generally we shall denote by $\rho(S)$ the value on S of a function ρ which is constant on S. We obtain

$$\int_{t_0}^{T}\int_S (e^{-2s\eta^{(1)}} + e^{-2s\eta^{(2)}})|\partial_t g|^2 d\sigma\ dt \leq \int_{t_0}^{T} (e^{-2s\eta^{(1)}(t,S)} + e^{-2s\eta^{(2)}(t,S)}) \int_S |\partial_t g|^2 d\sigma\ dt$$
$$\leq M'|\xi|^2 \int_{t_0}^{T} (e^{-2s\eta^{(1)}(t,S)} + e^{-2s\eta^{(2)}(t,S)})dt,$$

from trace inequalities and from Assumption 3.1, for $M' = C_{Tr}(1+\frac{c_{max}}{c_{min}})^2 M$, where $|\xi| = \sqrt{\xi_0^2+\xi_1^2}$, since $|\alpha| \leq (1+\frac{c_{max}}{c_{min}})|\xi|$ from Assumption 1.2. The constant C_{Tr} is the constant found in the trace estimates

$$\int_S |\partial_n\widetilde{\rho}|^2 \leq C_{Tr}|\rho|^2_{H^2(\Omega_i)}, \quad i=0,1,$$

if $\rho_{|\Omega_i} \in H^2(\Omega_i)$, $i=0,1$. Similarly, since $\varphi^{(i)}$, $i=1,2$, are constant on S, we have

$$\int_{t_0}^{T}\int_S (e^{-2s\eta^{(1)}}\varphi^{(1)^j} + e^{-2s\eta^{(2)}}\varphi^{(2)^j})\ |g|^2\ d\sigma\ dt$$
$$\leq M'|\xi|^2 \int_{t_0}^{T} (e^{-2s\eta^{(1)}}\varphi^{(1)^j} + e^{-2s\eta^{(2)}}\varphi^{(2)^j})(t,S)\ dt, \quad j \in \mathbb{N}.$$

With

$$w_k(s,\lambda) := \int_{t_0}^{T} (e^{-2s\eta^{(1)}}\varphi^{(1)^k} + e^{-2s\eta^{(2)}}\varphi^{(2)^k})(t,S)\ dt, \quad k \in \mathbb{N},$$

and

$$W_{i,j}(s,\lambda) := \int_{\Omega_{i,j}} (e^{-2s\eta^{(1)}}\varphi^{(1)\,\frac{3}{2}} + e^{-2s\eta^{(2)}}\varphi^{(2)\,\frac{3}{2}})(T',x)|\Delta\widetilde{y}(T',x)|^2 dx,$$

$i = 0, 1$, $j = 1, \dots, p_i$, we thus obtain, for $\lambda \geq \lambda_1$ and $s \geq s_1$,

$$\begin{aligned}
&\sum_{i=0,1}\sum_{j=1}^{p_i} |\xi_{i,j}|^2 \Big\{ (1 - CK'^2 s^{-\frac{3}{2}}\lambda^{-2}) W_{i,j}(s,\lambda) \\
&\qquad\qquad - CM'[s^{-\frac{1}{2}}\lambda^{-1} w_1(s,\lambda) + s^{-\frac{3}{2}}\lambda^{-2} w_4(s,\lambda) + s^{-\frac{7}{2}}\lambda^{-2} w_0(s,\lambda)] \Big\} \\
&\quad \leq C \int_{\Omega'} \left(e^{-2s\eta^{(1)}} {\varphi^{(1)}}^{\frac{3}{2}} + e^{-2s\eta^{(2)}} {\varphi^{(2)}}^{\frac{3}{2}} \right) (T', x) |c\Delta u(T', .)|^2 dx \\
&\quad + C s^{-\frac{1}{2}} \lambda^{-1} \int_{t_0}^{T}\!\!\int_{\gamma} (e^{-2s\eta^{(1)}} \varphi^{(1)} + e^{-2s\eta^{(2)}} \varphi^{(2)}) \, |\partial_n v|^2 \, d\sigma \, dt. \qquad (3.8)
\end{aligned}$$

To obtain a stability result we need to prove that the coefficients for $|\xi_{i,j}|^2$, $i = 0, 1$, $j = 1, \dots, p_i$, can be made positive. To do so we need to understand the behavior of the integrals $w_k(s,\lambda)$ and $W_{i,j}(s,\lambda)$ as s and λ become large.

We first establish the asymptotic behavior of $w_k(s,\lambda)$. We set

$$w_k^{(i)}(s,\lambda) := \int_{t_0}^{T} e^{-2s\eta^{(i)}(t,S)} {\varphi^{(i)}}^k(t,S) \, dt, \quad k \in \mathbb{N}.$$

LEMMA 3.5. *The following estimates holds:*

$$\begin{aligned}
w_k^{(i)}(s,\lambda) = e^{-2s\eta^{(i)}(T',S)} \, {\varphi^{(i)}}^k(T',S) \Bigg\{ &\frac{\sqrt{\pi} s^{-\frac{1}{2}}}{\sqrt{\phi''(T')}\sqrt{e^{\lambda\overline{\beta}} - e^{\lambda\beta^{(i)}(S)}}} \\
&+ \mathcal{O}\left(\frac{s^{-\frac{3}{2}}}{(e^{\lambda\overline{\beta}} - e^{\lambda\beta^{(i)}(S)})^{\frac{3}{2}}} \right) \Bigg\}, \qquad (3.9)
\end{aligned}$$

with $\phi(t) = \frac{1}{(T-t)(t-t_0)}$, *for* $s > s_2 > 0$ *and* $\lambda > \lambda_2 > 0$.

Proof. Let $T^{(2)}$ and $T^{(3)}$ be such that $t_0 < T^{(2)} < T' < T^{(3)} < T$. We choose $\chi_1 \in \mathscr{C}_c^\infty([t_0, T^{(2)}))$, $\chi_2 \in \mathscr{C}_c^\infty((t_0, T))$, and $\chi_3 \in \mathscr{C}_c^\infty((T^{(3)}, T])$, all three nonnegative, such that $\chi_1 + \chi_2 + \chi_3 = 1$ and $\chi_1 = 1$ in a neighborhood of t_0, $\chi_2 = 1$ in a neighborhood of T', and $\chi_3 = 1$ in a neighborhood of T. With this partition of unity we break $w_k^{(i)}$ into three pieces: $w_k^{(i)} = w_k^{(i,1)} + w_k^{(i,2)} + w_k^{(i,3)}$ with

$$w_k^{(i,j)}(s,\lambda) := \int_{t_0}^{T} e^{-2s\eta^{(i)}(t,S)} {\varphi^{(i)}}^k(t,S) \, \chi_j(t) \, dt, \quad j = 1, 2, 3.$$

The first and the third term are treated similarly. Let $s_2 > 0$ and $\lambda_2 > 0$. We set $\tau(s,\lambda,S) = s(e^{\lambda\overline{\beta}} - e^{\lambda\beta^{(i)}(S)})$. We observe

$$\begin{aligned}
w_k^{(i,1)}(s,\lambda) &= e^{k\lambda\beta^{(i)}(S)} \int_{t_0}^{T} e^{-2s\eta^{(i)}(t,S)} \phi^k(t) \, \chi_1(t) \, dt \\
&\leq e^{k\lambda\beta^{(i)}(S)} e^{-2(s-s_2)\eta^{(i)}(T^{(2)},S)} \int_{t_0}^{T} e^{-2s_2\eta^{(i)}(t,S)} \phi^k(t) \, \chi_1(t) \, dt \\
&\leq C(s_2,\lambda_2) e^{k\lambda\beta^{(i)}(S)} e^{-2(s-s_2)\eta^{(i)}(T^{(2)},S)} \\
&\leq C e^{k\lambda\beta^{(i)}(S)} e^{-2s\eta^{(i)}(T',S)} e^{-2s(\eta^{(i)}(T^{(2)},S) - \eta^{(i)}(T',S))} \\
&= e^{k\lambda\beta^{(i)}(S)} e^{-2s\eta^{(i)}(T',S)} \mathcal{O}(\tau(s,\lambda,S)^{-l}),
\end{aligned}$$

for all $l \in \mathbb{N}$, if $s > s_2$ and $\lambda > \lambda_2$, since

$$s(\eta^{(i)}(T^{(2)}, S) - \eta^{(i)}(T', S)) = \tau(s, \lambda, S)(\phi(T^{(2)}) - \phi(T')),$$

and $\phi(T^{(2)}) - \phi(T') > 0$.

For the second term $w_k^{(i,2)}$ we write

$$\begin{aligned} w_k^{(i,2)}(s,\lambda) &:= e^{k\lambda\beta^{(i)}(S)} \int_{t_0}^{T} e^{-2\tau(s,\lambda,S)\phi(t)} \phi^k(t)\ \chi_2(t)\ dt \\ &= e^{k\lambda\beta^{(i)}(S)} e^{-2\tau(s,\lambda,S)\phi(T')} \int_{t_0}^{T} e^{-2\tau(s,\lambda,S)(\phi(t)-\phi(T'))} \phi^k(t)\ \chi_2(t)\ dt. \end{aligned}$$

We then apply the following stationary phase formula [15, Theorem 7.7.5] in one dimension:

$$\begin{aligned} &\left| \int u(y) e^{i\omega f(y)} dy - e^{i\omega f(y_0)} \left(\frac{\omega f''(y_0)}{2\pi i} \right)^{-\frac{1}{2}} \sum_{l<l_1} \omega^{-l} L_l u \right| \\ &\qquad \le C\omega^{-l_1-\frac{1}{2}} \sum_{|\alpha|\le 2l_1} \sup |\partial^\alpha u|, \qquad \omega > 0,\ u \in \mathscr{C}_c^\infty, \end{aligned}$$

where L_l is a differential operator of order $2l$ evaluated at T', $L_0 u = u(T')$. (To obtain the additional factor $-\frac{1}{2}$ in the r.h.s. as compared to the formula given in [15, Theorem 7.7.5], simply write the formula to the order $l_1 + 1$.) This formula is valid if $\mathrm{Im}(f) \ge 0$, $\mathrm{Im}(f(y_0)) = 0$, $f'(y_0) = 0$, $f''(y_0) \ne 0$, and $f'(y) \ne 0$ in $\mathrm{supp}(u) \setminus \{y_0\}$. Here the phase function $f(t) = i(\phi(t) - \phi(T'))$ is imaginary (note that $\phi''(T') > 0$), $\omega = 2\tau(s, \lambda, S)$, and $u = \phi^k \chi_2$. The stationary point is $t = T'$. With $l_1 = 1$, we obtain

$$\begin{aligned} &\left| \int_{t_0}^{T} e^{-2\tau(s,S)(\phi(t)-\phi(T'))} \phi^k(t)\ \chi_2(t)\ dt - \phi^k(T') \sqrt{\frac{\pi}{\phi''(T')}} \tau(s,\lambda,S)^{-\frac{1}{2}} \right| \\ &\qquad \le C\tau(s,\lambda,S)^{-3/2}. \end{aligned}$$

This yields (3.9). ∎

To achieve our goal we also need an estimation from below for the terms $W_{i,j}(s,\lambda)$. We set

$$W_{k,j}^{(i)}(s,\lambda) := \int_{\Omega_{k,j}} e^{-2s\eta^{(i)}(T',x)} \varphi^{(i)\,\frac{3}{2}}(T', x)\ |\Delta\widetilde{y}(T', x)|^2 dx,$$

$k = 0, 1$, $j = 1, \ldots, p_k$, $i = 1, 2$. For the terms $W_{0,j}^{(i)}(s,\lambda)$, we have the following.

LEMMA 3.6. *Let $\varepsilon > 0$. We have*

$$W_{0,j}^{(i)}(s,\lambda) \ge C_{s_2,i,j} \frac{r^2 |S_j|}{s\lambda} e^{-2s\eta^{(i)}(T',S)} (\varphi^{(i)}(T', S))^{\frac{1}{2}} e^{-\lambda\varepsilon}, \quad i = 1, 2,\ j = 1, \ldots, p_0,$$

for $s \ge s_2 > 0$, $\lambda > 0$, and where $S_j = \bigcup_{k=1,\ldots,p_1} S_{jk}$.

Proof. In the proof, we shall write β in place of $\beta^{(i)}$, etc. Taking δ sufficiently small, we start by choosing a small neighborhood W of S_j in $\overline{\Omega}_0$ *globally* parametrized by $(\sigma, y) \in [0, \delta] \times S_j$ (see the proof of Lemma A.6 in the appendix). In fact, we

can choose the coordinates and the small neighborhood of S_j such that $\sigma = cst$ corresponds to level sets for the function β (use $\nabla\beta$ for the vector field v in the proof of Lemma A.6). Note that in the neighborhood W the function β decreases with σ.

Estimating from below the Jacobian[1] originating from the change of variables and observing that the integrand is constant w.r.t. y, we obtain

$$\begin{aligned} W_{0,j}(s,\lambda) &\geq Cr^2|S_j| \int_0^\delta e^{-2s\eta(T',\sigma)}\varphi^{\frac{3}{2}}(T',\sigma)d\sigma \\ &= Cr^2|S_j|e^{-2s\eta(T',S)} \int_0^\delta e^{-2s(\eta(T',\sigma)-\eta(T',S))}\varphi^{\frac{3}{2}}(T',\sigma)d\sigma. \end{aligned}$$

We now use the change of variables $\sigma' = \eta(T',\sigma) - \eta(T',S) \geq 0$, which yields

$$W_{0,j}(s,\lambda) \geq Cr^2|S_j|\lambda^{-1}e^{-2s\eta(T',S)} \int_0^{\delta'} e^{-2s\sigma'}\varphi^{\frac{1}{2}}(T',\sigma)|\partial_\sigma\beta|^{-1}d\sigma',$$

where $\delta' = \eta(T',\delta) - \eta(T',S)$. We can find in W a positive lower bound for $(\partial_\sigma\beta)^{-1}$ independent of δ, i.e., the size of W. We thus obtain

$$\begin{aligned} W_{0,j}(s,\lambda) &\geq Cr^2|S_j|\lambda^{-1}e^{-2s\eta(T',S)}\varphi^{\frac{1}{2}}(T',\delta) \int_0^{\delta'} e^{-2s\sigma'}d\sigma' \\ &\geq Cr^2|S_j|s^{-1}\lambda^{-1}e^{-2s\eta(T',S)}\varphi^{\frac{1}{2}}(T',\delta) \int_0^{s\delta'} e^{-2\sigma'}d\sigma' \\ &\geq C'(s)r^2|S_j|s^{-1}\lambda^{-1}e^{-2s\eta(T',S)}\varphi^{\frac{1}{2}}(T',\delta), \end{aligned}$$

with $C'(s)$ increasing with s. Observe now that

$$\begin{aligned} \varphi(T',\delta) &= \frac{e^{\lambda\beta(\delta)}}{(T-T')(T'-t_0)} = \frac{e^{\lambda\beta(S)}}{(T-T')(T'-t_0)}e^{\lambda(\beta(\delta)-\beta(S))} \\ &= \varphi(T',S)e^{\lambda(\beta(\delta)-\beta(S))}. \end{aligned}$$

Choosing δ sufficiently small such that $\frac{1}{2}(\beta(S)-\beta(\delta)) \leq \varepsilon$ thus yields the desired result. □

With the previous lemmas we can now prove that the coefficient of $|\xi_{0,j}|^2$, $j = 1,\ldots,p_0$, in (3.8) can be made positive. This requires taking *both* λ and s sufficiently large.

PROPOSITION 3.7. *Let $1 \leq j \leq p_0$. There exists $\lambda_{2,j} \geq \lambda_1$ such that if $\lambda \geq \lambda_{2,j}$, then for s sufficiently large*

$$\begin{aligned} A_{0,j} &= (1 - CK'^2 s^{-\frac{3}{2}}\lambda^{-2})W_{0,j}(s,\lambda) \\ &\quad - CM'[s^{-\frac{1}{2}}\lambda^{-1}w_1(s,\lambda) + s^{-\frac{3}{2}}\lambda^{-2}w_4(s,\lambda) + s^{-\frac{7}{2}}\lambda^{-2}w_0(s,\lambda)] \geq C(s,\lambda) > 0, \end{aligned}$$

with $C(s,\lambda) = C(s,\lambda,r,K,M,c_{max},c_{min},j)$.

Proof. It suffices to prove the result for $w_k^{(i)}(s,\lambda)$ and $W_{0,j}^{(i)}(s,\lambda)$. We shall write β in place of $\beta^{(i)}$, etc. We take s sufficiently large such that $(1-CK'^2 s^{-\frac{3}{2}}\lambda^{-2}) \geq c_0 > 0$.

[1]Note that the estimation from below of the Jacobian is independent from the size of the neighborhood W.

From Lemmas 3.9 and 3.6, for $\varepsilon > 0$ we obtain, for s sufficiently large,

$$A_{0,j} \geq -s^{-2}\nu(\lambda)C_1 e^{-2s\eta(T',S)} + \frac{1}{s\lambda}e^{-2s\eta(T',S)}\left[\frac{C_0}{2}C_\varepsilon r^2|S_j|(\varphi(T',S))^{\frac{1}{2}}e^{-\lambda\varepsilon} - CM'\frac{\sqrt{\pi}\varphi(T',S)}{\sqrt{\phi''(T')}\sqrt{e^{\lambda\overline{\beta}}-e^{\lambda\beta(S)}}}\right], \tag{3.10}$$

where ν is a bounded function. We first treat the second term in the previous expression. Note that this term originates from the estimate from below for $W_{0,j}^{(i)}$ and the estimate of $s^{-\frac{1}{2}}\lambda^{-1}w_1^{(i)}$ by Lemma 3.5. The other terms in $A_{0,j}$ and the remaining part of the estimation of $s^{-\frac{1}{2}}\lambda^{-1}w_1^{(i)}$ are lumped into the first term of (3.10).

Now choose $\varepsilon < \frac{1}{2}(\overline{\beta} - \beta(S))$ (recall that $\overline{\beta} > \beta(S)$ because $m > 1$). Then since $\overline{\beta} > \beta(S)$ we have

$$\frac{\varphi(T',S)}{\sqrt{e^{\lambda\overline{\beta}}-e^{\lambda\beta(S)}}} = o((\varphi(T',S))^{\frac{1}{2}}e^{-\lambda\varepsilon})$$

for λ large. Thus the second term can be made positive for λ, say $\lambda = \lambda_{2,j}$, sufficiently large.

Once λ is fixed larger than $\lambda_{2,j}$, the first term in (3.10) can be made positive by taking s sufficiently large. □

We now prove that the coefficient of $|\xi_{1,j}|^2$, $j = 1, \ldots, p_1$, in (3.8) can be made positive. Here, the parameter λ is not of use.

PROPOSITION 3.8. *Let $1 \leq j \leq p_1$. Let $\lambda \geq \lambda_1$. Then for s sufficiently large*

$$\begin{aligned} A_{1,j} &= (1 - CK'^2 s^{-\frac{3}{2}}\lambda^{-2})W_{1,j}(s,\lambda) \\ &\quad - CM'[s^{-\frac{1}{2}}\lambda^{-1}w_1(s,\lambda) + s^{-\frac{3}{2}}\lambda^{-2}w_4(s,\lambda) + s^{-\frac{7}{2}}\lambda^{-2}w_0(s,\lambda)] \geq C(s,\lambda) > 0, \end{aligned}$$

with $C(s,\lambda) = C(s,\lambda,r,K,M,c_{max},c_{min})$.

Proof. It suffices to prove the result for $w_k^{(i)}(s,\lambda)$ and $W_{1,j}^{(i)}(s,\lambda)$. We shall write β in place of $\beta^{(i)}$, etc. We take s sufficiently large such that $(1 - CK'^2 s^{-\frac{3}{2}}\lambda^{-2}) \geq C_0 > 0$.

We first write

$$e^{-2s\eta(t,S)} = e^{-2s\eta(T',S)}e^{-2s(\eta(t,S)-\eta(T',S))}$$

and observe that for $s \geq s_0 > 0$

$$\int_{t_0}^{T} e^{-2s(\eta(t,S)-\eta(T',S))}\varphi^k(t,S)dt \leq L(s_0,\lambda,k),$$

for some positive $L(s_0,\lambda,k)$. From Lemma 2.1, there exists $V \Subset \Omega_{1,j}$ such that $\inf_{x\in V}\beta > \beta(S)$. Then with

$$\eta_{max}^{T',V} = \sup_{x\in V}\frac{e^{2\lambda K^{(i)}} - e^{\lambda\beta^{(i)}(x)}}{(T'-t_0)(T-T')}$$

we have $-\eta(T',x) \geq -\eta_{max}^{T',V} > -\eta(T',S)$, for $x \in V$ and $s > 0$. These observations yield

$$W_{1,j}(s) \geq r^2\int_V e^{-2s\eta(T',x)}\varphi^{\frac{3}{2}}(T',x)dx \geq C(\lambda)r^2|V|e^{-2s\eta_{max}^{T',V}}$$

and

$$w_k(s) \leq L(s_0, \lambda, k) e^{-2s\eta(T',S)},$$

which implies the result. □

With (3.8) and Propositions 3.7 and 3.8, recalling that $v = u_t = \partial_t(y - \widetilde{y})$, we have thus obtained the following stability result.

THEOREM 3.9. *Let γ be a subset of the boundary Γ of an open set Ω of $\mathbb{R}^n$ that satisfies Assumption* 1.5*, and let γ satisfy Assumption* 1.4*. We assume that the diffusion coefficients c and $\widetilde{c}$ satisfy Assumptions* 1.2 *and* 3.4 *and $c_0 - c_1 \geq \Delta > 0$. Let y_0, $\widetilde{y}_0$ in $L^2(\Omega)$ and let y, $\widetilde{y}$ be solutions to* (3.1)–(3.2) *satisfying Assumption* 3.1*. Then there exists a constant C,*

$$C = C(\Omega, T, t_0, \gamma, S, \mathcal{O}^{(1)}, \mathcal{O}^{(2)}, M, K, r, c_{min}, c_{max}, \Delta),$$

such that

(3.11)
$$\sum_{i=0,1} \sum_{j=1}^{p_i} |c_{ij} - \widetilde{c}_{ij}|^2 \leq C|\partial_n(\partial_t y - \partial_t \widetilde{y})|^2_{L^2((0,T)\times\gamma)} + C|\Delta y(T',.) - \Delta\widetilde{y}(T',.)|^2_{L^2(\Omega')}.$$

We shall see in Proposition 4.5 below that we can achieve the regularity properties and estimates of Assumption 3.1.

Remark 3.10. Observe than in the statement of Theorem 3.9 the initial condition y_0 and $\widetilde{y}_0$ need not be equal (see systems (3.1)–(3.2)).

Remark 3.11. If the position of the interface S is known, we can improve the result of Theorem 3.9 by locally observing the solutions y and $\widetilde{y}$ at time T'. Let $\omega = \omega_0 \cup \omega_1$, with ω_1 a neighborhood of γ in $\overline{\Omega}_1$ and ω_0 a neighborhood of S in $\overline{\Omega}_0$. We can, in fact, relax Assumption 3.1 with the following: Let $r > 0$; then the solution $\widetilde{y}$ is such that $|\Delta\widetilde{y}(T')| \geq r > 0$ in ω. Then, we obtain a stability estimate by solely observing y and $\widetilde{y}$ at time T' on ω in place of Ω:

(3.12)
$$\sum_{i=0,1} \sum_{j=1}^{p_i} |c_{ij} - \widetilde{c}_{ij}|^2 \leq C|\partial_n(\partial_t y - \partial_t \widetilde{y})|^2_{L^2((0,T)\times\gamma)} + C|\Delta y(T',.) - \Delta\widetilde{y}(T',.)|^2_{L^2(\omega)}.$$

We briefly sketch the proof, as it closely follows the exposition of the proof of estimate (3.11) in this section. It relies on the following lemma.

LEMMA 3.12. *Let $s_3 > 0$ and $\lambda_3 > 0$. There exists $C > 0$ such that for all $s \geq s_3$ and $\lambda \geq \lambda_3$ we have*

$$\int_{\Omega_i} \left(e^{-2s\eta^{(1)}} {\varphi^{(1)}}^{\frac{3}{2}} + e^{-2s\eta^{(2)}(T',x)} {\varphi^{(2)}}^{\frac{3}{2}} \right)(T',x)dx$$
$$\leq C \int_{\omega_i} \left(e^{-2s\eta^{(1)}} {\varphi^{(1)}}^{\frac{3}{2}} + e^{-2s\eta^{(2)}(T',x)} {\varphi^{(2)}}^{\frac{3}{2}} \right)(T',x)dx, \qquad i = 0, 1.$$

Proof. Note that β reaches its maximum in $\overline{\Omega}_1$ on γ. Set $A = \sup_{\Omega_1 \setminus \omega_1} \beta$, let $\varepsilon > 0$ be sufficiently small such that $B = A + \varepsilon < \sup_{\Omega_1} \beta$, and define $\widetilde{\omega}_1 = \{x \in \Omega_1;\ A + \varepsilon < \beta(x)\}$. Then $\widetilde{\omega}_1 \subset \omega_1$, with nonempty interior, and we set $K = \frac{|\Omega_1 \setminus \widetilde{\omega}_1|}{|\widetilde{\omega}_1|}$.

We define $\eta_B^{(i)}$ and $\varphi_B^{(i)}$ to be equal to $\eta^{(i)}$ and $\varphi^{(i)}$ at time T' and $\beta(x)$ replaced by B. Then for all $s \geq s_3$ and $\lambda \geq \lambda_3$ we have

$$\int_{\Omega_1 \setminus \widetilde{\omega}_1} e^{-2s\eta_B^{(i)}} {\varphi_B^{(i)}}^{\frac{3}{2}} dx \leq K \int_{\widetilde{\omega}_1} e^{-2s\eta_B^{(i)}} {\varphi_B^{(i)}}^{\frac{3}{2}} dx, \quad i = 1, 2.$$

We conclude by observing that, for $i = 1, 2$,

$$\begin{aligned} e^{-2s\eta_B^{(i)}} &\leq e^{-2s\eta^{(i)}(T',x)}, \quad \varphi_B^{(i)} \leq \varphi^{(i)}(T', x), \quad x \in \widetilde{\omega}_1, \\ e^{-2s\eta^{(i)}(T',x)} &\leq e^{-2s\eta_B^{(i)}}, \quad \varphi^{(i)}(T', x) \leq \varphi_B^{(i)}, \quad x \in \Omega_1 \setminus \widetilde{\omega}_1. \end{aligned}$$

The same method can be applied for the second set of integrals on Ω_0 and ω_0. ∎

Continuation of Remark 3.11. In the statement of Lemma 3.3 we can replace the integration domain, Ω', in the l.h.s. of the estimate by ω.

To reach an equation of the form of (3.8) with the volume integrals computed[2] over ω in place of Ω we write the following counterpart to (3.7):

$$\begin{aligned} &\sum_{i=0,1} \int_{t_0}^{T} \int_{\Omega_i} (e^{-2s\eta^{(1)}} + e^{-2s\eta^{(2)}}) \, |\xi \Delta \partial_t \widetilde{y}|^2 \, dx \, dt \\ &\qquad \leq K'^2 \sum_{i=0,1} \sum_{j=1}^{p_i} |\xi_{i,j}|^2 \int_{\Omega_{i,j} \cap \omega} \left(e^{-2s\eta^{(1)}(T',x)} + e^{-2s\eta^{(2)}(T',x)} \right) |\Delta \widetilde{y}(T', x)|^2 dx \end{aligned}$$

by Lemma 3.12. The result of Lemma 3.6 remain unchanged as ω_0 is a neighborhood of S in $\overline{\Omega}_0$. In the proof of Proposition 3.8 we can choose the open set V to be in ω_1.

Remark 3.13. Note that if we assume that $y(T', .) = \widetilde{y}(T', .)$, then the stability estimate becomes

$$\sum_{i=0,1} \sum_{j=1}^{p_i} |c_{ij} - \widetilde{c}_{ij}|^2 \leq C |\partial_n (\partial_t y - \partial_t \widetilde{y})|^2_{L^2((0,T)\times\gamma)}.$$

Such an additional assumption is sometimes made, e.g., in [16].

With Theorem 3.9 we have the following uniqueness result.

COROLLARY 3.14. *Under the same assumptions as in Theorem* 3.9 *and if*

$$\begin{aligned} &\partial_n (\partial_t (y - \widetilde{y}))(t, x) = 0 \quad \text{in } (t_0, T) \times \gamma, \\ &\Delta y(T', x) - \Delta \widetilde{y}(T', x) = 0 \quad \text{in } \Omega', \end{aligned}$$

then $c = \widetilde{c}$. *Furthermore,* $y_0 = \widetilde{y}_0$.

Proof. The second assertion remains to be proved. If $c = \widetilde{c}$, then $u = y - \widetilde{y} \in \mathscr{D}_A$, with $A = \nabla \cdot (c \nabla(.))$ (see the appendix), is the solution to

$$\begin{cases} \partial_t u - \nabla \cdot (c \nabla u) = 0 & \text{in } (0, T) \times \Omega, \\ u = 0 & \text{on } (0, T) \times \Gamma, \\ u(0, x) = u_0(x) & \text{in } \Omega, \end{cases}$$

with $u_0 = y_0 - \widetilde{y}_0$. Thus $u = S(t)u_0$. We have $\Delta(u)(T') = 0$ in Ω'. Thus $\nabla \cdot (c \nabla u)_{|_{\Omega'}}(T') = 0$. Since $u(T') \in \mathscr{D}_A$ we have $u(T') = 0$. Since the semigroup $S(t)$

[2]Including the definition of $W_{i,j}$, $i = 0, 1$, $j = 1, \ldots, p_i$.

generated by $-\nabla\cdot(c\nabla(.))$ is analytic by Proposition 5, we obtain that $S(t)u_0 = 0$ for all $t > 0$. The continuity in $t = 0^+$ yields $u_0 = 0$. □

If we make further assumptions on the initial conditions y_0 and $\widetilde{y}_0$, we can in fact obtain a stability result for these initial conditions as well. This is the subject of section 5.

Remark 3.15. In the stability result obtained here, we have made the choice to make some of the measurements on part of the boundary $(0,T)\times\Gamma$. Derivation of a Carleman estimate, as in [8], with an r.h.s. with an "observation" in an inner volume $(0,T)\times\omega$ of $(0,T)\times\Omega_1$ would yield a stability estimate like (3.11) with $|\partial_t y - \partial_t\widetilde{y}|_{L^2((t_0,T)\times\omega)}$ in the r.h.s.

Remark 3.16. Observe that in place of Assumption 3.4 we could have assumed solely that the difference $\xi = c - \widetilde{c}$ is piecewise-constant. Then, we would replace $c\Delta u$ by $\nabla\cdot(c\nabla u)$ in the r.h.s. of (3.8). This would yield a stability estimate of the form

$$\sum_{i=0,1}\sum_{j=1}^{p_i}|c_{ij}-\widetilde{c}_{ij}|^2 \le C|\partial_n(\partial_t y-\partial_t\widetilde{y})|^2_{L^2((0,T)\times\gamma)} + C|\nabla y(T',.)-\nabla\widetilde{y}(T',.)|^2_{(L^2(\Omega'))^n} + C|\Delta y(T',.)-\Delta\widetilde{y}(T',.)|^2_{L^2(\Omega')}. \tag{3.13}$$

In Remark 4.7, we show that Assumption 3.1 can be fulfilled, in the case where ξ is piecewise-constant, if $|\nabla\widetilde{c}|$ is sufficiently small and $\widetilde{c}$ sufficiently regular.

Remark 3.17. Here we have set the r.h.s. of the parabolic equations in (3.1) and (3.2) to zero. With a nonvanishing r.h.s., say $f(t,x)$, the same stability result holds once the hypotheses of Assumption 3.1 are fulfilled, since (3.3) is preserved. Assuming that the function $f(t,x)$ is bounded and sufficiently regular, say of class $\mathscr{C}^k$, with k sufficiently large, we can obtain the same results as in Proposition 4.5 of section 4, by a slight modification of the argumentation. In this case also, the class of solutions y and $\widetilde{y}$ satisfying Assumption 3.1 is nonempty.

Remark 3.18. The stability result obtained here can be extended to a more complicated geometry, for instance in the case of three or more embedded materials. We can consider some open sets $V_0,\dots,V_m$ such that

$$V_0 \Subset V_1 \Subset \cdots \Subset V_m = \Omega$$

and then set $\Omega_i := V_i\setminus\overline{V}_{i-1}$, $i = 1,\dots,m$, and $\Omega_0 = V_0$. We then assume the diffusion coefficient to be piecewise-constant, constant on each connected component of Ω_i, and assume a monotony condition across interfaces as in Assumption 1.2. With geometric conditions on the open sets Ω_i, $i = 0,\dots,m$, similar to those made on Ω_0 and Ω_1 here we can then prove a Carleman estimate of the form (2.5). The method exposed in this section then yields a stability result as in Theorem 3.9. For more details see our complete preprint [5]. The results of section 5 can be extended similarly.

4. Existence of solutions y, $\widetilde{y}$ satisfying Assumption 3.1. We propose a possible choice of boundary condition h and of initial condition $\widetilde{y}_0$ to achieve the particular properties for the solutions y and $\widetilde{y}$ listed in Assumption 3.1 needed in the proof of Theorem 3.9.

We shall denote by $S(t)$ (resp., $\widetilde{S}(t)$) the analytic semigroup generated by unbounded operator A (resp., $\widetilde{A}$) formally defined by $-\nabla\cdot(c\nabla(.))$ (resp., $-\nabla\cdot(\widetilde{c}\nabla(.))$) on $L^2(\Omega)$ with domain (see the appendix)

$$\mathscr{D}_A = \{u\in H_0^1(\Omega);\ \nabla\cdot(c\nabla u)\in L^2(\Omega)\}$$
$$(\text{resp., } \mathscr{D}_{\widetilde{A}} = \{u\in H_0^1(\Omega);\ \nabla\cdot(\widetilde{c}\nabla u)\in L^2(\Omega)\}).$$

The convention we use here is $S(t) = e^{-tA}$.

LEMMA 4.1. *Let $r > 0$ and let $\widetilde{c} \in L^\infty(\Omega)$. There exist $\widetilde{y}_0 \in \mathscr{D}_{\widetilde{A}}$ and $\chi : [0,T] \to \mathbb{R}$ such that the solution to*

$$\begin{cases} \partial_t \widetilde{y} - \nabla \cdot (\widetilde{c} \nabla \widetilde{y}) = 0 & \text{in } (0,T) \times \Omega, \\ \widetilde{y}(t,x) = \chi(t) & \text{on } (0,T) \times \Gamma, \\ \widetilde{y}(t,.) - \chi(t) \in \mathscr{D}_{\widetilde{A}}, & 0 < t \leq T, \\ \widetilde{y}(0) = \widetilde{y}_0 \end{cases} \tag{4.1}$$

satisfies $|\nabla \cdot (\widetilde{c} \nabla \widetilde{y})(T')| \geq r > 0$ a.e. The function χ can be chosen such that χ' is a positive constant.

Proof. Observe that $p(t,x) = \widetilde{y}(t,x) - \chi(t)$ is the solution to

$$\begin{cases} \partial_t p - \nabla \cdot (\widetilde{c} \nabla p) = -\chi'(t) & \text{in } (0,T) \times \Omega, \\ p(t,x) = 0 & \text{on } (0,T) \times \Gamma, \\ p(t,.) \in \mathscr{D}_{\widetilde{A}}, & 0 < t \leq T, \\ p(0,x) = \widetilde{y}_0(x) - \chi(0) = p_0 \in L^2(\Omega) \end{cases} \tag{4.2}$$

and is thus given by Duhamel's formula [21]

$$p(t) = \widetilde{S}(t) p_0 - \int_0^t \widetilde{S}(t-s) \chi'(s) ds. \tag{4.3}$$

In fact, we choose χ of the form $\chi(t) = -\rho t$, where ρ is a negative constant. We also choose $\widetilde{y}_0$ such that $p_0 = \widetilde{y}_0 \in \mathscr{D}_{\widetilde{A}}$ and $\nabla \cdot (\widetilde{c} \nabla \widetilde{y}_0) \geq r_0 > r$ a.e. in Ω (choose $f \in L^2(\Omega)$ such that $f > r_0$, and solve the elliptic problem $\nabla \cdot (\widetilde{c} \nabla \widetilde{y}_0) = f$ for $\widetilde{y}_0$ in $H_0^1(\Omega)$). We choose ρ such that $-r_0 < \rho \leq -r < 0$.

The solution p to (4.2) is unique in $\mathscr{C}^1([0,T], L^2(\Omega)) \cap \mathscr{C}^0([0,T], \mathscr{D}_{\widetilde{A}})$ and given by (4.3) [7, Theorem 3 and following Remark 2, section XVII B.1]. Denoting by $\mathbf{1}$ the function identically equal to 1 on Ω, we find

$$p(t) = \widetilde{S}(t) p_0 + \rho \int_0^t \widetilde{S}(s) \mathbf{1} ds,$$

which yields $q := -\widetilde{A} p + \rho \mathbf{1} := \widetilde{S}(t)(\nabla \cdot (\widetilde{c} \nabla p_0) + \rho \mathbf{1})$ [21, Theorem 1.2.4]. Hence q is the solution to

$$\begin{cases} \partial_t q - \nabla \cdot (\widetilde{c} \nabla q) = 0 & \text{in } (0,T) \times \Omega, \\ q(t,x) = 0 & \text{on } (0,T) \times \Gamma, \\ q(t,.) \in \mathscr{D}_{\widetilde{A}}, & 0 < t \leq T, \\ q(0,x) = q_0(x) := \nabla \cdot (\widetilde{c} \nabla p_0) + \rho \mathbf{1}. \end{cases}$$

We now apply the maximum principle (which is valid for L^∞ diffusion coefficients) [6, proof of Theorem IX.3], which reads for the time interval $[0, T']$

$$\operatorname*{ess\,inf}_{\overline{Q}_{T'}} \mathrm{q} \geq \min(0, \operatorname*{ess\,inf}_{\Omega} \mathrm{q}_0) = 0, \qquad \mathrm{Q}_{\mathrm{T}'} = (0, \mathrm{T}') \times \Omega.$$

This yields $\nabla \cdot (\widetilde{c} \nabla p)(T', x) \geq -\rho \geq r > 0$ a.e. ∎

LEMMA 4.2. *Let $l > n/2$ and $l \geq 2$. Let $\widetilde{c} \in L^\infty(\Omega)$ be such that $\widetilde{c}_{|\Omega_i}$ is $\mathscr{C}^{l-1}(\overline{\Omega}_i)$, $i = 0, 1$. Let Ω be such that S and $\partial\Omega$ are of class $\mathscr{C}^l$. Let $\widetilde{y}_0 \in \mathscr{D}_{\widetilde{A}}$ and the function*

$\chi : [0.T] \to \mathbb{R}$, *such that* χ' *is constant, both be chosen according to Lemma* 4.1. *Then* $\nabla\cdot(\widetilde{c}\nabla\widetilde{y})_{|\Omega_i} \in \mathscr{C}^k((0,T], L^\infty(\Omega_i))$, $i = 0, 1$, *for all* $k \in \mathbb{N}$. *Let* $\varepsilon > 0$; *then* $\nabla\cdot(\widetilde{c}\nabla\widetilde{y})_{|\Omega_i}$, $i = 0, 1$, *remain in a bounded domain of* $\mathscr{C}^k([\varepsilon, T], L^\infty(\Omega_i))$ *for all* $k \in \mathbb{N}$, *uniformly w.r.t.* $\widetilde{c}$ *and* $\widetilde{y}_0$, *for* $0 < c_{min} \le \widetilde{c} \le c_{max}$ *and* $\nabla \cdot (\widetilde{c}\nabla\widetilde{y}_0)$ *in a bounded domain of* $L^2(\Omega)$.

Proof. We use the notation of the proof of Lemma 4.1. We set $p(t,x) = \widetilde{y}(t,x) - \chi(t)$ and observe that $q := -\widetilde{A}p + \rho\mathbf{1}$ is the solution to

$$\begin{cases} \partial_t q - \nabla \cdot (\widetilde{c}\nabla q) = 0 & \text{in } (0,T)\times\Omega, \\ q(t,x) = 0 & \text{on } (0,T)\times\Gamma, \\ q(t,.) \in \mathscr{D}_{\widetilde{A}}, & 0 < t \le T, \\ q(0,x) = q_0(x) := \nabla \cdot (\widetilde{c}\nabla p_0) + \rho\mathbf{1}. \end{cases}$$

From Corollary 5 we have that $q_{|(0,T]\times\Omega_i} \in \mathscr{C}^k((0,T]; H^l(\Omega_i))$, $i = 0, 1$, for all $k \in \mathbb{N}$. Since $l > n/2$, the space $H^l(\Omega_i)$ is continuously embedded in $L^\infty(\Omega_i)$, which yields the result. The last statement follows from Remark 5. ∎

Remark 4.3. In the case of $n = 2, 3$, which concerns most of the applications, we choose $m = 2$. The condition on S, $\partial\Omega$ and the coefficients $\widetilde{c}_{|\Omega_i}$ in the previous lemma are then the default ones assumed in the introduction.

Let the function χ, such that $\rho = -\chi'$ is constant, be chosen according to Lemma 4.1. We then have the following regularity property.

LEMMA 4.4. *Let* $c, \widetilde{c} \in L^\infty(\Omega)$ *be such that* $c_{|\Omega_i}, \widetilde{c}_{|\Omega_i}$ *are* $\mathscr{C}^1(\overline{\Omega}_i)$, $i = 0, 1$; $0 < c_{min} \le c, \widetilde{c} \le c_{max}$; *and* y_0, $\widetilde{y}_0 \in L^2(\Omega)$ *remain in a bounded domain of* $L^2(\Omega)$. *The solutions* $\widetilde{y}$ *and* y *to*

$$\begin{cases} \partial_t \widetilde{y} - \nabla \cdot (\widetilde{c}\nabla\widetilde{y}) = 0 & \textit{in } (0,T)\times\Omega, \\ \widetilde{y}(t,x) = \chi(t) & \textit{on } (0,T)\times\Gamma, \\ \textit{transmission conditions (TC1)}, \\ \widetilde{y}(0) = \widetilde{y}_0, \end{cases}$$

$$\begin{cases} \partial_t y - \nabla \cdot (c\nabla y) = 0 & \textit{in } (0,T)\times\Omega, \\ y(t,x) = \chi(t) & \textit{on } (0,T)\times\Gamma, \\ \textit{transmission conditions (TC1)}, \\ y(0) = y_0 \end{cases}$$

belong to $\mathscr{C}^k((0,T], H^1(\Omega))$ *and are such that* $\widetilde{y}_{|\Omega_i}, y_{|\Omega_i} \in \mathscr{C}^k((0,T], H^2(\Omega_i))$, $i = 0, 1$, *for all* $k \in \mathbb{N}$. *Let* $\varepsilon > 0$; *then for all* $k \in \mathbb{N}$, $\widetilde{y}_{|\Omega_i}, y_{|\Omega_i}$ *remain in a bounded domain of* $\mathscr{C}^k((\varepsilon, T], H^2(\Omega_i))$, $i = 0, 1$, *uniformly w.r.t.* $\widetilde{c}$, y_0, *and* $\widetilde{y}_0$.

Proof. We work out the proof for y. We define $p(t,x) = y(t,x) - \chi(t)$. The function p is the solution to

$$\begin{cases} \partial_t p - \nabla \cdot (c\nabla p) = \rho & \text{in } (0,T)\times\Omega, \\ p(t,x) = 0 & \text{on } (0,T)\times\Gamma, \\ p(t,.) \in \mathscr{D}_A, & 0 < t \le T, \\ p(0) = p_0 = y_0 - \chi(0) \in L^2(\Omega). \end{cases} \tag{4.4}$$

It suffices to prove the result for p. Since ρ is constant, the (mild) solution to (4.4) is a classical solution [21, Theorem 4.3.2]. We prove below that $p \in \mathscr{C}^k((0,T], \mathscr{D}_A)$,

$k > 0$. Thus $p \in \mathscr{C}^k((0,T], L^2(\Omega))$. Since $D_A \subset H_0^1$ with continuous injection, then $p \in \mathscr{C}^k((0,T], H_0^1(\Omega))$. By Proposition 5 the maps $p \mapsto p_{|\Omega_i}$, $i = 0, 1$, are continuous from $\mathscr{D}_A$ into $H^2(\Omega_i)$. Thus $p_{|\Omega_i} \in \mathscr{C}^k((0,T], H^2(\Omega_i))$.

The solution p is given by

$$p(t) = S(t)p_0 + \rho \int_0^t S(s)\mathbf{1}ds,$$

where $\mathbf{1}$ is the function identically equal to 1 on Ω. The first term $p_1 = S(t)p_0$ in $\mathscr{C}^k((0,T], \mathscr{D}_{A^l})$ for all $k, l > 0$ by Proposition 5. For the second term $p_2 = \rho \int_0^t S(s)\mathbf{1}ds$ we have (see [21, Theorem 1.2.4]) $-Ap_2 = \rho(S(t)\mathbf{1} - \mathbf{1})$. Thus $Ap_2 \in \mathscr{C}^k((0,T], L^2(\Omega))$, i.e., $p_2 \in \mathscr{C}^k((0,T], \mathscr{D}_A)$, for all $k > 0$. The boundedness statement follows from Remark 5. □

With the proposed initial condition $\widetilde{y}_0$ and boundary condition $h(t,x) = \chi(t)$, we have thus obtained the following regularity and boundedness properties.

PROPOSITION 4.5. *Let $r > 0$. Let $y_0 \in L^2(\Omega)$. Let Ω be such that S and $\Gamma = \partial\Omega$ are of class $\mathscr{C}^l$, and let $\widetilde{c} \in L^\infty(\Omega)$ be such that $\widetilde{c}_{|\Omega_i}$ is $\mathscr{C}^{l-1}(\overline{\Omega}_i)$, $i = 0, 1$, with $> n/2$, $l \geq 2$. There exists $h(t,x) \in \mathscr{C}([0,T] \times \Gamma)$ and an initial condition $\widetilde{y}_0 \in \mathscr{D}_{\widetilde{A}}$ such that the solutions y, $\widetilde{y}$ to systems (3.1)–(3.2) satisfy the following:*

1. $\nabla \cdot (\widetilde{c}\nabla\widetilde{y})(T') \geq r > 0$;
2. $\nabla \cdot (\widetilde{c}\nabla\widetilde{y})_{|\Omega_i} \in \mathscr{C}^k([t_0,T], L^\infty(\Omega_i))$, $i = 0, 1$, *for all* $k \in \mathbb{N}$;
3. y, $\widetilde{y} \in \mathscr{C}^k([t_0,T], L^2(\Omega)) \cap \mathscr{C}^k((t_0,T], H^1(\Omega))$ *for all* $k \in \mathbb{N}$;
4. $y_{|\Omega_i}, \widetilde{y}_{|\Omega_i} \in \mathscr{C}^k([t_0,T], H^2(\Omega_i))$, $i = 0, 1$, *for all* $k \in \mathbb{N}$.

The restrictions $\widetilde{y}_{|\Omega_i}$ remain in a bounded domain of $\mathscr{C}^k([t_0,T], H^2(\Omega_i))$ and $\nabla \cdot (\widetilde{c}\nabla\widetilde{y})_{|\Omega_i}$ in a bounded domain of $\mathscr{C}^k([t_0,T], L^\infty(\Omega_i))$, uniformly w.r.t. $\widetilde{c}$ and $\widetilde{y}_0$ if $0 < c_{min} \leq \widetilde{c} \leq c_{max}$ and $\nabla \cdot (\widetilde{c}\nabla\widetilde{y}_0)$ remain in a bounded domain of $L^2(\Omega)$.

With Proposition 4.5 we observe that Assumption 3.1 can be fulfilled in the framework of Assumption 3.4 when $c_{min} \leq \widetilde{c} \leq c_{max}$ and $\widetilde{y}_0 \in \mathscr{D}_{\widetilde{A}}$ such that $\nabla \cdot (\widetilde{c}\nabla\widetilde{y}_0)$ remain in a bounded domain of $L^2(\Omega)$ for properly chosen boundary conditions $h(t,x)$.

Remark 4.6. Observe that we could simply assume that $\widetilde{y}_0 \in L^2(\Omega)$ and design the boundary condition $h(t,x)$ to reach a proper state in $\mathscr{D}_{\widetilde{A}}$ in a finite time $t_1 < t_0$. This can be achieved as the parabolic equation we study here is null-controllable, i.e., exactly controllable to the trajectories [8].

Remark 4.7. In the case where we assume only that ξ is piecewise-constant, then from Proposition 4.5.1, we can obtain Assumption 3.1.1, in the case where $|\nabla\widetilde{c}|$ is sufficiently small, since, from the proof of Lemma 4.2, we also find that $\nabla\widetilde{y}_{|\Omega_i}$, $i = 0, 1$, remain in bounded domains of $\mathscr{C}^k([\varepsilon,T], (L^\infty(\Omega_i))^n)$. Note that we need to assume a sufficiently large regularity on the coefficient $\widetilde{c}$ in the proof of Lemma 4.2.

5. Uniqueness and stability estimate for the initial conditions. In this section we closely follow the method of [22]. We shall assume the following.

ASSUMPTION 5.1. *Let $r_0 > 0$. The initial conditions y_0 and $\widetilde{y}_0$ satisfy*

1. *y_0 is in a bounded domain of $\mathscr{D}_A$;*
2. *$\widetilde{y}_0$ is in a bounded domain of $\mathscr{D}_{\widetilde{A}}$;*
3. *$\nabla \cdot (\widetilde{c}\nabla\widetilde{y}_0) \geq r_0$;*
4. *y, $\widetilde{y}$ are in a bounded domain of $\mathscr{C}^1([0,T], L^2(\Omega))$, where y and $\widetilde{y}$ are the solutions to (3.1)–(3.2).*

Observe that Assumption 5.1.4 implies that $\widetilde{y}_{|\Omega_i}, y_{|\Omega_i}$ are in a bounded domain of $\mathscr{C}([0,T], H^2(\Omega_i))$, $i = 0, 1$, by Proposition A.3.

We define $\widetilde{z} = \partial_t \widetilde{y} \in \mathscr{C}([0,T], L^2(\Omega))$, and thus $\widetilde{z}(0)$ is well defined in $L^2(\Omega)$. We introduce w the solution to

$$\begin{cases} \partial_t w - \nabla \cdot (c \nabla w) = 0 & \text{in } (0,T) \times \Omega, \\ w(t,x) = \partial_t h(t,x) & \text{on } (0,T) \times \Gamma, \\ \text{transmission conditions (TC1)} & \text{on } (0,T) \times S, \\ w(0) = \widetilde{z}(0), \end{cases} \tag{5.1}$$

and we further assume the following.

ASSUMPTION 5.2. *The functions $\widetilde{z}$, w are in a bounded domain of $L^2(0,T,H^1(\Omega))$.*

We also assume that the diffusion coefficients c and $\widetilde{c}$ are piecewise-constant (Assumption 3.4).

Observe that if we choose the boundary condition $h(t,x) = -\rho t$ for $0 < t \leq T$ according to the proof of Lemma 4.1 (with $0 < r < r_0$), then the above assumption are fulfilled. In fact, the results of section 4 show that Assumption 3.1 is then satisfied. In addition, Assumptions 5.1.4 and 5.2 are fulfilled by the following lemma.

LEMMA 5.3. *If $h(t,x) = -\rho t$, then the solutions y, $\widetilde{y}$ to* (3.1)–(3.2) *and w to* (5.1) *satisfy Assumptions* 5.1.4 *and* 5.2.

Proof. We prove $w \in L^2(0,T,H^1(\Omega))$. The proof is the same for $\widetilde{z}$. Let $p(t,x) = w(t,x) - \partial_t h(t,x) = w(t,x) + \rho$. Then p satisfies

$$\begin{cases} \partial_t p - \nabla \cdot (c \nabla p) = 0 & \text{in } (0,T) \times \Omega, \\ p(t,x) = 0 & \text{on } (0,T) \times \Gamma, \\ \text{transmission conditions (TC1)} & \text{on } (0,T) \times S, \\ p(0) = \widetilde{z}(0) + \rho \in L^2(\Omega). \end{cases}$$

We thus have the usual energy estimate

$$\frac{1}{2}|p(t)|^2_{L^2(\Omega)} + \int_0^t \int_\Omega c |\nabla p|^2 dt dx = \frac{1}{2}|p(0)|^2_{L^2(\Omega)},$$

and p, and thus w, is in $L^2(0,T,H^1(\Omega))$ and remains in a bounded domain of this space if $c_{min} \leq c \leq c_{max}$ and $\widetilde{y}_0$ remains in a bounded domain of $\mathscr{D}_{\widetilde{A}}$.

To prove that y is in a bounded domain of $\mathscr{C}^1([0,T], L^2(\Omega))$ (the proof is the same for $\widetilde{y}$), we set $p(t,x) = y(t,x) + \rho t$ and observe that $q := -Ap + \rho \mathbf{1}$ is $\mathscr{C}([0,T], L^2(\Omega))$ and thus $p \in \mathscr{C}([0,T], \mathscr{D}_A)$. Then $p \in \mathscr{C}^1([0,T], L^2(\Omega))$. ∎

Define v_1 and v_2 that satisfy

$$\begin{cases} \partial_t v_1 - \nabla \cdot (c \nabla v_1) = \nabla \cdot ((c - \widetilde{c}) \nabla \partial_t \widetilde{y}) & \text{in } (0,T) \times \Omega', \\ v_1 = 0 & \text{on } (0,T) \times \Gamma, \\ \text{transmission conditions (TC2)} & \text{on } (0,T) \times S, \\ v_1(0) = 0, \end{cases}$$

$$\begin{cases} \partial_t v_2 - \nabla \cdot (c \nabla v_2) = 0 & \text{in } (0,T) \times \Omega', \\ v_2 = 0 & \text{on } (0,T) \times \Gamma, \\ \text{transmission conditions (TC1)} & \text{on } (0,T) \times S, \\ v_2(0) = \partial_t (y - \widetilde{y})(0). \end{cases}$$

Observe that $\partial_t (y - \widetilde{y})(0)$ is well defined and in a bounded domain of $L^2(\Omega)$ by Assumption 5.1.

With an argument of logarithmic convexity we have

$$|v_2(t)|_{L^2(\Omega)} \leq K^{1-t/T'} \, |v_2(T')|^{t/T'}_{L^2(\Omega)}, \quad 0 \leq t \leq T'. \tag{5.2}$$

Such an estimate makes use of the convexity of $F(t) = \ln(|v_2(t)|^2_{L^2(\Omega)})$ and $|v_2(0)|_{L^2(\Omega)} \leq K$ (see [20, section 2.3] for further details).

We now prove the following lemma.

LEMMA 5.4. *There exists $C > 0$ such that*

$$|v_1(t)|_{L^2(\Omega)} \leq C|c - \widetilde{c}|^{\frac{1}{2}}_{L^\infty(\Omega)}, \quad 0 \leq t \leq T.$$

Note that v_1 satisfies transmission condition (TC2) and thus does not belong to $\mathscr{D}_A$. We thus cannot use some argument of regularity w.r.t. the source term from parabolic theory for v_1.

Proof. First observe that $v_1 = w - \widetilde{z}$. From Assumption 5.2, w and $\widetilde{z}$ remain in a bounded domain of $L^2(0,T,H^1(\Omega))$. Now $\partial_t(w - \widetilde{z}) - \nabla \cdot (c\nabla w - \widetilde{c}\nabla\widetilde{z}) = 0$, which after multiplication by $w - \widetilde{z}$, integration over Ω, and an integration by parts, yields

$$\begin{aligned} 0 &= \frac{1}{2}\partial_t \int_\Omega |w - \widetilde{z}|^2 \, dx + \int_\Omega (c\nabla w - \widetilde{c}\nabla\widetilde{z}) \cdot (\nabla w - \nabla\widetilde{z}) \, dx \\ &= \frac{1}{2}\partial_t \int_\Omega |w - \widetilde{z}|^2 \, dx + \int_\Omega (c - \widetilde{c})\nabla w \cdot (\nabla w - \nabla\widetilde{z}) \, dx + \int_\Omega \widetilde{c}|\nabla w - \nabla\widetilde{z}|^2 \, dx \end{aligned}$$

since $\nabla \cdot (c\nabla w - \widetilde{c}\nabla\widetilde{z}) \in L^2(\Omega)$ by the definitions of $\mathscr{D}_A$ and $\mathscr{D}_{\widetilde{A}}$ (see the appendix). We thus obtain

$$\begin{aligned} \frac{1}{2}\partial_t \int_\Omega |w - \widetilde{z}|^2 \, dx &\leq \left| \int_\Omega (c - \widetilde{c})\nabla w \cdot (\nabla w - \nabla\widetilde{z}) \, dx \right| \\ &\leq |c - \widetilde{c}|_{L^\infty(\Omega)} \, |\nabla w|_{L^2(\Omega)} \, |\nabla w - \nabla\widetilde{z}|_{L^2(\Omega)}. \end{aligned}$$

Integrating over $(0,t)$ yields the result. ∎

As in section 3, we define $v = \partial_t(y - \widetilde{y})$ and observe that $v = v_1 + v_2$. We thus have

$$|v(t)|_{L^2(\Omega)} \leq |v_1(t)|_{L^2(\Omega)} + |v_2(t)|_{L^2(\Omega)} \leq C(|c - \widetilde{c}|^{\frac{1}{2}}_{L^\infty(\Omega)} + |v_2(T')|^{t/T'}_{L^2(\Omega)}), \quad 0 \leq t \leq T',$$

and

$$\begin{aligned} |v_2(T')|_{L^2(\Omega)} &\leq |v(T')|_{L^2(\Omega)} + |v_1(T')|_{L^2(\Omega)} \\ &\leq C(|(\Delta y - \Delta\widetilde{y})(T')|_{L^2(\Omega)} + |c - \widetilde{c}|_{L^\infty(\Omega)} + |c - \widetilde{c}|^{\frac{1}{2}}_{L^\infty(\Omega)}), \end{aligned}$$

making use of

$$v = \nabla \cdot (c\nabla u) + \nabla((c - \widetilde{c}) \cdot \nabla\widetilde{y}) = c(\Delta y - \Delta\widetilde{y}) + (c - \widetilde{c})\Delta\widetilde{y} \quad \text{in } \Omega',$$

where Assumption 3.4 was applied. This yields

$$|v(t)|_{L^2(\Omega)} \leq C\left(|c - \widetilde{c}|^{\frac{1}{2}}_{L^\infty(\Omega)} + \nu^{t/T'}\right), \quad 0 \leq t \leq T',$$

with $\nu = |(y-\widetilde{y})(T')|_{H^2(\Omega')} + |c-\widetilde{c}|_{L^\infty(\Omega)} + |c-\widetilde{c}|^{\frac{1}{2}}_{L^\infty(\Omega)}$. Because of the regularity of v w.r.t. time t we now have

$$|y_0-\widetilde{y}_0|_{L^2(\Omega)} = |u_0|_{L^2(\Omega)} = \left|\int_0^{T'} v(t)dt - u(T')\right|_{L^2(\Omega)}$$
$$\leq \int_0^{T'} |v(t)|_{L^2(\Omega)}dt + |y(T')-\widetilde{y}(T')|_{L^2(\Omega)},$$

which gives

$$|y_0-\widetilde{y}_0|_{L^2(\Omega)} \leq C\left(T'\frac{\nu-1}{\ln(\nu)} + T'|c-\widetilde{c}|^{\frac{1}{2}}_{L^\infty(\Omega)}\right) + |y(T')-\widetilde{y}(T')|_{L^2(\Omega)}.$$

We thus have

$$|y_0-\widetilde{y}_0|_{L^2(\Omega)} \leq C\left(\frac{\nu-1}{\ln(\nu)} + C'\nu\right).$$

Observing that $x < \frac{x-1}{\ln(x)}$ for $x \in (0,1)$, we obtain

$$|y_0-\widetilde{y}_0|_{L^2(\Omega)} \leq C\frac{\nu-1}{\ln(\nu)} \quad \text{if } \nu < 1.$$

With Theorem 3.9 we obtain

$$|c-\widetilde{c}|^2_{L^\infty(\Omega)} \leq C|\partial_n(\partial_t y - \partial_t\widetilde{y})|^2_{L^2((0,T)\times\gamma)} + C|\Delta y(T',.) - \Delta\widetilde{y}(T',.)|^2_{L^2(\Omega')}. \tag{5.3}$$

When the r.h.s. of (5.3) is sufficiently small, we obtain

$$0 < \nu \leq C\left(|\partial_n(\partial_t y - \partial_t\widetilde{y})|^2_{L^2((0,T)\times\gamma)} + |\Delta y(T',.) - \Delta\widetilde{y}(T',.)|^2_{L^2(\Omega')}\right)^{\frac{1}{4}} < 1.$$

In that case

$$\ln(\nu) \leq C'\ln(C|(y-\widetilde{y})(T')|^2_{H^2(\Omega')} + C|\partial_n(\partial_t y - \partial_t\widetilde{y})|^2_{L^2((0,T)\times\gamma)}) < 0,$$

and thus

$$\frac{\nu-1}{\ln(\nu)} \leq C'(\nu-1)/\ln(C|(y-\widetilde{y})(T')|^2_{H^2(\Omega')} + C|\partial_n(\partial_t y - \partial_t\widetilde{y})|^2_{L^2((0,T)\times\gamma)}).$$

We thus obtain the following stability theorem for the initial conditions.

THEOREM 5.5. *Under the hypothesis of Theorem* 3.9*, in addition to Assumptions* 5.1 *and* 5.2 *there exist some constants* $C, C' > 0$*,*

$$C = C(\Omega, T, t_0, \gamma, S, \mathcal{O}^{(1)}, \mathcal{O}^{(2)}, M, K, r, c_{min}, c_{max}, \Delta),$$
$$C' = C'(\Omega, T, t_0, \gamma, S, \mathcal{O}^{(1)}, \mathcal{O}^{(2)}, M, K, r, c_{min}, c_{max}, \Delta),$$

such that

$$|y_0-\widetilde{y}_0|_{L^2(\Omega)} \leq C'/\left|\ln\left(C|(y-\widetilde{y})(T')|_{H^2(\Omega')} + C|\partial_n(\partial_t y - \partial_t\widetilde{y})|_{L^2((0,T)\times\gamma)}\right)\right|$$

for $|(y-\widetilde{y})(T')|_{H^2(\Omega')} + |\partial_n(\partial_t y - \partial_t\widetilde{y})|_{L^2((0,T)\times\gamma)}$ *sufficiently small.*

Appendix. Basic regularity properties. Let $\mathcal{A}$ be formally defined by $-\nabla \cdot (c\nabla(.))$ on $L^2(\Omega)$. The diffusion coefficient c is first assumed to be in $L^\infty(\Omega)$ and such that $c(x) \geq \alpha > 0$ for all $x \in \Omega$. We denote by A the unbounded operator with domain

$$\mathscr{D}_A = \{u \in H_0^1(\Omega);\ \nabla \cdot (c\nabla u) \in L^2(\Omega)\},$$

defined by $A(u) = -\nabla \cdot (c\nabla(u))$ for $u \in \mathscr{D}_A$.

PROPOSITION A.1. *Let $u_0 \in L^2(\Omega)$. There exists a unique u such that*

$$u \in \mathscr{C}([0,T]; L^2(\Omega)) \cap \mathscr{C}^1(]0,T]; L^2(\Omega)) \cap \mathscr{C}(]0,T]; \mathscr{D}_A)$$

and

$$\begin{cases} \partial_t u - \nabla \cdot (c\nabla(u)) = 0, \quad t \in]0,T], \\ u(0) = u_0, \end{cases} \tag{A.1}$$

for $T > 0$ (T can be chosen to be ∞). Furthermore, $u \in L^2(0,T; H_0^1(\Omega))$ and $u \in \mathscr{C}^k(]0,T]; \mathscr{D}_{A^l})$ for all $k, l \in \mathbb{N}$.

If $u_0 \in \mathscr{D}_A$, then

$$u \in \mathscr{C}^1([0,T]; L^2(\Omega)) \cap \mathscr{C}([0,T]; \mathscr{D}_A).$$

PROPOSITION A.2. *The semigroup $S(t)$ generated by the unbounded operator A on $L^2(\Omega)$ is analytic.*

We now give further regularity properties when placed in the geometrical situation studied in this article, that is, if the diffusion coefficient c is piecewise $\mathscr{C}^1$ and discontinuous across some $\mathscr{C}^2$ interface S. We use the notation set in the main text of the article.

PROPOSITION A.3. *Let the diffusion coefficient, c, be such that $c_{|\Omega_i} \in \mathscr{C}^1(\overline{\Omega}_i)$, $i = 0, 1$. If $p \in \mathscr{D}_A$, then $p_{|\Omega_i} \in H^2(\Omega_i)$, $i = 0, 1$. Furthermore $|p_{|\Omega_i}|_{H^2(\Omega_i)} \leq C|\nabla \cdot (c\nabla p)|_{L^2(\Omega)}$.*

COROLLARY A.4. *Let $m \in \mathbb{N}$, and let $c_{|\Omega_i} \in \mathscr{C}^{m+1}(\Omega_i)$, $i = 0, 1$, and S and $\partial\Omega$ be of class $\mathscr{C}^{m+2}$. Then if $u_0 \in L^2(\Omega)$, the solution u to* (A.1) *is such that $u_{|\Omega_i} \in \mathscr{C}^k((0,T], H^{m+2}(\Omega_i))$, $i = 0, 1$, for all $k \in \mathbb{N}$.*

Remark A.5. Let $\varepsilon > 0$. With the notation of the above corollary, observe that the map

$$\begin{aligned} L_{\varepsilon,i} : L^2(\Omega) &\to \mathscr{C}^k([\varepsilon, T], H^{m+2}(\Omega_i)), \\ u_0 &\mapsto (t \mapsto u_{|\Omega_i}(t)), \end{aligned}$$

is continuous for $i = 0, 1$, since [6, Theorem VII.7]

$$\begin{aligned} |u(t)|_{L^2(\Omega)} &\leq |u_0|_{L^2(\Omega)}, \\ |\partial_t u(t)|_{L^2(\Omega)} = |\nabla \cdot (c\nabla u(t))|_{L^2(\Omega)} &\leq \left|\frac{1}{t}\right| u_0|_{L^2(\Omega)}. \end{aligned}$$

We finish with the following lemma, which is needed in section 3.

LEMMA A.6. *There exists a neighborhood W of S globally parametrized with $(\sigma, y) \in]-\epsilon, \epsilon[\times S$.*

Proof. In a small neighborhood of S we can extend the unit normal vector n to S into a $\mathscr{C}^2$ vector field v. In an even smaller neighborhood U of S we can assume that v is such that $|v| \geq a > 0$. If we integrate this vector field, we find that there is $\epsilon > 0$ such that the flow χ_σ of this $\mathscr{C}^2$ vector field over the interval $]-\epsilon, \epsilon[$ is confined in U, since S is compact. For $y \in S$, the orientation of the unit normal (see above) is such that $\chi_\sigma(y) \in \Omega_1$ for $\sigma \in]-\epsilon, 0[$, and $\chi_0(y) = y$ and $\chi_\sigma(y) \in \Omega_0$ for $\sigma \in]0, \epsilon[$.

Now set

$$W = \{\chi_\sigma(y);\ y \in S \text{ and } \sigma \in\ ]-\epsilon, \epsilon[\,\},$$

which is an open neighborhood of S. Note that if x is in W, then there exist a unique $y \in S$ and a unique $\sigma \in\ ]-\epsilon, \epsilon[$ such that $x = \chi_\sigma(y)$. □

Acknowledgments. The authors wish to thank M. Cristofol, Y. Dermenjian, and O. Poisson for numerous discussions on the subject of this article. The authors also thank F. Boyer for helpful comments. The authors also wish to thank an anonymous reviewer for his extremely detailed comments and corrections, which improved the exposition of the article.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] L. Baudouin and J.-P. Puel, *Uniqueness and stability in an inverse problem for the Schrödinger equation*, Inverse Problems, 18 (2002), pp. 1537–1554.
[3] A. Benabdallah, Y. Dermenjian, and J. Le Rousseau, *Carleman estimates for the one-dimensional heat equation with a discontinuous coefficient and applications*, C. R. Mécanique, 334 (2006), pp. 582–586.
[4] A. Benabdallah, Y. Dermenjian, and J. Le Rousseau, *Carleman estimates for the one-dimensional heat equation with a discontinuous coefficient and applications to controllability and an inverse problem*, J. Math. Anal. Appl., (2007), to appear.
[5] A. Benabdallah, P. Gaitan, and J. Le Rousseau, *Stability of Discontinuous Diffusion Coefficients and Initial Conditions in an Inverse Problem for the Heat Equation*, preprint, LATP, Universités de Marseille, Marseille, France, 2005; available online from http://www.cmi.univ-mrs.fr/∼jlerous/publications.html.
[6] H. Brezis, *Analyse Fonctionnelle*, Masson, Paris, 1983.
[7] R. Dautray and J.-L. Lions, *Analyse Mathématique et Calcul Numérique pour les Sciences et les Techniques, Vol.* 7, Masson, Paris, 1984.
[8] A. Doubova, A. Osses, and J.-P. Puel, *Exact controllability to trajectories for semilinear heat equations with discontinuous diffusion coefficients*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 621–661.
[9] H. Egger, H. W. Engl, and M. V. Klibanov, *Global uniqueness and Hölder stability for recovering a nonlinear source term in a parabolic equation*, Inverse Problems, 21 (2005), pp. 271–290.
[10] E. Fernández-Cara and S. Guerrero, *Global Carleman inequalities for parabolic systems and application to controllability*, SIAM J. Control Optim., 45 (2006), pp. 1395–1446.
[11] A. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes vol. 34, Seoul National University, Korea, 1996.
[12] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[13] L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1963.
[14] L. Hörmander, *The Analysis of Linear Partial Differential Operators, Vol.* IV, Springer-Verlag, New York, Berlin, 1985.
[15] L. Hörmander, *The Analysis of Linear Partial Differential Operators, Vol.* I, 2nd ed., Springer-Verlag, New York, Berlin, 1990.
[16] O. Imanuvilov and M. Yamamoto, *Lipschitz stability in inverse problems by Carleman estimate*, Inverse Problems, 14 (1998), pp. 1229–1245.
[17] V. Isakov, *Carleman type estimates in an anisotropic case and applications*, J. Differential Equations, 105 (1993), pp. 217–238.
[18] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer-Verlag, Berlin, 1998.
[19] M. V. Klibanov, *Global uniqueness of a multidimensional inverse problem for a nonlinear parabolic equation by a Carleman estimate*, Inverse Problems, 20 (2004), pp. 1003–1032.
[20] L. E. Payne, *Improperly Posed Problems in Partial Differential Equations*, CBMS-NSF Reg. Conf. Ser. Appl. Math. 22, SIAM, Philadelphia, 1975.
[21] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[22] M. Yamamoto and J. Zou, *Simultaneous reconstruction of the initial temperature and heat radiative coefficient*, Inverse Problems, 17 (2001), pp. 1181–1202.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1882–1896

NONLINEAR AIMD CONGESTION CONTROL AND CONTRACTION MAPPINGS*

URIEL G. ROTHBLUM† AND ROBERT SHORTEN‡

Abstract. This papers analyzes a class of nonlinear *additive-increase multiplicative-decrease* (AIMD) protocols that are widely deployed in communication networks. It is demonstrated that the use of these protocols guarantees that the system has a unique stable outcome to which it converges geometrically under all starting points. The development is based on a contraction argument and the derivation of explicit bounds on the contraction coefficient of corresponding operators in terms of the network parameters. In particular, bounds on the corresponding rate of convergence are obtained, improving upon known bounds for standard (linear) AIMD networks.

Key words. AIMD congestion control, high-speed networking, contraction mappings

AMS subject classifications. 90B18, 68M12, 68M10, 47H10

DOI. 10.1137/050646226

1. Introduction.

Traffic generated by the *transmission control protocol* (TCP) accounts for 85%–95% of all traffic on today's Internet [8]. TCP, in congestion avoidance mode, is based primarily on Chiu and Jain's [6] *additive-increase multiplicative-decrease* (AIMD) paradigm for decentralized allocation of a shared resource (e.g., bandwidth) among competing users. The AIMD paradigm is based upon a network of users competing for the available resource by using two basic strategies; they probe for their share of the available resource by utilizing more and more of the resource (the additive-increase (AI) stage) and then instantaneously downscale their utilization-rates in a multiplicative fashion when notified (simultaneously) that capacity was reached (the multiplicative-decrease stage). With some minor modifications, the AIMD algorithm has served the networking community well over the past two decades and it continues to provide the basic building block upon which today's Internet communication is built.

From a mathematical perspective, the dynamics of networks in which the AIMD algorithm is deployed have been studied extensively in the networking, computer science, and mathematics literature; for example, see [5, 10, 19, 24, 25, 26, 27, 29] and references therein. In these papers, some fundamental properties have been established for systems that are controlled by the AIMD algorithm, both in a deterministic and in a stochastic setting; see, for example, [1, 2, 3]. In particular, it has been shown that (with a fixed number of users) such networks possess unique stable equilibria to which the system converges geometrically under all starting points. However, recently in the context of designing high-speed communication networks, several authors have suggested basic modifications to the AIMD algorithm; for example, see [7, 9, 11, 13, 14, 20, 28]. One idea underlying these modifications is to allow the employment of more aggressive probing for available bandwidth by replacing the

*Received by the editors November 29, 2005; accepted for publication (in revised form) March 17, 2007; published electronically November 21, 2007. This work was supported by Science Foundation Ireland grant 04/IN3/I460.

http://www.siam.org/journals/sicon/46-5/64622.html

†Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 32000, Israel (rothblum@ie.technion.ac.il).

‡The Hamilton Institute, National University of Ireland, Maynooth, Maynooth, Co. Kildare, Ireland (robert.shorten@may.ie).

linear-in-time increase of probing that is a feature of TCP with nonlinear growth. We refer to the resulting algorithms as nonlinear AIMD (NAIMD) algorithms. While the modifications appear minor from an algorithmic viewpoint, they result in networks with dynamic properties different than those employing the basic (linear) AIMD; see [4]. Remarkably, despite increasing deployment of these algorithms (e.g., a high-speed TCP algorithm is implemented as part of the Linux operating system), many basic questions pertaining to the behavior of such networks have not yet been addressed (with the notable exception of [20, 21]). Our contribution here is to use contraction arguments to prove convergence of a large class of congestion control protocols that are currently being explored for deployment on the Internet. These include HTCP, High-speed TCP, and many others [7, 13, 14]. In doing this, we model the network of TCP flows not by using a fluid approximation, or by placing assumptions of the distribution of times between congestion notifications, but rather as a system of interacting agents that are coupled together via a capacity constraint. This approach is in contrast to much of the prior work on the topic [15, 16, 17, 18], a large amount of which makes use of either a fluid assumption or assumptions on the stochastic nature of the drop process. A key advantage of modeling the interactions between the AIMD flows in this manner is that we capture not only the network asymptotics, but also the dynamic properties of the network, such as network convergence rate, using only standard linear algebraic arguments.

The objective of the current paper is to examine and study a class of NAIMD algorithms. Under the assumption of user-synchronization, we show that the corresponding networks always possess a unique equilibrium to which the system converges geometrically. Specifically, we observe resource-utilizations at instances when the system becomes saturated (congested) and demonstrate that the transition between consecutive saturation instances is governed by a *contraction* transformation. The aforementioned properties of equilibria then follow from standard results about contractions (see section 3). Our development is based on the derivation of explicit bounds on the contraction coefficient in terms of the network parameters, yielding bounds on the corresponding rate of convergence; in particular, we improve previously obtained bounds for standard AIMD algorithms.

Our results are important for a number of reasons. Network congestion control represents one of the most important problems in decentralized control, both from a practical and a theoretical perspective. Networks in which AIMD-like algorithms are deployed form the backbone of today's Internet; yet surprisingly, many properties of such networks remain unexplored. In particular, the manner in which the NAIMD parameters of each of the users affect the existence and uniqueness of the network equilibria, the nature of the network equilibria, and the rate of convergence to the equilibrium state, as well as the sensitivity of the equilibrium state to changes in network parameters (the ability to control the network fairness), is a very important issues that has yet to be adequately addressed by the research community in a general setting. An equally important consideration in the design of networks is whether new protocols can co-exist with standard AIMD networks without completely starving standard AIMD sources of the available resource. Our results in this paper present a partial solution to many of these questions and represent a first step in addressing these basic system-theoretic issues.

The outline of the paper is as follows. We describe standard and nonlinear variants of AIMD in section 2. In section 3 we introduce preliminaries, in particular about contractions. We use operators to formulate the evolution of the resource-utilization

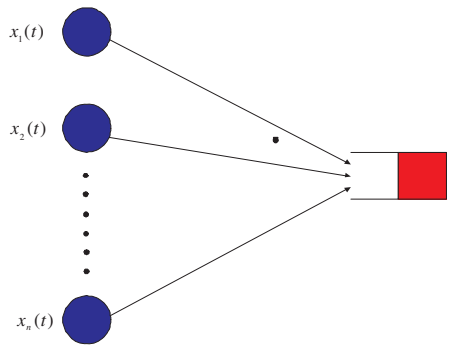


FIG. 1. *n-player system.*

in section 4 and prove our main results in section 5. Section 6 is devoted to final conclusions.

2. Preamble: AIMD congestion control. In their original paper [6], Chiu and Jain consider a system in which n-users compete for a resource having limited availability per unit time, e.g., bandwidth in communication networks. The users' actions consist of (continuously) probing the availability of the resource by submitting requests for its use—these requests are satisfied whenever global capacity is not exceeded. The situation is depicted in Figure 1, with $x_i(t)$ representing the number of units of the resource that user $i = 1, \ldots, n$ is using at time $t \geq 0$. A key assumption in the model formulated by Chiu and Jain is the assertion that the users do not communicate directly with each other. Further, users are provided information about the availability of the resource only when the collective utilization of the resource exceeds some capacity constraint. At such time-instances, referred to as *congestion events*, all users are instantly and simultaneously informed through a binary feedback. The users are assumed to respond instantly to these notifications of congestion by decentralized downscaling of their individual utilization-rates. Given this basic setting, the problem is then to develop an algorithm that produces probing strategies for the users so that each user will infer his or her "fair" share of the shared resource in a decentralized manner.

Comment. In the current paper we focus on the *synchronized* problem, referring to simultaneous notification of congestion to all users to which they all respond. In *unsynchronized systems*, the signal about system-saturation is not transmitted simultaneously to all users. While synchronization is not valid in many real communication networks, the study of such systems is important for two reasons. First, it represents an important first step towards the understanding of more general systems. Second, synchronization appears to be a common feature of high-speed communication networks [28], and consequently the understanding of the behavior of such networks may be of merit in some practical situations.

Linear AIMD congestion control. The AIMD algorithm of Chiu and Jain describes probing strategies that evolve in cycles, each cycle having two phases. The first phase of the cycles is instantaneous. It occurs when capacity is reached, users are notified, and each user responds by downscaling his or her utilization-rate (abruptly) by a multiplicative factor. This phase is called the *multiplicative-decrease* (MD) phase. During the second phase of a cycle, each user increases the utilization-rate linearly until congestion is reached again, at which point the first phase of the next cycle is

entered. The second phase is called the *additive-increase* (AI) phase. The utilization-rate at the end of the AI phase is then the initial transmission-rate for the next cycle. (The expression AIMD, rather than MDAI, is used because of historic reasons.)

Denote the share of the collective resource allocated to player i at time t by $x_i(t)$ and let $x(t) = [x_1(t), \ldots, x_n(t)]^T$. The capacity constraint requires that $\sum_{i=1}^{n} x_i(t) < C$, with C as the total capacity of the resource available to the entire system. Without loss of generality, we may and will assume that units of the resource are normalized so that $C = 1$.

The kth cycle begins at a time t^k at which the global utilization of the resource reaches capacity. The instantaneous decrease of the utilization-rate of player i during the MD phase of the kth cycle is expressed by

$$x_i(t_+^k) = \beta_i x_i(t^k), \tag{1}$$

where t_+^k is an instance after t^k and β_i is a constant in the open interval $(0, 1)$. During the AI phase of the kth cycle, the utilization-rate of user i evolves according to

$$x_i(t) = x_i(t_+^k) + \alpha_i(t - t^k), \tag{2}$$

where t_+^k is as before and α_i is a positive constant. The $(k+1)$st cycle begins at time t^{k+1}, which equals the time t at which the right-hand side of (2) reaches capacity.

A typical trajectory of the utilization-rates in a system with two users that apply AIMD with $\beta_1 = 0.25, \beta_2 = 0.5$, and $\alpha_1 = \alpha_2 = 1$ is depicted in Figure 2. In this figure, points on the line $\{x \in \mathbb{R}^2 : x_1, x_2 \geq 0\,, x_1 + x_2 = 1\}$ represent the utilization-rates at congestion events.

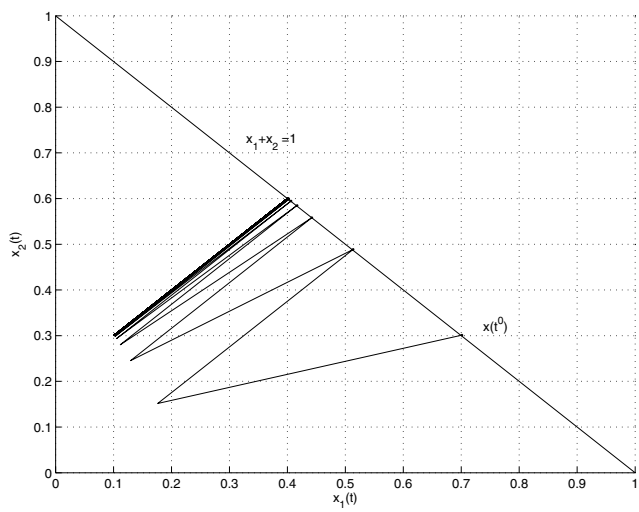


FIG. 2. *Utilization-rates under AIMD.*

Motivated by trajectories of the form illustrated in Figure 2, we find that a convenient framework in which to study the implication of the AIMD algorithm is to consider the utilization-rates at congestion events that occur at times $t^1, t^2, \ldots$. Combining (1) and (2), we see that the evolution of the utilization-rate of user i between the kth and $(k+1)$st congestion points is given by

$$x_i(t^{k+1}) = \beta_i x_i(t^k) + \alpha_i(t^{k+1} - t^k). \tag{3}$$

This avenue of investigation is explored in [5, 26], where it is shown that the transformation of the utilization-rates between consecutive congestion points is linear. Specifically, as $\sum_{i=1}^{n} x_i(t^k) = \sum_{i=1}^{n} x_i(t^{k+1}) = 1$, we have that

$$1 = \sum_{i=1}^{n} x_i(t^{k+1}) = \sum_{i=1}^{n} \beta_i x_i(t^k) + \left(\sum_{i=1}^{n} \alpha_i \right) (t^{k+1} - t^k)$$

and

$$t^{k+1} - t^k = \frac{1 - \sum_{i=1}^{n} \beta_i x_i(t^k)}{\sum_{i=1}^{n} \alpha_i} = \frac{\sum_{i=1}^{n} (1 - \beta_i) x_i(t^k)}{\sum_{i=1}^{n} \alpha_i};$$

substituting this expression into (3), we have that

$$x(t^{k+1}) = A x(t^k), \tag{4}$$

where

$$A = \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_n \end{bmatrix} + \frac{1}{\sum_{j=1}^{n} \alpha_i} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \begin{bmatrix} 1 - \beta_1, & \ldots & , 1 - \beta_n \end{bmatrix}. \tag{5}$$

The following facts are deducible from the above explicit form of A (see [5]):

(a) (Synchronized) AIMD systems have a unique fixed point.

(b) The vector $x(t^k)$ asymptotically approaches the ray generated by the vector $[\frac{\alpha_1}{1-\beta_1}, \ldots, \frac{\alpha_n}{1-\beta_n}]^T$; convergence is geometric at a rate that equals the second largest modulus of an eigenvalue of the matrix A.

(c) By renumbering the users in decreasing order of their multiplicative factor β_i, that is, having $0 < \beta_n \leq \cdots \leq \beta_2 \leq \beta_1 < 1$, the second largest eigenvalue of A is bounded below by β_2 and above by β_1.

Nonlinear AIMD congestion control. We next describe a nonlinear variant of the basic AIMD algorithm, which we call the NAIMD. Specifically, NAIMD coincides with the standard (linear) AIMD, except that in the AI phase the increase in the utilization-rate of each user i is dictated by a nonlinear function of time that we denote by $a_i(.)$. So, (2) has to be modified by replacing the (multiplying constant) α_i by (the function) a_i, while (1) remains unchanged. The evolution of the utilization-rate of user i between the kth and $(k+1)$st congestion points is then given by

$$x_i(t^{k+1}) = \beta_i x_i(t^k) + a_i(t^{k+1} - t^k), \tag{6}$$

which replaces (3).

It has been recently shown by several authors that some choices of the $a_i(\cdot)$ lead to poor dynamic properties, including the lack of stable utilization-rates; see [4, 11]). The analysis presented here, together with the analysis and modeling framework in [12], constitutes an important first step in modeling high-speed networks that employ AIMD-like protocols. In [12] the authors use a "product of matrices" approach to establish conditions that guarantee the stability of the network where the growth functions are functions of the current state. Here we use contraction mapping arguments to establish the important result of unconditional stability of the network when the growth functions are functions of time since the last notification of congestion.

In particular, in the current paper we restrict our attention to the application of NAIMD with functions $a_i(.)$ that are nondecreasing. A number of authors have already shown that corresponding strategies offer an attractive framework for designing congestion control that is suitable for deployment in high-speed networks [13]. As in the linear case, one may consider the NAIMD only at the congestion points over which it defines an operator that maps the utilization-rates between consecutive points. Our principal contribution is the introduction of a contraction mapping approach to show that this class of operators is well behaved and has properties that are parallel to those of standard (linear) AIMD.

The next example illustrates the evolution of the utilization-rates in a 2-user system in which NAIMD is applied and to which our forthcoming results apply.

Example 1. Consider a system with two users that apply NAIMD with $\beta_1 = 0.25, \beta_2 = 0.5, a_1(t) = t$, and $a_2(t) = t^2$ for all $t \geq 0$. The trajectory of the utilization-rates in this system is depicted in Figure 3.

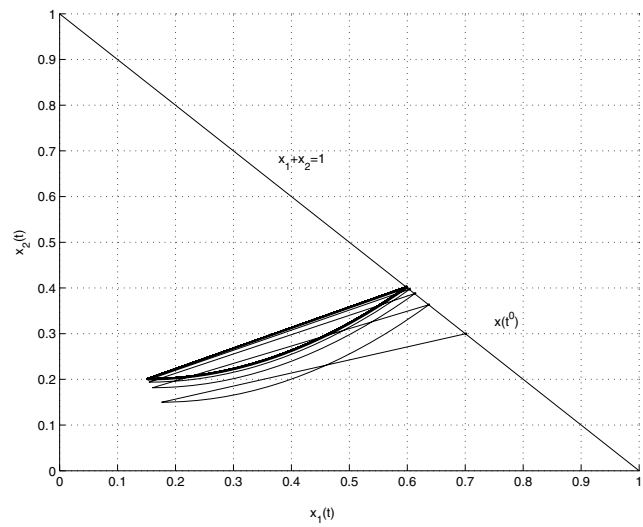


FIG. 3. *Utilization-rates under NAIMD.*

3. Notation and preliminaries. Throughout, we use notation $\mathbb{R}$ for the set of real numbers, $\mathbb{R}^n$ for the n-dimensional real Euclidean space, and $\mathbb{R}^{n\times n}$ for the space of $n \times n$ matrices with real entries. As usual, subscripts are used to denote coordinates of vectors and matrices, the notation $\leq$ and $<$ is used, respectively, for the coordinatewise weak and strict order over vectors and matrices, and $||\,.\,||_1$ denotes the ℓ_1 norm on $\mathbb{R}^n$. Given a vector β in $\mathbb{R}^n$, we use the notation D_β for the diagonal $n \times n$ matrix, whose diagonal elements are $\beta_1, \ldots, \beta_n$. Given two vectors u and v in $\mathbb{R}^n$, their *Hadamard product* $(u_1v_1, \ldots, u_nv_n)^T = D_u v \in \mathbb{R}^n$ will be denoted $u \circ v$. Given a convex set $C \subseteq \mathbb{R}^n$, its *tangential hull* $\{\alpha(x - y) : x, y \in C \text{ and } \alpha \in \mathbb{R}\}$ will be denoted $\text{tng}(C)$ (it is the smallest hyperplane containing C, translated to the origin).

With n as a given positive integer, we let e^i denote the ith unit vector and we let e denote the vector $(1, \ldots, 1)^T \in \mathbb{R}^n$. Also, we let S denote the *unit simplex* $\{x \in \mathbb{R}^n : x \geq 0\,, \sum_{i=1}^n x_i = 1\}$; in particular, $\text{tng}\,(S) = \{z \in \mathbb{R}^n : e^T z = 0\}$.

Given an operator $\mathcal{A}$ on a compact subset C of $\mathbb{R}^n$, we denote the ℓ_1 *operator*

norm of $\mathcal{A}$ on C by $||\mathcal{A}||_1^C$, that is,

$$||\mathcal{A}||_1^C = \max_{x,y\in C, x\neq y} \frac{||\mathcal{A}(x)-\mathcal{A}(y)||_1}{||x-y||_1}. \tag{7}$$

When C is convex and $\mathcal{A}$ is a linear operator represented by a matrix A (that is, $\mathcal{A}(x) = Ax$ for each $x \in C$), we have that

$$||\mathcal{A}||_1^C = \max_{z\in \mathrm{tng}(C), z\neq 0} \frac{||Az||_1}{||z||_1} = \max_{z\in \mathrm{tng}(C), ||z||_1=1} ||Az||_1. \tag{8}$$

We say that $\mathcal{A}$ is a *contraction on C with respect to the ℓ_1* norm if $||\mathcal{A}||_1^C < 1$. In this case, the Banach fixed point theorem assures us that $\mathcal{A}$ has a unique fixed point, say, x^*, and that the iterates of $\mathcal{A}$ on any arbitrarily selected point $x \in C$ converge geometrically to x^* at a rate that is bounded by $\gamma \equiv ||\mathcal{A}||_1^C$; specifically, we have that

$$||\mathcal{A}^n(x) - x^*||_1 \leq \frac{\gamma^n}{1-\gamma}||x - Ax||_1 \quad \text{for } n = 0, 1, \ldots. \tag{9}$$

To verify (9) with $n = 0$ observe that $||x - x^*||_1 \leq ||x - Ax||_1 + ||Ax - Ax^*||_1 \leq ||x-Ax||_1+\gamma||x-x^*||_1$, implying that $||x-x^*||_1 \leq (1-\gamma)^{-1}||x-Ax||_1$. For arbitrary $n \geq 0$, we then have that $||\mathcal{A}^n(x) - x^*||_1 = ||\mathcal{A}^n(x) - \mathcal{A}^n(x^*)||_1 \leq \gamma^n||x - x^*||_1 \leq \frac{\gamma^n}{1-\gamma}||x - Ax||_1$. The significance of (9) is in the fact that it allows one to bound the proximity to the unique fixed point of iterates of $\mathcal{A}$ on an arbitrary point in terms of the computable quantity $(1 - \gamma)^{-1}||x - Ax||_1$ and the geometrically decreasing term γ^n.

4. Operator-formulation. As stated in the preamble, we assume throughout that the capacity of the resource is normalized to 1. Our first undertaking is to formally model the evolution of the utilization-rates in an n-user system that is governed by the nonstandard (that is, not necessarily linear) variant of AIMD with transition over each cycle expressed by (6). The data we have at hand is then a vector $\beta \in \mathbb{R}^n$ and functions $a_i(\cdot) : [0, \infty) \to [0, \infty)$ for $i = 1, \ldots, n$ that are assumed to have the following properties:

(i) $0 < \beta < e$.
(ii) For $i = 1, \ldots, n$, $a_i(\cdot)$ is nondecreasing and continuous and has $a_i(0) = 0$.
(iii) $\sum_i a_i(\cdot)$ is strictly increasing and its range includes $[0, 1]$.

For each $t \in [0, \infty)$, let $a(t) = [a_1(t), \ldots, a_n(t)]^T$. Also, let $g(\cdot) : [0, \infty) \to [0, \infty)$ be the function that maps each $0 \leq t < \infty$ into $g(t) = e^T a(t)$. The assumptions about the $a_i(\cdot)$'s assure us that the function $g(\cdot)$ is strictly increasing and continuous and that its range includes $[0, 1]$. These properties of $g(\cdot)$ assure us that it has a continuous, strictly increasing inverse on $[0, 1]$, which we denote $g^{-1}(\cdot)$.

Consider a cycle k in which the initial utilization-rates of the users are $x_1, \ldots, x_n$, respectively, and let $x = (x_1, \ldots, x_n)^T$. As the total resource capacity is normalized to 1, we have that $x \in S$. Let t_x be the duration of a phase in which the initial utilization-rates are given by x. It then follows that, with

$$x_i = x_i(t^k) \text{ and } t_x = t^{k+1} - t^k,$$

by (6) the transition of the utilization-rate over a cycle which starts with utilization-rates represented by $x \in S$ is expressed by

$$\mathcal{A}(x) = (\beta \circ x) + a(t_x). \tag{10}$$

Notification of congestion is transmitted when the cumulative transmission-rate reaches the unit capacity, implying that

$$1 = \sum_{i=1}^{n} \mathcal{A}(x)_i = e^T(\beta \circ x) + e^T a(t_x).$$

So,

$$t_x = g^{-1}[1 - e^T(\beta \circ x)] \tag{11}$$

(the expression $g^{-1}[1-e^T(\beta\circ x)]$ is well defined because $0 < e^T(\beta\circ x) = \sum_{i=1}^{n} \beta_i x_i < 1$). Substituting (11) into (10), we conclude that

$$\mathcal{A}(x) = (\beta \circ x) + a\Big(g^{-1}[1 - e^T(\beta \circ x)]\Big). \tag{12}$$

Observing that for $x \in S$, $\mathcal{A}(x) \geq (\beta \circ x) \geq 0$ and

$$\begin{aligned} e^T\mathcal{A}(x) &= e^T(\beta \circ x) + e^T a\Big(g^{-1}[1 - e^T(\beta \circ x)]\Big) \\ &= e^T(\beta \circ x) + [1 - e^T(\beta \circ x)] = 1, \end{aligned}$$

we have that $\mathcal{A}$ maps S into S; that is, $\mathcal{A}$ is an operator on S.

The operator $\mathcal{A}$ expresses the transitions of the utilization-rates over the phases AIMD. Specifically, denote the vector of utilization-rates at the beginning of phase $k = 1, 2, \ldots$ by x^k. We then have from (6) and the above derivation that

$$x^{k+1} = \mathcal{A}(x^k). \tag{13}$$

We next study the operator $\mathcal{A}$ with the goal of understanding the evolution of the utilization-rate when NAIMD is followed. In particular, we will establish that $\mathcal{A}$ is a contraction with respect to the ℓ_1 norm on S, ensuring that $\mathcal{A}$ has a stable unique fixed point and that iterates of $\mathcal{A}$ applied to an arbitrary initial point converge to that fixed point.

5. Main results.

5.1. Existence and uniqueness of fixed points.

Our first result shows that the operator $\mathcal{A}$ has a unique fixed point.

THEOREM 5.1. *There exists a unique $t^* > 0$ satisfying $e^T(I - D_\beta)^{-1}a(t^*) = 1$; further, for this t^*,*

$$x^* \equiv (I - D_\beta)^{-1}a(t^*) \tag{14}$$

is a unique fixed point of the operator $\mathcal{A}$.

Proof. Evidently, a vector $\bar{x} \in S$ is a fixed point of the operator $\mathcal{A}$ defined by (10) if and only if

$$(I - D_\beta)\bar{x} = \bar{x} - (\beta \circ \bar{x}) = a(t_{\bar{x}}) \tag{15}$$

(with $t_{\bar{x}}$ defined by (11)). As $0 < \beta < e$, $I - D_\beta$ is invertible and its inverse is the diagonal matrix whose diagonal elements are $(1-\beta_1)^{-1}, \ldots, (1-\beta_n)^{-1}$; in particular, $(I - D_\beta)^{-1} = \sum_{k=0}^{\infty} D_\beta^k \geq I$ and (15) is equivalent to

$$\bar{x} = (I - D_\beta)^{-1}a(t_{\bar{x}}). \tag{16}$$

As $e^T x = 1$ for each $x \in S$, we conclude that if $\bar{x}$ is a fixed point of $\mathcal{A}$, then $\bar{x}$ must satisfy (16) and

$$e^T (I - D_\beta)^{-1} a(t_{\bar{x}}) = 1. \tag{17}$$

We next construct a point $x^* \in S$ that uniquely satisfies (16); it will then follow from the above paragraph that x^* is a unique fixed point of $\mathcal{A}$. Our construction is motivated by (17).

Let $h(\cdot) : [0, \infty) \to [0, \infty)$ be the real-valued function mapping each $t \geq 0$ into $h(t) = e^T (I - D_\beta)^{-1} a(t)$. The properties of $a(\cdot)$ assure us that $h(\cdot)$ is continuous, $h(0) = 0$, and for all $t > t' > 0$, $h(t) = e^T (I - D_\beta)^{-1} a(t) \geq e^T a(t) = g(t) > 0$ and $h(t) - h(t') = e^T (I - D_\beta)^{-1} [a(t) - a(t')] \geq e^T [a(t) - a(t')] > 0$. So, $h(\cdot)$ is continuous and strictly increasing and its domain includes $[0, 1]$ (recall that the domain of $g(\cdot)$ includes $[0, 1]$). These properties of $h(\cdot)$ assure us that there is a unique $t^* > 0$ satisfying $h(t^*) = 1$. With x^* given by (14), we then have that $x^* = (I - D_\beta)^{-1} a(t^*) \geq a(t^*) \geq 0$ and $e^T x^* = e^T (I - D_\beta)^{-1} a(t^*) = 1$, so $x^* \in S$. Further,

$$g(t^*) = e^T a(t^*) = e^T (I - D_\beta) x^* = e^T x^* - e^T D_\beta x^* = 1 - e^T (\beta \circ x^*),$$

implying that $t^* = g^{-1}[1 - e^T (\beta \circ x^*)] = t_{x^*}$. Substituting $t^* = t_{x^*}$ into (14) implies that $x^* = (I - D_\beta)^{-1} a(t_{x^*})$, verifying that $x^* \in S$ satisfies (16)—the characterizing condition for fixed points of $\mathcal{A}$. We conclude that x^* is a fixed point of $\mathcal{A}$. Further, (17) shows that any fixed point $\bar{x}$ of $\mathcal{A}$ must satisfy $h(t_{\bar{x}}) = 1 = h(t^*)$; as $h(\cdot)$ is strictly increasing, we then have that $t_{\bar{x}} = t^*$ and (16) implies that $\bar{x} = (I - D_\beta)^{-1} a(t_{\bar{x}}) = (I - D_\beta)^{-1} a(t^*) = x^*$. □

Theorem 5.1 and its proof do not show that the operator $\mathcal{A}$ is a contraction; consequently, they do not imply that the unique fixed point of $\mathcal{A}$ has the useful properties of fixed points of contractions, i.e., successive approximation and stability. However, (14) provides an explicit representation of the unique fixed point of $\mathcal{A}$ in terms of a (unique) solution of the equation $e^T (I - D_\beta)^{-1} a(t) = 1$, generally not available from contraction arguments. Further, we observe that as the function mapping each nonnegative t into $e^T (I - D_\beta)^{-1} a(t)$ is strictly increasing, a solution to $e^T (I - D_\beta)^{-1} a(t) = 1$ can be approximated by bisection, yielding an efficient computational method to approximate x^*.

We next illustrate the unique fixed point of the system described in Example 1.

Example 1 (continued). For the data of Example 1, $e^T (I - D_\beta)^{-1} a(t) = \frac{4}{3} t + 2t^2$ for each $t \geq 0$, $t^* = .45$, and $x(t^*)^T = (\frac{4}{3} t^*, 2(t^*)^2) = (.6, .4)$; the latter is consistent with Figure 3.

5.2. Contractions.

A. The linear case. In the linear case we have a vector $a \in \mathbb{R}^n \setminus \{0\}$ satisfying $a \geq 0$ such that $a(t) = at$ for all $t \geq 0$. In this case, $g(t) = e^T at$ for each $t \geq 0$ and $g^{-1}(y) = y / e^T a$ for each $0 \leq y \leq 1$, so (12) yields the following representation for the operation of $\mathcal{A}$ on $x \in S$:

$$\begin{aligned} \mathcal{A}(x) &= \beta \circ x + a \frac{1 - e^T (\beta \circ x)}{e^T a} = D_\beta x + a \frac{e^T x - e^T D_\beta x}{e^T a} \\ &= \left[D_\beta + \frac{a e^T (I - D_\beta)}{e^T a} \right] x \end{aligned} \tag{18}$$

(in addition to (12), the above uses the facts that $\beta \circ x = D_\beta x$ and $e^T x = 1$ for each

$x \in S$). Equation (18) shows that $\mathcal{A}$ is linear on S, represented by the matrix $[D_\beta + \frac{ae^T(I-D_\beta)}{e^T a}] \in \mathbb{R}^{n\times n}$. In particular, we see that $\mathcal{A}$ is invariant under the multiplication of a by a positive scalar. Thus, without loss of generality, we can and will assume that $e^T a = 1$, in which case (18) simplifies to

$$\mathcal{A}(x) = [D_\beta + ae^T(I - D_\beta)]x. \tag{19}$$

Let

$$A = [D_\beta + ae^T(I - D_\beta)]. \tag{20}$$

As $e^T a = 1$,

$$e^T A = e^T[D_\beta + ae^T(I - D_\beta)] = e^T D_\beta + e^T ae^T - e^T ae^T D_\beta = e^T,$$

and as $0 < \beta < e$, A is nonnegative. So, A is column-stochastic. In particular, its Perron–Frobenius eigenvalue is 1 with e^T as a corresponding left-eigenvector.

Our next goal is to show that $\mathcal{A}$ is a contraction on S with respect to the ℓ_1 norm; in fact, we derive an explicit bound on the ℓ_1 operator norm of $\mathcal{A}$. It will be convenient to enumerate the indices so that

$$1 > \beta_1 \geq \beta_2 \geq \cdots \geq \beta_n > 0. \tag{21}$$

THEOREM 5.2. *Let $a \in S$ and $\beta \in \mathbb{R}^n$ satisfy* (21). *Then*

$$||\mathcal{A}||_1^S = \beta_1(1 - a_1) + \beta_2 a_1 \leq \beta_1 < 1. \tag{22}$$

Proof. As $\mathcal{A}$ is linear with representing $n \times n$ matrix $[D_\beta + ae^T(I - D_\beta)]$ (see (19)), we have from (8) that

$$||\mathcal{A}||_1^S = \max_{z\in \mathrm{tng}(S), ||z||_1=1} ||[D_\beta + ae^T(I - D_\beta)]z||_1. \tag{23}$$

The right-hand side of (23) is the maximum of a convex function of z over the polytope $\{z \in \mathbb{R}^n : e^T z = 0, ||z||_1 = 1\}$ and is attained at one of the vertices of that polytope; these vertices have the representation $\frac{1}{2}(e^i - e^j)$, where $i, j = 1, \ldots, n$ and $i \neq j$. Thus,

$$||\mathcal{A}||_1^S = \max_{1\leq j<i\leq n} \frac{1}{2}||[D_\beta + ae^T(I - D_\beta)](e^i - e^j)||_1. \tag{24}$$

We observe that for $s, u = 1, \ldots, n$,

$$\left\{[D_\beta + ae^T(I - D_\beta)]e^s\right\}_u = \begin{cases} \beta_s + a_s(1 - \beta_s) & \text{if } u = s, \\ a_u(1 - \beta_s) & \text{if } u \neq s\ . \end{cases}$$

Recalling (21) and the normalization condition $e^T a = 1$, we conclude that for

$i, j = 1, \ldots, n$ and $i > j$ we have that $\beta_j \geq \beta_i$ and

$$\begin{aligned}
&\left\| [D_\beta + ae^T(I - D_\beta)] \Big(\frac{e^i - e^j}{2} \Big) \right\|_1 \\
&\quad = \frac{1}{2} \Big(|\beta_i + a_i(1 - \beta_i) - a_i(1 - \beta_j)| + |\beta_j + a_j(1 - \beta_j) - a_j(1 - \beta_i)| \\
&\qquad + \sum_{u \neq i,j} |a_u(1 - \beta_i) - a_u(1 - \beta_j)| \Big) \\
&\quad = \frac{1}{2} \Big(|\beta_i + a_i(\beta_j - \beta_i)| + |\beta_j(1 - a_j) + a_j\beta_i| + \sum_{u \neq i,j} |a_u(\beta_j - \beta_i)| \Big). \\
&\quad = \frac{1}{2} \Big([\beta_i + a_i(\beta_j - \beta_i)] + [\beta_j(1 - a_j) + a_j\beta_i] + \sum_{u \neq i,j} a_u(\beta_j - \beta_i) \Big) \\
&\quad = \frac{1}{2} \Big(\beta_i + (1 - a_j)(\beta_j - \beta_i) + [\beta_j(1 - a_j) + a_j\beta_i] \Big) \\
&\quad = \beta_j(1 - a_j) + \beta_i a_j. \qquad (25)
\end{aligned}$$

In particular, the right-hand side of (25) lies between β_j and β_i. It now follows from (21) that the maximum on the right-hand side of (24) equals $\beta_1(1 - a_1) + \beta_2 a_1$, attained when $j = 1$ and $i = 2$. Finally, the inequalities on the right-hand side of (22) are immediate from (21). □

The following immediate corollary of Theorem 5.2 demonstrates that (synchronized) AIMD results in a unique stable point to which the system converges geometrically from all potential starting points.

COROLLARY 5.3. *In the linear case, $\mathcal{A}$ has a unique fixed point x^* to which the iterates of $\mathcal{A}$ on any arbitrarily selected point in S converge geometrically; in particular,* (9) *is satisfied for every $x \in S$.*

Proof. The conclusions follow from Theorem 5.2 and the standard properties of contractions. □

Let $\tau(A)$ denote the second largest modulus of A's eigenvalues. Theorem 5.2 yields the following bound on $\tau(A)$.

COROLLARY 5.4.

$$\tau(A) \leq ||\mathcal{A}||_1^S = \beta_1(1 - a_1) + \beta_2 a_1 \leq \beta_1 < 1. \qquad (26)$$

Proof. As the matrix A defined by (20) is column-stochastic, the main result of [23] implies that for every norm $|| \cdot ||$, the matrix norm of A over the intersection of the $|| \cdot ||$-unit ball and the hyperplane $\{z \in R^n : e^T z = 0\}$ is an upper bound on $\tau(A)$. As $\text{tng}(S) = \{z \in R^n : e^T z = 0\}$, we have that the right-hand side of (23) is a specification of this bound when $|| \cdot ||$ is the ℓ_1 norm, yielding the first inequality of (26). The remaining parts of (26) follow from Theorem 5.2. □

The dynamics of network-transmission under the (synchronized) linear version of AIMD was studied in [5]. The principal contribution of that paper was showing that the matrix $A' \equiv [D_\beta + \frac{ae^T(I - D_\beta)}{e^T a}]$ is diagonally similar to the sum of a real symmetric matrix and a real rank-1 perturbation. The following facts are then easily deduced:

(a) A' is diagonally similar to a (real) positive diagonal matrix.
(b) Except for the Perron eigenvalue, all of the eigenvalues of A' lie in the interval $[\beta_n, \beta_1]$.
(c) If all the β_i's are distinct and $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A', then $0 < \beta_n < \lambda_{n-1} < \beta_{n-2} < \cdots < \lambda_2 < \beta_1 < \lambda_1 = 1$.

(d) $\beta_2 \le \tau(A) \le \beta_1$ (and in the inequalities are strict when the β_i's are distinct).

Corollary 5.4 established that $\tau(A) \le \beta_1(1-a_1) + \beta_2 a_1$; when $\beta_1 \ne \beta_2$, this is a tighter upper bound on $\tau(A)$ than is available from (d). The following example demonstrates the fact that we derive an improved bound.

Example 2. Consider a system with three users that are governed by standard (linear) AIMD with $\beta_1 = 0.8, \beta_2 = 0.7, \beta_3 = 0.2$, and $\alpha_1 = \alpha_2 = \alpha_2 = 1$. From (5), the matrix A is then given by

$$A = \begin{bmatrix} 0.8 & 0 & 0 \\ 0 & 0.7 & 0 \\ 0 & 0 & 0.5 \end{bmatrix} + \frac{1}{3}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} [0.2 \;\; 0.3 \;\; 0.8].$$

From the results in [5] stated in (d) we have that the eigenvalue of A with the second largest modulus, say λ_2, satisfies $0.7 \le \lambda_2 \le 0.8$. From Corollary 5.4 we have the sharper upper bound:

$$\lambda_2 \le \beta_1(1-a_1) + \beta_2 a_1 = 0.7667.$$

The eigenvalues of A are $1.0000, 0.7549$, and 0.5785, respectively, demonstrating the virtue of the bound on λ_2 derived in Theorem 5.2 over the bound from (d).

We conclude the discussion of the linear version of AIMD with a specification of Theorem 5.2 to the case where the vector a is a unit vector—a case that will prove useful in analyzing networks under the nonlinear variant of AIMD.

COROLLARY 5.5. *If $a = e^i$ for some $i = 1, \ldots, n$ and $\beta \in \mathbb{R}^n$ satisfies* (21)*, then $||\mathcal{A}||_1^S = \beta_1$ when $i = 1$ and $||\mathcal{A}||_1^S = \beta_1$ when $i \ne 1$.*

B. The general (nonlinear) case. We next consider the general AIMD case in which $a(\cdot)$ is not necessarily linear, and we use our earlier results for the linear case to demonstrate that, in the more general case as well, the operator $\mathcal{A}$ is a contraction on S.

Given a vector $a \in \mathbb{R}^n$, we will use the notation $\mathcal{A}^a$ for the linear operator corresponding to the case where $a(t) = at$ for each $t \ge 0$.

THEOREM 5.6. *The operator $\mathcal{A}$ satisfies*

$$||\mathcal{A}||_1^S \le \max_{i=1,\ldots,n} ||\mathcal{A}^{e^i}||_1^S = \beta_1 < 1. \tag{27}$$

Proof. Consider any two points $x, y \in S$. Without loss of generality, assume that $e^T(\beta \circ x) \le e^T(\beta \circ y)$. As g^{-1} is (strictly) increasing, it follows that

$$t_x = g^{-1}[1 - e^T(\beta \circ x)] \ge g^{-1}[1 - e^T(\beta \circ y)] = t_y. \tag{28}$$

As $a(\cdot)$ is nondecreasing, we conclude that $a(t_x) \ge a(t_y)$, and therefore

$$||a(t_x) - a(t_y)||_1 = e^T[a(t_x) - a(t_y)] = g(t_x) - g(t_y) = e^T(\beta \circ y) - e^T(\beta \circ x). \tag{29}$$

We next observe (from (10)) that

$$\mathcal{A}(x) - \mathcal{A}(y) = \beta \circ (x - y) + a(t_x) - a(t_y). \tag{30}$$

From (30) and (29) we conclude that

$$\begin{aligned} ||\mathcal{A}(x) - \mathcal{A}(y)||_1 &\le ||\beta \circ (x-y)||_1 + ||a(t_x) - a(t_y)||_1 \\ &= ||\beta \circ (x-y)||_1 + e^T(\beta \circ y) - e^T(\beta \circ x). \end{aligned} \tag{31}$$

As $\sum_i x_i = \sum_i y_i = 1$, there exists an index i^* with $x_{i^*} \geq y_{i^*}$. We next consider the linear case determined by the vector $a = e^{i^*}$ and the vector β unchanged. Specifically, we let $\mathcal{A}^{e^{i^*}}$ be the linear operator that corresponds to the matrix $[D_\beta + e^{i^*}e^T(I - D_\beta)]$. Let t_x^* and t_y^* be defined by (11) with respect to this (linear) case. As $e^T e^{i^*} = 1$, we have that $t_x^* = 1 - e^T(\beta \circ x)$, $t_y^* = 1 - e^T(\beta \circ y)$, and

$$t_x^* - t_y^* = e^T(\beta \circ y) - e^T(\beta \circ x) \geq 0. \tag{32}$$

Similar to the derivation of (30), we get that

$$\mathcal{A}^*(x) - \mathcal{A}^*(y) = \beta \circ (x - y) + e^{i^*}(t_x^* - t_y^*). \tag{33}$$

As $x_{i^*} \geq y_{i^*}$ and $t_x^* \geq t_y^*$, we conclude from (33), (32), and (31) that

$$\begin{aligned} ||\mathcal{A}^*(x) - \mathcal{A}^*(y)||_1 &= [\beta_{i^*}(x_{i^*} - y_{i^*}) + (t_x^* - t_y^*)] + \sum_{u \neq i^*} |\beta_u(x_u - y_u)| \\ &= \sum_u |\beta_u(x_u - y_u)| + (t_x^* - t_y^*) \\ &= ||\beta \circ (x - y)||_1 + e^T(\beta \circ y) - e^T(\beta \circ x) \\ &\geq ||\mathcal{A}(x) - \mathcal{A}(y)||_1, \end{aligned} \tag{34}$$

implying that

$$||\mathcal{A}(x) - \mathcal{A}(y)||_1^S \leq \max_i ||\mathcal{A}^{e^i}||_1^S = \beta_1 \tag{35}$$

(the last equality following from Corollary 5.5). ∎

Equation (27) demonstrates that β_1 is a bound on the contraction coefficient of $\mathcal{A}$—it is remarkable that this bound is independent of the $a_i(\cdot)$'s.

As for the linear case, we get the following immediate corollary of Theorem 5.6—it demonstrates that (synchronized) NAIMD results is a unique stable point to which the system converges geometrically from every potential starting point.

COROLLARY 5.7. *In the general (nonlinear) case, $\mathcal{A}$ has a unique fixed point x^* to which the iterates of $\mathcal{A}$ on any potential starting point converge geometrically; in particular,* (9) *is satisfied for every $x \in S$.*

Proof. The conclusions follow from Theorem 5.6 and the standard properties of contractions. ∎

6. Conclusions. In this paper we have presented, for the first time, a proof of stability for a class of NAIMD algorithms. Bounds on the rate of convergence are given for these algorithms and simulation results are given to illustrate the efficacy of our results. Future work will report on the behavior of unsynchronized networks. In this context the recent results presented in [30, 31] are likely to prove useful.

Acknowledgments. The authors are grateful to Fabian Wirth for constructive comments. RS also thanks Chris King, Douglas Leith, and Mehmet Akar for discussions related to the material in this paper.

REFERENCES

[1] F. BACCELLI AND D. HONG, *AIMD, fairness and fractal scaling of TCP traffic*, in Proceedings of the 21st Annual IEEE Conference on Computer Communications (INFOCOM), IEEE Computer Society, Los Alamitos, CA, 2002, pp. 229–238.

[2] E. Altman, E. El-Azouzi, D. Ros, and B. Tuffin, *Loss policies for competing TCP/IP connections*, Comput. Networks, 50 (2006), pp. 1799–1815.

[3] E. Altman, D. Barman, B. Tuffin, and M. Vojnovic, *Parallel TCP sockets: Simple model, throughput and validation*, in Proceedings of the 25th Annual IEEE Conference on Computer Communications (INFOCOM), IEEE Computer Society, Los Alamitos, CA, 2006, pp. 1–12.

[4] E. Altman, K. Avrachenkov, and B. Prabhu, *Fairness in MIMD congestion control algorithms*, in Proceedings of the 24th Annual IEEE Conference on Computer Communications (INFOCOM), IEEE Computer Society, Los Alamitos, CA, 2005, pp. 1350–1361.

[5] A. Berman, R. Shorten, and D. Leith, *Positive matrices associated with synchronised communication networks*, Linear Algebra Appl., 393 (2004), pp. 47–54.

[6] D. Chiu and R. Jain, *Analysis of the increase/decrease algorithms for congestion avoidance in computer networks*, J. Comput. Networks, 17 (1989), pp. 1–14.

[7] S. Floyd, *HighSpeed TCP for Large Congestion Windows*, Paper IETF RFC 3649, Experimental, ICSI Center for Internet Research, Berkeley, CA, 2003. Available online at http://www.icir.org/floyd/hstep.html.

[8] Z. Hhoa, S. Darbha, and A. Reddy, *A method for estimating the proportion of nonresponsive traffic at a router*, IEEE/ACM Trans. Networking, 12 (2004), pp. 708–718.

[9] S. Jin, L. Guo, I. Matta, and A. Bestavros, *A Spectrum of TCP-Friendly Window-Based Congestion Control Algorithms*, Technical report TR-2001-015, Boston University, Boston, MA, 2001.

[10] F. Kelly, A. Maulloo, and D. Tan, *Rate control in communication networks: Shadow prices, proportional fairness, and stability*, J. Oper. Res. Soc., 49 (1998), pp. 237–252.

[11] T. Kelly, *On Engineering a Stable and Scalable TCP Variant*, Technical report CUED/F-INFENG/TR.435, Engineering Department, Cambridge University, Cambridge, UK, 2002.

[12] C. King, R. Shorten, F. Wirth, and M. Akar, *Growth conditions for the global stability of highspeed communication networks,* IEEE Trans. Automat. Control, to appear.

[13] D. J. Leith and R. N. Shorten, *H-TCP protocol for high-speed long-distance networks*, in Proceedings of the 2nd Workshop on Protocols for Fast Long Distance Networks, Argonne, Canada, 2004.

[14] D. J. Leith and R. N. Shorten, *H-TCP: TCP Congestion Control for High Bandwidth-Delay Product Path*, IETF Internet Draft, Internet Engineering Task Force, 2005. Available online at http://www.hamilton.ie/net/draft-leith-tcp-htcp-00.txt.

[15] E. Altman, K. Avratchenkov, C. Barakat, and R. Nunez Queija, *TCP modeling in the presence of nonlinear window growth*, in Proceedings of the ITC-17, Salvador da Bahia, Brazil, 2001.

[16] E. Altman, K. Avrachenkov, A. A. Kherani, and B. J. Prabhu, *Performance analysis and stochasitc stability of congestion control protocols*, in Proceedings of the 24th Annual IEEE Conference on Computer Communications (INFOCOM), IEEE Computer Society, Los Alamitos, CA, 2005, pp. 1316–1327.

[17] E. Altman, K. Avratchenkov, C. Barakat, and R. Nunez Queija, *State-Dependent M/G/*1 *Type Queueing Analysis for Congestion Control in Data Networks*, Report PNA-R0005, CWI, Amsterdam, 2000.

[18] A. Budhiraja, F. Hernandez-Campos, V. G. Kulkarni, and F. D. Smith, *Stochastic differential equation for TCP window size: Analysis and experimental validation*, Probab. Engrg. Inform. Sci., 18 (2004), pp. 111–140.

[19] A. Leizarowitz, R. Stanojevic, and R. Shorten, *Towards an analysis and design framework for communication networks with Markovian dynamics*, IEEE Proc. Control Theory Appl., 153 (2006), pp. 506–519.

[20] N. R. Sastry and S. S. Lam, *CYRF: A theory of window-based unicast congestion control*, IEEE/ACM Trans. Networking, 13 (2005), pp. 330–342.

[21] D. Bansal and H. Balakrishnan, *Binomial congestion control algorithms*, in Proceedings of the 20th Annual IEEE Conference on Computer Communications (INFOCOM), IEEE Computer Society, Los Alamitos, CA, 2001, pp. 631–640.

[22] U. G. Rothblum, *Explicit solutions to optimization problems on the intersections of the unit ball of the l_1 and l_∞ norms with a hyperplane*, SIAM J. Algebraic Discrete Methods, 5 (1984), pp. 619–632.

[23] U. G. Rothblum and C. P. Tan, *Upper bounds on the maximum modulus of subdominant eigenvalues of nonnegative matrices*, Linear Algebra Appl., 66 (1985), pp. 45–86.

[24] R. Shorten, D. J. Leith, J. Foy, and R. Kilduff, *Analysis and design of congestion control in synchronized communication networks*, Automatica, 41 (2005), pp. 725–730.

[25] R. Shorten, C. King, F. Wirth, and D. Leith, *Modelling TCP congestion control dynamics in drop-tail environments,* Automatica J. IFAC, 43 (2007), pp. 441–449.
[26] R. N. Shorten, F. Wirth, and D. J. Leith, *A positive systems model of TCP-like congestion control: Asymptotic results*, IEEE/ACM Trans. Networking, 14 (2006), pp. 616–629.
[27] R. Srikant, *The Mathematics of Internet Congestion Control*, Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 2004.
[28] L. Xu, K. Harfoush, and I. Rhee, *Binary increase congestion control for fast long-distance networks*, in Proceedings of the 23rd Annual IEEE Conference on Computer Communications (INFOCOM), IEEE Computer Society, Los Alamitos, CA, 2004, pp. 2414–2524.
[29] F. Wirth, R. Stanojević, R. Shorten, and D. Leith, *Stochastic equilibria of AIMD communication networks*, SIAM J. Matrix Anal., 28 (2006), pp. 703–723.
[30] G. Last, *Ergodicity properties of stress release, repairable system and workload models*, Adv. Appl. Probab., 36 (2004), pp. 471–498.
[31] O. Boxma, D. Perry, W. Stadje, and S. Zacks, *A Markovian growth-collapse model*, Adv. Appl. Probab., 38 (2006), pp. 221–243.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 5, pp. 1897–1921

BOLZA PROBLEMS WITH DISCONTINUOUS LAGRANGIANS AND LIPSCHITZ-CONTINUITY OF THE VALUE FUNCTION*

ANDREA DAVINI†

Abstract. We study the local Lipschitz-continuity of the value function v associated with a Bolza problem in the presence of a Lagrangian $L(x,q)$, convex and uniformly superlinear in q, but only Borel-measurable in x. Under these assumptions, the associated integral functional is not lower semicontinuous with respect to the suitable topology which ensures the existence of minimizers, so all results known in the literature fail to apply. Yet, the Lipschitz regularity of v does not depend on the existence of minimizers. In fact, it is enough to control the derivatives of quasi-minimal curves, but the problem is nontrivial due to the general growth conditions assumed here on $L(x,\cdot)$. We propose a new approach, based on suitable reparameterization arguments, to obtain suitable a priori estimates on the Lipschitz constants of quasi minimizers. As a consequence of our analysis, we derive the Lipschitz-continuity of v and a compactness result for value functions associated with sequences of locally equibounded discontinuous Lagrangians.

Key words. Bolza problems, value function, Hamilton–Jacobi equations

AMS subject classifications. 49J45, 49K99, 49J53

DOI. 10.1137/060654311

1. Introduction.

1.1. Description of the problem and main results.

A typical issue in partial differential equations is that of proving the local Lipschitz-continuity in $(0,+\infty)\times\mathbb{R}^N$ of the *value function*

$$v(t,x) := \inf\left\{ u(\gamma(0)) + \int_0^t L(s,\gamma(s),\dot\gamma(s))\,\mathrm{d}s \ :\ \gamma\in W^{1,1}\big([0,t],\mathbb{R}^N\big),\ \gamma(t)=x \right\}$$

associated with a Lagrangian $L:[0,+\infty)\times\mathbb{R}^N\times\mathbb{R}^N\to(-\infty,+\infty]$ and to a possibly discontinuous *initial cost* $u:\mathbb{R}^N\to(-\infty,+\infty]$. This is an important step when one is interested in showing that v is a solution, in a suitable generalized sense, of the equation

$$\partial_t u + H(x,Du) = 0 \qquad \text{in } (0,+\infty)\times\mathbb{R}^N, \tag{1}$$

where H is the Hamiltonian associated with L through the Fenchel transform. When L is continuous, it is well known that v is a solution of (1) in the *viscosity sense* (see, e.g., [3, 4, 26]). For discontinuous (and autonomous) Lagrangians, a PDE interpretation of the value function has been provided by Dal Maso and Frankowska in [18, 19]. By making use of the so-called contingent derivatives, these authors prove that v satisfies the Hamilton–Jacobi equation, in a suitable generalized sense, and characterize it as the unique solution of the associated Cauchy problem with initial datum u when the latter is lower semicontinuous.

The study of discontinuous Hamilton–Jacobi equations is a field which is gaining attention from the viewpoints of both theory and applications; see [6, 8, 10, 11, 12,

*Received by the editors March 14, 2006; accepted for publication (in revised form) April 2, 2007; published electronically November 21, 2007.
http://www.siam.org/journals/sicon/46-5/65431.html
†Dipartimento di Matematica, Università di Roma "La Sapienza," P.le Aldo Moro 2, 00185 Roma, Italy (davini@mat.uniroma1.it).

13, 28, 30, 33, 34]. It is related to the study of geodesic distances, some discontinuous control problems, combustion phenomena in nonhomogeneous media, and geometric optic propagation in the presence of layers; see [5, 23, 27]. The analysis we will develop here supplies the tools for proving representation formulas for generalized solutions of time-dependent measurable Hamilton–Jacobi equations in the spirit of [12]. This issue will be discussed in the forthcoming paper [9].

The Lipschitz-continuity of v is strictly related to the regularity of solutions to the *Bolza problem*,

$$\min\left\{u(\gamma(0)) + \int_0^t L(s,\gamma(s),\dot\gamma(s))\,\mathrm{d}s \;:\; \gamma \in W^{1,1}\big([0,t],\mathbb{R}^N\big),\ \gamma(t)=x\right\}, \tag{2}$$

and to the possibility of finding some a priori estimates on the Lipschitz constants of minimizers. Clearly, any solution of (2) is also a Lagrangian minimizer with respect to its boundary conditions.

The study of regularity properties of Lagrangian minimizers is a classical topic in the calculus of variations; see, for instance, [7, 15]. The first results were obtained by Tonelli in 1915 [35] and the early 1920s [36] for real-valued smooth Lagrangians $L(s,x,q)$, coercive and strictly convex in q. More recently, Tonelli's results have been generalized by Clarke and Vinter [17] to the case of measurable, locally bounded integrands $L(s,x,q)$ which are locally Lipschitz in x, convex, and uniformly superlinear in q. By using the tools of nonsmooth analysis, the classical Euler–Lagrange necessary condition is expressed in terms of a differential inclusion.

The autonomous case has been widely studied. The results of [17] have been extended by Ambrosio, Ascenzi, and Buttazzo in [2] to the case of a locally bounded Lagrangian $L(x,q)$, convex and uniformly superlinear in q. In [18], Dal Maso and Frankowska succeeded in proving the same results without assuming any convexity in q. They also obtained some uniform estimates on the Lipschitz constant of the minimizers, which are used to prove that the associated value function v is locally Lipschitz in $(0,+\infty)\times\mathbb{R}^N$, provided problem (2) admits solutions for every $(t,x)\in(0,+\infty)\times\mathbb{R}^N$.

Here we will be concerned with the case of a Borel-measurable Lagrangian $L:\mathbb{R}^N\times\mathbb{R}^N\to\mathbb{R}$, locally bounded with respect to (x,q), convex, and uniformly superlinear in q. The growth conditions assumed on L can be restated in the following equivalent form:

$$\alpha(|q|)\le L(x,q)\le\beta(x,|q|)\qquad\text{for every } (x,q)\in\mathbb{R}^N\times\mathbb{R}^N,$$

where $\alpha(\cdot)$ and $\beta(x,\cdot)$ are superlinear functions from $[0,+\infty)$ to $\mathbb{R}$, with β locally bounded on $\mathbb{R}^N\times[0,+\infty)$ (cf. Lemma 2.3). The model example of Lagrangians included in this class are of the form

$$L(x,q)=F(q)+n(x),$$

with $F(\cdot)$ convex and superlinear, and $n(\cdot)$ Borel-measurable and bounded.

The main result we prove is the local Lipschitz-continuity of the value function v in $(0,+\infty)\times\mathbb{R}^N$. Several Lipschitz-regularity results for the value function associated with a discontinuous Lagrangian, depending on the continuity properties enjoyed by the initial cost, are given in section 4. Moreover, a compactness result holding for sequences of value functions is derived in subsection 4.2 as a consequence of what was proved in [20] (cf. Theorem 3.21). This kind of result essentially relies on the fact that

all the Lipschitz estimates we provide do not depend explicitly on the Lagrangian but only on the way $L(x,q)$ grows when $|q| \to +\infty$, i.e., on the functions α, β.

We remark that all results of the paper hold, with the obvious changes of notation, if $\mathbb{R}^N$ is replaced by a connected, smooth Riemannian manifold $\mathcal{M}$ without boundary. Proofs can be rephrased by using local coordinates. When $\mathcal{M}$ is compact, some additional information on the Lipschitz continuity of the value function is deduced.

With respect to the literature quoted above, the key new point in this paper consists of dealing with a case when the minimizers of the Bolza problem do not exist in general. This gives rise to serious technical difficulties, since all arguments known in the literature exploit the existence of minimizers to derive information on their Lipschitz constants, and this in turn gives the desired regularity of v via a rather standard argument. The same reasoning, however, works as soon as we provide suitable integral estimates on the derivatives of quasi-optimal curves for $v(t,x)$ (i.e., curves that realize the value $v(t,x)$ in (2), up to an addition of a suitably small positive constant) depending with some uniformity on $(t,x) \in (0,+\infty) \times \mathbb{R}^N$. The problem is, however, nontrivial due to the general growth conditions assumed here on $L(x,\cdot)$, in particular to the fact that functions $\alpha(\cdot), \beta(x,\cdot)$ may have different growths for $|q| \to +\infty$.

The novelty of our approach relies on an unusual way of employing the DuBois–Raymond condition, which motivates the introduction of a distinct family of Lipschitz curves parameterized in a special way (cf. Definition 3.12). The core of our arguments consists of proving that, in the formula defining $v(t,x)$, it is not restrictive to consider only curves belonging to this family (see section 3). Once this is established, it is rather easy to obtain the a priori estimates on the Lipschitz constants of quasi-optimal curves for $v(t,x)$ that are needed to derive the desired regularity of the value function.

The analysis outlined above is carried out through suitable reparameterization techniques which use, in an essential way, the fact that L is autonomous and convex in q. The argument on which they are based was originally introduced in [22] and subsequently developed in [21] in the case of a continuous Lagrangian, but its use for the kind of problems studied herein seems new. A substantial effort is furthermore made to extend the techniques to the measurable setting and to gather the necessary information needed in the case at hand.

We end this discussion by mentioning that a possible alternative way to attack the problem would be to find a relaxed formulation of (2) in order to apply the results of [18]. The difficulty here is proving that the relaxation of the functional $\gamma \mapsto \int_0^t L(\gamma, \dot\gamma)\,\mathrm{d}s$ admits an integral representation on $W^{1,1}\big([0,t],\mathbb{R}^N\big)$. The results proved in [1] ensure that this approach actually works if the Lagrangian enjoys the following growth conditions in q:

$$|q|^p \le L(x,q) \le \Lambda\,(1+|q|^p) \qquad \text{for every } (x,q) \in \mathbb{R}^N \times \mathbb{R}^N$$

for some $p > 1$ and $\Lambda > 0$. If we were to extend such results to the more general cases considered here, we would encounter difficulties similar to the ones previously described. As a matter of fact, a technical adaptation of the arguments here employed allows us to generalize the results of [1] to a wider class of abstract and integral functionals of autonomous type that includes, in particular, the ones considered in this paper. This issue is specifically studied in [20].

1.2. Strategy of the proof. To study the problem, we find it convenient to introduce the function

$$S(y,x,t) := \inf\left\{\int_0^t L(\gamma,\dot\gamma)\,\mathrm{d}s \,:\, \gamma\in W^{1,1}\big([0,t],\mathbb{R}^N\big),\ \gamma(0)=y,\ \gamma(t)=x\right\} \tag{3}$$

defined for every $(y,x,t)\in\mathbb{R}^N\times\mathbb{R}^N\times(0,+\infty)$, and to express v in the following equivalent form:

$$v(t,x)=\inf_{y\in\mathbb{R}^N}\Big(u(y)+S(y,x,t)\Big)\qquad\text{for every } (t,x)\in(0,+\infty)\times\mathbb{R}^N,$$

where u is a function from $\mathbb{R}^N$ to $(-\infty,+\infty]$ which is either uniformly continuous or bounded from below. To get the required regularity of v, it is enough to prove that S is locally Lipschitz in $\mathbb{R}^N\times\mathbb{R}^N\times(0,+\infty)$. This would immediately follow from [18] if we were able to prove that minimizing curves for $S(y,x,t)$ exist for every $(y,x,t)\in\mathbb{R}^N\times\mathbb{R}^N\times(0,+\infty)$. Unfortunately, this need not be true in our case. In fact, the superlinearity of $L(x,\cdot)$ and the Dunford–Pettis theorem (see [7, Chapter 2]) actually imply that every minimizing sequence for $S(y,x,t)$ admits a subsequence uniformly converging to some limit curve, but the lack of continuity of L does not guarantee that the associated integral functional is lower semicontinuous for the convergence at hand (classical results by Olech [29] and Ioffe [25] ensure that this is true if the Lagrangian is lower semicontinuous in x and convex in q), so the standard *direct method of the calculus of variations* fails to apply (see [7]).

Yet, existence of minimizers is not necessary to derive the desired regularity for S. Indeed, a fairly standard argument (see, for instance, [18, Proof of Theorem 4.4]) shows that S is locally Lipschitz as soon as we provide some a priori estimates on the Lipschitz constants of quasi minimizers for $S(y,x,t)$, with some uniformity with respect to $(y,x,t)\in\mathbb{R}^N\times\mathbb{R}^N\times(0,+\infty)$.[1] When the existence of a minimizer γ is postulated, as in [18], or ensured by the assumptions made on L, as in [17, 19], these can be derived from the fact that γ satisfies the DuBois–Raymond necessary condition, namely, that there exists a constant $a\in\mathbb{R}$ such that

$$L(\gamma(s),\dot\gamma(s))=\langle\dot\gamma(s),p\rangle-a\qquad\text{for every } p\in\partial_q L(\gamma(s),\dot\gamma(s)) \tag{4}$$

for almost every $s\in[0,t]$. Using the superlinearity of $L(x,\cdot)$, it is then easy to show that a is locally bounded with respect to (y,x,t), and this provides the desired control on the Lipschitz constant of γ.

Even if this reasoning cannot be applied in our case, we nevertheless notice that condition (4), which is crucial for obtaining the desired estimates, only provides information on the parameterization of the curve: When γ is action-minimizing, its parameterization must obey an optimality condition.

The idea we develop here is to separate the issue of parameterization from that of minimizing the action. This is achieved by first considering a minimization problem

[1] An *ε-minimizer* for $S(y,x,t)$ is a curve $\gamma\in W^{1,1}\big([0,t],\mathbb{R}^N\big)$ with $\gamma(0)=y$, $\gamma(t)=x$ such that

$$\int_0^t L(\gamma(s),\dot\gamma(s))\,\mathrm{d}s < S(y,x,t)+\varepsilon.$$

We say that γ is a *quasi minimizer* or it is *quasi optimal* for $S(y,x,t)$ if it is an ε-minimizer with $\varepsilon>0$ suitably small.

with fixed support as follows: We fix a Lipschitz curve $\gamma : [0, \ell] \to \mathbb{R}^N$ parameterized by the arc-length, the *support*, and we try to solve the problem

$$\min \left\{ \int_0^t L(\xi, \dot{\xi})\, \mathrm{d}s \,:\, \xi \in [\gamma]_t \right\} \tag{5}$$

for every $t > 0$, where $[\gamma]_t$ denotes the family of absolutely continuous curves $\xi : [0, t] \to \mathbb{R}^N$ obtained through a reparameterization of γ.[2] Here, the crucial remark is the following: Any solution of (5) satisfies (4) for some $a \in \mathbb{R}$; conversely, any $\xi \in [\gamma]_t$ satisfying (4) for some $a \in \mathbb{R}$ is a solution to (5) (cf. the proof of Theorem 3.16).

However, the existence of minimizers of (5) is not clear. The idea exploited here is to introduce the notion of *a-Lagrangian parameterization* for a curve ξ (cf. Definition 3.12), which amounts to requiring that ξ satisfy (4). Then we consider the multifunction $T_\gamma(\cdot)$ defined on $\mathbb{R}$ by

$$T_\gamma(a) := \{t > 0 \,:\, [\gamma]^\flat(a, t) \text{ is nonempty}\} \qquad \text{for every } a \in \mathbb{R},$$

where $[\gamma]^\flat(a, t)$ denotes the subset of $[\gamma]_t$ consisting of a-Lagrangian bi-Lipschitz reparameterizations of γ, and we remark that, by what we previously observed, the relation $t \in T_\gamma(a)$ implies that problem (5) admits a solution in $[\gamma]^\flat(a, t)$. Our attention is then addressed to establishing the relevant properties of the multifunction $T_\gamma(\cdot)$, with particular interest in its range $\bigcup_{a\in\mathbb{R}} T_\gamma(a)$ (see Proposition 3.13). When this coincides with $(0, +\infty)$, we conclude that problem (5) is solvable for every $t > 0$. In particular, (5) has a minimizer belonging to $[\gamma]^\flat(a, t)$ for some $a \in \mathbb{R}$, and its Lipschitz constant can be estimated by some $\kappa_a \in \mathbb{R}$ only depending on a and on the kind of growth conditions assumed on L. However, our analysis reveals that the range of $T_\gamma(\cdot)$ may actually be a bounded interval of the form $(0, T)$. In this instance, a solution to (5) exists if $t \le T$. For $t > T$, the minimum in (5) is only an infimum, in general; nevertheless, we are able to prove that this value can be obtained by minimizing the action over the family of κ_{c_γ}-Lipschitzian reparameterizations of γ, where κ_{c_γ} is a positive constant that can be estimated in terms of the growth conditions assumed on L (see Theorem 3.16).

This information is used to obtain the needed priori estimates on the Lipschitz constants of quasi minimizers (see Lemma 3.2): Since any absolutely continuous curve from $[0, t]$ to $\mathbb{R}^N$ belongs to $[\gamma]_t$ for a suitable choice of the Lipschitz curve $\gamma : [0, \ell] \to \mathbb{R}^N$ (cf. Lemma 3.11), we can always assume that a quasi minimizer for $S(y, x, t)$ is κ_a-Lipschitz continuous for some $a \in \mathbb{R}$. By using the superlinearity of $L(x, \cdot)$, we see that the constant a is last estimated with some uniformity with respect to (y, x, t).

1.3. Plan of the article. Section 2 contains the main notation and assumptions, together with some well-known propositions that will be needed in the rest of the paper.

The properties of the function S are studied in section 3. In subsection 3.1 some preliminary results are collected. The definition of a-Lagrangian reparameterization and the reparameterization arguments are presented in subsection 3.2. Here, the main properties of the multifunction $T_\gamma(\cdot)$ are established and used to study an action-minimization problem with fixed support. The information gathered is then exploited to derive the required priori estimates on the Lipschitz constants of quasi minimizers

[2] That is, $\xi = \gamma \circ \varphi$ on $[0, t]$ for some absolutely continuous map $\varphi : [0, t] \to [0, \ell]$ surjective and nondecreasing (cf. Definition 3.9).

(cf. Lemma 3.2), which is all we need to prove Theorem 3.1. To simplify the exposition, the Lagrangian L is initially assumed locally bounded in q, uniformly with respect to x. The consequent extension to the case of Lagrangians locally bounded in (x, q) is easily derived in subsection 3.3 via a localization argument (see Theorem 3.19). Here a sequential compactness result for locally equibounded discontinuous Lagrangians, established in [20], is recalled for later use.

The main results of the paper are derived in section 4 as a simple application of the preceding analysis. Subsection 4.1 contains several Lipschitz-regularity results for the value function associated with a discontinuous Lagrangian, depending on the continuity properties assumed on the initial cost. An extension to the case when $\mathbb{R}^N$ is replaced by a compact and connected smooth Riemannian manifold $\mathcal{M}$ without boundary is also provided. Lastly, subsection 4.2 contains a compactness result for the value functions associated with sequences of locally equibounded discontinuous Lagrangians.

2. Notation and standing assumptions. We list below the symbols used throughout this paper:

N	an integer number
$B_r(x)$	the open ball in $\mathbb{R}^N$ of radius r centered at x
B_r	the open ball in $\mathbb{R}^N$ of radius r centered at 0
$\mathbb{S}^{N-1}$	the $(N-1)$-dimensional unitary sphere of $\mathbb{R}^N$
$\mathcal{H}^k$	the k-dimensional Hausdorff measure
$\langle \cdot, \cdot \rangle$	the scalar product in $\mathbb{R}^N$
$[u]$	the integer part of $u \in \mathbb{R}$
$\mathbb{R}_+$	the set of nonnegative real numbers
$\mathcal{P}(\mathbb{R}_+)$	the family of subsets of $\mathbb{R}_+$
$\mathrm{UC}(\mathbb{R}^N)$	the space of uniformly continuous real functions on $\mathbb{R}^N$
$\mathrm{Lip}(\mathbb{R}^N)$	the space of Lipschitz-continuous real functions on $\mathbb{R}^N$

Given a subset U of $\mathbb{R}^k$, we denote by $\overline{U}$ its closure. We furthermore say that U is *compactly contained* in a subset V of $\mathbb{R}^k$ if $\overline{U}$ is compact and contained in V. If E is a Lebesgue measurable subset of $\mathbb{R}^k$, we denote by $|E|$ its k-dimensional Lebesgue measure, and we say that E is *negligible* whenever $|E| = 0$. The characteristic function of E is denoted by χ_E. We say that a property holds *almost everywhere* (*a.e.*) on $\mathbb{R}^k$ if it holds up to a negligible subset of $\mathbb{R}^k$. The Euclidean norm of $u \in \mathbb{R}^k$ is denoted by $|u|$. Given a measurable vector-valued function $f : E \to \mathbb{R}^m$, we write $\|f\|_\infty$ to mean $(\sum_{i=1}^k \|f_i\|^2_{L^\infty(E)})^{1/2}$, where f_i and $\|f_i\|_{L^\infty(E)}$ denote the ith component of f and the L^∞-norm of f_i, respectively. Given $X \subseteq \mathbb{R}^k$, we will denote by $\mathcal{B}(X)$ the family of all Borel subsets of X. A multifunction Γ from X to compact subsets of $\mathbb{R}$ is said to be *Borel-measurable* (cf. [14]) if

$$\{x \in X : \Gamma(x) \cap U \neq \emptyset\} \in \mathcal{B} \qquad \text{for every open set } U \subseteq \mathbb{R}.$$

We say that Γ is *upper semicontinuous* at x if, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\Gamma(z) \subseteq \Gamma(x) + (-\varepsilon, \varepsilon) \qquad \text{for all } z \in B_\delta(x) \cap X.$$

When $k = 1$, we say that Γ is *nondecreasing on* X if

$$\sup \Gamma(x) \leq \inf \Gamma(y) \qquad \text{for every } x, y \in X \text{ with } x < y.$$

We say that Γ is *nonincreasing on* X if the multifunction $-\Gamma(\cdot)$ is nondecreasing on X. A function $g : \mathbb{R}^k \to (-\infty, +\infty]$ will be called *superlinear* if

$$\lim_{|x|\to+\infty} \frac{g(x)}{|x|} = +\infty.$$

For a convex function f from $\mathbb{R}^k$ to $\mathbb{R}$, we will denote by $\partial f(x)$ the *subdifferential* of f at x, defined as

$$\partial f(x) := \{p \in \mathbb{R}^k \,:\, f(y) \geq f(x) + \langle p, y - x\rangle \quad \text{for every } y \in \mathbb{R}^k\}.$$

The set $\partial f(x)$ is closed and convex, and the multifunction $x \mapsto \partial f(x)$ is upper semicontinuous on $\mathbb{R}^k$. We furthermore have the following (see [31]).

PROPOSITION 2.1. *Let $f : \mathbb{R}^k \to \mathbb{R}$ be convex. Then f is locally Lipschitz in $\mathbb{R}^k$. More precisely, for every $x_0 \in \mathbb{R}^k$ and $r, \delta > 0$, we have*

$$|f(x) - f(y)| \leq |x - y|\, \frac{2}{\delta} \sup_{B_{r+\delta}(x_0)} |f| \qquad \textit{for every } x, y \in B_r(x_0).$$

In particular, $\partial f(x) \subset (2 \sup_{B_{r+1}} |f|)\, \overline{B}_1$ for every $x \in B_r$.

Given a function $f : \mathbb{R}^k \to \mathbb{R}$, we define its conjugate $f^* : \mathbb{R}^k \to (-\infty, +\infty]$ as follows:

$$f^*(x) := \sup_{y\in\mathbb{R}^k} \{\langle x, y\rangle - f(y)\} \qquad \text{for every } x \in \mathbb{R}^k.$$

We record for later use the following well-known facts (cf. [31, Theorem 23.5]).

PROPOSITION 2.2. *Let $f : \mathbb{R}^k \to \mathbb{R}$ be superlinear and convex. Then f^* is locally bounded and convex on $\mathbb{R}^k$. Moreover,*

$$f(x) = f^{**}(x) := \sup_{y\in\mathbb{R}^k} \{\langle x, y\rangle - f^*(y)\} \qquad \textit{for every } x \in \mathbb{R}^k.$$

The following conditions on $x, x^* \in \mathbb{R}^k$ are equivalent to each other:

(i) $f(x) + f^*(x^*) \leq \langle x, x^*\rangle$;
(ii) $f(x) + f^*(x^*) = \langle x, x^*\rangle$;
(iii) $x^* \in \partial f(x)$;
(iv) $x \in \partial f^*(x^*)$.

By a modulus we mean a nondecreasing function from $\mathbb{R}_+$ to $\mathbb{R}_+$, vanishing and continuous at 0. We denote by $W^{1,1}\big([0,t], \mathbb{R}^N\big)$ the space of absolutely continuous curves from the interval $[0,t]$ to $\mathbb{R}^N$. We recall that a curve $\gamma : [a,b] \to \mathbb{R}^N$ is said to be *parameterized by the arc-length* if $|\dot\gamma(s)| = 1$ for almost every $s \in (a,b)$. Throughout the paper, α, β always denote two functions from $\mathbb{R}_+$ to $\mathbb{R}_+$ that are convex, nondecreasing, and superlinear.

We will denote by L a function from $\mathbb{R}^N \times \mathbb{R}^N$ to $\mathbb{R}$ so as to satisfy the following assumptions:

(L1) L is Borel-measurable on $\mathbb{R}^N \times \mathbb{R}^N$;
(L2) $\alpha(|q|) \leq L(x,q) \leq \beta(|q|)$ for all $(x,q) \in \mathbb{R}^N \times \mathbb{R}^N$;
(L3) $L(x,\cdot)$ is convex for every $x \in \mathbb{R}^N$.

By (L2), it is not restrictive to assume, up to adding a constant to it, that L is positive. This will be systematically done in what follows. We also point out that the second inequality in (L2) is equivalent to requiring that

$$\sup\{L(x,q) \,:\, (x,q) \in \mathbb{R}^N \times B_R\} < +\infty \qquad \text{for any } R > 0.$$

In fact, the following holds.

LEMMA 2.3. *Let U be an open subset of $\mathbb{R}^k$ and $L : U \times \mathbb{R}^k \to \mathbb{R}_+$ such that*

$$\sup\{L(x,q) \,:\, x \in U,\, |q| \le n\} < +\infty \qquad \textit{for every } n \in \mathbb{N}.$$

Then there exists a function $\beta : \mathbb{R}_+ \to \mathbb{R}_+$, convex and nondecreasing, such that

$$L(x,q) \le \beta(|q|) \qquad \textit{for every } (x,q) \in U \times \mathbb{R}^k.$$

Proof. Set

$$a_n := \sup\{L(x,q) \,:\, x \in U,\, |q| \le n\} \qquad \text{for each } n \in \mathbb{N},$$

and

$$f(h) := \sum_{n=1}^{\infty} a_n\, \chi_{[n-1,n)}(h) \qquad \text{for every } h \ge 0.$$

As $L(x,q) \le f(|q|)$ for every $(x,q) \in U \times \mathbb{R}^k$, it will be enough to prove the statement for f. For each $n \in \mathbb{N}$, choose $m_n := \max\{2a_n/(n-1),\, a_n - a_{n-1}\}$ and set

$$\beta(h) := \sup_{n\in\mathbb{N}}\{\, a_{n+1} + m_{n+1}(h-n)\} \qquad \text{for every } h \ge 0.$$

By definition, each map $h \mapsto a_{n+1} + m_{n+1}(h-n)$ is greater than or equal to f on $[n-1,n)$ and less than 0 on $[0,n/2)$, and hence

$$f(h) \le \beta(h) < +\infty \qquad \text{for every } h \ge 0.$$

The remainder of the assertion follows, as β is the supremum of a family of convex and increasing functions. ∎

To any L satisfying assumptions (L1)–(L3) we associate the function $S(y,x,t)$, defined in (3), for every $(y,x,t) \in \mathbb{R}^N \times \mathbb{R}^N \times (0,+\infty)$. It is easy to check that the function S enjoys the following inequalities:

$$t\,\alpha\left(\frac{|y-x|}{t}\right) \le S(y,x,t) \le t\,\beta\left(\frac{|y-x|}{t}\right) \qquad \text{on } \mathbb{R}^N \times \mathbb{R}^N \times (0,+\infty). \tag{6}$$

Later on in the paper, condition (L2) will be relaxed to cover the case of a Lagrangian L locally bounded with respect to (x,q), i.e., such that

$$\sup\{L(x,q) \,:\, (x,q) \in B_R \times B_R\} < +\infty \qquad \text{for any } R > 0.$$

By Lemma 2.3, this amounts to replacing condition (L2) with the following:

(L2)′ $\alpha(|q|) \le L(x,q) \le \beta_n(|q|)$ for all $(x,q) \in B_n \times \mathbb{R}^N$ and $n \in \mathbb{N}$, where $(\beta_n)_{n\in\mathbb{N}}$ is a family of convex, nondecreasing, and superlinear functions from $\mathbb{R}_+$ to $\mathbb{R}_+$.

3. The key results. The goal of our analysis is to prove the local Lipschitz continuity of the function S associated via (3) with a Lagrangian satisfying assumptions (L1), (L2), (L3). As previously noticed, condition (L2) amounts to requiring that the function $L(x,\cdot)$ be superlinear and locally bounded on $\mathbb{R}^N$, uniformly with respect to x. The precise statement of the result that we will establish is the following.

THEOREM 3.1. *Let $L : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}_+$ be an autonomous Lagrangian satisfying conditions* (L1)–(L3). *Then the associated function S defined in* (3) *is locally Lipschitz in $\mathbb{R}^N \times \mathbb{R}^N \times (0, +\infty)$. More precisely, for every $M > 0$ there exists $K = K(M, \alpha, \beta)$ such that*

$$S \quad \text{is } K\text{-Lipschitz continuous in } \overline{C_M},$$

where $C_M := \{(y, x, t) \in \mathbb{R}^N \times \mathbb{R}^N \times (0, +\infty) \,:\, |x - y| < M\, t\}$.

The consequent extension to the case of Lagrangians locally bounded with respect to (x, q) will be easily derived via a localization argument at the end of this section; see Theorem 3.19.

Theorem 3.1 can be proved via a rather standard argument as soon as we derive some a priori estimates on the Lipschitz constant of quasi-optimal curves parameterized in $[0, t]$ and connecting y to x for every $(y, x, t) \in C_M$. This information can be derived from the following lemma.

LEMMA 3.2. *Let $x, y \in \mathbb{R}^N$ and $t > 0$ such that $S(y, x, t) < M\, t$. Then there exists a constant $\kappa = \kappa(M, \alpha, \beta)$ such that*

$$S(y, x, t) = \inf \left\{ \int_0^t L(\xi, \dot\xi)\, \mathrm{d}s \,:\, \xi(0) = y,\ \xi(t) = x,\ \|\dot\xi\|_\infty \le \kappa \right\}.$$

The proof of Lemma 3.2 is quite delicate and relies on a careful analysis of the role played by reparameterizations. It will be carried out in the next two subsections. Before that, let us show how Lemma 3.2 can be used to prove Theorem 3.1.

Proof of Theorem 3.1. For a fixed $M > 0$, choose (y_1, t_1, x_1) and (y_2, t_2, x_2) in C_M, and set

$$h := |t_1 - t_2| + |x_1 - x_2| + |y_1 - y_2|, \qquad s_0 := \frac{t_1 - t_2}{2} + h.$$

Since C_M is convex, it suffices to prove the statement locally, namely for small values of h. Choose $h < t_2/2$ so that $s_0 < t_1/2$. Fix $\varepsilon > 0$ and let $\gamma_1 \in W^{1,1}\big([0, t_1], \mathbb{R}^N\big)$ be an ε-minimizer connecting y_1 to x_1. As $S(y_1, x_1, t_1) < t_1\, \beta(M)$, by Lemma 3.2 we can assume $\|\dot\gamma\|_\infty \le \kappa$ for some constant $\kappa = \kappa(M, \alpha, \beta)$. Choose $u_1, v_1 \in \mathbb{R}^N$ so that

$$\gamma_1(s_0) = y_2 + h u_1, \qquad \gamma_1(t_1 - s_0) = x_2 + h v_1,$$

and note that $|u_1|, |v_1| < 1 + 2\kappa$. Define a curve $\gamma_2 : [0, t_2] \to \mathbb{R}^N$ connecting y_2 to x_2 by setting

$$\gamma_2(s) := \begin{cases} y_2 + s u_1 & \text{if } s \in [0, h], \\ \gamma_1(s_0 + s - h) & \text{if } s \in [h, t_2 - h], \\ x_2 + (t_2 - s) v_1 & \text{if } s \in [t_2 - h, t_2]. \end{cases}$$

Recalling that L is positive, we get

$$\begin{aligned} S(y_2, x_2, t_2) - S(y_1, x_1, t_1) &\le \int_0^{t_2} L(\gamma_2, \dot\gamma_2)\, \mathrm{d}s - \int_0^{t_1} L(\gamma_1, \dot\gamma_1)\, \mathrm{d}s \ + \varepsilon \\ &\le \int_0^{h} L(\gamma_2, u_1)\, \mathrm{d}s + \int_{t_2 - h}^{t_2} L(\gamma_2, u_2)\, \mathrm{d}s \ + \varepsilon \le 2\, \beta(1 + 2\kappa)\, h \ + \varepsilon, \end{aligned}$$

so, setting $\widetilde K := 2\, \beta(1 + 2\kappa)$, we obtain

$$S(y_2, x_2, t_2) - S(y_1, x_1, t_1) \le \widetilde K\, (|t_1 - t_2| + |x_1 - x_2| + |y_1 - y_2|) + \varepsilon.$$

As ε is arbitrary, the conclusion follows at once by interchanging the roles of (y_1, t_1, x_1) and (y_2, t_2, x_2) and by setting $K := \sqrt{2N + 1}\, \widetilde K$. ∎

3.1. Preliminary tools. Let $H : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ be the Hamiltonian associated with L through the Fenchel transform, namely,

$$H(x,p) := \max_{q\in\mathbb{R}^N} \{\langle p,q\rangle - L(x,q)\}.$$

The function H is Borel-measurable, and $H(x,\cdot)$ is convex and superlinear for every $x \in \mathbb{R}^N$. For every $a \in \mathbb{R}$, set

$$\sigma_a(x,q) := \max\{\langle q,p\rangle \,:\, H(x,p) \le a\} \qquad \text{for every } q \in \mathbb{R}^N,\ a \in \mathbb{R},$$

where we agree that $\sigma_a(x,q) = -\infty$ whenever $a < -L(x,0) = \min_{\mathbb{R}^N} H(x,\cdot)$.

PROPOSITION 3.3. *For any $a \in \mathbb{R}$, the following properties hold:*

(i) *$\sigma_a(x,\lambda q) = \lambda\,\sigma_a(x,q)$ for every $(x,q) \in \mathbb{R}^N \times \mathbb{R}^N$ and $\lambda > 0$;*

(ii) *$L(x,q) \ge \sigma_a(x,q) - a$ for every $(x,q) \in \mathbb{R}^N \times \mathbb{R}^N$.*

Proof. Assertion (i) is clear by definition. To prove (ii), we recall that L is the Fenchel transform of H (cf. Proposition 2.2), and hence

$$L(x,q) = \max_{p\in\mathbb{R}^N} \{\langle p,q\rangle - H(x,p)\} \ge \max_{H(x,p)\le a} \{\langle p,q\rangle - H(x,p)\} \ge \sigma_a(x,q) - a, \tag{7}$$

as claimed. □

For any $(x,q) \in \mathbb{R}^N \times \mathbb{R}^N$ and $a \in \mathbb{R}$, we set

$$\Lambda_a(x,q) := \{\,\lambda \in [0,+\infty) \,:\, L(x,\lambda q) = \sigma_a(x,\lambda q) - a\} \tag{8}$$

and

$$\underline{\lambda}_a(x,q) := \inf \Lambda_a(x,q), \qquad \overline{\lambda}_a(x,q) := \sup \Lambda_a(x,q).$$

We agree that $\underline{\lambda}_a(x,q) = \overline{\lambda}_a(x,q) = 0$ whenever $\Lambda_a(x,q) = \emptyset$, that is, when $a < -L(x,0)$.

We define the following functions:

$$\alpha_*(u) := \max_{\lambda\in\mathbb{R}} \{u\lambda - \alpha(|\lambda|)\}, \quad \beta_*(u) := \max_{\lambda\in\mathbb{R}} \{u\lambda - \beta(|\lambda|)\} \qquad \text{for every } u \in \mathbb{R},$$

and we remark that they are convex and superlinear as $\alpha(|\cdot|)$ and $\beta(|\cdot|)$ are also. For every $a \in \mathbb{R}$, set

$$R_a := \max\{|u| \,:\, \beta_*(u) \le a\} \tag{9}$$

and

$$\kappa_a := 2\max\{\alpha_*(u) \,:\, |u| \le R_a + 1\}. \tag{10}$$

The following compactness result holds.

LEMMA 3.4. $\Lambda_a(x,q) \subseteq [0,\kappa_a]$ *for every* $(x,q) \in \mathbb{R}^N \times \mathbb{S}^{N-1}$.

Proof. From the definition of α_* and β_*, we obtain

$$\beta_*(|p|) \le H(x,p) \le \alpha_*(|p|) \qquad \text{for all } (x,p) \in \mathbb{R}^N \times \mathbb{R}^N; \tag{11}$$

in particular,

$$\{p \in \mathbb{R}^N \,:\, H(x,p) \le a\} \subseteq \overline{B}_{R_a} \qquad \text{for every } x \in \mathbb{R}^N. \tag{12}$$

Pick up $(x,q)\in\mathbb{R}^N\times\mathbb{S}^{N-1}$. From (7) and Proposition 2.2 we derive that $\lambda\in\Lambda_a(x,q)$ if and only if $\lambda\,q\in\partial_p H(x,p)$ for some $H(x,p)\le a$. In particular,

$$\Lambda_a(x,q)\subseteq\left\{|v|\,:\,v\in\partial_p H(x,p)\quad\text{for some } p\in\overline{B}_{R_a}\right\},$$

and the conclusion follows at once in view of (11) and Proposition 2.1. □

We now fix $(x,q)\in\mathbb{R}^N\times\mathbb{S}^{N-1}$ and examine the properties of the multifunction $a\mapsto\Lambda_a(x,q)$. Proposition 2.2 yields that

$$L(x,\lambda\,q)=\langle p,\lambda\,q\rangle-H(x,p)\qquad\text{for any } p\in\partial_q L(x,\lambda\,q) \tag{13}$$

for any given $\lambda\in\mathbb{R}$. In view of Proposition 3.3(ii), we infer that $\lambda\in\Lambda_a(x,q)$ if and only if $a\in H(x,\partial_q L(x,\lambda\,q))$.

We start by considering the set-valued map $A(\lambda):=H(x,\partial_q L(x,\lambda\,q))$ on $[0,+\infty)$, which is the inverse of $a\mapsto\Lambda_a(x,q)$, in the sense of set-valued analysis (see [32, Chapter 5]). Indeed, note that

$$\Lambda_a(x,q)=\{\,\lambda\in[0,+\infty)\,:\,a\in H(x,\partial_q L(x,\lambda\,q))\}. \tag{14}$$

PROPOSITION 3.5. *Let $A(\cdot)$ as above. The following facts hold:*

(i) *For any $\lambda\in\mathbb{R}$,*

$$A(\lambda)=[\underline{a}(\lambda),\overline{a}(\lambda)]\quad\textit{for some } -L(x,0)\le\underline{a}(\lambda)\le\overline{a}(\lambda)<+\infty.$$

Moreover,

$$A(0)=\{\,-L(x,0)\},\qquad\lim_{\lambda\to+\infty}\underline{a}(\lambda)=+\infty.$$

(ii) *The set-valued map $A(\cdot)$ is upper semicontinuous on $[0,+\infty)$. In particular, $\underline{a}(\cdot)$ is lower semicontinuous and $\overline{a}(\cdot)$ is upper semicontinuous on $[0,+\infty)$.*

(iii) *The set-valued map $\lambda\mapsto A(\lambda)$ is nondecreasing on $[0,+\infty)$.*

(iv) $\bigcup_{\lambda\ge0}A(\lambda)=[-L(x,0),+\infty)$.

Proof. The function $f(\lambda):=L(x,\lambda\,q)$ is convex and superlinear, and hence so is its conjugate f^*. We claim that $A(\lambda)=f^*(\partial f(\lambda))$ for every $\lambda\ge0$. Indeed, by Proposition 2.2 we know that

$$f^*(\partial f(\lambda))=\lambda\,\partial f(\lambda)-f(\lambda)\qquad\text{for any } \lambda\ge0.$$

By classical results of nonsmooth analysis (cf. [16, Theorem 2.3.10]), we also know that $\partial f(\lambda)=\langle\partial_q L(x,\lambda\,q),q\rangle$, and hence the above equality becomes

$$f^*(\partial f(\lambda))=\langle\partial_q L(x,\lambda\,q),\lambda\,q\rangle-L(x,\lambda\,q)\qquad\text{for any } \lambda\ge0,$$

and the right-hand side term coincides with $A(\lambda)$ by (13), as claimed.

Let us now prove the above stated properties of $A(\cdot)$. As f is convex, its subdifferential $\partial f(\lambda)$ is a compact interval of $\mathbb{R}$, so the same is true for $A(\lambda)$. The equality $A(0)=\{-L(x,0)\}$ is an immediate consequence of (13), while the other assertion follows by the superlinearity of f^* and f. That proves (i). The upper semicontinuity of $A(\cdot)$ comes from the fact that the multifunction $\lambda\mapsto\partial f(\lambda)$ is upper semicontinuous and f^* is continuous. The remainder of (ii) follows by definition of $\underline{a}(\cdot)$, $\overline{a}(\cdot)$.

Let us prove (iii). Since f^* and f are convex, the multimappings $u\mapsto\partial f^*(u)$ and $\lambda\mapsto\partial f(\lambda)$ are nondecreasing on $\mathbb{R}$. By superlinearity, we get, in particular,

$$\bigcup_{\lambda\ge0}\partial f(\lambda)=[\underline{u}(0),+\infty)\qquad\text{with } \underline{u}(0)\in\partial f(0).$$

By duality (cf. Proposition 2.2), $0 \in \partial f^*(\underline{u}(0))$, so the monotonicity of $\partial f^*(\cdot)$ yields that f^* is nondecreasing on $[\underline{u}(0), +\infty)$. Item (iv) comes from (ii) and (iii). ∎

Example 3.6. Take a Lagrangian of the form $L(x,q) = F(q) + n(x)$ for every $(x,q) \in \mathbb{R}^N \times \mathbb{R}^N$, with $F(\cdot)$ convex and superlinear, and $n(\cdot)$ Borel-measurable and bounded. We have

$$A(\lambda) = F^*\left(\partial F(\lambda\, q)\right) - n(x) \qquad \text{for every } \lambda \geq 0$$

for any fixed $(x,q) \in \mathbb{R}^N \times \mathbb{S}^{N-1}$. When, for instance, $F(q) = |q|^2/2$, it reduces to

$$A(\lambda) = |\lambda\, q|^2/2 - n(x).$$

We use this information to prove a result that will be crucial in our future analysis.

PROPOSITION 3.7. *Let $(x,q) \in \mathbb{R}^N \times \mathbb{S}^{N-1}$. The following facts hold:*

(i) *For any $a \geq -L(x,0)$, we have*

$$\Lambda_a(x,q) = [\underline{\lambda}_a(x,q), \overline{\lambda}_a(x,q)] \qquad \textit{for some} \quad 0 \leq \underline{\lambda}_a(x,q) \leq \overline{\lambda}_a(x,q) < +\infty.$$

Moreover,

$$\underline{\lambda}_{-L(x,0)}(x,q) = 0, \qquad \lim_{a \to +\infty} \underline{\lambda}_a(x,q) = +\infty.$$

(ii) *The set-valued map $a \mapsto \Lambda_a(x,q)$ is upper semicontinuous and nondecreasing on $[-L(x,0), +\infty)$.*

(iii) *$\underline{\lambda}_a(x,q) = \sup_{b<a} \overline{\lambda}_b(x,q)$ for any $a > -L(x,0)$ and $\overline{\lambda}_a(x,q) = \inf_{b>a} \underline{\lambda}_b(x,q)$ for any $a \geq -L(x,0)$.*

(iv) *$\underline{\lambda}_a(x,q) \geq \frac{a+L(x,0)}{2R_a}$ for any $a > -L(x,0)$, with R_a defined by (9).*

Proof. We recall that $\Lambda_a(x,q) = \{\lambda \geq 0 : a \in A(\lambda)\}$. The monotonicity and coercivity properties of the set-valued map $a \mapsto \Lambda_a(x,q)$ are consequences of Proposition 3.5, while the equality $\underline{\lambda}_{-L(x,0)}(x,q) = 0$ is apparent by definition of (8). In particular, $\Lambda_a(x,q)$ is a bounded interval for any $a \geq -L(x,0)$.

To prove the upper semicontinuity of $a \mapsto \Lambda_a(x,q)$, we need to show that, for each pair of sequences $(a_n)_n$ and $(\lambda_n)_n$ such that $a_n \to a \in \mathbb{R}$, $\lambda_n \to \lambda \in \mathbb{R}$, and $\lambda_n \in \Lambda_{a_n}(x,q)$ for every $n \in \mathbb{N}$, we have $\lambda \in \Lambda_a(x,q)$. This easily follows by the upper semicontinuity of $A(\cdot)$ (in fact, it is equivalent; cf. [32, Theorem 5.7]). In particular, this implies that $\Lambda_a(x,q)$ is closed for any $a \geq -L(x,0)$.

Assertion (iii) immediately follows from the monotone and semicontinuous character of the map $a \mapsto \Lambda_a(x,q)$.

Let us prove (iv). Choose $a > -L(x,0)$ and set $\lambda := \underline{\lambda}_a(x,q)$. By Proposition 3.3(ii) we get

$$\sigma_a(x, \lambda q) = L(x, \lambda q) + a \geq \sigma_{-L(x,0)}(x, \lambda q) + a + L(x,0),$$

and hence by (12),

$$a + L(x,0) \leq \lambda \left(\sigma_a(x,q) - \sigma_{-L(x,0)}(x,q)\right) \leq \lambda \left(R_a + R_{-L(x,0)}\right) |q|,$$

and the statement follows as $R_{-L(x,0)} < R_a$ by definition. ∎

Example 3.8. Let $L(x,q) := |q|^2/2 + n(x)$ for every $(x,q) \in \mathbb{R}^N \times \mathbb{R}^N$, with $n(\cdot)$ Borel-measurable and bounded. For any fixed $(x,q) \in \mathbb{R}^N \times \mathbb{S}^{N-1}$, we have

$$\Lambda_a(x,q) = \left\{ \frac{1}{|q|} \sqrt{2\left(a + n(x)\right)} \right\} \qquad \text{for every } a \geq -n(x).$$

3.2. Optimal reparameterizations. Let us now consider a Lipschitz curve γ defined on a bounded interval $J := [0, \ell]$.

DEFINITION 3.9. *A curve ξ defined on a bounded interval $[0,t]$ is said to be a* reparameterization *of γ if there exists an absolutely continuous map $\varphi : [0,t] \to [0,\ell]$, surjective and nondecreasing, such that*

$$\xi = \gamma \circ \varphi \qquad \text{on } [0,t].$$

We furthermore say that ξ is a (bi)-Lipschitz reparameterization *of γ if φ is a (bi)-Lipschitz homeomorphism.*

Remark 3.10. For reasons that will be clear soon, we want to allow a reparameterization to remain stopped at a point for a specific amount of time. This accounts for the choice of the unusual definition given above.

We introduce the following notation:

$$\begin{aligned}[\gamma]_t &:= \{\xi \in W^{1,1}\big([0,t],\mathbb{R}^N\big) \,:\, \xi \text{ is a reparameterization of } \gamma\}\\ [\gamma]_t^\flat &:= \{\xi \in W^{1,1}\big([0,t],\mathbb{R}^N\big) \,:\, \xi \text{ is a bi-Lipschitz reparameterization of } \gamma\}.\end{aligned}$$

The following lemma comes from classical results of analysis in metric spaces (see, e.g., [24, section VII.2]). We give a proof for the reader's convenience.

LEMMA 3.11. *Let $\xi \in W^{1,1}\big([0,t],\mathbb{R}^N\big)$. Then there exists a Lipschitz curve γ, defined on a bounded interval $[0,\ell]$, such that $\xi \in [\gamma]_t$. We can furthermore assume that γ is parameterized by the arc-length.*

Proof. Let $\varphi(s) := \int_0^s |\dot\xi(\varsigma)|\,\mathrm{d}\varsigma$ for every $s \in [0,t]$, and set $\ell := \varphi(t)$. Clearly, the map $\varphi : [0,t] \to [0,\ell]$ is absolutely continuous, surjective, and nondecreasing. We claim that the statement holds true with $\gamma(s) := \xi(\varphi^{-1}(s))$ for every $s \in [0,\ell]$.

Indeed, it is easy to see that γ is well defined. Now choose a pair of points a, b in $[0,\ell]$ with $a < b$. By the monotone character of φ, we have $\varphi^{-1}(a) = [A_-, A_+]$, $\varphi^{-1}(b) = [B_-, B_+]$ for some $A_- \le A_+ < B_- \le B_+$. Moreover,

$$|\gamma(b) - \gamma(a)| \le \mathcal{H}^1\big(\gamma([a,b])\big) = \mathcal{H}^1\big(\xi([A_-,B_+])\big) = \int_{A_-}^{B_+} |\dot\xi(\varsigma)|\,\mathrm{d}\varsigma = b - a,$$

which yields that γ is 1-Lipschitz continuous. From the fact that $\int_0^\ell |\dot\gamma(\varsigma)|\,\mathrm{d}\varsigma = \mathcal{H}^1\big(\gamma([0,\ell])\big) = \ell$, we finally get that γ is parameterized by the arc-length. ∎

A further step in the analysis is carried out by picking up some special reparameterizations of the curve γ.

DEFINITION 3.12. *Let ξ be a curve defined on a bounded interval $[0,t]$ and $a \in \mathbb{R}$. We say that ξ has an a-*Lagrangian parameterization *if*

$$L(\xi(s),\dot\xi(s)) = \sigma_a(\xi(s),\dot\xi(s)) - a \qquad \text{for a.e. } s \in [0,t].$$

For any $a \in \mathbb{R}$ and $t > 0$, we define

$$\begin{aligned}[\gamma](a,t) &:= \{\xi \in [\gamma]_t \,:\, \xi \text{ has an } a\text{-Lagrangian parameterization}\},\\ [\gamma]^\flat(a,t) &:= \{\xi \in [\gamma]_t^\flat \,:\, \xi \text{ has an } a\text{-Lagrangian parameterization}\}.\end{aligned}$$

Now assume γ is parameterized by the arc-length, and let

$$c_\gamma := \operatorname{ess\,sup}_{s\in J} \; -L\left(\gamma(s),0\right).$$

We define a multifunction $T_\gamma : (c_\gamma, +\infty) \to \mathcal{P}(\mathbb{R}_+)$ by setting

$$T_\gamma(a) := \{t > 0 \,:\, [\gamma]^\flat(a,t) \text{ is nonempty}\}.$$

The properties of the multifunction $T_\gamma(\cdot)$ are stated below.

PROPOSITION 3.13. *Let γ and $T(\cdot) := T_\gamma(\cdot)$ as above. The following facts hold:*

(i) *For any $a > c_\gamma$, $T(a)$ is a compact interval in $(0, +\infty)$, namely,*

$$T(a) := [\underline{T}\,(a), \overline{T}\,(a)] \qquad \text{for some} \quad \overline{T}\,(a) \geq \underline{T}\,(a) > 0.$$

(ii) *The multifunction $\underline{T}(\cdot)$ is nondecreasing and upper semicontinuous on $(c_\gamma, +\infty)$. Moreover, $\inf_{a>c_\gamma} \overline{T}\,(a) = 0$.*

(iii) *Let $\underline{T}(c_\gamma) := \sup_{a>c_\gamma} \overline{T}(a)$. If $\underline{T}(c_\gamma)$ is finite, then $[\gamma](c_\gamma, \underline{T}(c_\gamma)) \neq \emptyset$.*

In particular, for any $0 < t \leq \underline{T}(c_\gamma)$ with $t < +\infty$, there exists $a \geq c_\gamma$ such that γ admits an a-Lagrangian Lipschitz reparameterization on $[0,t]$.

We first prove an auxiliary lemma.

LEMMA 3.14. *Let $\gamma : [0,\ell] \to \mathbb{R}^N$ be a Lipschitz curve parameterized by the arc-length and $a \in \mathbb{R}$. The following facts hold:*

(i) *For every $t > 0$ and $\xi \in [\gamma]_t$, the map $\sigma_a(\xi(\cdot), \dot\xi(\cdot))$ is Lebesgue-measurable on $[0,t]$, and*

$$\int_0^t \sigma_a(\xi(s), \dot\xi(s))\,\mathrm{d}s = \int_0^\ell \sigma_a(\gamma(s), \dot\gamma(s))\,\mathrm{d}s. \tag{15}$$

(ii) *The maps $\underline{\lambda}_a(\gamma(\cdot), \dot\gamma(\cdot))$, $\overline{\lambda}_a(\gamma(\cdot), \dot\gamma(\cdot))$ are Lebesgue-measurable on $[0,\ell]$.*

Proof. Take $t > 0$ and $\xi \in [\gamma]_t$. Since the map $s \mapsto (\xi(s), \dot\xi(s))$ is Lebesgue-measurable, in order to prove (i) it is enough to show that the function σ_a is Borel-measurable on $\mathbb{R}^N \times \mathbb{R}^N$. To this aim, let $(p_n)_n$ and $(\lambda_n)_n$ be dense sequences in $\mathbb{R}^N$ and $(0, +\infty)$, respectively. The Borel-measurable character of σ_a follows at once as we have

$$\sigma_a(x,q) = \inf_k \left(\sup_n \left\{ \langle p_n, q\rangle\, \vartheta_{E_n^k}(x) \right\} \right) \qquad \text{for every } (x,q) \in \mathbb{R}^N \times \mathbb{R}^N,$$

where $E_n^k := \{x \in \mathbb{R}^N \,:\, H(x,p_n) \leq a + 1/k\}$ and $\vartheta_{E_n^k}(\cdot)$ denotes the function identically 1 on E_n^k and $-\infty$ elsewhere. Equality (15) is a consequence of the fact that $\sigma_a(x,\cdot)$ is positively 1-homogeneous.

Let us prove (ii). Since the map $s \mapsto (\gamma(s), \dot\gamma(s))$ takes values in $\mathbb{R}^N \times \mathbb{S}^{N-1}$ for a.e. $s \in [0,\ell]$, it suffices to show that the functions $\underline{\lambda}_a$, $\overline{\lambda}_a$ are Borel-measurable on $\mathbb{R}^N \times \mathbb{S}^{N-1}$. Let us show the statement for $\underline{\lambda}_a$. For each $n \in \mathbb{N}$, set

$$F_n := \{(x,q) \in \mathbb{R}^N \times \mathbb{S}^{N-1} \,:\, H(x, \partial_q L(x, \lambda_n\, q)) \cap (-\infty, a) \neq \emptyset\},$$

which is Borel-measurable for the multifunction $(x,q) \mapsto H(x, \partial_q L(x, \lambda_n\, q))$ that is also. The assertion follows for we have $\underline{\lambda}_a(x,q) = \sup_n \lambda_n\, \chi_{F_n}(x,q)$ on $\mathbb{R}^N \times \mathbb{S}^{N-1}$, in view of (14) and Proposition 3.7. The analogous statement for $\overline{\lambda}_a$ can be proved in a similar way. ∎

Proof of Proposition 3.13. (i) Fix $a > c_\gamma$ and set

$$\underline{\lambda}_a(\varsigma) := \underline{\lambda}_a(\gamma(\varsigma), \dot\gamma(\varsigma)), \quad \overline{\lambda}_a(\varsigma) := \overline{\lambda}_a(\gamma(\varsigma), \dot\gamma(\varsigma)) \qquad \text{for a.e. } \varsigma \in [0,\ell].$$

Let

$$\underline{T}(a) := \int_0^\ell \frac{1}{\overline{\lambda}_a(\varsigma)}\,\mathrm{d}\varsigma, \quad \overline{T}(a) := \int_0^\ell \frac{1}{\underline{\lambda}_a(\varsigma)}\,\mathrm{d}\varsigma.$$

Such quantities are well-defined, positive real values, thanks to Proposition 3.7(iv) and to the measurable character of $\underline{\lambda}_a(\cdot)$, $\overline{\lambda}_a(\cdot)$. To show that they belong to $T(a)$, we will prove the existence of two curves $\underline{\gamma}_a$, $\overline{\gamma}_a$, defined on $\big[0,\underline{T}(a)\big]$ and $\big[0,\overline{T}(a)\big]$, respectively, which are a-Lagrangian bi-Lipschitz reparameterizations of γ. To this aim, let us define

$$\underline{f}_a(s) := \int_0^s \frac{1}{\overline{\lambda}_a(\varsigma)}\,\mathrm{d}\varsigma, \quad \overline{f}_a(s) := \int_0^s \frac{1}{\underline{\lambda}_a(\varsigma)}\,\mathrm{d}\varsigma \qquad \text{for any } s \in [0,\ell]$$

and set

$$\underline{\varphi}_a := (\underline{f}_a)^{-1}, \qquad \overline{\varphi}_a := \big(\overline{f}_a\big)^{-1},$$

defined on $\big[0,\underline{T}(a)\big]$ and $\big[0,\overline{T}(a)\big]$, respectively. As

$$\dot{\underline{\varphi}}_a(\tau) = \overline{\lambda}_a(\underline{\varphi}_a(\tau)), \qquad \dot{\overline{\varphi}}_a(\tau) = \underline{\lambda}_a(\overline{\varphi}_a(\tau)) \qquad \text{for a.e. } \tau,$$

we immediately derive that $\underline{\varphi}_a$ and $\overline{\varphi}_a$ are order-preserving bi-Lipschitz diffeomorphisms. Let us set

$$\underline{\gamma}_a := \gamma \circ \underline{\varphi}_a \quad \text{on } \big[0,\underline{T}(a)\big], \qquad \overline{\gamma}_a := \gamma \circ \overline{\varphi}_a \quad \text{on } \big[0,\overline{T}(a)\big].$$

Since

$$\dot{\underline{\gamma}}_a(\cdot) := \overline{\lambda}_a(\underline{\varphi}_a(\cdot))\,\dot{\gamma}(\underline{\varphi}_a(\cdot)) \qquad \text{a.e. on } \big[0,\underline{T}(a)\big]$$

and

$$\dot{\overline{\gamma}}_a(\cdot) := \underline{\lambda}_a(\overline{\varphi}_a(\cdot))\,\dot{\gamma}(\overline{\varphi}_a(\cdot)) \qquad \text{a.e. on } \big[0,\overline{T}(a)\big],$$

we conclude that the curves $\underline{\gamma}_a$ and $\overline{\gamma}_a$ have an a-Lagrangian parameterization by the very definition of $\overline{\lambda}_a$ and $\underline{\lambda}_a$.

In order to prove that $[\underline{T}(a),\overline{T}(a)] \subseteq T(a)$, we will show that

$$\delta\,\underline{T}(a) + (1-\delta)\,\overline{T}(a) \in T(a) \qquad \text{for any } \delta \in (0,1). \tag{16}$$

Fix $\delta \in (0,1)$ and set

$$\delta(\varsigma) := \frac{\delta\,\overline{\lambda}_a(\varsigma)}{\delta\,\overline{\lambda}_a(\varsigma) + (1-\delta)\,\underline{\lambda}_a(\varsigma)}, \qquad \lambda(\varsigma) := \delta(\varsigma)\,\underline{\lambda}_a(\varsigma) + (1-\delta(\varsigma))\,\overline{\lambda}_a(\varsigma)$$

for almost every $\varsigma \in [0,\ell]$, and

$$f(s) := \int_0^s \frac{1}{\lambda(\varsigma)}\,\mathrm{d}\varsigma \quad \text{for } s \in [0,\ell], \qquad \varphi := f^{-1} \quad \text{on } [0,f(\ell)].$$

Since $\delta(\varsigma) \in [0,1]$ for almost every $\varsigma \in [0,\ell]$, we get that $\lambda_a(\varsigma) \in \Lambda_a(\gamma(\varsigma),\dot{\gamma}(\varsigma))$ for almost every $\varsigma \in [0,\ell]$; in particular, φ is an order-preserving bi-Lipschitz diffeomorphism. Arguing as above, we see that the curve $\gamma_a := \gamma \circ \varphi$ is an a-Lagrangian

bi-Lipschitz reparameterization of γ on $\big[0, f(\ell)\big]$, so $f(\ell) \in T(a)$. Now it is easy to check, by definition of $\delta(\cdot)$, that $f(\ell) = \delta\,\overline{T}\,(a) + (1-\delta)\,\underline{T}\,(a)$. That proves (16) as δ was arbitrarily chosen in $(0,1)$.

Let us now prove that $T(a) \subseteq [\underline{T}\,(a), \overline{T}\,(a)]$. Let $T \in T(a)$ and $\widetilde{\gamma} := \gamma \circ \varphi$ be an a-Lagrangian reparameterization of γ for some order-preserving bi-Lipschitz diffeomorphism $\varphi : [0,T] \to [0,\ell]$. Then

$$\dot{\varphi}(\tau) \in \Lambda_a\left(\gamma(\varphi(\tau)), \dot{\gamma}(\varphi(\tau))\right) \qquad \text{for a.e. } \tau \in [0,T].$$

Let $f := \varphi^{-1}$. We have

$$T = f(\ell) = \int_0^\ell \dot{f}(\varsigma)\,\mathrm{d}\varsigma = \int_0^\ell \frac{1}{\dot{\varphi}(f(\varsigma))}\,\mathrm{d}\varsigma,$$

and since $\dot{\varphi}(f(\varsigma)) \in \Lambda_a(\gamma(\varsigma), \dot{\gamma}(\varsigma)) = [\underline{\lambda}_a(\varsigma), \overline{\lambda}_a(\varsigma)]$ for a.e. $\varsigma \in [0,\ell]$, we clearly get $T \in [\underline{T}\,(a), \overline{T}\,(a)]$.

(ii) Let $b > a > c_\gamma$. Then $\underline{\lambda}_b(\varsigma) \geq \overline{\lambda}_a(\varsigma)$ for almost every $\varsigma \in [0,\ell]$, and hence $\overline{T}(b) \leq \underline{T}(a)$. This proves that $T(\cdot)$ is a nonincreasing multifunction. To prove that $T(\cdot)$ is upper semicontinuous on $(c_\gamma, +\infty)$, it will be enough to show that

$$\underline{T}(a) = \sup_{b>a} \overline{T}(b), \qquad \overline{T}(a) = \inf_{b<a} \underline{T}(b) \qquad \text{for any } a > c_\gamma.$$

This actually follows as a simple application of the monotone convergence theorem and by the monotonicity properties of $\underline{\lambda}_a$, $\overline{\lambda}_a$ (cf. Proposition 3.7(iii)). The last assertion holds by definition of $\underline{T}(a)$ since $\sup_{a>c_\gamma} \overline{\lambda}_a(\varsigma) = +\infty$ for almost every $\varsigma \in [0,\ell]$.

(iii) Let $\underline{T}(c_\gamma)$ be finite. Arguing as in (i), we may find a nonincreasing sequence of Borel-measurable maps $\lambda_n : [0,\ell] \to [0,+\infty)$ such that, for each $n \in \mathbb{N}$,

$$T_n = \int_0^\ell \frac{1}{\lambda_n(\varsigma)}\,\mathrm{d}\varsigma \qquad \text{and} \qquad \lambda_n(\varsigma) \in \Lambda_{c_\gamma+1/n}(\gamma(\varsigma), \dot{\gamma}(\varsigma)) \qquad \text{for a.e. } \varsigma \in [0,\ell],$$

with $\sup_n T_n = \underline{T}(c_\gamma)$. Set

$$\lambda(\varsigma) = \inf_n \lambda_n(\varsigma) \qquad \text{for every } \varsigma \in [0,\ell].$$

Then $\lambda(\cdot)$ is measurable and $\lambda(\varsigma) \in \Lambda_{c_\gamma}(\gamma(\varsigma), \dot{\gamma}(\varsigma))$ for almost every $\varsigma \in [0,\ell]$. Moreover, the monotone convergence theorem yields

$$\underline{T}(c_\gamma) = \sup_{n\in\mathbb{N}}\ T_n = \sup_{n\in\mathbb{N}} \int_0^\ell \frac{1}{\lambda_n(\varsigma)}\,\mathrm{d}\varsigma = \int_0^\ell \frac{1}{\lambda(\varsigma)}\,\mathrm{d}\varsigma;$$

in particular, the map

$$f(s) := \int_0^s \frac{1}{\lambda(\varsigma)}\,\mathrm{d}\varsigma$$

is increasing and absolutely continuous on $[0,\ell]$. A c_γ-Lagrangian Lipschitz reparametrization of γ defined on $[0, \underline{T}(c_\gamma)]$ can now be obtained by setting $\widetilde{\gamma} := \gamma \circ \varphi$ with $\varphi := (f)^{-1}$ on $[0, \underline{T}(c_\gamma)]$. Lastly, the fact that the multifunction is upper semicontinuous, monotone, and convex-set-valued implies that

$$\bigcup_{a>c_\gamma} T(a) = \big(0, \underline{T}(c_\gamma)\big),$$

and this is enough to obtain the remainder of the statement. □

Example 3.15. Let $L(x,q) := |q|^2/2 + n(x)$ for every $(x,q) \in \mathbb{R}^N \times \mathbb{R}^N$, with $n(\cdot)$ Borel-measurable and bounded. Let $\gamma : [0,\ell] \to \mathbb{R}^N$ be a curve parameterized by the arc-length. Then $c_\gamma = \operatorname{ess\,sup}_{s\in[0,\ell]} -n(\gamma(s))$ and (cf. Example 3.8)

$$T_\gamma(a) = \int_0^\ell \frac{1}{\sqrt{2\,(a + n(\gamma(s)))}}\,\mathrm{d}s \qquad \text{for every } a > c_\gamma.$$

We now seek an optimal reparameterization of γ on the interval $[0,t]$ for any given $t \in (0,+\infty)$. For that, it suffices that $\underline{T}_\gamma(c_\gamma) = +\infty$ by Proposition 3.13, but this need not be true in general, not even in the simple case considered in Example 3.15. However, even in the case $\underline{T}_\gamma(c_\gamma) < +\infty$ we are able to derive an estimate on the Lipschitz constants of quasi-optimal reparameterizations. This is a crucial step in our study.

THEOREM 3.16. *Let $\gamma : [0,\ell] \to \mathbb{R}^N$ be a Lipschitz curve parameterized by the arc-length. Then, for every $t \in (0,+\infty)$, there exists $a \geq c_\gamma$ such that*

$$\inf_{\xi\in[\gamma]_t} \int_0^t L(\xi,\dot\xi)\,\mathrm{d}s = \inf_{\xi\in[\gamma]_t} \left\{ \int_0^t L(\xi,\dot\xi)\,\mathrm{d}s \,:\, \|\dot\xi\|_\infty \leq \kappa_a \right\} = \int_0^\ell \sigma_a(\gamma,\dot\gamma)\,\mathrm{d}s - a\,t,$$

with κ_a given by (10). *The above infimum is a minimum whenever $t \leq \underline{T}_\gamma(c_\gamma)$ and is, in particular, attained by some curve belonging to $[\gamma]^\flat(a,t)$ with $a > c_\gamma$ when $t < \underline{T}_\gamma(c_\gamma)$.*

Proof. By Proposition 3.3 and Lemma 3.14, we get

$$\int_0^t L(\xi,\dot\xi)\,\mathrm{d}s \geq \int_0^t \Big(\sigma_a(\xi,\dot\xi) - a\Big)\,\mathrm{d}s = \int_0^\ell \sigma_a(\gamma,\dot\gamma)\,\mathrm{d}s - a\,t \tag{17}$$

for any $a \geq c_\gamma$ and $\xi \in [\gamma]_t$, and (17) is an equality whenever $\xi \in [\gamma](a,t)$. The assertion for $t \leq \underline{T}_\gamma(c_\gamma)$ follows by Proposition 3.13 and Lemma 3.4.

Let us now assume $t > \underline{T}_\gamma(c_\gamma)$ and set $h := t - \underline{T}_\gamma(c_\gamma)$. Let $\xi \in [\gamma](c_\gamma, \underline{T}_\gamma(c_\gamma))$. By definition of c_γ, there exists, for each $n \in \mathbb{N}$, $s_n \in \big(0, \underline{T}_\gamma(c_\gamma)\big)$ such that

$$c_\gamma + L(\xi(s_n),0) < \frac{1}{n}.$$

To ease notation, we will write c_n in place of $-L(\xi(s_n),0)$. We define

$$\xi_n(s) := \begin{cases} \xi(s) & \text{if } s \in (0,s_n], \\ \xi(s_n) & \text{if } s \in [s_n, s_n+h], \\ \xi(s-h) & \text{if } s \in [s_n+h, t). \end{cases}$$

We have

$$\int_0^t L(\xi_n,\dot\xi_n)\,\mathrm{d}s = \int_0^{\underline{T}_\gamma(c_\gamma)} L(\xi,\dot\xi)\,\mathrm{d}s - h\,c_n = \int_0^{\underline{T}_\gamma(c_\gamma)} \sigma_{c_\gamma}(\xi,\dot\xi)\,\mathrm{d}s - \underline{T}_\gamma(c_\gamma)\,c_\gamma$$
$$-h\,c_n = \int_0^\ell \sigma_{c_\gamma}(\gamma,\dot\gamma)\,\mathrm{d}s - c_\gamma\,t + h\,\big(c_\gamma - c_n\big) < \int_0^\ell \sigma_{c_\gamma}(\gamma,\dot\gamma)\,\mathrm{d}s - c_\gamma\,t + \frac{h}{n}.$$

Taking (17) into account, we derive

$$\int_0^\ell \sigma_{c_\gamma}(\gamma,\dot\gamma)\,\mathrm{d}s - c_\gamma\,t \leq \int_0^t L(\xi_n,\dot\xi_n)\,\mathrm{d}s < \int_0^\ell \sigma_{c_\gamma}(\gamma,\dot\gamma)\,\mathrm{d}s - c_\gamma\,t + \frac{h}{n},$$

and we conclude by letting $n \to +\infty$. □

Remark 3.17. If in Theorem 3.16 the Lagrangian is assumed lower semicontinuous in x, we can furthermore say that, for every $t > 0$,

$$\inf_{\xi \in [\gamma]_t} \int_0^t L(\xi, \dot\xi)\,\mathrm{d}s = \min\left\{ \int_0^t L(\xi, \dot\xi)\,\mathrm{d}s \,:\, \xi \in [\gamma](a,t) \right\} \tag{18}$$

for some constant $a \geq c_\gamma$. This can be proved by considering, in place of $T_\gamma(\cdot)$, the set-valued map defined as

$$T^*_\gamma(a) := \{t > 0 \,:\, [\gamma](a,t) \text{ is nonempty}\}$$

for every $a \geq c^*_\gamma := \sup_{s\in J} -L\,(\gamma(s), 0)$. The multifunction $T^*_\gamma(\cdot)$ agrees with $T_\gamma(\cdot)$ on $(c^*_\gamma, +\infty)$. Indeed, the inequality

$$\underline{\lambda}_a(\gamma(s), \dot\gamma(s)) \geq \frac{a - c^*_\gamma}{2R_a} \qquad \text{for a.e. } s \in [0, \ell],$$

which holds by Proposition 3.7, implies that $[\gamma](a,t) = [\gamma]^\flat(a,t)$ for every $a > c^*_\gamma$ and $t > 0$ (cf. the argument that $T(a) \subseteq [\underline{T}\,(a), \overline{T}\,(a)]$ in the proof of Proposition 3.13). On the other hand, we always have

$$T^*_\gamma(c^*_\gamma) = [\underline{T}_\gamma(c^*_\gamma), +\infty) \tag{19}$$

when $T(c^*_\gamma)$ is finite, and that is enough to get the statement in view of (17).

To prove (19), let ξ be a curve belonging to $[\gamma](c^*_\gamma, \underline{T}(c^*_\gamma))$ (which does exist by Proposition 3.13) and take $s_0 \in [0, \underline{T}_\gamma(c^*_\gamma)]$ such that $L(\xi(s_0), 0) = -c^*_\gamma$. Such an s_0 always exists by the upper semicontinuity of $-L(\gamma(\cdot), 0)$ on $[0, \ell]$. For every $h > 0$, define $\xi_h : [0, \underline{T}(c^*_\gamma) + h] \to \mathbb{R}^N$ as

$$\xi_h(s) := \begin{cases} \xi(s) & \text{if } s \in [0, s_0], \\ \xi(s_0) & \text{if } s \in [s_0, s_0 + h], \\ \xi(s - h) & \text{if } s \in [s_0 + h, \underline{T}_\gamma(c^*_\gamma) + h]. \end{cases}$$

It is easily seen that ξ_h is a c^*_γ-Lagrangian reparameterization of γ. This shows that

$$\underline{T}_\gamma(c^*_\gamma) + h \in T^*_\gamma(c^*_\gamma) \qquad \text{for every } h > 0,$$

as claimed.

Remark 3.18. The argument described in Remark 3.17 above actually shows that (18) holds whenever the map $s \mapsto L(\gamma(s), 0)$ attains its infimum on $[0, \ell]$, for instance, when it is lower semicontinuous.

With the aid of the results obtained so far, we can now prove Lemma 3.2.

Proof of Lemma 3.2. Choose $\overline{n} \in \mathbb{N}$ such that $M/\alpha(\overline{n}) < 1/2$ and set

$$A = A(\overline{n}) := \max\{\alpha_*(u) \,:\, |u| \leq 2\,\beta(\overline{n} + 1)\},$$

where $\alpha_*(u) := \max_{\lambda\in\mathbb{R}}\{\lambda u - \alpha(|\lambda|)\}$. We claim that the statement holds with $\kappa := \kappa_A$ defined according to (10). Indeed, pick a curve $\xi \in W^{1,1}\big([0,t], \mathbb{R}^N\big)$ such that

$$\int_0^t L(\xi, \dot\xi)\,\mathrm{d}s < M\,t,$$

and let $\gamma : [0, \ell] \to \mathbb{R}^N$ be a Lipschitz curve, parameterized by the arc-length, such that $\xi \in [\gamma]_t$, according to Lemma 3.11. In view of Theorem 3.16, up to choosing a different ξ in $[\gamma]_t$ without increasing the action, we can always assume that either $\|\dot\xi\|_\infty \le \kappa_{c_\gamma}$ or $\xi \in [\gamma]^\flat(a,t)$ for some $a > c_\gamma$. In the first case, we note that

$$c_\gamma \le \alpha_*(0)$$

for $-L(x,0) \le -\alpha(0) \le \alpha_*(0)$ for every $x \in \mathbb{R}^N$. The claim follows by the definition of κ_A as $\alpha_*(0) \le A$.

Let us instead assume that ξ belongs to $[\gamma]^\flat(a,t)$ for some $a > c_\gamma$. In particular, $|\dot\xi(s)| \neq 0$ a.e. on $[0,t]$. Set $J := \{s \in [0,t] \,:\, 0 < |\dot\xi(s)| < \overline{n}\}$. We have

$$M\,t > \int_0^t L(\xi,\dot\xi)\,\mathrm{d}s \ge \int_0^t \alpha(|\dot\xi|)\,\mathrm{d}s \ge \alpha(\overline{n})\,\big|[0,t]\setminus J\,\big|,$$

and hence $|J| > t/2$. Pick up a differentiability point $\overline{s} \in J$ for ξ. By the fact that ξ has an a-Lagrangian parameterization, we derive that

$$a \in H\left(\xi(\overline{s}), \partial_q L(\xi(\overline{s}), \dot\xi(\overline{s}))\right);$$

in particular, $a \le A$ by Proposition 2.1. As $|\dot\xi(s)| \in \Lambda_a(\xi(s), \dot\xi(s)/|\dot\xi(s)|)$ for a.e. $s \in [0,t]$, the claim follows by Lemma 3.4 since $\kappa_a \le \kappa_A$ by definition (10). □

3.3. Further extensions. Let us now consider a Lagrangian $L : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}_+$ which satisfies in place of (L2), condition (L2)′ for some family $(\beta_n)_{n\in\mathbb{N}}$ of convex, nondecreasing, and superlinear functions from $\mathbb{R}_+$ to $\mathbb{R}_+$; i.e., which is uniformly superlinear in q and locally bounded in $\mathbb{R}^N \times \mathbb{R}^N$. It is easy to generalize Theorem 3.1 as follows.

THEOREM 3.19. *Let $L : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}_+$ be an autonomous Lagrangian satisfying conditions* (L1), (L2)′, (L3). *Then the associated function S defined through* (3) *is locally Lipschitz in $\mathbb{R}^N \times \mathbb{R}^N \times (0,+\infty)$. More precisely, for every $M, r > 0$ there exists a constant $K = K\big(M, r, \alpha, (\beta_n)_n\big)$ such that*

$$S \quad \textit{is } K\textit{-Lipschitz continuous in } \overline{C_M(r)},$$

where $C_M(r) := \{(y,t,x) \in B_r \times B_r \times (0,r) \,:\, |x-y| < M\,t\}$.

Proof. For every $n \in \mathbb{N}$, let us denote by S_n the function associated with the Lagrangian $L_n(x,q) := L(x,q)\,\chi_{B_n}(x) + \beta_n(|q|)\,\chi_{\mathbb{R}^N\setminus B_n}(x)$ through (3). We claim that, for every $M, r > 0$, there exists an index $k = k\big(M, r, \alpha, (\beta_n)_n\big)$ such that

$$S = S_k \quad \text{on } C_M(r). \tag{20}$$

Clearly, that is enough to conclude by Theorem 3.1.

Let us fix $M, r > 0$. We first notice that

$$S(y,x,t) < r\,\beta_m\,(M) \qquad \text{for any } (y,t,x) \in C_M(r), \tag{21}$$

where $m := [r] + 1$. Let γ be a curve in $W^{1,1}\big([0,t], \mathbb{R}^N\big)$ connecting y to x such that γ is quasi optimal for $S(y,x,t)$. By (21), it is not restrictive to assume that

$$\int_0^t L(\gamma,\dot\gamma)\,\mathrm{d}s < r\,\beta_m\,(M)\,,$$

in particular,

$$\int_0^t |\dot\gamma|\,\mathrm{d}s < r\Big(\alpha_1 + \beta_m(M)\Big),$$

with $\alpha_1 > 0$ such that $\alpha(|q|) \geq |q| - \alpha_1$ for any $q \in \mathbb{R}^N$. As γ has end-points lying in B_r, we deduce that γ is entirely contained in the open ball B_k with

$$k := \big[r\big(1 + \alpha_1 + \beta_m(M)\big)\big] + 1.$$

Thus

$$S(y,x,t) = \inf\left\{\int_0^t L(\gamma,\dot\gamma)\,\mathrm{d}s \,:\, \gamma(0) = y,\ \gamma(t) = x,\ \gamma([0,t]) \subset B_k\right\}$$

for every $(y,t,x) \in C_M(r)$, and claim (20) follows at once as L coincides with L_k on $B_k \times \mathbb{R}^N$. □

Remark 3.20. Theorem 3.19 still holds if, in place of condition (L3), L satisfies the following weaker convexity assumption:

(L3)′ For every $t_1 < t_2$ in $\mathbb{R}$ and $\gamma \in W^{1,\infty}\big((t_1,t_2),\mathbb{R}^N\big)$,

$$\lambda \mapsto L(\gamma(s), \lambda\,\dot\gamma(s)) \quad \text{is convex on } \mathbb{R} \qquad \text{for a.e. } s \in (t_1,t_2).$$

The reparameterization techniques and the approach described above can be in fact adapted to this setting. The lack of convexity of L in q gives rise to some technical difficulties. For instance, it is no longer true that L is the Fenchel transform of H. These obstructions can be overcome by computing the Fenchel transform of L along straight lines of any fixed direction, and by accordingly modifying the definition of σ_a. For the details, see [20].

We conclude this section by recording a result proved in [20] that we will need later. Let us denote by $\mathcal{L} = \mathcal{L}\big(\alpha,(\beta_n)_n\big)$ the family of Lagrangians satisfying assumptions (L1), (L2)′, (L3)′, where α and β_n, $n \in \mathbb{N}$, are fixed, convex, nondecreasing, and superlinear functions from $\mathbb{R}_+$ to $\mathbb{R}_+$. Let $\Sigma = \Sigma\big(\alpha,(\beta_n)_n\big)$ be the space of functions S on $\mathbb{R}^N \times \mathbb{R}^N \times (0,+\infty)$ associated via (3) with Lagrangians belonging to $\mathcal{L}$. We endow Σ with the metric induced by the uniform convergence on the compact subset of $\mathbb{R}^N \times \mathbb{R}^N \times (0,+\infty)$. The following result holds (see [20]).

THEOREM 3.21. *The space of functions Σ is compact; i.e., every sequence $(S_k)_k$ in Σ admits a subsequence which converges to some element S of Σ, locally uniformly in $\mathbb{R}^N \times \mathbb{R}^N \times (0,+\infty)$.*

Remark 3.22. The Lagrangian L associated with the limit function S via (3) can be obtained by "differentiation" as follows:

$$L(x,q) = \lim_{h\to 0^+} \frac{S(x,x+hq,h)}{h} \qquad \text{for every } (x,q) \in \mathbb{R}^N \times \mathbb{R}^N. \tag{22}$$

As proved in [20], L is continuous in q for every x and convex for almost every x. However, the convexity of $L(x,\cdot)$ for every $x \in \mathbb{R}^N$ is not ensured, even if the approximating functions S_k are associated with Lagrangians convex in q.

4. Main theorems.

4.1. Lipschitz-regularity of the value function. We now use the information gathered so far to prove some regularity properties of the value function $v : (0,+\infty)\times\mathbb{R}^N \to \mathbb{R}$ defined as

$$v(t,x) := \inf\left\{ u(\gamma(0)) + \int_0^t L(\gamma,\dot\gamma)\,\mathrm{d}s \;:\; \gamma\in W^{1,1}\big([0,t],\mathbb{R}^N\big),\ \gamma(t)=x\right\}, \tag{23}$$

where $u:\mathbb{R}^N\to(-\infty,+\infty]$, $u\not\equiv+\infty$, and L is a Lagrangian satisfying assumptions (L1)–(L3). The above formula can be equivalently restated as

$$v(t,x) = \inf_{y\in\mathbb{R}^N}\Big(u(y)+S_y(t,x)\Big), \tag{24}$$

where $S_y(t,x)$ stands for the function $S(y,x,t)$ associated with L via (3). In what follows, $\operatorname{dom} g$ will denote the effective domain of the function $g:\mathbb{R}^N\to[-\infty,+\infty]$, i.e., the subset of $\mathbb{R}^N$ where g is finite valued; g^+ will denote the positive part of g, namely, $g^+(x) := \max\{g(x),0\}$ for every $x\in\mathbb{R}^N$. We will also use the following notation:

$$\|g\|_{\ell^\infty(E)} := \sup_{x\in E}|g(x)|, \qquad E\subseteq\mathbb{R}^N,$$

or, more simply, $\|g\|_{\ell^\infty}$ when $E=\mathbb{R}^N$.

THEOREM 4.1. *Let v be defined by* (23) *for some $L:\mathbb{R}^N\times\mathbb{R}^N\to\mathbb{R}_+$ satisfying conditions* (L1)–(L3)*. The following facts hold:*

(i) *If $u\not\equiv+\infty$, then*

$$\lim_{t\to0^+}\|\,(v(t,\cdot)-u)^+\,\|_{\ell^\infty(\operatorname{dom}u)}=0, \tag{25}$$

in particular, $\lim_{t\to0^+}v(t,x)\le u(x)$ for every $x\in\mathbb{R}^N$;

(ii) *if u is bounded from below, then $v(t,x)$ is locally Lipschitz in $(0,+\infty)\times\mathbb{R}^N$;*

(iii) *if u is either bounded or in $\mathrm{UC}(\mathbb{R}^N)$, then, for any $t_0>0$, there exists a constant $K_{t_0}=K(t_0,u,\alpha,\beta)$ such that*

$$v(t,x) \qquad \text{is } K_{t_0}\text{-Lipschitz in} \quad [t_0,+\infty)\times\mathbb{R}^N;$$

(iv) *if $u\in\mathrm{Lip}(\mathbb{R}^N)$, then there exists a constant $K=K(u,\alpha,\beta)$ such that*

$$v(t,x) \qquad \text{is } K\text{-Lipschitz in} \quad [0,+\infty)\times\mathbb{R}^N.$$

Proof. Pick a point $x_0\in\operatorname{dom}(u)$ and plug $y=x_0$ into the expression on the right-hand side of (24). We get

$$v(t,x)\le u(x_0)+t\,\beta\left(\frac{|x-x_0|}{t}\right). \tag{26}$$

Inequality (26) with $x_0=x$ immediately gives (25) whenever $x\in\operatorname{dom}(u)$.

Let us now assume that u is bounded from below. By inequality (26), any $y\in\mathbb{R}^N$ which is t-optimal for $v(t,x)$ satisfies

$$u(y)+S_y(t,x)\le u(x_0)+t\left(\beta\left(\frac{|x-x_0|}{t}\right)+1\right),$$

and that yields, for $y \neq x$,

$$\frac{\alpha\left(\frac{|x-y|}{t}\right)}{\frac{|x-y|}{t}} \leq \frac{u(x_0)-u(y)}{|x-y|} + \left(\beta\left(\frac{|x-x_0|}{t}\right)+1\right)\frac{t}{|x-y|}. \tag{27}$$

When (t,x) varies in an open set U compactly contained in $(0,+\infty)\times\mathbb{R}^N$, inequality (27) is certainly false if $(y,t,x) \notin \overline{C_M}$ for a suitably large $M = M(U, \|(-u)^+\|_{\ell^\infty}, \alpha, \beta)$. Hence

$$v(t,x) = \inf_y \left\{ u(y) + S_y(t,x) \,:\, (y,t,x) \in \overline{C_M} \right\} \qquad \text{for every } (t,x) \in U,$$

and assertion (ii) follows, as v is the infimum of a family of equi-Lipschitz functions, by Theorem 3.1.

Items (iii) and (iv) can be proved analogously by replacing x_0 with x in (27) and by choosing as U the sets $(t_0,+\infty)\times\mathbb{R}^N$ and $(0,+\infty)\times\mathbb{R}^N$, respectively. For the case $u \in \mathrm{UC}(\mathbb{R}^N)$, we also use the fact that, for any such u, there exists $\varepsilon > 0$ such that

$$|u(x)-u(y)| < \frac{|x-y|}{\varepsilon} \qquad \text{for every } x,y \in \mathbb{R}^N \text{ with } |x-y| > \varepsilon. \qquad \square$$

Let us now assume that the Lagrangian L satisfies, in place of (L2), condition (L2)′ for some family $(\beta_n)_{n\in\mathbb{N}}$ of convex, nondecreasing, and superlinear functions from $\mathbb{R}_+$ to $\mathbb{R}_+$. We provide the following generalization of Theorem 4.1.

THEOREM 4.2. *Let v be defined by* (23) *for some $L : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}_+$ satisfying conditions* (L1), (L2)′, (L3). *The following facts hold:*

(i) *If $u \not\equiv +\infty$, then*

$$\lim_{t\to 0^+} \| \left(v(t,\cdot)-u\right)^+ \|_{\ell^\infty(B_r \cap \mathrm{dom}\, u)} = 0 \qquad \textit{for every } r > 0,$$

in particular, $\lim_{t\to 0^+} v(t,x) \leq u(x)$ *for every $x \in \mathbb{R}^N$;*

(ii) *if u is either bounded from below or in* $\mathrm{UC}(\mathbb{R}^N)$, *then $v(t,x)$ is locally Lipschitz in $(0,+\infty)\times\mathbb{R}^N$. More precisely, for every open set U compactly contained in $(0,+\infty)\times\mathbb{R}^N$, there exists a constant $K = K\big(U,u,\alpha,(\beta_n)_n\big)$ such that*

$$v(t,x) \qquad \textit{is } K\textit{-Lipschitz in } U.$$

The proof is omitted, for it can be easily recovered by arguing as above and by using Theorem 3.19 in place of Theorem 3.1.

Lastly, we want to point out that all results of this paper can be easily extended to the case when $\mathbb{R}^N$ is replaced by a connected smooth Riemannian manifold $\mathcal{M}$ without boundary. In this case, the Lagrangian L is defined on the tangent bundle $T\mathcal{M}$ of $\mathcal{M}$ and satisfies assumptions (L1), (L2) or (L2)′, (L3), with $T\mathcal{M}$, $\|\cdot\|_x$, and $\mathcal{M}$ in place of $\mathbb{R}^N\times\mathbb{R}^N$, $|\cdot|$ and $\mathbb{R}^N$, respectively.[3] When $\mathcal{M}$ is compact, Theorem 4.1 can be partially improved as follows.

PROPOSITION 4.3. *Let $\mathcal{M}$ be a compact and connected smooth Riemannian manifold without boundary and $L : T\mathcal{M} \to \mathbb{R}_+$ an autonomous Lagrangian satisfying conditions* (L1)–(L3), *with $T\mathcal{M}$, $\|\cdot\|_x$, and $\mathcal{M}$ in place of $\mathbb{R}^N\times\mathbb{R}^N$, $|\cdot|$, and $\mathbb{R}^N$,*

[3] $\|\cdot\|_x$ denotes the Riemannian norm on $T_x\mathcal{M}$ for every $x \in \mathcal{M}$.

respectively. Let v be defined by (23) with $u \not\equiv +\infty$ and bounded from below. Then, for any $t_0 > 0$, there exists a constant $K_{t_0} = K(t_0, \alpha, \beta)$ such that

$$v(t,x) \qquad \text{is } K_{t_0}\text{-Lipschitz in} \quad [t_0, +\infty) \times \mathcal{M}.$$

Remark 4.4. Our point is that the constant K_{t_0} appearing above is independent of the initial cost u, provided v is a well-defined real function on $(0, +\infty) \times \mathcal{M}$. This is actually guaranteed by the conditions $u \not\equiv +\infty$ and $\inf_{\mathcal{M}} u > -\infty$.

Proof. Let us denote by $\delta_{\mathcal{M}}$ the distance on $\mathcal{M}$ induced by its Riemannian metric. For any $M > 0$, set

$$C_M := \{(y, x, t) \in \mathcal{M} \times \mathcal{M} \times (0, +\infty) \,:\, \delta_{\mathcal{M}}(x, y) < M\, t\}.$$

Since $\mathcal{M}$ is compact, for any $t_0 > 0$ there exists M_{t_0}, depending on t_0 and on the diameter of $\mathcal{M}$ only, such that $\mathcal{M} \times \mathcal{M} \times [t_0, +\infty) \subset C_{M_{t_0}}$. Hence

$$v(t,x) = \inf_y \left\{ u(y) + S_y(t,x) \,:\, (y,x,t) \in C_{M_{t_0}} \right\} \quad \text{for any } (t,x) \in [t_0, +\infty) \times \mathcal{M},$$

and the assertion follows, as v is the infimum of a family of equi-Lipschitz functions by Theorem 3.1. ∎

4.2. A compactness result for value functions. Let $\mathcal{L} = \mathcal{L}\big(\alpha, (\beta_n)_n\big)$ be defined as in subsection 3.3, and let $(L_k)_k$ be a sequence of Lagrangians belonging to $\mathcal{L}$ and convex in q. For each $k \in \mathbb{N}$, let

$$v_k(t,x) := \inf \left\{ u_k(\gamma(0)) + \int_0^t L_k(\gamma(s), \dot\gamma(s))\, \mathrm{d}s \;:\; \gamma \in W^{1,1}\big([0,t], \mathbb{R}^N\big),\ \gamma(t) = x \right\}$$

for every $(t,x) \in (0, +\infty) \times \mathbb{R}^N$, where u_k is a function from $\mathbb{R}^N$ to $(-\infty, +\infty]$ with $u_k \not\equiv +\infty$. With the aid of Theorem 3.21, we can prove the following result.

THEOREM 4.5. *Let $(v_k)_k$ be defined as above, and suppose one of the following conditions holds:*

(a) *The functions u_k are equibounded from below on $\mathbb{R}^N$;*

(b) *the functions u_k are equiuniformly continuous on $\mathbb{R}^N$.*

Then, up to subsequences, $(v_k)_k$ locally uniformly converges on $(0, +\infty) \times \mathbb{R}^N$ to the function v defined as

$$v(t,x) := \inf \left\{ u_*(\gamma(0)) + \int_0^t L(\gamma(s), \dot\gamma(s))\, \mathrm{d}s \;:\; \gamma \in W^{1,1}\big([0,t], \mathbb{R}^N\big),\ \gamma(t) = x \right\},$$

where L is a Lagrangian belonging to $\mathcal{L}$ and u_* is the function defined as

$$u_*(x) := \inf \left\{ \liminf_k u_k(x_k) \,:\, x_k \to x \right\} \quad \textit{for every } x \in \mathbb{R}^N.$$

Proof. Let us denote by S_k the function associated with L_k via (3). By Theorem 3.21 we know that, up to subsequences, S_k converge to S, locally uniformly on $\mathbb{R}^N \times \mathbb{R}^N \times (0, +\infty)$. Let L be the element of $\mathcal{L}$ derived from S via (22). For every $M, r > 0$, the functions S_k are equi-Lipschitz continuous. Moreover, for every open set U compactly contained in $(0, +\infty) \times \mathbb{R}^N$ there exists a constant M independent of k such that

$$v_k(t,x) = \inf_y \left\{ u_k(y) + S_k(y,x,t) \,:\, (y,t,x) \in \overline{C_M(r)} \right\} \quad \text{for every } (t,x) \in U,$$

where r is a sufficiently large positive number such that $U \subset (0,r) \times B_r$. To see this, argue as in the proof of Theorem 4.1 and note that M can be estimated in terms of $\sup_k \|(-u_k)^+\|_{\ell^\infty}$ or of the continuity modulus shared by the functions $(u_k)_k$. In particular, the functions v_k are equi-Lipschitz continuous on U. By the Ascoli–Arzelà theorem, the proof then reduces to showing that

$$\lim_k v_k(t,x) = v(t,x) \qquad \text{for every } (t,x) \in U.$$

Let us first prove that

$$v(t,x) \geq \limsup_k v_k(t,x) \qquad \text{for every } (t,x) \in U. \tag{28}$$

Choose $y \in \mathbb{R}^N$ and let $y_k \to y$ such that $u_k(y_k)$ converge to $u_*(y)$. We have

$$u_*(y) + S(y,x,t) = \lim_k u_k(y_k) + S_k(y_k,x,t) \geq \limsup_k v_k(t,x),$$

and (28) follows by taking the infimum of the above inequality for all $y \in \mathbb{R}^N$. Next, let us prove that

$$\liminf_k v_k(t,x) \geq v(t,x) \qquad \text{for every } (t,x) \in U. \tag{29}$$

For each $k \in \mathbb{N}$, take y_k such that $(y_k,t,x) \in \overline{C_M(r)}$ and

$$v_k(t,x) + \frac{1}{k} \geq u_k(y_k) + S_k(y_k,x,t).$$

By possibly considering a subsequence, we can assume that $(y_k)_k$ converges to some point $y \in \mathbb{R}^N$. We infer that

$$\liminf_k v_k(t,x) \geq \liminf_k u_k(y_k) + S(y,x,t) \geq u_*(y) + S(y,x,t),$$

and (29) follows. □

Acknowledgment. The author wishes to thank the anonymous referee who has carefully read the article, pointed out many errors and misprints in previous versions of the paper, and whose critical remarks have been helpful in improving the presentation.

REFERENCES

[1] M. Amar, G. Bellettini, and S. Venturini, *Integral representation of functionals defined on curves of $W^{1,p}$*, Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998), pp. 193–217.

[2] L. Ambrosio, O. Ascenzi, and G. Buttazzo, *Lipschitz regularity for minimizers of integral functionals with highly discontinuous integrands*, J. Math. Anal. Appl., 142 (1989), pp. 301–316.

[3] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations. With appendices by Maurizio Falcone and Pierpaolo Soravia*, Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 1997.

[4] G. Barles, *Solutions de viscosité des équations de Hamilton–Jacobi*, Math. Appl. (Berlin) 17, Springer-Verlag, Paris, 1994.

[5] G. Barles and J. M. Roquejoffre, *Large time behaviour of fronts governed by eikonal equations*, Interfaces Free Bound., 5 (2003), pp. 83–102.

[6] A. Briani and A. Davini, *Monge solutions for discontinuous Hamiltonians*, ESAIM Control Optim. Calc. Var., 11 (2005), pp. 229–251.

[7] G. Buttazzo, M. Giaquinta, and S. Hildebrandt, *One-Dimensional Variational Problems. An Introduction*, Oxford Lecture Ser. Math. Appl., 15, The Clarendon Press, Oxford University Press, New York, 1998.

[8] L. Caffarelli, M. G. Crandall, M. Kocan, and A. Swiech, *On viscosity solutions of fully nonlinear equations with measurable ingredients,* Comm. Pure Appl. Math., 49 (1996), pp. 365–397.

[9] F. Camilli, A. Davini, and A. Siconolfi, *Lax-type formulas for time-dependent measurable Hamilton–Jacobi equations,* in preparation.

[10] F. Camilli, *An Hopf-Lax formula for a class of measurable Hamilton–Jacobi equations*, Nonlinear Anal., 57 (2004), pp. 265–286.

[11] F. Camilli and A. Siconolfi, *Hamilton–Jacobi equations with measurable dependence on the state variable*, Adv. Differential Equations, 8 (2003), pp. 733–768.

[12] F. Camilli and A. Siconolfi, *Time-dependent measurable Hamilton–Jacobi equations*, Comm. Partial Differential Equations, 30 (2005), pp. 813–847.

[13] F. Camilli and A. Siconolfi, *Effective Hamiltonian and homogenization of measurable eikonal equations*, Arch. Ration. Mech. Anal., 183 (2007), pp. 1–20.

[14] G. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions*, Springer-Verlag, Berlin, 1977.

[15] L. Cesari, *Optimization Theory and Applications*, Springer-Verlag, New York, 1983.

[16] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[17] F. H. Clarke and R. B. Vinter, *Regularity properties of solutions to the basic problem in the calculus of variations*, Trans. Amer. Math. Soc., 289 (1985), pp. 73–98.

[18] G. Dal Maso and H. Frankowska, *Autonomous integral functionals with discontinuous nonconvex integrands: Lipschitz regularity of minimizers, DuBois–Reymond necessary conditions, and Hamilton–Jacobi equations*, Appl. Math. Optim., 48 (2003), pp. 39–66.

[19] G. Dal Maso and H. Frankowska, *Value functions for Bolza problems with discontinuous Lagrangians and Hamilton–Jacobi inequalities*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 369–393.

[20] A. Davini, *Integral representation of abstract functionals of autonomous type*, Proc. Roy. Soc. Edinburgh Sect. A, to appear.

[21] A. Davini and A. Siconolfi, *A generalized dynamical approach to the large time behavior of solutions of Hamilton–Jacobi equations*, SIAM J. Math. Anal., 38 (2006), pp. 478–502.

[22] A. Fathi and A. Siconolfi, *PDE aspects of Aubry–Mather theory for continuous convex Hamiltonians*, Calc. Var. Partial Differential Equations, 22 (2005), pp. 185–228.

[23] G. N. Galbraith, *Extended Hamilton–Jacobi characterization of value functions in optimal control*, SIAM J. Control Optim., 39 (2000), pp. 281–305.

[24] M. Giaquinta and G. Modica, *Analisi Matematica*, Vol. 3, Pitagora Editrice, Bologna, 2000.

[25] A. D. Ioffe, *On the lower semicontinuity of integral functionals.* I, SIAM J. Control Optim., 15 (1977), pp. 521–538.

[26] P. L. Lions, *Generalized Solutions of Hamilton–Jacobi Equations*, Res. Notes in Math. 69, Pitman (Advanced Publishing Program), Boston, MA, London, 1982.

[27] G. Namah and J. M. Roquejoffre, *The "hump" effect in solid propellant combustion*, Interfaces Free Bound., 2 (2000), pp. 449–467.

[28] R. T. Newcomb and J. Su, *Eikonal equations with discontinuities*, Differential Integral Equations, 8 (1995), pp. 1947–1960.

[29] C. Olech, *Weak lower semicontinuity of integral functionals. Existence theorem issue*, J. Optim. Theory Appl., 19 (1976), pp. 3–16.

[30] D. Ostrov, *Solutions of Hamilton–Jacobi equations and scalar conservation laws with discontinuous space-time dependence*, J. Differential Equations, 182 (2002), pp. 51–77.

[31] R. T. Rockafellar, *Convex Analysis*, Princeton Mathematical Series 28, Princeton University Press, Princeton, NJ, 1970.

[32] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. [Fundamental Principles of Mathematical Sciences] 317, Springer-Verlag, Berlin, 1998.

[33] P. Soravia, *Boundary value problems for Hamilton–Jacobi equations with discontinuous Lagrangian*, Indiana Univ. Math. J., 51 (2002), pp. 451–477.

[34] T. Strömberg, *On viscosity solutions of irregular Hamilton–Jacobi equations*, Arch. Math. (Basel) 81 (2003), pp. 678–688.

[35] L. Tonelli, *Sur une méthode directe du calcul des variations*, Rend. Circ. Mat. Palermo, 39 (1915), pp. 223–264.

[36] L. Tonelli, *Fondamenti di Calcolo delle Variazioni*, Vols. 1 & 2, Zanichelli, Bologna, 1921, 1923.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 1923–1941

EXISTENCE AND NONEXISTENCE RESULTS OF AN OPTIMAL CONTROL PROBLEM BY USING RELAXED CONTROL*

HONGWEI LOU†

Abstract. Relaxed controls have proved to be very useful in studying the existence of optimal controls in optimal control theory. Many positive results have been obtained in the literature. However, negative results have also made their rare appearances. The optimal control problem considered in this paper looks quite simple. Yet, by treating such a problem, we can get interesting results, substantiating our idea as to whether an optimal control exists or not. In our opinion, the method used in the paper can be applied to more generalized cases.

Key words. optimal control, existence, nonexistence, relaxed control

AMS subject classifications. 49J20, 49J45

DOI. 10.1137/050628386

1. Introduction. Since the introduction by Young [37] and McShane [25] in the late 1930s, generalized curves have been widely used in existence theories for calculus of variations and optimal control theory, especially for problems that lack convex conditions.

In optimal control theory, generalized curves were transformed as relaxed controls by Gamkrelidze [15], McShane [26], and Warga [35]; see also Warga [36]. Thanks to the useful Filippov lemma (see [13] and Corollary 2.26 in Chapter 3 of [18]), it becomes quite convenient to prove that, under some proper conditions, there exists a (classical) control which has the same effect as an optimal relaxed control, leading to the existence of an optimal (classical) control. It is our opinion that relaxed control is the most important tool for studying the existence of optimal controls when a convex condition such as Cesari's condition (see [18], for example) is not assumed.

To the best of our knowledge, the first result on the existence of optimal classical controls without assuming convexity conditions was established by Neustadt [28] for (finite-dimensional) linear systems. Various later results in finite-dimensional spaces were obtained by Artstein [2], Balder [3], [4], [5], [6], Berliocchi and Lasry [8], Cellina and Colombo [9], Cesari [10], Colombo and Goncharov [11], Marcellini [23], Mariconda [24], Olech [29], and Raymond [31], [32], to mention a few. On the other hand, in infinite-dimensional cases, fewer results can be found in the literature. The readers are referred to Flores-Bazán and Perrotta [14] and Suryanarayana [34] for hyperbolic equations and to Lou [19], [20], [22] and Raymond [33] for elliptic and parabolic cases.

For general results concerning the existence of optimal relaxed controls, see [1], [12], and [30] and the references cited therein.

Concerning the nonexistence results, the following is a typical counterexample which has been mentioned in many books; see, for example, [36, Ch. 3, p. 246]. Other similar examples can be found in [10, Ch. 9, p. 321] and [12, Ch. 2, p. 51].

*Received by the editors April 2, 2005; accepted for publication (in revised form) April 12, 2007; published electronically November 28, 2007. This work was supported by NSFC (10671040), FANEDD (200522), and NCET (06-0352).

http://www.siam.org/journals/sicon/46-6/62838.html

†School of Mathematical Sciences, Fudan University, Shanghai 200433, China, and Key Laboratory of Mathematics for Nonlinear Sciences (Fudan University), Ministry of Education, Shanghai, China (hwlou@fudan.edu.cn).

EXAMPLE 1.1. *Let* $U = [-1, 1]$ *(or* $U = \{-1, 1\}$*),*

$$\mathcal{U} = \Big\{ v(\cdot) : [0,1] \to U \;\Big|\; v(\cdot) \text{ measurable} \Big\},$$

$$\frac{dy(t)}{dt} = u(t), \qquad t \in [0,1],$$

and

$$I(u(\cdot)) = \int_0^1 \Big(y^2(t) - u^2(t) \Big)\, dt.$$

Then it is easy to see that

$$I(u(\cdot)) > -1 = \inf_{v(\cdot)\in\mathcal{U}} I(v(\cdot)) \quad \forall\, u(\cdot) \in \mathcal{U}. \tag{1.1}$$

Thus

$$I(\bar{u}(\cdot)) = \inf_{v(\cdot)\in\mathcal{U}} I(v(\cdot))$$

has no solution. That is, optimal control does not exist.

We point out that, in the literature, rare general results were found concerning the nonexistence of optimal controls by using relaxed control.

Now, we would like to state the problems that we are going to study in this paper. To this end, let $\Omega \subset \mathbb{R}^n$ be a bounded domain with a smooth boundary $\partial\Omega$, let $M > 0$ be a given constant, and let[1]

$$\mathcal{U} = \Big\{ v(\cdot) : \Omega \to [0, M] \;\Big|\; v(\cdot) \text{ measurable} \Big\}.$$

For a positive integer m and $p \in [1, +\infty)$, denote by $W^{m,p}(\Omega)$ the usual Sobolev space, i.e.,

$$W^{m,p}(\Omega) = \big\{ f \in L^p(\Omega) \;\big|\; \forall\, |\rho| \le m,\, \partial^\rho f \in L^p(\Omega) \big\}$$

endowed with norm

$$\|f\|_{W^{m,p}(\Omega)} = \sum_{|\rho|\le m} \|\partial^\rho f\|_{L^p(\Omega)},$$

where $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$ is called a multi-index, which is an n-tuple of nonnegative integers ρ_i, $|\rho| = \sum_{i=1}^n \rho_i$, and the derivatives $\partial^\rho f = \partial_{x_1}^{\rho_1} \ldots \partial_{x_n}^{\rho_n} f$ are taken in a weak sense. Denote by $W_0^{m,p}(\Omega)$ the closure of $C_0^\infty(\Omega)$ in $W^{m,p}(\Omega)$. In this paper, we consider the following controlled system:

$$\begin{cases} -\triangle\, y(x) = u(x) & \text{in } \Omega, \\ y|_{\partial\Omega} = 0, \end{cases} \tag{1.2}$$

[1]Naturally, $\mathcal{U}$ is in fact a set of equivalence classes. Two measurable functions $u(\cdot), v(\cdot) : \Omega \to [0, M]$ will appear as the same element of $\mathcal{U}$ if $u(x) = v(x)$ a.e. Ω. The definition of $\mathcal{R}(\Omega, U)$ in section 2 is similar.

where the control $u(\cdot)$ belongs to $\mathcal{U}$ and the corresponding state $y(\cdot)$ is the solution of (1.2). Throughout this paper, by a solution $y(\cdot)$ of (1.2) corresponding to control $u(\cdot)$, we mean $y(\cdot) \in W_0^{1,2}(\Omega)$ and

$$\int_\Omega \nabla y(x) \cdot \nabla \varphi(x)\, dx = \int_\Omega u(x)\varphi(x)\, dx \quad \forall\, \varphi \in C_0^\infty(\Omega).$$

We would like to mention that under the assumptions of this paper, any weak solution $y(\cdot)$ must be in $W^{2,2}(\Omega)$, and therefore, it is necessarily a strong solution.

Next, we introduce the following cost functional:

$$J(u(\cdot)) = \int_\Omega F\big(y(x), u(x)\big)\, dx. \tag{1.3}$$

Then our optimal control problem can be stated as follows.

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

$$J(\bar{u}(\cdot)) = \inf_{u(\cdot)\in\mathcal{U}} J(u(\cdot)). \tag{1.4}$$

Any control $\bar{u}(\cdot) \in \mathcal{U}$ satisfying (1.4) is called an optimal (classical) control.

In the rest of this paper, some general existence and nonexistence results for optimal controls will be presented by means of relaxed controls.

2. Relaxation. In this section, we would like to introduce and study the relaxation of our Problem (C). To begin with, let $U = [0, M]$, denote by $\mathcal{M}_+^1(U)$ the set of all probability measures in U, and denote by $\mathcal{R}(\Omega, U)$ the set of all measurable $\mathcal{M}_+^1(U)$-valued functions on Ω. Clearly, $\sigma(\cdot) \in \mathcal{R}(\Omega, U)$ if and only if

$$\sigma(x) \in \mathcal{M}_+^1(U) \quad \text{a.e. } x \in \Omega,$$

and

$$x \mapsto \int_U h(v)\sigma(x)(dv) \text{ is measurable} \quad \forall\, h \in C(U),$$

where $C(U)$ denotes the space of continuous functions on U. Let $C(U)^*$ and $L^1(\Omega; C(U))^*$ be the dual spaces of $C(U)$ and $L^1(\Omega; C(U))$ with the weak star topology, respectively. We regard $\mathcal{M}_+^1(U)$ and $\mathcal{R}(\Omega, U)$ as subspaces of $C(U)^*$ and $L^1(\Omega; C(U))^*$, respectively, by setting

$$\theta(h) \triangleq \int_U h(v)\theta(dv) \quad \forall\, \theta \in \mathcal{M}_+^1(U), \quad h \in C(U) \tag{2.1}$$

and

$$\sigma(g) \triangleq \int_\Omega dx \int_U g(x, v)\sigma(x)(dv) \quad \forall\, \sigma \in \mathcal{R}(\Omega, U), \quad g \in L^1(\Omega; C(U)), \tag{2.2}$$

where (2.2) is well defined by [36, Theorem IV.1.6, p. 266]. Thus,

$$\sigma_k \to \sigma \quad \text{in } \mathcal{R}(\Omega, U)$$

means that

$$\int_\Omega dx \int_U h(x,v)\sigma_k(x)(dv) \to \int_\Omega dx \int_U h(x,v)\sigma(x)(dv)$$
$$\forall\, h \in L^1(\Omega; C(U)).$$

Now, we state our optimal relaxed control problem corresponding to Problem (C) as follows.

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

$$J(\bar\sigma(\cdot)) = \inf_{\sigma(\cdot)\in\mathcal{R}(\Omega,U)} J(\sigma(\cdot)), \tag{2.3}$$

where

$$J(\sigma(\cdot)) \stackrel{\triangle}{=} \int_\Omega dx \int_U F\big(y(x), v\big)\sigma(x)(dv), \tag{2.4}$$

and $y(\cdot)$ is the state corresponding to relaxed control $\sigma(\cdot) \in \mathcal{R}(\Omega, U)$; namely, it is the solution of the following:

$$\begin{cases} -\triangle\, y(x) = \displaystyle\int_U v\sigma(x)(dv) \quad \text{in } \Omega, \\ y|_{\partial\Omega} = 0. \end{cases} \tag{2.5}$$

Note that $\mathcal{U}$ can be imbedded into $\mathcal{R}(\Omega, U)$ by identifying each $u(\cdot) \in \mathcal{U}$ with the Dirac measure-valued function $\delta_{u(\cdot)} \in \mathcal{R}(\Omega, U)$. Moreover, $J(\delta_{u(\cdot)})$ defined by (2.4) coincides with $J(u(\cdot))$ defined by (1.3). Thus, notation $J(\sigma(\cdot))$ would not cause any confusion. On the other hand, it is known that $\mathcal{U}$ is dense in $\mathcal{R}(\Omega, U)$; i.e., for any $\sigma(\cdot) \in \mathcal{R}(\Omega, U)$, there exists a sequence $u_k(\cdot)$ in $\mathcal{U}$ such that

$$\delta_{u_k(\cdot)} \to \sigma(\cdot) \quad \text{in } \mathcal{R}(\Omega, U);$$

see, for example, Lemma 2.4 in [21] (where the result is more general); see also [36].

By the density of $\mathcal{U}$ in $\mathcal{R}(\Omega, U)$, one can easily get that, under some weak assumptions (for example, when $F(\cdot,\cdot)$ is continuous),

$$\inf_{\sigma(\cdot)\in\mathcal{R}(\Omega,U)} J(\sigma(\cdot)) = \inf_{u(\cdot)\in\mathcal{U}} J(u(\cdot)). \tag{2.6}$$

Thus, an optimal control of Problem (C) must be an optimal relaxed control of Problem (R). Furthermore, for an optimal relaxed control $\bar\sigma(\cdot)$, if there exists a $\bar u : \Omega \to U$ such that

$$\bar\sigma(x) = \delta_{\bar u(x)} \quad \text{a.e. } x \in \Omega,$$

then $\bar u(\cdot)$ must be measurable, that is, $\bar u(\cdot) \in \mathcal{U}$ (we simply say that $\bar\sigma(\cdot) \in \mathcal{U}$ in this case). Thus, $\bar u(\cdot)$ must be an optimal control of Problem (C). In other words, if Problem (R) has an optimal relaxed control $\bar\sigma(\cdot) \in \mathcal{R}(\Omega, U)$ such that $\operatorname{supp}\bar\sigma(x)$ (the support of $\bar\sigma(x)$) is a singleton of U for almost all $x \in \Omega$, then Problem (C) admits at least one optimal classical control, whereas, if every optimal relaxed control of Problem (R) is not an element of $\mathcal{U}$, then Problem (C) admits no optimal control.

Now, we recall that for (1.2)–(1.3), if $F(\cdot,\cdot)$ is continuous on $\mathbb{R}\times U$, then Cesari's condition is equivalent to $F(y,\cdot)$ being convex. For a general system, the definition

of Cesari's condition is a little bit more complicated, and we refer the readers to [7], [10], and [18]. As an example, we consider a system governed by ordinary differential equations:

$$\frac{dy}{dt} = f(t, y(t), u(t)). \tag{2.7}$$

The cost functional has the form

$$I(u(\cdot)) = \int_{t_0}^{T} f^0(t, y(t), u(t))\, dt,$$

where $y(t) \in \mathbb{R}^m$ and

$$u(\cdot) \in \mathcal{U} = \Big\{ v(\cdot) : [t_0, T] \to U \;\Big|\; v(\cdot) \text{ measurable} \Big\}.$$

Then the definition of Cesari's condition[2] is that for almost all $t \in [t_0, T]$, the set

$$\mathcal{E}(t, y) \triangleq \bigcup_{u \in U} \Big\{ (z, z^0) \in \mathbb{R}^m \times \mathbb{R} \;\Big|\; z = f(t, y, u),\ z^0 \geq f^0(t, y, u) \Big\}$$

has the Cesari property at every $y \in \mathbb{R}^m$:

$$\mathcal{E}(t, y) = \bigcap_{\delta > 0} \overline{\text{co}} \left(\bigcup_{|\tilde{y} - y| < \delta} \mathcal{E}(t, \tilde{y}) \right).$$

For the characterization of the above condition, see Lemma 4.1 of [4], Proposition 2.18 of [5], and Chapter 3 of [18].

One of the most important advantages of relaxed control is that optimal relaxed control exists under relatively weaker conditions. The following lemma is crucial in deriving existence of optimal relaxed controls.

LEMMA 2.1. *Suppose U is a compact metric space. Then $\mathcal{R}(\Omega, U)$ is convex and sequentially compact.*

For a proof of the above lemma, see Warga [36, Theorem IV.2.1, p. 272].

Now we will state the existence and necessary condition for an optimal relaxed control. It is a corollary of Theorems 3.2 and 4.1 in [19].

PROPOSITION 2.2. *Let $M > 0$ and $U = [0, M]$. Suppose that $F \in C^1(\mathbb{R} \times U)$. Then Problem* (R) *admits at least one optimal relaxed control. Let $\bar{\sigma}(\cdot)$ be an optimal relaxed control and $\bar{y}(\cdot)$ be the state corresponding to $\bar{\sigma}(\cdot)$, i.e.,*

$$\begin{cases} -\triangle\, \bar{y}(x) = \displaystyle\int_U v \bar{\sigma}(x)(dv) & \text{in } \Omega, \\ \bar{y}|_{\partial\Omega} = 0. \end{cases} \tag{2.8}$$

Define $\bar{\psi}(\cdot)$ to be the solution of

$$\begin{cases} -\triangle\, \bar{\psi}(x) = -\displaystyle\int_U F_y\big(\bar{y}(x), v\big)\, \bar{\sigma}(x)(dv) & \text{in } \Omega, \\ \bar{\psi}|_{\partial\Omega} = 0 \end{cases} \tag{2.9}$$

[2] Generally, we can define Cesari's condition for a general metric set U.

and

$$H(y,w,\psi) \stackrel{\triangle}{=} w\psi - F(y,w). \tag{2.10}$$

Then

$$\operatorname{supp}\bar{\sigma}(x) \subseteq \Big\{ w \in U \;\Big|\; H\big(\bar{y}(x),w,\bar{\psi}(x)\big) = \max_{v\in U} H\big(\bar{y}(x),v,\bar{\psi}(x)\big) \Big\} \quad a.e.\ x\in\Omega. \tag{2.11}$$

We mention that when $\bar{\sigma}(\cdot)$ is a classical control, i.e., $\bar{\sigma}(\cdot) = \delta_{\bar{u}(\cdot)}$ with $\bar{u}(\cdot)\in\mathcal{U}$, (2.11) becomes

$$H\big(\bar{y}(x),\bar{u}(x),\bar{\psi}(x)\big) = \max_{v\in U} H\big(\bar{y}(x),v,\bar{\psi}(x)\big) \quad \text{a.e. } x\in\Omega. \tag{2.12}$$

This is just the usual maximum condition.

Now we conclude this section by presenting a useful lemma. The set $\{x\in\Omega \mid \varphi(x)\in E\}$ will be denoted by $\{\varphi\in E\}$ for simplicity.

LEMMA 2.3. *Let C be a constant. If $\varphi\in W^{m,1}(\Omega)$, $m\geq 1$, then*

$$\partial^\rho\varphi(x) = 0 \quad a.e.\ on\ \{\varphi = C\} \quad \forall\, 1\leq|\rho|\leq m.$$

For the case $m=1$, the above result can be found in Morrey, Jr. [27, p. 69]. See also Kinderlehrer and Stampacchia [17, Ch. 2]. The remaining cases can be obtained easily by induction.

By Lemma 2.3, if $\varphi(\cdot)\in W^{2,1}(\Omega)$, then we have

$$-\triangle\varphi(x) = 0 \quad \text{a.e. on } \{\varphi = C\} \tag{2.13}$$

for any constant C.

3. Main results. In this section, we will present our main results. The proofs are based on careful analysis of the necessary conditions of the optimal relaxed pairs $(\bar{y}(\cdot),\bar{\sigma}(\cdot))$. We introduce the following standing assumptions.

(P1) The function $F\in C^3(\mathbb{R}\times[0,M])$ such that

$$\begin{array}{l} F_{yy}(y,u)\leq 0, \quad F_{yu}(y,u)\geq 0, \quad F_{uu}(y,u)<0 \\ F_{yyy}(y,u)\geq 0, \quad F_{yyu}(y,u)\geq 0 \end{array} \qquad \forall\, y\in\mathbb{R}, \quad u\in[0,M].$$

(P2)

$$F(0,M) - F(0,0) < 0, \qquad F_y(0,0)\leq 0.$$

It is possible to set other conditions different from (P1)–(P2). We will see that the case we choose is nontrivial and interesting.

We now state our main result.

THEOREM 3.1. *Let $M>0$ and* (P1)–(P2) *hold. Define $(z(\cdot),\zeta(\cdot))$ to be the solution of*

$$\begin{cases} -\triangle z(x) = M \quad in\ \Omega, \\ -\triangle\zeta(x) = -F_y\big(z(x),M\big) \quad in\ \Omega, \\ z|_{\partial\Omega} = \zeta|_{\partial\Omega} = 0, \end{cases} \tag{3.1}$$

and let

$$\ell = \inf_{x\in\Omega}\left[\zeta(x) - \frac{F\big(z(x),M\big) - F\big(z(x),0\big)}{M}\right]. \tag{3.2}$$

Then

(i) *if* $\ell \geq 0$, *Problem* (C) *admits at least one optimal control;*

(ii) *if* $\ell < 0$ *and*

$$F_{yy}(y,u) < 0 \quad \forall\, y \in \mathbb{R}, \quad u \in [0,M], \tag{3.3}$$

Problem (C) *admits no optimal control.*

Proof. Consider Problem (R) corresponding to Problem (C). By Proposition 2.2, there exists a $\bar\sigma(\cdot) \in \mathcal{R}(\Omega,U)$ satisfying (2.3). Moreover, let $\bar y(\cdot)$, $\bar\psi(\cdot)$, and $H(\cdot,\cdot,\cdot)$ be defined by (2.8), (2.9), and (2.10), respectively. Then we have (2.11). Further, it follows from (2.8), the L^p-estimate of the elliptic equation, and

$$\left|\int_U v\bar\sigma(x)(dv)\right| \leq M$$

that

$$\bar y(\cdot) \in W^{2,p}(\Omega) \quad \forall\, p \in [1,+\infty). \tag{3.4}$$

Thus,

$$\bar y(\cdot) \in C^{1,\alpha}(\overline{\Omega}) \quad \forall\, \alpha \in (0,1) \tag{3.5}$$

by the Sobolev imbedding theorem (see [16, Ch. 7], for example). Consequently, by the continuity of F_y, we have

$$\big|F_y(\bar y(x),v)\big| \leq C \quad \forall\, x \in \Omega, \quad v \in U,$$

for some constant $C > 0$. In particular,

$$\left|\int_U F_y(\bar y(x),v)\bar\sigma(x)(dv)\right| \leq C.$$

Therefore, similarly to (3.4)–(3.5), we can get

$$\bar\psi(\cdot) \in W^{2,p}(\Omega) \cap C^{1,\alpha}(\overline{\Omega}) \quad \forall\, p \in [1,+\infty), \quad \alpha \in (0,1). \tag{3.6}$$

Now, let us make an observation on (2.11). By (P1),

$$\frac{\partial^2}{\partial u^2}H(y,u,\psi) = -F_{uu}(y,u) > 0 \quad \forall\,(y,u,\psi) \in \mathbb{R}\times U\times\mathbb{R}. \tag{3.7}$$

Thus, it follows from (2.11) that

$$\begin{aligned}\operatorname{supp}\bar\sigma(x) &\subseteq \begin{cases} \{M\} & \text{if } H(\bar y(x),M,\bar\psi(x)) > H(\bar y(x),0,\bar\psi(x)),\\ \{0\} & \text{if } H(\bar y(x),M,\bar\psi(x)) < H(\bar y(x),0,\bar\psi(x)),\\ \{0,M\} & \text{if } H(\bar y(x),M,\bar\psi(x)) = H(\bar y(x),0,\bar\psi(x)) \end{cases}\\ &= \begin{cases} \{M\} & \text{if } \bar\varphi(x) > h\\ \{0\} & \text{if } \bar\varphi(x) < h\\ \{0,M\} & \text{if } \bar\varphi(x) = h \end{cases} \qquad \text{a.e. } x \in \Omega,\end{aligned} \tag{3.8}$$

where

$$h = \frac{F(0,M) - F(0,0)}{M} \tag{3.9}$$

and

$$\bar{\varphi}(x) = \bar{\psi}(x) - \frac{F\big(\bar{y}(x), M\big) - F\big(\bar{y}(x), 0\big)}{M} + h. \tag{3.10}$$

We have

$$\bar{\varphi}|_{\partial\Omega} = 0, \tag{3.11}$$

and by (P2),

$$h < 0. \tag{3.12}$$

Moreover, by (3.4), (3.6), and (P1), one has

$$\bar{\varphi}(\cdot) \in W^{2,p}(\Omega) \cap C^{1,\alpha}(\overline{\Omega}) \quad \forall\, p \in [1, +\infty), \quad \alpha \in (0,1). \tag{3.13}$$

By (3.11)–(3.12), $\{\bar{\varphi} > h\}$ has a positive measure. Since

$$\int_U v\bar{\sigma}(x)(dv) \geq 0 \quad \text{in } \Omega$$

and (3.8) implies

$$\int_U v\bar{\sigma}(x)(dv) = M > 0 \quad \text{a.e. } x \in \{\bar{\varphi} > h\},$$

we must have

$$\bar{y}(x) > 0 \quad \text{in } \Omega \tag{3.14}$$

by the strong maximum principle for elliptic partial differential equations (see [16, Ch. 3]).

On the other hand, (3.8) implies

$$\operatorname{supp} \bar{\sigma}(x) = \{0\} \quad \text{a.e. } x \in \{\bar{\varphi} < h\}. \tag{3.15}$$

A crucial property of $\bar{\varphi}(\cdot)$ is that $\Omega_1 \equiv \{\bar{\varphi} < h\}$ has a zero measure. Otherwise, suppose that Ω_1 has a positive measure. Then it is an open set since $\bar{\varphi}(\cdot)$ is continuous by (3.13). Then (3.11) and $h < 0$ lead to

$$\bar{\varphi}|_{\partial\Omega_1} = h. \tag{3.16}$$

Consequently, by (2.8)–(2.9), (3.15), and (P1)–(P2), we obtain

$$
\begin{aligned}
(3.17)\quad -\triangle\bar{\varphi}(x) &= -\triangle\bar{\psi}(x) + \triangle\,\frac{F\big(\bar{y}(x),M\big)-F\big(\bar{y}(x),0\big)}{M}\\
&= -\int_U F_y\big(\bar{y}(x),v\big)\,\bar{\sigma}(x)(dv) + \frac{F_y\big(\bar{y}(x),M\big)-F_y\big(\bar{y}(x),0\big)}{M}\triangle\bar{y}(x)\\
&\quad + \frac{F_{yy}\big(\bar{y}(x),M\big)-F_{yy}\big(\bar{y}(x),0\big)}{M}\,|\nabla\bar{y}(x)|^2\\
&= -F_y\big(\bar{y}(x),0\big) + \frac{F_{yy}\big(\bar{y}(x),M\big)-F_{yy}\big(\bar{y}(x),0\big)}{M}\,|\nabla\bar{y}(x)|^2\\
&\ge -F_y\big(\bar{y}(x),0\big)\\
&\ge -F_y(0,0)\ge 0 \quad \text{a.e. } x\in\Omega_1.
\end{aligned}
$$

Combining the above with (3.16) and the weak maximum principle for elliptic partial differential equations, we get that

$$\bar{\varphi}(x)\ge h \quad \text{a.e. } x\in\Omega_1.$$

This is a contradiction. Therefore, Ω_1 must be of zero measure, and

$$\bar{\varphi}(x)\ge h \quad \text{a.e. } x\in\Omega. \tag{3.18}$$

With the above preparation, we are now ready to prove our main results.

(i) Let $\ell\ge 0$. It is enough to prove that there is a classical control $\bar{u}(\cdot)\in\mathcal{U}$ such that

$$J(\bar{u}(\cdot))\le J(\bar{\sigma}(\cdot)). \tag{3.19}$$

In fact, by (3.7) and (3.18), one has

$$H(\bar{y}(x),M,\bar{\psi}(x)) = \max_{v\in U} H(\bar{y}(x),v,\bar{\psi}(x)) \quad \text{a.e. } x\in\Omega.$$

Combining the above with (2.10)–(2.11), we see that for almost all $x\in\Omega$,

$$w\bar{\psi}(x) - F\big(\bar{y}(x),w\big) = M\bar{\psi}(x) - F\big(\bar{y}(x),M\big) \quad \forall\, w\in\operatorname{supp}\bar{\sigma}(x).$$

Thus,

$$
\begin{aligned}
(3.20)\quad J(\bar{\sigma}(\cdot)) &= \frac{1}{2}\int_\Omega dx\int_U 2F\big(\bar{y}(x),v\big)\,\bar{\sigma}(x)(dv)\\
&= \frac{1}{2}\int_\Omega dx\int_U \big[(v-M)\bar{\psi}(x) + F\big(\bar{y}(x),M\big) + F\big(\bar{y}(x),v\big)\big]\,\bar{\sigma}(x)(dv)\\
&= \frac{1}{2}\int_\Omega dx\int_U \big[-\bar{y}(x)F_y\big(\bar{y}(x),v\big) - M\bar{\psi}(x) + F\big(\bar{y}(x),M\big)\\
&\qquad + F\big(\bar{y}(x),v\big)\big]\,\bar{\sigma}(x)(dv);
\end{aligned}
$$

here we have used the equality

$$
\begin{aligned}
&\int_\Omega dx\int_U v\bar{\psi}(x)\,\bar{\sigma}(x)(dv)\\
&\quad = \int_\Omega (-\triangle\bar{y}(x))\bar{\psi}(x)\,dx\\
&\quad = \int_\Omega \bar{y}(x)(-\triangle\bar{\psi}(x))\,dx\\
&\quad = -\int_\Omega dx\int_U \bar{y}(x)F_y\big(\bar{y}(x),v\big)\,\bar{\sigma}(x)(dv).
\end{aligned}
$$

Now, let $V = \{0, M\}$ and $\mathcal{R}(\Omega, V)$ be defined similarly to $\mathcal{R}(\Omega, U)$. Consider the relaxed system

$$\begin{cases} -\triangle\, y(x) = \displaystyle\int_V v\sigma(x)(dv) \quad \text{in } \Omega, \\ -\triangle\, \psi(x) = -\displaystyle\int_V F_y\big(y(x), v\big)\sigma(x)(dv) \quad \text{in } \Omega, \\ y|_{\partial\Omega} = \psi|_{\partial\Omega} = 0 \end{cases} \tag{3.21}$$

and the cost functional

$$\begin{aligned} J^*(\sigma(\cdot)) = \frac{1}{2}\int_\Omega dx \int_V \big[&-y(x)F_y\big(y(x), v\big) \\ &- M\psi(x) + F\big(y(x), M\big) + F\big(y(x), v\big)\big]\,\sigma(x)(dv). \end{aligned} \tag{3.22}$$

Thus $\bar\sigma(\cdot) \in \mathcal{R}(\Omega, V)$ and

$$J^*(\bar\sigma(\cdot)) = J(\bar\sigma(\cdot)). \tag{3.23}$$

Since $\ell \geq 0$, by Lemma 3.3, which will be proved below, we have

$$J^*(\delta_{\bar u(\cdot)}) \leq J^*(\sigma(\cdot)) \quad \forall\, \sigma(\cdot) \in \mathcal{R}(\Omega, V)$$

with

$$\bar u(\cdot) \equiv M.$$

In particular,

$$J(\bar u(\cdot)) = J^*(\delta_{\bar u(\cdot)}) \leq J^*(\bar\sigma(\cdot)) = J(\bar\sigma(\cdot)).$$

The first equality in the above can be verified directly, almost by following the same idea for proving (3.23).

Thus, we have proved that $\bar u(\cdot)$ is an optimal classical control for Problem (C). As a matter of fact, the proof of Lemma 3.3 will show that

$$\bar\sigma(x) = \delta_{\bar u(x)} \quad \text{a.e. } x \in \Omega.$$

(ii) Let $\ell < 0$. We will prove that $\operatorname{supp}\bar\sigma(x)$ is not a singleton in a positive measure set. By (3.8), $\operatorname{supp}\bar\sigma \subseteq \{0, M\}$ for almost all $x \in \Omega$. Thus, we need to show that $\big\{\operatorname{supp}\bar\sigma = \{0, M\}\big\}$ has positive measure, which can be proved if we prove that $E_M \equiv \big\{\operatorname{supp}\bar\sigma \neq \{M\}\big\}$ has positive measure and $E_0 \equiv \big\{\operatorname{supp}\bar\sigma = \{0\}\big\}$ has zero measure.

If E_M has zero measure, by (2.8)–(2.9) and (3.1), we have

$$\bar\psi(x) = \zeta(x) \quad \text{in } \Omega.$$

Thus, it follows from (3.2) and (3.10) that

$$\inf_{x\in\Omega} \bar\varphi(x) = \ell + h < h.$$

This contradicts (3.18). Therefore, E_M has a positive measure.

By (3.8) and (3.18), it holds that

$$\bar{\varphi}(x) = h \quad \text{a.e. } x \in E_0.$$

Thus, by (2.13) and (3.13),

$$-\triangle\, \bar{\varphi}(x) = 0 \quad \text{a.e. } x \in E_0.$$

On the other hand, noting that (3.17) holds on E_0, we get

$$-\triangle\, \bar{\varphi}(x) \geq -F_y\big(\bar{y}(x), 0\big) > -F_y(0,0) \geq 0 \quad \text{a.e. } x \in E_0$$

by (3.3), (3.14), and (P2).

Thus, we see that $0 > 0$ a.e. on E_0. That is, E_0 has a zero measure.

To summarize, we have that $\big\{\operatorname{supp} \bar{\sigma} = \{0, M\}\big\}$ has a positive measure. In other words, an optimal relaxed control of Problem (R) will not be an element of $\mathcal{U}$. Therefore, Problem (C) does not have optimal controls. □

Now we present a lemma used in the proof of the above theorem. Let

$$\mathbf{f}(\cdot) = \begin{pmatrix} f^1(\cdot) \\ f^2(\cdot) \end{pmatrix}, \qquad \mathbf{y}(\cdot) = \begin{pmatrix} y^1(\cdot) \\ y^2(\cdot) \end{pmatrix}. \tag{3.24}$$

Consider the relaxed system

$$\begin{cases} -\triangle\, \mathbf{y} = \displaystyle\int_V \mathbf{f}(\mathbf{y}(x), v)\sigma(x)(dv) \quad \text{in } \Omega, \\ \mathbf{y}|_{\partial\Omega} = 0 \end{cases} \tag{3.25}$$

and the corresponding cost functional

$$\widehat{J}(\sigma(\cdot)) = \int_\Omega dx \int_V f^0(\mathbf{y}(x), v)\sigma(x)(dv). \tag{3.26}$$

We have the following proposition.

PROPOSITION 3.2. *Let $V = \{0, M\}$ and $f^i(\cdot) \in C^1(\mathbb{R}^3)$ $(i = 0, 1, 2)$. Consider* (3.25)–(3.26). *If*

$$\frac{\partial f^1}{\partial y^1}(y^1, y^2, v) \leq 0, \quad \frac{\partial f^1}{\partial y^2}(y^1, y^2, v) = 0, \quad \frac{\partial f^2}{\partial y^2}(y^1, y^2, v) \leq 0, \tag{3.27}$$

then there exists at least one relaxed control $\bar{\sigma}(\cdot)$ minimizing $\widehat{J}(\cdot)$ over $\mathcal{R}(\Omega, V)$. Moreover, let $\bar{\mathbf{y}}(\cdot)$ be the solution of (3.25) *with $\sigma(\cdot)$ replaced by $\bar{\sigma}(\cdot)$, and let $\mathbf{\Psi}(\cdot)$ be the solution of*

$$\begin{cases} -\triangle\, \mathbf{\Psi}(x) = \displaystyle\int_V \big\{\mathbf{f}_{\mathbf{y}}\big(\bar{\mathbf{y}}(x), v\big)\mathbf{\Psi}(x) - f^0_{\mathbf{y}}\big(\bar{\mathbf{y}}(x), v\big)\big\}\bar{\sigma}(x)(dv) \quad \textit{in } \Omega, \\ \mathbf{\Psi}|_{\partial\Omega} = 0, \end{cases} \tag{3.28}$$

where

$$\mathbf{f}_{\mathbf{y}} = \begin{pmatrix} \frac{\partial f^1}{\partial y^1} & \frac{\partial f^2}{\partial y^1} \\ \frac{\partial f^1}{\partial y^2} & \frac{\partial f^2}{\partial y^2} \end{pmatrix}.$$

Then

$$\text{supp}\,\bar{\sigma}(x) \subseteq \Big\{ w \in V \;\Big|\; \widehat{H}(x,w) = \max_{v\in V} \widehat{H}(x,v) \Big\} \quad \textit{a.e. } x \in \Omega, \tag{3.29}$$

where

$$\widehat{H}(x,v) = \langle \mathbf{\Psi}(x), \mathbf{f}(\bar{\mathbf{y}}(x),v)\rangle - f^0(\bar{\mathbf{y}}(x),v).$$

The above proposition still holds if V is replaced by a compact subset of $\mathbb{R}$. It can be proved similarly to Theorems 3.2 and 4.1 in [19]. We omit the proof here. Note that (3.27) is mainly used to guarantee the existence and uniqueness of a solution to (3.25), while Proposition 2.2 can be regarded as a special case of Proposition 3.2.

LEMMA 3.3. *Let $V = \{0, M\}$ and* (P1)–(P2) *hold. Consider the relaxed system* (3.21)–(3.22). *Let ℓ be defined by* (3.1)–(3.2) *and $\ell \geq 0$. Then*

$$J^*(\delta_{\bar{u}(\cdot)}) \leq J^*(\sigma(\cdot)) \quad \forall\, \sigma(\cdot) \in \mathcal{R}(\Omega, V),$$

where

$$\bar{u}(x) \equiv M.$$

Proof. By Proposition 3.2, one can find a relaxed control $\bar{\sigma}(\cdot)$ minimizing $J^*(\cdot)$ over $\mathcal{R}(\Omega, V)$. Moreover,

$$\text{supp}\,\bar{\sigma}(x) \subseteq \Big\{ w \in V \;\Big|\; H^*(x,w) = \max_{v\in V} H^*(x,v) \Big\} \quad \text{a.e. } x \in \Omega, \tag{3.30}$$

where

$$\begin{aligned} H^*(x,u) = \Psi(x)u - \big(z(x) - \bar{y}(x)\big) F_y\big(\bar{y}(x),u\big) - F\big(\bar{y}(x),u\big) \\ + M\bar{\psi}(x) - F\big(\bar{y}(x),M\big) \end{aligned}$$

with $\bar{y}(\cdot)$ solving the following equation:

$$\begin{cases} -\triangle\, \bar{y}(x) = \displaystyle\int_V v\bar{\sigma}(x)(dv) & \text{in } \Omega, \\ -\triangle\, \bar{\psi}(x) = -F_y\big(\bar{y}(x),v\big)\bar{\sigma}(x)(dv) & \text{in } \Omega, \\ \bar{y}|_{\partial\Omega} = \bar{\psi}|_{\partial\Omega} = 0, \end{cases} \tag{3.31}$$

where $z(\cdot)$ is defined by (3.1) and $\Psi(\cdot)$ is the solution of

$$\begin{cases} -\triangle\, \Psi(x) = \displaystyle\int_V F_{yy}\big(\bar{y}(x),v\big)\bar{\sigma}(x)(dv)\,\big(\bar{y}(x) - z(x)\big) \\ \qquad\qquad - F_y\big(\bar{y}(x),M\big) \qquad\qquad \text{in } \Omega, \\ \Psi|_{\partial\Omega} = 0. \end{cases} \tag{3.32}$$

We claim that

$$\text{supp}\,\bar{\sigma}(x) = \{M\} \quad \text{a.e. } x \in \Omega. \tag{3.33}$$

Otherwise, both

$$\widetilde{E} \triangleq \Big\{ x \in \Omega \;\Big|\; \int_V v\bar{\sigma}(x)(dv) < M \Big\}$$

and

$$E \triangleq \Big\{ x \in \Omega \;\Big|\; H^*(x,M) \le H^*(x,0) \Big\} = \Big\{ x \in \Omega \;\Big|\; \Psi(x) \le \Phi(x) \Big\}$$

have positive measures (see (3.30)), where

$$\Phi(x) \triangleq \frac{F_y\big(\bar y(x), M\big) - F_y\big(\bar y(x), 0\big)}{M} \big(z(x) - \bar y(x)\big) + \frac{F\big(\bar y(x), M\big) - F\big(\bar y(x), 0\big)}{M}. \tag{3.34}$$

Since

$$\int_V v \bar\sigma(x)(dv) \le M \quad \text{in } \Omega,$$

while the strict inequality holds on $\widetilde{E}$, by the strong maximum principle for elliptic equations we obtain (see (3.1) and (3.31))

$$\bar y(x) < z(x) \quad \text{a.e. } x \in \Omega. \tag{3.35}$$

Thus, by (P1) it holds that (see (3.1) and (3.32))

(3.36)

$$\begin{aligned}
&-\triangle(\Psi - \zeta) \\
&= \int_V F_{yy}(\bar y, v)\bar\sigma(x)(dv)\,(\bar y - z) - F_y(\bar y, M) + F_y(z, M) \\
&= \left[\int_V F_{yy}(\bar y, v)\bar\sigma(x)(dv) - \int_0^1 F_{yy}(\bar y + t(z-\bar y), M)\,dt\right](\bar y - z) \\
&= \int_V \bar\sigma(x)(dv) \int_0^1 dt \int_0^1 \big[F_{yyy}\big(\bar y + st(z-\bar y), v + s(M-v)\big)(z-\bar y)^2 \\
&\qquad\qquad + F_{yyu}\big(\bar y + st(z-\bar y), v + s(M-v)\big)(M-v)(z-\bar y)\big]\,ds \\
&\ge 0.
\end{aligned}$$

Therefore,

$$\Psi(x) \ge \zeta(x) \quad \text{a.e. } x \in \Omega. \tag{3.37}$$

On the other hand (note (3.34)),

$$\begin{aligned}
&\frac{F\big(z(x), M\big) - F\big(z(x), 0\big)}{M} - \Phi(x) \\
&= \big(z(x) - \bar y(x)\big)^2 \int_0^1 dt \int_0^1 ds \int_0^1 F_{yu}\big(\bar y(x) + \alpha t(z(x) - \bar y(x)), sM\big)\,d\alpha \\
&\ge 0.
\end{aligned}$$

Thus, it follows from (3.2) and $\ell \ge 0$ that

$$\begin{aligned}
\Psi(x) &\ge \zeta(x) \ge \zeta(x) - \ell \\
&\ge \frac{F\big(z(x), M\big) - F\big(z(x), 0\big)}{M} \ge \Phi(x) \quad \text{in } \Omega.
\end{aligned}$$

Consequently,

$$E \subseteq \Big\{ x \in \Omega \;\Big|\; \Psi(x) = \zeta(x) \Big\} \triangleq E_0$$

and

$$E \subseteq \left\{ x \in \Omega \;\middle|\; \zeta(x) = \frac{F\big(z(x), M\big) - F\big(z(x), 0\big)}{M} \right\} \triangleq E_1.$$

Since E has a positive measure, so do E_0 and E_1.

By (3.36) and the strong maximum principle for elliptic equations, E_0 has a positive measure, which means that

$$\Psi(x) \equiv \zeta(x) \quad \text{in } \Omega. \tag{3.38}$$

Thus, by (3.38), we can get from (3.36) and (3.35) that, for almost all $x \in \Omega$,

$$\begin{array}{l} F_{yyy}\big(\bar{y}(x) + st(z(x) - \bar{y}(x)), v + s(M - v)\big) = 0 \\ F_{yyu}\big(\bar{y}(x) + st(z(x) - \bar{y}(x)), v + s(M - v)\big) = 0 \end{array} \quad \forall\, s, t \in [0,1], \quad v \in \operatorname{supp} \bar{\sigma}(x)$$

since in (3.36), both of the two terms before the symbol "$\geq$" are nonnegative.

Noting that

$$\widetilde{E} = \Big\{ x \;\Big|\; \operatorname{supp} \bar{\sigma}(x) \supseteq \{0\} \Big\}$$

and that it has a positive measure, one must have

$$F_{yyy}(y,u) = 0, \quad F_{yyu}(y,u) = 0 \quad \forall\, y \in [0, \max z], \quad u \in [0, M]. \tag{3.39}$$

On the other hand, for any $\sigma(\cdot) \in \mathcal{R}(\Omega, V)$, the corresponding solution $y(\cdot; \sigma(\cdot))$ of (3.21) satisfies

$$0 \leq y(x; \sigma(\cdot)) \leq z(x) \quad \forall\, x \in \Omega.$$

Thus, combining the above with (3.39), we can suppose that

$$F_{yyy}(y,u) = 0, \quad F_{yyu}(y,u) = 0 \quad \forall\, y \in \mathbb{R}, \quad u \in [0, M], \tag{3.40}$$

without loss of generality. Therefore,

$$F(y,u) = Ay^2 + g(u)y + h(u) \quad \forall\, y \in \mathbb{R}, \quad u \in [0, M],$$

where A is a constant and $g(\cdot), h(\cdot) \in C^3[0, M]$. Thus, by (3.1)

$$\begin{aligned} &-\triangle\, \zeta(x) \\ &\quad = -F_y(z(x), M) \\ &\quad = -2Az(x) - g(M) \quad \text{in } \Omega. \end{aligned} \tag{3.41}$$

Further, it follows from Lemma 2.3 that

$$\begin{aligned} &-\triangle\, \zeta(x) \\ &\quad = -\triangle\, \frac{F\big(z(x), M\big) - F\big(z(x), 0\big)}{M} \\ &\quad = -\triangle \left[\frac{g(M) - g(0)}{M} z(x) + \frac{h(M) - h(0)}{M} \right] \\ &\quad = g(M) - g(0) \quad \text{a.e. } x \in E_1. \end{aligned}$$

Therefore,

$$-2Az(x) = 2g(M) - g(0) \quad \text{a.e. } x \in E_1. \tag{3.42}$$

Case 1. $A \neq 0$. We have

$$z(x) = \frac{g(0) - 2g(M)}{2A} \quad \text{a.e. } x \in E_1.$$

Consequently, by (2.13) and $z \in W^{2,2}(\Omega)$, it holds that

$$-\triangle z(x) = 0 \quad \text{a.e. } x \in E_1.$$

Thus, combining the above with (3.1), we get

$$M = 0 \quad \text{a.e. } x \in E_1.$$

This contradicts that E_1 has a positive measure.

Case 2. $A = 0$. In this case, (3.41) and (3.42) imply

$$-\triangle \zeta(x) = -g(M) \quad \text{in } \Omega$$

and

$$g(0) = 2g(M). \tag{3.43}$$

Thus,

$$\zeta(x) = -\frac{g(M)}{M} z(x) \quad \text{in } \Omega.$$

Therefore, by (3.43) and the definition of E_1,

$$\begin{aligned} -\frac{g(M)}{M} z(x) &= \zeta(x) \\ &= \frac{F\big(z(x), M\big) - F\big(z(x), 0\big)}{M} \\ &= \frac{g(M) - g(0)}{M} z(x) + \frac{h(M) - h(0)}{M} \\ &= -\frac{g(M)}{M} z(x) + \frac{h(M) - h(0)}{M} \quad \text{a.e. } x \in E_1. \end{aligned}$$

Consequently,

$$h(M) - h(0) = 0,$$

since E_1 has a positive measure. This contradicts

$$h(M) - h(0) = F(0, M) - F(0, 0) < 0.$$

The above two cases show that no matter what A is, we always end up with a contradiction. Hence, (3.33) must hold, namely,

$$\bar{\sigma}(x) = \delta_M = \delta_{\bar{u}(x)} \quad \text{a.e. } x \in \Omega,$$

proving our lemma. □

Next, for the case that $F_{yy}F_{yu} \geq 0$, we have the following result.

THEOREM 3.4. *Let $M > 0$, $F \in C^3(\mathbb{R} \times [0, M])$, and*

$$F_{uu}(y, u) < 0 \quad \forall\, y \in \mathbb{R}, \quad u \in [0, M].$$

Moreover, suppose that one of the following conditions holds:

(Q1) $F_y(0, 0) \leq 0$ *and* $\forall\, y \geq 0$, $u \in [0, M]$,

$$\begin{aligned} &F_{yy}(y, u) \leq 0, \quad F_{yu}(y, u) \leq 0, \quad F_{yyu}(y, u) \geq 0, \\ &F_{yy}(y, u) + F_{yu}(y, u) < 0. \end{aligned}$$

(Q2) $F_y(0, 0) \geq 0$ *and* $\forall\, y \geq 0$, $u \in [0, M]$,

$$\begin{aligned} &F_{yy}(y, u) \geq 0, \quad F_{yu}(y, u) \geq 0, \quad F_{yyu}(y, u) \leq 0, \\ &F_{yy}(y, u) + F_{yu}(y, u) > 0. \end{aligned}$$

Then Problem (C) *admits at least one optimal control. In fact, any optimal relaxed control of Problem* (R) *is an optimal classical control of Problem* (C).

Proof. Without loss of generality, we suppose that (Q1) holds.

Similar to that in the proof of Theorem 3.1, the relaxed problem (R) admits an optimal relaxed control $\bar{\sigma}(\cdot) \in \mathcal{R}(\Omega, U)$. What we need to prove is that for almost all $x \in \Omega$, $\operatorname{supp} \bar{\sigma}(x)$ is a singleton. To this end, let $\bar{y}(\cdot)$, $\bar{\psi}(\cdot)$, $\bar{\varphi}(\cdot)$, and h be defined as in the proof of Theorem 3.1 (see (2.8)–(2.9) and (3.9)–(3.10)). Then (3.8) holds and thus we need only prove that for almost all $x \in \{\bar{\varphi} = h\}$, $\operatorname{supp} \bar{\sigma}(x)$ is a singleton.

We claim that for almost all $x \in \{\bar{\varphi} = h\}$,

$$\operatorname{supp} \bar{\sigma}(x) = \{0\}. \tag{3.44}$$

If it is not the case, then

$$E \triangleq \big\{x \;\big|\; \operatorname{supp} \bar{\sigma}(x) \neq \{0\}\big\} \cap \{\bar{\varphi} = h\}$$

has a positive measure, and consequently,

$$\bar{y}(x) > 0 \quad \forall\, x \in \Omega.$$

By (2.13), for almost all $x \in E$, it holds that

$$\begin{aligned} 0 = -\triangle \bar{\varphi}(x) &= -\triangle \bar{\psi}(x) + \triangle \frac{F\big(\bar{y}(x), M\big) - F\big(\bar{y}(x), 0\big)}{M} \\ &= -\int_U F_y\big(\bar{y}(x), v\big)\, \bar{\sigma}(x)(dv) + \frac{F_y\big(\bar{y}(x), M\big) - F_y\big(\bar{y}(x), 0\big)}{M} \triangle \bar{y}(x) \\ &\quad + \frac{F_{yy}\big(\bar{y}(x), M\big) - F_{yy}\big(\bar{y}(x), 0\big)}{M} |\nabla \bar{y}(x)|^2. \end{aligned} \tag{3.45}$$

Since (Q1) implies that all three terms in (3.45) are nonnegative, we get that for almost all $x \in E$,

$$\int_U F_y\big(\bar{y}(x), v\big)\, \bar{\sigma}(x)(dv) = 0, \tag{3.46}$$

$$\frac{F_y\big(\bar{y}(x), M\big) - F_y\big(\bar{y}(x), 0\big)}{M} \int_U v\, \bar{\sigma}(x)(dv) = 0. \tag{3.47}$$

Let $x \in E$ be a point satisfying (3.46)–(3.47). Then we have the following cases.

Case 1.

$$\frac{F_y\big(\bar{y}(x), M\big) - F_y\big(\bar{y}(x), 0\big)}{M} \neq 0; \tag{3.48}$$

then

$$\int_U v\,\bar{\sigma}(x)(dv) = 0,$$

and (3.44) follows.

Case 2.

$$\frac{F_y\big(\bar{y}(x), M\big) - F_y\big(\bar{y}(x), 0\big)}{M} = 0. \tag{3.49}$$

By (Q1), we must have

$$F_{yu}\big(\bar{y}(x), u\big) = 0 \quad \forall\, u \in [0, M]$$

and consequently,

$$F_{yy}(\bar{y}(x), 0) < 0.$$

Therefore, by (Q1) and the continuity of F_{yy}, we get

$$\begin{aligned} F_y\big(\bar{y}(x), v\big) &\leq F_y\big(\bar{y}(x), 0\big) \\ &= \int_0^1 F_{yy}\big(t\bar{y}(x), 0\big)\bar{y}(x)\,dt + F_y(0,0) \\ &< 0 \quad \forall\, v \in [0, M]. \end{aligned}$$

Combining the above with (3.46), we get (3.44).

Thus, (3.44) holds for almost all $x \in E$. This is a contradiction. Consequently, (3.44) holds for almost all $\{\bar{\varphi} = h\}$ and we conclude the proof. □

Now, let us look at the case that $(y, u) \mapsto F(y, u)$ is quadratic, i.e.,

$$F(y, u) = ay^2 + byu + cu^2.$$

The following is a summary of what we have obtained in the above for this special case:

(i) If $c \geq 0$, then Cesari's condition holds and Problem (C) admits at least one optimal control.

(ii) If $c < 0$, $a < 0$, $b > 0$, then (P1)–(P2) and (3.3) hold. Thus, we can solve Problem (C) by using Theorem 3.1. More precisely, let

$$\begin{cases} -\triangle\,\xi(x) = 1 & \text{in } \Omega, \\ \xi|_{\partial\Omega} = 0. \end{cases}$$

Then Problem (C) admits at least one optimal control if and only if

$$\max_{x\in\Omega} \xi(x) \leq -\frac{c}{2b}.$$

(iii) If $c < 0$ and $a = b = 0$, then Problem (C) is trivial. It admits an optimal control $\bar{u}(\cdot) \equiv M$.

(iv) If $c < 0$ and $ab \geq 0$, $a + b \neq 0$, then F satisfies the conditions of Theorem 3.4. Thus, any optimal relaxed control $\bar{\sigma}(\cdot)$ is in fact a classical control. Therefore, Problem (C) admits at least one optimal control in this case.

From the above, one can see how our results apply to the quadratic functional case.

Note that in establishing existence results without convexity conditions, most authors assumed that the state and the control are separated in the cost functional, i.e., the functionals take the form $g(x, y(x)) + h(x, u(x))$. On the other hand, Mariconda [24] considered the case when the state variable y and the control variable u are not separated in the integrand of the cost functional, but he assumed that the integrand is concave in the state variable y. As we have seen from our Theorems 3.1 and 3.4, such a concavity is neither sufficient nor necessary for guaranteeing the existence of an optimal control.

Finally, we point out that in some cases, conditions that we presented for guaranteeing the existence of an optimal control are not only sufficient but also necessary. (Consider the cases when $M > 0$, (P1)–(P2), and (3.3) hold. Then, Theorem 3.1 shows that Problem (C) admits at least one optimal control if and only if $\ell \geq 0$.)

Acknowledgment. The author would like to thank the anonymous referees for their helpful comments.

REFERENCES

[1] N. U. Ahmed, *Properties of relaxed trajectories for a class of nonlinear evolution equations on a Banach space*, SIAM J. Control Optim., 21 (1983), pp. 953–967.
[2] Z. Artstein, *On a variational problem*, J. Math. Anal. Appl., 45 (1974), pp. 404–415.
[3] E. J. Balder, *On a useful compactification for optimal control problems*, J. Math. Anal. Appl., 72 (1979), pp. 391–398.
[4] E. J. Balder, *A general approach to lower semicontinuity and lower closure in optimal control theory*, SIAM J. Control Optim., 22 (1984), pp. 570–598.
[5] E. J. Balder, *On seminormality of integral functionals and their integrands*, SIAM J. Control Optim., 24 (1986), pp. 95–121.
[6] E. J. Balder, *New existence results for optimal controls in the absence of convexity: The importance of extremality*, SIAM J. Control Optim., 32 (1994), pp. 890–916.
[7] L. D. Berkovitz, *Optimal Control Theory*, Springer-Verlag, New York, 1983.
[8] H. Berliocchi and J. M. Lasry, *Intégrandes normales et measures paramétrées en calcul des variations*, Bull. Soc. Math. France, 101 (1973), pp. 129–184.
[9] A. Cellina and G. Colombo, *On a classical problem of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 97–106.
[10] L. Cesari, *Optimization Theory and Applications, Problems with Ordinary Differential Equations*, Springer-Verlag, New York, 1983.
[11] G. Colombo and V. V. Goncharov, *Existence for nonconvex optimal problem with nonlinear dynamics*, Nonlinear Anal., 24 (1995), pp. 795–800.
[12] H. O. Fattorini, *Relaxed controls in infinite dimensional systems*, in Estimation and Control of Distributed Parameter Systems, Internat. Ser. Numer. Math. 100, Birkhäuser, Basel, 1991, pp. 115–128.
[13] A. F. Filippov, *On certain questions in the theory of optimal control*, SIAM J. Control Ser. A, 1 (1962), pp. 76–84.
[14] F. Flores-Bazán and S. Perrotta, *Nonconvex variational problems related to a hyperbolic equation*, SIAM J. Control Optim., 37 (1999), pp. 1751–1766.
[15] R. Gamkrelidze, *Principle of Optimal Control Theory*, Plenum Press, New York, 1978.
[16] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[17] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1981; reprinted as Classics Appl. Math. 31, SIAM, Philadelphia, 2000.
[18] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser Boston, Boston, MA, 1995.
[19] H. Lou, *Existence of optimal controls for semilinear elliptic equations without Cesari type conditions*, ANZIAM J., 45 (2003), pp. 115–131.
[20] H. Lou, *Existence of optimal controls for semlinear parabolic equations without Cesari-type conditions*, Appl. Math. Optim., 47 (2003), pp. 121–142.
[21] H. Lou, *Maximum principle of optimal control for degenerate quasi-linear elliptic equations*, SIAM J. Control Optim., 42 (2003), pp. 1–23.
[22] H. Lou, *Existence of optimal controls in the absence of Cesari-type conditions for semilinear elliptic and parabolic systems*, J. Optim. Theory Appl., 125 (2005), pp. 367–391.
[23] P. Marcellini, *Alcune osservazioni sull'exixtenza del minimo di integrali del calcolo delle variazioni senza ipotesi di convessità*, Rend. Mat. (6), 13 (1980), pp. 271–281.
[24] C. Mariconda, *A generalization of the Cellina–Colombo theorem for a class of non-convex variational problems*, J. Math. Anal. Appl., 175 (1993), pp. 514–552.
[25] E. J. McShane, *Generalized curves*, Duke Math. J., 6 (1940), pp. 513–536.
[26] E. J. McShane, *Relaxed controls and variational problems*, SIAM J. Control, 5 (1967), pp. 438–485.
[27] C. B. Morrey, Jr., *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, 1966.
[28] L. W. Neustadt, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.
[29] C. Olech, *Integrals of set-valued functions and linear optimal control problems*, in Colloque sur la Théorie Mathématique du Contrôle Optimal, Vander, Louvain, 1970, pp. 109–125.
[30] N. S. Papageorgiou, *Properties of the relaxed trajectories of evolution equations and optimal control*, SIAM J. Control Optim., 27 (1989), pp. 267–288.
[31] J. P. Raymond, *Conditions nécessaires et suffisantes d'existence de solutions en calcul des variations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 4 (1987), pp. 169–202.
[32] J. P. Raymond, *Existence theorems in optimal control theory without convexity assumptions*, J. Optim. Theory Appl., 67 (1990), pp. 109–132.
[33] J. P. Raymond, *Existence theorems without convexity assumptions for optimal control problems governed by parabolic and elliptic systems*, Appl. Math. Optim., 26 (1992), pp. 39–62.
[34] M. B. Suryanarayana, *Existence theorems for optimization problem concerning linear, hyperbolic partial differential equations without convexity conditions*, J. Optim. Theory Appl., 19 (1976), pp. 47–61.
[35] J. Warga, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.
[36] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[37] L. C. Young, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Sci. Lettres Varsovie, C. III, 30 (1937), pp. 212–234.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 1942–1971

NUMERICAL APPROXIMATIONS FOR NONZERO-SUM STOCHASTIC DIFFERENTIAL GAMES*

HAROLD J. KUSHNER†

Abstract. The Markov chain approximation method is a widely used and efficient family of methods for the numerical solution of many types of stochastic control problems in continuous time for reflected-jump-diffusion–type models. It converges under broad conditions, and it has been extended to zero-sum stochastic differential games. We apply the method to a class of nonzero stochastic differential games with a diffusion system model where the controls for the two players are separated in the dynamics and cost function. There have been successful applications of the algorithms, but convergence proofs have been lacking. It is shown that equilibrium values for the approximating chain converge to equilibrium values for the original process and that any equilibrium value for the original process can be approximated by an ϵ-equilibrium for the chain for arbitrarily small $\epsilon > 0$. The numerical method solves a stochastic game for a finite-state Markov chain.

Key words. stochastic differential games, nonzero-sum games, numerical methods, Markov chain approximations

AMS subject classifications. 60F17, 65C30, 65C40, 91A15, 91A23, 93E25

DOI. 10.1137/050647931

1. Introduction.

The aim of this paper is to extend the Markov chain approximation method to numerically solve nonzero-sum stochastic differential games. The method is widely used, robust, and relatively easy to implement. It covers the majority of stochastic control problems in continuous time, for controlled reflected-jump-diffusion–type models of recent interest, and converges under broad conditions. The method was extended to zero-sum stochastic differential games in [16, 17, 18], with the last two references concerned with the ergodic cost case, extending partial prior results such as those of [1, 22, 23]. There has been successful numerical work done on nonzero-sum differential games [12, 13], based on this procedure, but there do not seem to exist results concerning convergence. Works such as [5] are concerned with approximations to nonzero-sum games in normal form and do not apply to the system models or to the type of approximations that appear in our numerical approximations.

We will consider a discounted cost problem for a diffusion model in a compact set G with absorption on the boundary. The state space G and the boundary absorption are selected only to simplify the development so that we can concentrate on the issues that are unique to the nonzero-sum case. One can replace the boundary absorption by boundary reflection if the reflection directions satisfy the conditions in [19] or in [16]. We will work with two-player games. Any number of players can be dealt with, but we stick to two for notational simplicity. The nonzero-sum game is difficult because, as opposed to the zero-sum case, its players are not strictly competitive and have their own value functions, and the methods of proof used previously need considerable modification.

*Received by the editors December 18, 2005; accepted for publication (in revised form) April 16, 2007; published electronically November 28, 2007. This work was partially supported by Army Research Office contract DAAD19-99-1-0223 and National Science Foundation grant ECS 0097447.
http://www.siam.org/journals/sicon/46-6/64793.html

†Applied Mathematics Department, Lefschetz Center for Dynamical Systems, Brown University, Providence, RI 02912 (hjk@dam.brown.edu).

The idea of the Markov chain approximation method is to first approximate the controlled diffusion dynamics by a suitable Markov chain on a finite state space with a discretization parameter h, then approximate the cost functions. One solves the game problem for the simpler chain model, and then proves that the value functions associated with equilibrium or ϵ-equilibrium strategies for the chain converge to the value functions associated with equilibrium or ϵ_1-equilibrium strategies for the diffusion model, where $\epsilon_1 \to 0$ as $\epsilon \to 0$. The methods of proof are purely probabilistic; i.e., no PDE techniques are required, so knowledge of whatever PDEs yield the equilibrium values is not needed. The essential condition is a natural "local consistency" condition. Getting approximations that satisfy this condition is usually straightforward. Many methods are discussed in [19] and all of them are applicable to the game problem of interest here. Furthermore, the numerical approximations are represented as processes which are close to the original, which gives the method intuitive meaning. We are not concerned with algorithms for numerically solving the game for the chain model, but rather show only convergence of the solutions to the desired values as the discretization parameter goes to zero.

Let us comment briefly on some of the particular difficulties posed by the nonzero-sum problem. The convergence proof for the single-player case works roughly as follows. One solves for the optimal (cost minimizing) controls for the approximating chain, with optimal value function $V^h(x)$ for approximation parameter h and initial condition x. Then one interpolates the corresponding chain into a continuous-time process and shows that the weak-sense limit is a controlled diffusion. The limit value function $\liminf V^h(x)$ cannot be better than the optimal value $V(x)$ for the diffusion. To show that $\limsup V^h(x)$ cannot be greater than $V(x)$, one uses a particular $0 < \epsilon$-optimal control for the original process that can be applied to the chain for each value of h, and such that the weak-sense limit (process, control) is just the ϵ-optimal (process, control) for the original model. Since the cost values for the chain under this control are no less than $V^h(x)$, the limit is the ϵ-optimal diffusion process. Since ϵ is arbitrary, we have that $\limsup V^h(x) \le V(x)$. Hence $V^h(x) \to V(x)$.

The proof for the two-person zero-sum game in [16] is much harder, but the essential goal is similar. One has the advantages that the controls are determined by a minmax operation and that there is a single cost function, so that one player's gain is another's loss—properties that the nonzero-sum game does not have. One gets an ϵ-optimal *strategy* for player 2 (resp., for player 1), the maximizing (resp., minimizing) player, that is designed such that it can be applied to the approximating chain, no matter what the other players policy is, and such that in the limit, one has this strategy for player 2 (resp., for player 1) and some arbitrary control for player 1 (resp., for player 2). One uses these and the minmax relations to show that the lim inf of the sequence of upper values and lim sup of the sequence of lower values for the chain must be between the lower and upper values for the original problem. Then the uniqueness of the value (proved in [16]) for the original problem yields the desired result.[1]

Such an approach cannot be applied to the nonzero-sum game, where each player has its own value function and one seeks Nash equilibria instead of minmax = maxmin solutions. Furthermore, unique equilibria are not too common, and we are forced to look much more closely at the structure of the chain and (for the purposes of the proof, not for the numerics) try to approximate it so that it has a "diffusion" form with a

[1]The papers [17, 18] concern the ergodic cost function and use quite different methods and relaxed feedback controls.

driving process that does not heavily depend on the control, with minimal change in the values. This requires that we work with strong-sense solutions, rather than with the weak-sense solutions that were used in [19]. Unlike in the single-player problem, one must work with strategies and not simply controls, at least for one player at a time.

In section 2, the model and the cost functions for the players are defined, the boundary conditions are discussed, and a review of some background material is given. We also give a "uniform in the controls" discrete-time approximation that will be used in what follows. The convergence proof heavily depends on the fact that the players' strategies for the original diffusion process can be simplified (uniformly in the controls), with various approximations to the controls and with some players' controls delayed. The necessary results are developed in section 3. A particular representation of an ϵ-equilibrium *strategy*, in terms of a "smooth" conditional probability, depending only on selected samples of the driving Wiener process (and not on the entire Wiener process), is given in section 4. This representation of the strategy will be crucial to the proof. Various facts concerning the Markov chain approximation are collected in subsection 5.1. The reader is referred to [19] for a fuller treatment.

The numerical algorithms use the approximating Markov chain. But the proof of convergence requires that we approximate (uniformly in the controls or strategies) this approximating chain by a process that is driven by a particular martingale difference process which allows us to get analogues of the various approximation results (in section 3); i.e., to show that the probability law of the chain and the costs change little if the control process is approximated in various ways. These results are new and should be of broad use in dealing with numerical approximations. Theorem 6.1 in section 6 shows that an "approximate" equilibrium (value or strategy) for the diffusion is an "approximate" equilibrium (value or strategy) for the chain for a small discretization parameter h. If the ϵ-equilibrium value for the chain is unique for small $\epsilon > 0$, then the convergence proof is complete since an approximate equilibrium value for the chain is also one for the diffusion. If the value is not unique, then the proof of this last fact is more difficult, and we restrict our attention to the case where the diffusion coefficient does not depend on the state. This is done in Theorem 6.2, which is a consequence of Theorem 5.6, which, in turn, applies a strong approximation theorem to show that the discrete-time approximation to the diffusion and that for the interpolated chain are very close, uniformly in the controls.

2. The model. We consider systems of the following form where $x(t) \in \mathbb{R}^v$, Euclidean v-space,

$$x(t) = x(0) + \int_0^t \sum_{i=1}^{2} b^i(x(s), u_i(s))ds + \int_0^t \sigma(x(s))dw(s), \tag{2.1}$$

where player i, $i = 1, 2$, has control $u_i(\cdot)$, and $w(\cdot)$ is a standard vector-valued Wiener process. The control stops at the first time τ that the boundary of a set G is hit (it equals infinity if the boundary is never reached). More will be said about G following Condition A2.1 and in Condition A2.2. Let $\beta > 0$ and let E_x^u denote the expectation given the use of control $u(\cdot) = (u_1(\cdot), u_2(\cdot))$ and initial condition x. Then the cost function for player i is

$$W_i(u) = E_x^u \int_0^\tau e^{-\beta t} k_i(x(s), u_i(s))ds + E_x^u e^{-\beta\tau} g_i(x(\tau)). \tag{2.2}$$

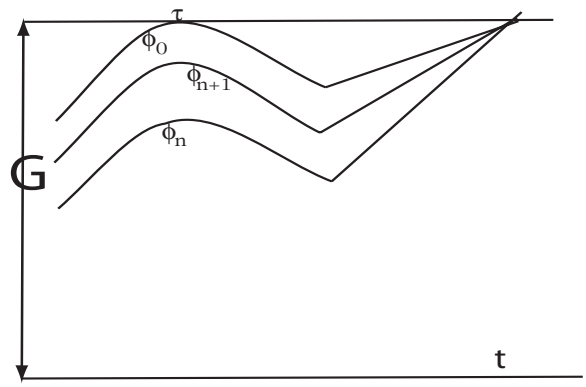


FIG. 1. *Continuity of first exit times.*

Define $b(\cdot) = b^1(\cdot) + b^2(\cdot)$, $k(\cdot) = k_1(\cdot) + k_2(\cdot)$. The following condition is assumed to hold.

CONDITION A2.1. *The functions $b^i(\cdot)$ and $\sigma(\cdot)$ are bounded, continuous, and Lipschitz continuous in x, uniformly in u. The controls $u_i(\cdot)$ for player i take values in U_i, a compact set in some Euclidean space, and the functions $k_i(\cdot)$ and $g_i(\cdot)$ are bounded and continuous.*

A control $u_i(\cdot)$ is said to be in $\mathcal{U}_i$, the set of admissible controls for player i, if it is measurable, nonanticipative with respect to $w(\cdot)$, and U_i-valued. Later we will introduce strategies and admissible relaxed controls. Part of the proof uses a weak convergence analysis as in [19], and to the extent possible we use the results of that reference. For S a topological space, let $D[S;0,\infty)$ denote the S-valued functions on $[0,\infty)$ that are right-continuous and have left-hand limits, and with the Skorkohod topology [7, 19] used. If $S = \mathbb{R}^v$, then we write $D[S;0,\infty) = D^v[0,\infty)$.

The first hitting time τ. Getting numerical solutions requires working in a bounded state space. Often the physics of the problem provide both a bounded state space and the proper boundary conditions. Otherwise, "numerical" boundaries are added. In any case, one needs to provide the necessary boundary conditions. These will be equivalent to either reflection or absorption at the boundary. Both are covered in [19]. Here, we chose boundary absorption, but the details that are unique to the nonzero-sum game problem would be the same in both cases.

The nature of the hitting time τ of the boundary of the set G poses a particular concern from the point of view of the convergence of the numerical algorithm. The proof of convergence generates a sequence of process approximations (continuous-time interpolations of the approximating chain), and the exit or boundary hitting time of this sequence has to converge, in an appropriate probabilistic sense, to the exit time of (2.1). In fact, no matter what the numerical procedure, something analogous must take place. In order to see the problem, refer to Figure 1. In the figure, the sequence of functions $\phi_n(\cdot)$ converges to the limit function $\phi_0(\cdot)$, but the sequence of first contact times (τ_n) of $\phi_n(\cdot)$ converges to a time τ_0, which is not the moment τ of first contact of $\phi_0(\cdot)$ with the boundary line ∂G of G. The problem in this case is that the limit function $\phi_0(\cdot)$ is tangent to ∂G at the time of first contact.

For our control problem, if the approximating costs are to converge to the costs for (2.1), (2.2), then we need to ensure, at least with probability one (w.p.1), that the paths of the limit $x(\cdot)$ are not "tangent" to ∂G at the moment τ of first

hitting the boundary. For $\phi(\cdot)$ in $D^v[0,\infty)$ (with the Skorokhod topology used), define the function $\hat{\tau}(\phi)$ with values in the compactified infinite interval $\overline{\mathbb{R}}^+ = [0,\infty]$ by $\hat{\tau}(\phi) = \infty$ if $\phi(t) \in G^0$, the interior of G, for all $t < \infty$; otherwise use

$$\hat{\tau}(\phi) = \inf\{t : \phi(t) \notin G^0\}.$$

In the example of Figure 1, $\hat{\tau}(\cdot)$ is not continuous at the path $\phi_0(\cdot)$.

If the $\phi_0(\cdot)$ in the figure were a sample path of a Wiener process $w(\cdot)$, then the probability would be zero that it is "tangent" to the boundary of G at the point of first contact. Indeed, w.p.1, the path would cross the line infinitely often in any small time interval after first contact. Hence, w.p.1, the first hitting times of any approximating sequence would converge to the first hitting time of $w(\cdot)$. The situation is similar if the Wiener process were replaced by the solution to a stochastic differential equation with a uniformly positive definite covariance matrix $a(x) = \sigma(x)\sigma'(x)$ if the boundary ∂G of G is "smooth." The following condition will be used. Note that the condition on the stopping time can be assured to hold if the randomized stopping approximation discussed below is used.

CONDITION A2.2. *For a continuous real-valued function $\Phi(\cdot)$ on $\mathbb{R}^v$, define $G = \{x : \Phi(x) \le 0\}$, and suppose that it is the closure of its interior $\{x : \Phi(x) < 0\}$. For each initial condition and control, the function $\hat{\tau}(\cdot)$ is continuous (as a map from $D^v[0,\infty)$ to the compactified interval $[0,\infty]$) w.p.1 relative to the measure induced by the solution to* (2.1).

The tangency problem would be a concern with any numerical method, since they all depend on some sort of approximation. For example, the convergence theorems for the classical finite difference methods for elliptic and parabolic equations generally use a nondegeneracy condition on $a(x)$ in order to (implicitly) guarantee Condition A2.2.

The verification of Condition A2.2 for the case where $a(x)$ is degenerate is more complicated, and one needs to work with the particular structure of the individual case. The boundary can often be divided into several pieces, where we are able to treat each piece separately. For example, there might be a segment where a "directional nondegeneracy" of $a(x)$ guarantees the almost sure continuity of the exit times of the paths which exit on that segment, plus a segment where the direction of the drift gives a similar guarantee, plus a segment on which escape is not possible, and finally a "remaining" segment. Frequently, the last "complementary" set is a finite set of points or a curve of dimension lower than that of the boundary. Special considerations concerning these points can often resolve the issue there. An important class of such a degenerate example is illustrated in [14, pp. 64–66]. In that two-dimensional example, G is the symmetric square box centered about the origin, the system is ($x = (x_1, x_2)$)

$$\begin{aligned} dx_1 &= x_2 dt, \\ dx_2 &= u dt + dw, \end{aligned}$$

and the control $u(\cdot)$ is bounded. The above cited "complementary set" is just the two points which are the intersections of the horizontal axis with the boundary, and these points can be taken care of by a test such as that in Theorem 6.1 of [21].[2]

Randomized stopping. An alternative to Condition A2.2. The boundaries in control problems are often not fixed precisely. For example, they might be introduced simply to bound the state space. The original control problem might be defined in

[2]See also [19, p. 280], where it is shown that the Girsanov transformation can play a useful role in the verification of Condition A2.2.

an unbounded space, but the space is then truncated for numerical reasons. Even if there is a given "target set," it is often not necessary to fix it too precisely. Such considerations give us some freedom to slightly vary the boundary. The "randomized stopping" alternative discussed next exploits these ideas and ensures Condition A2.2. Under randomized stopping, the probability of stopping at time t (if the process has not yet been stopped) goes to unity at that time as $x(t)$ approaches ∂G. This can be formalized as follows [19].

For some small $\epsilon > 0$, let $\bar{\lambda}(\cdot) > 0$ be a continuous function on the set $N_\epsilon(\partial G) \cap G^0$, where $N_\epsilon(\partial G)$ is the ϵ-neighborhood of the boundary and G^0 is the interior of G. Let $\bar{\lambda}(x) \to \infty$ as x converges to ∂G. Then stop $x(\cdot)$ at time t with stopping rate $\bar{\lambda}(x(t))$ and stopping cost $g_i(x(t))$ for player i. Randomized stopping is equivalent to adding an additional (and state-dependent) discount factor which is active near the boundary.

Relaxed controls $r_i(\cdot)$. In control theory, when working with problems concerning convergence of sequences or approximations, it is usual to use the so-called relaxed controls in lieu of ordinary controls. They are used for theoretical purposes only, i.e., to get approximation and convergence proofs. Suppose that for some filtration $\{\mathcal{F}_t, t < \infty\}$, standard vector-valued $\mathcal{F}_t$-Wiener process $w(\cdot)$, and for $i = 1, 2$, $r_i(\cdot)$ is a measure on the Borel sets of $U_i \times [0, \infty)$ such that $r_i(U_i \times [0, t]) = t$ and the process $r_i(A \times [0, \cdot])$ is measurable and nonanticipative for each Borel set $A \subset U_i$. Then $r_i(\cdot)$ is said to be an *admissible relaxed control* for player i with respect to $w(\cdot)$ [8, 19]. Abusing notation slightly, we use $\mathcal{U}_i$ for the set of admissible relaxed controls as well as for the set of admissible ordinary controls $u_i(\cdot)$. If the Wiener process and filtration are obvious or unimportant, we simply say that $r_i(\cdot)$ is an admissible relaxed control or a relaxed control for player i. For Borel sets $A \subset U_i$, we will write $r_i(A \times [t_0, t_1]) = r_i(A, [t_0, t_1])$, and write $r_i(A, t_1)$ if $t_0 = 0$. Define $U = U_1 \times U_2$ and $\mathcal{U} = \mathcal{U}_1 \times \mathcal{U}_2$. Henceforth $\{\mathcal{F}_t\}$ will denote a filtration such that $w(\cdot)$ is an $\mathcal{F}_t$-standard Wiener process and $r(\cdot)$ is admissible for the $r(\cdot)$ of concern.

For almost all (ω, t) and each Borel $A \subset U_i$, one can define the left derivative[3]

$$r_i'(A, t) = \lim_{\delta \to 0} \frac{r_i(A, t) - r_i(A, t - \delta)}{\delta}.$$

Without loss of generality, we can suppose that the limit exists for all (ω, t). Then for all (ω, t), $r_i'(\cdot, t)$ is a probability measure on the Borel sets of U_i, and for any bounded Borel set B in $U_i \times [0, \infty)$,

$$r_i(B) = \int_0^\infty \int_{U_i} I_{\{(\alpha_i, t) \in B\}} r_i'(d\alpha_i, t) dt.$$

An ordinary control $u_i(\cdot)$ can be represented in terms of the relaxed control $r_i(\cdot)$ that is defined by its derivative, which takes the form $r_i'(A, t) = I_A(u_i(t))$, where $I_A(u_i)$ is unity if $u_i \in A$ and is zero otherwise. The weak topology [19] will be used on the space of admissible relaxed controls. Relaxed controls are commonly used in control theory to prove existence and approximation theorems, since any sequence of relaxed controls has a weakly convergent subsequence. The use of relaxed controls does not change the range of values of the cost functions.

Define the "product" relaxed control $r(\cdot)$ by its derivative $r'(\cdot, t) = r_1'(\cdot, t) \times r_2'(\cdot, t)$. Thus $r(\cdot)$ is a product measure, with marginals $r_i(\cdot), i = 1, 2$. We will usually write $r(\cdot) = (r_1(\cdot), r_2(\cdot))$ without ambiguity. The pair $(w(\cdot), r(\cdot))$ is called an *admissible*

[3] "Left" is used because we need the derivative to be nonanticipative.

pair if each $r_i(\cdot)$ is admissible with respect to $w(\cdot)$. In relaxed control terminology, (2.1) and (2.2) are written as

$$\begin{aligned} x(t) &= x(0) + \sum_{i=1}^{2} \int_0^t \int_{U_i} b^i(x(s), \alpha_i) r_i'(d\alpha_i, s) ds + \int_0^t \sigma(x(s)) dw(s) \\ &= x(0) + \int_0^t \int_{U_i} b(x(s), \alpha_i) r'(d\alpha_i, s) ds + \int_0^t \sigma(x(s)) dw(s), \end{aligned} \tag{2.3}$$

$$W_i(x, r) = E_x^r \int_0^\tau e^{-\beta t} \int_{U_i} k_i(x(s), \alpha_i) r_i'(d\alpha_i, s) ds + E_x^r e^{-\beta\tau} g_i(x(\tau)). \tag{2.4}$$

The drift terms can be written as, e.g., $\int_0^t \int_U b(x(s), \alpha) r'(d\alpha, s) ds$.

A discrete-time system. We will also need the discrete-time form

$$\begin{aligned} x^\Delta(n\Delta + \Delta) = x^\Delta(n\Delta) &+ \int_{n\Delta}^{n\Delta+\Delta} \int_U b(x^\Delta(n\Delta), \alpha) r'(d\alpha, s) ds \\ &+ \sigma(x^\Delta(n\Delta))[w(n\Delta + \Delta) - w(\Delta)]. \end{aligned} \tag{2.5}$$

We can define the continuous-time interpolation $x^\Delta(\cdot)$ either by $x^\Delta(t) = x^\Delta(n\Delta)$ for $t \in [n\Delta, n\Delta + \Delta)$, or as (on the same interval)

$$x^\Delta(t) = x^\Delta(n\Delta) + \int_{n\Delta}^t \int_U b(x^\Delta(n\Delta), \alpha) r'(d\alpha, s) ds + \int_{n\Delta}^t \sigma(x^\Delta(n\Delta)) dw(t), \tag{2.6}$$

where it is assumed that $r(t, \cdot)$ is adapted to $\mathcal{F}_{n\Delta-}$ for $t \in [n\Delta, n\Delta + \Delta)$. The associated cost function $W_i^\Delta(x, r)$ is (2.4) with $x^\Delta(\cdot)$ replacing $x(\cdot)$. Let $r^\Delta(\cdot), r(\cdot)$ be admissible relaxed controls with respect to $w(\cdot)$ with $r^\Delta(\cdot) \to r(\cdot)$ w.p.1 (in the weak topology) and $r^\Delta(\cdot)$ adapted as above. Then, as $\Delta \to 0$, the sequence of solutions $\{x^\Delta(\cdot)\}$ also converges w.p.1, uniformly on any bounded time interval, and the limit $(x(\cdot), r(\cdot), w(\cdot))$ solves (2.3). By Condition A2.2, the first hitting times of the boundary also converge w.p.1 to that of the limit. The costs converge as well. The analogous result holds if the randomized stopping alternative is used.

Randomized and relaxed controls. For the discrete-time system (2.5) or (2.6), the relaxed control can be approximated by a randomized ordinary control (which relates the relaxed control to randomized strategies) as follows. Let $r(\cdot)$ be a relaxed control that is admissible with respect to $w(\cdot)$. Let $\tilde{u}_{i,n}^\Delta$ be a random variable with the distribution $r_{i,n}^\Delta(\cdot) = E_{n\Delta}\left[r_i\left(\cdot, [n\Delta, n\Delta + \Delta]\right)\right]/\Delta$, where $E_{n\Delta}$ denotes the conditional expectation given $\mathcal{F}_{n\Delta-}$. Set $\tilde{u}_n^\Delta = (\tilde{u}_{1,n}^\Delta, \tilde{u}_{2,n}^\Delta)$, define its continuous-time interpolation (with intervals Δ) $\tilde{u}^\Delta(\cdot)$, and define $\tilde{x}^\Delta(0) = x^\Delta(0) = x(0)$ and

$$\tilde{x}^\Delta(n\Delta + \Delta) = \tilde{x}^\Delta(n\Delta) + \Delta b(\tilde{x}^\Delta(n\Delta), \tilde{u}_n^\Delta) + \sigma(\tilde{x}^\Delta(n\Delta))[w(n\Delta + \Delta) - w(n\Delta)]. \tag{2.7}$$

Let $\tilde{x}^\Delta(t)$ denote the continuous-time interpolation. Define $r_n^\Delta(\cdot) = r_{1,n}^\Delta(\cdot) r_{2,n}^\Delta(\cdot)$, and let $r^\Delta(\cdot)$ be the relaxed control with derivative $r_n^\Delta(\cdot)$ on $[n\Delta, n\Delta + \Delta)$. Then we have the following result, where $r^\Delta(\cdot)$ is used for $x^\Delta(\cdot)$ in (2.6). The theorem implies that in the continuous limit, randomized controls turn into relaxed controls.

THEOREM 2.1. *Assume Condition* A2.1 *and use* $r_n^\Delta(\cdot)$ *in* (2.5) *and* (2.6). *Then for any* $T < \infty$,

$$\lim_{\Delta \to 0} \sup_{x(0) \in G} \sup_{r \in \mathcal{U}} E \sup_{t \le T} \left| x^\Delta(t) - x(t) \right|^2 = 0, \tag{2.8a}$$

$$\lim_{\Delta \to 0} \sup_{x(0) \in G} \sup_{r \in \mathcal{U}} E \sup_{t \le T} \left| x^\Delta(t) - \tilde{x}^\Delta(t) \right|^2 = 0. \tag{2.8b}$$

Under the additional Condition A2.2*, the costs for* (2.5) *and* (2.7) *converge* (*uniformly in* $x(0), r(\cdot)$) *to those for* (2.3) *as well.*

Comment on the proof. Define $\delta x_n^\Delta = x^\Delta(n\Delta) - \tilde{x}^\Delta(n\Delta)$. Then

$$\begin{aligned}\delta x_{n+1}^\Delta = \delta x_n^\Delta + \Delta \int_U \left[b(x^\Delta(n\Delta),\alpha) - b(\tilde{x}^\Delta(n\Delta),\alpha)\right] r_n^\Delta(d\alpha) \\ + \left[\sigma(x^\Delta(n\Delta)) - \sigma(\tilde{x}^\Delta(n\Delta))\right][w(n\Delta+\Delta) - w(n\Delta)] + N_n^\Delta,\end{aligned}$$

where

$$N_n^\Delta = \Delta\left[\int_U b(\tilde{x}^\Delta(n\Delta),\alpha) r_n^\Delta(d\alpha) - b(\tilde{x}^\Delta(n\Delta),\tilde{u}_n^\Delta)\right]$$

is an $\mathcal{F}_{n\Delta}$- martingale difference by the definition of $\tilde{u}_n^\Delta(\cdot)$ via the conditional distribution given $\mathcal{F}_{n\Delta}$. Also $E_{n\Delta}|N_n^\Delta|^2 = O(\Delta^2)$. The proof of the uniform (in the control and initial condition) convergence to zero of $|x^\Delta(\cdot) - \tilde{x}^\Delta(\cdot)|$ and of the differences between the integrals

$$E\int_0^t e^{-\beta t} k(\tilde{x}^\Delta(s),\tilde{u}^\Delta(s))ds, \quad E\int_0^t\int_U e^{-\beta t} k(x^\Delta(s),\alpha) r^{\Delta,\prime}(d\alpha,s)ds$$

can then be completed by using the Lipschitz condition and this martingale and conditional variance property. This implies (2.8b). An analogous argument can be used to get (2.8a) for each $r(\cdot)$ and $x(0)$. The facts that Condition A2.2 holds for (2.3) and that (2.8) holds imply that the stopping times for $x^\Delta(\cdot), \tilde{x}^\Delta(\cdot)$ converge to those for (2.3) as well, for each $x(0)$ and $r(\cdot)$.

The uniformity in (2.8b) and in the convergence of the costs can be proved by an argument by contradiction that goes roughly as follows. Suppose, for example, that the uniformity in (2.8b) does not hold. Then, for intervals Δ_m and relaxed controls $r^m(\cdot), m = 1, 2, \ldots$, define $r_n^{m,\Delta_m}(\cdot)$ as $r_n^\Delta(\cdot)$ was, but based on $r^m(\cdot)$, and let $r^{m,\Delta_m}(\cdot)$ denote the interpolation of the associated relaxed control. Let $\Delta_m \to 0$. Let $x^m(\cdot)$ solve (2.3) and $x^{m,\Delta_m}(\cdot)$ solve (2.6), both under $r^m(\cdot)$. Let $\tilde{x}^{m,\Delta_m}(\cdot)$ solve (2.7) under $r^{m,\Delta_m}(\cdot)$. Suppose that, for some $T < \infty$, $\limsup_{m\to\infty} E\sup_{t\le T}|x^{m,\Delta_m}(t) - \tilde{x}^{m,\Delta_m}(t)|^2 > 0$.

Take an arbitrary weakly convergent subsequence of $x^m(\cdot), x^{m,\Delta_m}(\cdot)$, $\tilde{x}^{m,\Delta_m}(\cdot)$, $r^m(\cdot)$, $r^{m,\Delta_m}(\cdot)$, $w(\cdot)$, also indexed by m and with the (weak-sense) limit denoted by $x(\cdot), \hat{x}(\cdot), \tilde{x}(\cdot), r(\cdot), \hat{r}(\cdot), \hat{w}(\cdot)$. Then it is easy to show that $x(\cdot) = \hat{x}(\cdot) = \tilde{x}(\cdot)$ and $r(\cdot) = \hat{r}(\cdot)$, that $\hat{w}(\cdot)$ is a standard Wiener process, that $x(\cdot), \hat{x}(\cdot), \tilde{x}(\cdot), \hat{r}(\cdot)$ are nonanticipative with respect to $\hat{w}(\cdot)$, and that the limit set satisfies (2.3). Assume, without loss of generality, that Skorokhod representation is used [7, 19] so that we can suppose that the original and limit processes are all defined on the same probability space and that convergence is w.p.1 in the Skorokhod topology. Then

$$\lim_{m\to\infty} E\sup_{t\le T}\left|\tilde{x}^{m,\Delta_m}(t) - \hat{x}(t)\right|^2 = 0$$

and

$$\lim_{m\to\infty} E\sup_{t\le T}\left|x^{m,\Delta_m}(t) - \hat{x}(t)\right|^2 = 0,$$

a contradiction to the assertion that the uniformity in $x(0)$ and $r(\cdot)$ in (2.8b) does not hold. □

3. Approximating the controls. The convergence proofs will require the use of special approximations to the general ordinary or relaxed controls, and the necessary approximations are developed in this section and in Theorem 4.1.

For each admissible relaxed control $r(\cdot)$ and $\epsilon > 0$, let $r_i^\epsilon(\cdot)$ be admissible relaxed controls with respect to the same filtration and Wiener process $w(\cdot)$, with derivatives $r_i^{\epsilon,\prime}(\cdot)$, and that satisfy

$$\lim_{\epsilon\to 0} \sup_{r_i\in\mathcal{U}_i} E \sup_{t\le T} \left| \int_0^t \int_{U_i} \phi_i(\alpha_i) \left[r_i'(d\alpha_i, s) - r_i^{\epsilon,\prime}(d\alpha_i, s) \right] ds \right| = 0, \quad i = 1, 2, \tag{3.1}$$

for each bounded and continuous real-valued nonrandom function $\phi_i(\cdot)$ and each $T < \infty$. Let $x(\cdot)$ and $x^\epsilon(\cdot)$ denote the solutions to (2.3) corresponding to $r(\cdot)$ and $r^\epsilon(\cdot)$, resp., with the same $w(\cdot)$ used, but perhaps different initial conditions. In particular, define $x^\epsilon(\cdot)$ by

$$x^\epsilon(t) = x^\epsilon(0) + \int_0^t \int_U b(x^\epsilon(s), \alpha) r^{\epsilon,\prime}(d\alpha, s) ds + \int_0^t \sigma(x^\epsilon(s)) dw(s). \tag{3.2}$$

The processes $x(\cdot)$ and $x^\epsilon(\cdot)$ depend on $r(\cdot)$ and $r^\epsilon(\cdot)$, resp., but this dependence is suppressed in the notation. The next theorem shows that the solution $x(\cdot)$ is continuous in the controls in the sense that (3.3) below holds, and that the costs corresponding to $r(\cdot)$ and $r^\epsilon(\cdot)$ are arbitrarily close for small ϵ, uniformly in $r(\cdot)$.

THEOREM 3.1. *Assume Condition* A2.1. *Let* $(r(\cdot), r^\epsilon(\cdot))$ *satisfy* (3.1) *for each bounded and continuous* $\phi_i(\cdot), i = 1, 2$, *and* $T < \infty$. *Define* $\delta x^\epsilon(t) = x^\epsilon(t) - x(t)$. *Then for each* t,

$$\lim_{\epsilon\to 0} \sup_{x(0), x^\epsilon(0): |x^\epsilon(0) - x(0)| \to 0} \sup_{r\in\mathcal{U}} E \sup_{s\le t} |\delta x^\epsilon(s)|^2 = 0. \tag{3.3}$$

Under the additional Condition A2.2,

$$\lim_{\epsilon\to 0} \sup_{x(0), x^\epsilon(0): |x^\epsilon(0) - x(0)| \to 0} \sup_{r\in\mathcal{U}} |W_i(x, r) - W_i(x, r^\epsilon)| = 0, \quad i = 1, 2. \tag{3.4}$$

Comments on the proof. The proof is very similar to that of Theorem 2.1, and we comment only on the use of (3.1). We can write

$$\begin{aligned} \delta x^\epsilon(t) = \delta x^\epsilon(0) &+ \int_0^t \int_U \left[b(x^\epsilon(s), \alpha) - b(x(s), \alpha) \right] r'(d\alpha, s) ds \\ &+ \int_0^t \left[\sigma(x^\epsilon(s)) - \sigma(x(s)) \right] dw(s) \\ &+ \int_0^t \int_U b(x^\epsilon(s), \alpha) \left[r^{\epsilon,\prime}(d\alpha, s) - r'(d\alpha, s) \right] ds. \end{aligned} \tag{3.5}$$

It will be seen that the sup over any finite time interval of the absolute value of the last term of (3.5) goes to zero in mean square, by virtue of (3.1). For small $\lambda > 0$, that term can be rewritten as (modulo $O(\lambda)$)

$$\begin{aligned} &\sum_{l=0}^{[t/\lambda]-1} \int_{l\lambda}^{l\lambda+\lambda} \int_U b(x^\epsilon(l\lambda), \alpha) \left[r^{\epsilon,\prime}(d\alpha, s) - r'(d\alpha, s) \right] ds \\ &+ \sum_{l=0}^{[t/\lambda]-1} \int_{l\lambda}^{l\lambda+\lambda} \left[b(x^\epsilon(s), \alpha) - b(x^\epsilon(l\lambda), \alpha) \right] \left[r^{\epsilon,\prime}(d\alpha, s) - r'(d\alpha, s) \right] ds. \end{aligned} \tag{3.6}$$

Here $[t/\lambda]$ denotes the integer part of t/λ. As $\lambda \to 0$ the expectation of the square of the sup over any finite interval of the last term goes to zero, uniformly in $r(\cdot), r^\epsilon(\cdot), x(0)$, $x^\epsilon(0)$, whether or not (3.1) holds, since

$$\lim_{\lambda\to 0} \sup_r \sup_\epsilon \sup_{l\lambda\le t} E \sup_{s\le\lambda} |x^\epsilon(l\lambda+s) - x^\epsilon(l\lambda)|^2 = 0. \tag{3.7}$$

Assumption (3.1) can be used to show that the same uniform limit in mean square holds for the first term of (3.6) for any λ, as $\epsilon \to 0$. The proof of (3.3) is a consequence of these facts and the Lipschitz condition. The convergence of the costs is a consequence of the convergence of the paths and controls, and an argument concerning the convergence of the stopping times such as used in Theorem 2.1. □

Finite-valued and piecewise-constant approximations $r^\epsilon(\cdot)$ in (3.1). Now some approximations of subsequent interest will be described. It will be seen that we can confine our attention to control processes that are just piecewise-constant and finite-valued ordinary admissible controls. Consider the following discretization of the U_i. Let $U_i \in \mathbb{R}^{c_i}$, Euclidean c_i-space. Given $\mu > 0$, partition $\mathbb{R}^{c_i}$ into disjoint (hyper)cubes $\{R_i^{\mu,l}\}$ with diameters μ. The boundaries can be assigned to the subsets in any way. Define $U_i^{\mu,l} = U_i \cap R_i^{\mu,l}$ for the finite number (p_i^μ) of nonempty intersections. Choose a point $\alpha_i^{\mu,l} \in U_i^{\mu,l}$. Now, given admissible $(r_1(\cdot), r_2(\cdot))$, define the approximating admissible relaxed control $r_i^\mu(\cdot)$ on the control value space $U_i^\mu = \{\alpha_i^{\mu,l}, l \le p_i^\mu\}$ by its derivative as $r_i^{\mu,\prime}(\alpha_i^{\mu,l}, t) = r_i'(U_i^{\mu,l}, t)$. Denote the set of such controls by $\mathcal{U}_i(\mu)$. The following theorem is a consequence of Theorem 3.1. A version can also be found in [16].

THEOREM 3.2. *Assume Conditions* A2.1–A2.2*, and the above approximation of $r_i(\cdot)$ by $r_i^\mu(\cdot) \in \mathcal{U}_i(\mu), i = 1, 2$. Then* (3.1) *and Theorem* 3.1 *hold for μ replacing ϵ, no matter what the $\{U_i^{\mu,l}, \alpha_i^{\mu,l}\}$ are. The same result holds if we approximate only one of the $r_i(\cdot)$.*

Finite-valued, piecewise-constant, and "delayed" approximations. The proofs of convergence depend on showing that the cost changes little if the control actions of any player are discretized in time and are slightly delayed. Let $r_i^\mu(\cdot) \in \mathcal{U}_i(\mu)$, where the control value space for player i is U_i^μ. Let $\Delta > 0$. Define the "backward" differences $\Delta_{i,k}^{\mu,l} = r_i^\mu(\alpha_i^{\mu,l}, k\Delta) - r_i^\mu(\alpha_i^{\mu,l}, k\Delta - \Delta)$, $l \le p_i^\mu$, $k = 1, \ldots$. Define the piecewise-constant ordinary controls $u_i^{\mu,\Delta}(\cdot) \in \mathcal{U}_i(\mu)$ on the interval $[k\Delta, k\Delta + \Delta)$ by

$$u_i^{\mu,\Delta}(t) = \alpha_i^{\mu,l} \text{ for } t \in \left[k\Delta + \sum_{\nu=1}^{l-1} \Delta_{i,k}^{\mu,\nu},\ k\Delta + \sum_{\nu=1}^{l} \Delta_{i,k}^{\mu,\nu}\right). \tag{3.8}$$

Note that on $[k\Delta, k\Delta + \Delta)$, $u_i^{\mu,\Delta}(\cdot)$ takes the value $\alpha_i^{\mu,l}$ on a time interval of length $\Delta_{i,k}^{\mu,l}$. Note also that the $u_i^{\mu,\Delta}(\cdot)$ are "delayed," in that the values of $r_i(\cdot)$ on $[k\Delta - \Delta, k\Delta)$ determine the values of $u_i^{\mu,\Delta}(\cdot)$ on $[k\Delta, k\Delta + \Delta)$. Thus $u_i^{\mu,\Delta}(t), t \in [k\Delta, k\Delta + \Delta)$, is $\mathcal{F}_{k\Delta-}$-measurable. Let $r_i^{\mu,\Delta}(\cdot)$ denote the relaxed control representation of $u_i^{\mu,\Delta}(\cdot)$, with time derivative $r_i^{\mu,\Delta,\prime}(\cdot)$. Let $\mathcal{U}_i(\mu, \delta)$ denote the subset of $\mathcal{U}_i(\mu)$ that are ordinary controls that are constant on the intervals $[l\delta, l\delta + \delta), l = 0, 1, \ldots$.

The intervals $\Delta_{i,k}^{\mu,l}$ in (3.8) are just real numbers. For later use, it is important to have them be some multiple of some small $\delta > 0$, where Δ/δ is an integer. Consider one method of doing this. Divide $[k\Delta, k\Delta + \Delta)$ into Δ/δ subintervals of length δ each. Working in order $l = 1, 2 \ldots$, to each value $\alpha_i^{\mu,l}$ first assign (the integer part) $[\Delta_{i,k}^{\mu,l}/\delta]$

successive subintervals of length δ. The total fraction of time that is unassigned on any bounded time interval will go to zero as $\delta \to 0$, and how control values are assigned to them will have little effect. However, for specificity for future use consider the following method. The unassigned length for value $\alpha_i^{\mu,l}$ is $L_{i,k}^{\mu,\delta,l} = \Delta_{i,k}^{\mu,l} - [\Delta_{i,k}^{\mu,l}/\delta]\delta$, $i \le p_i^\mu$. Define the sum $S_{i,k}^{\mu,\delta} = \sum_l L_{i,k}^{\mu,\delta,l}$, which must be an integral multiple of δ. Then assign each unassigned δ-interval at random with value $\alpha_{i,k}^{\mu,l}$ chosen with probability $L_{i,k}^{\mu,\delta,l}/S_{i,k}^{\mu,\delta}$. By Theorem 2.1, this assignment and randomization approximates the original relaxed control.

Let $\mathcal{U}_i(\mu,\delta,\Delta)$ denote the set of such controls. If $u_i^{\mu,\delta,\Delta}(\cdot)$ is obtained from $r_i(\cdot)$ in this way, then it is a function of $r_i(\cdot)$, but this functional dependence will be omitted in the notation. Let $r_i^{\mu,\Delta,\delta,\prime}(\cdot)$ denote the time derivative of $r_i^{\mu,\Delta,\delta}(\cdot)$. As stated in the next theorem, which is a consequence of Theorem 3.1, for fixed μ and small δ, $u_i^{\mu,\delta,\Delta}(\cdot)$ well approximates the effects of $u_i^{\mu,\Delta}(\cdot)$ and $r_i(\cdot)$, uniformly in $r_i(\cdot)$ and $\{\alpha_i^{\mu,l}\}$. In particular, (3.1) holds in the sense that for each $\mu > 0$, $\Delta > 0$, and bounded and continuous $\phi_i(\cdot)$, for $i = 1, 2$,

$$\lim_{\delta\to 0} \sup_{r_i\in\mathcal{U}_i} E \sup_{t\le T} \left| \int_0^t \int_{U_i} \phi_i(\alpha_i) \left[r_i^{\mu,\delta,\Delta,\prime}(d\alpha_i, s) - r_i^{\mu,\Delta,\prime}(d\alpha_i, s) \right] ds \right| = 0. \tag{3.9}$$

THEOREM 3.3. *Assume Conditions* A2.1–A2.2. *Let $r_i(\cdot) \in \mathcal{U}_i, i = 1, 2$. Given $(\mu,\delta,\Delta) > 0$, approximate as discussed above the theorem to get $r_i^{\mu,\delta,\Delta}(\cdot) \in \mathcal{U}_i(\mu,\delta,\Delta)$. Then* (3.1) *holds for $r_i^{\mu,\delta,\Delta}(\cdot)$ and (μ,δ,Δ) replacing $r_i^\epsilon(\cdot)$ and ϵ, respectively. Also,* (3.9) *holds. In particular, given $\epsilon > 0$, there are $\mu_\epsilon > 0, \delta_\epsilon > 0, \Delta_\epsilon > 0$, and $\kappa_\epsilon > 0$, such that for $\mu \le \mu_\epsilon, \delta \le \delta_\epsilon, \Delta \le \Delta_\epsilon$, and $\delta/\Delta \le \kappa_\epsilon$,*

$$\sup_x \sup_{r_1} \sup_{r_2} \left| W_i(x, r_1, r_2) - W_i(x, r_1, u_2^{\mu,\delta,\Delta}) \right| \le \epsilon. \tag{3.10}$$

The expression (3.10) *holds with the indices* 1 *and* 2 *interchanged or if both controls are approximated.*

Consider the discrete-time system (2.5) *with either the interpolation that is piecewise-constant or* (2.6). *Then the $\mu_\epsilon > 0, \delta_\epsilon > 0, \Delta_\epsilon > 0$, and $\kappa_\epsilon > 0$, can be defined so that*

$$\sup_x \sup_{r_1} \sup_{r_2} \left| W_i(x, r_1, r_2) - W_i^\Delta(x, r_1, u_2^{\mu,\delta,\Delta}) \right| \le \epsilon. \tag{3.11}$$

The expression (3.11) *holds with the indices* 1 *and* 2 *interchanged or if both controls are approximated and/or further delayed by Δ.*

Note on the initial values of the controls. Since the controls are delayed by Δ, we can assign the values on the initial interval $[0, \Delta]$ in any way. Let the values $u_i(l\delta), l\delta \le \Delta$, be in U_i^μ and fixed for $i = 1, 2$.

4. Equilibria and approximations. *Elliott–Kalton strategies.* The classical definition of strategy as used in differential games for models such as (2.1) or (2.3) is that of Elliott and Kalton [6, 9]. A strategy $c_1(\cdot)$ for player 1 is a mapping from $\mathcal{U}_2$ to $\mathcal{U}_1$ with the following property. If admissible controls $r_2(\cdot)$ and $\tilde{r}_2(\cdot)$ satisfy $r_2(s) = \tilde{r}_2(s)$ for $s \le t$, then $c_1(r_2)(s) = c_1(\tilde{r}_2)(s), s \le t$, and with an analogous definition for player 2 strategies. Let $\mathcal{C}_i$ denote the set of such strategies or mappings for player i. An Elliott–Kalton strategy is a generalization of a feedback control. The current control action that it yields for any player is a function only of the past

control actions and does not otherwise depend on the form of the strategy of the other player.

A pair $\bar{c}_i(\cdot) \in \mathcal{C}_i, i = 1, 2$, is said to be an ϵ-equilibrium strategy pair if for all admissible controls $r_i(\cdot), i = 1, 2$,[4]

$$\begin{aligned} W_1(x, \bar{c}_1, \bar{c}_2) &\geq W_1(x, r_1, \bar{c}_2) - \epsilon, \\ W_2(x, \bar{c}_1, \bar{c}_2) &\geq W_2(x, \bar{c}_1, r_2) - \epsilon. \end{aligned} \tag{4.1}$$

The notation $W_1(x, c_1, c_2)$ implies that each player i uses its strategy $c_i(\cdot)$. When writing $W_i(x, c_1, c_2)$, it is assumed that the associated process is well defined. This will be the case here, since Theorem 3.3 implies that it is sufficient to restrict attention to strategies whose control functions are piecewise-constant and finite-valued and can depend only on slightly delayed values of the other players' control realizations. If (4.1) holds with $\epsilon = 0$, then we have an equilibrium strategy pair. The controls can be either ordinary or relaxed. The notation $W_2(x, c_1, r_2)$ implies that player 1 uses its strategy $c_1(\cdot)$ and player 2 uses the relaxed control $r_2(\cdot)$.

The above definition of strategy does not properly allow for randomized controls, where the realized responses given by the strategy of a player to a fixed control process of the other player might differ, depending on the random choices that it makes. So we also allow randomized strategies that have the form of the second line in (4.2) below for either one or both of the players. Theorem 2.1 shows the connection between relaxed and randomized controls.

We will require the following assumption.[5]

ASSUMPTION A4.1. *For each small $\epsilon > 0$ there is an ϵ-equilibrium Elliott–Kalton strategy $(\bar{c}_1^\epsilon(\cdot), \bar{c}_2^\epsilon(\cdot))$ under which the solution to* (2.1) *or* (2.3) *is well defined.*

The following approximation theorem will be a key item in the development.

THEOREM 4.1. *Assume Conditions* A2.1 *and* A2.2. *Given $\epsilon_1 > 0$, there are positive numbers μ, δ, Δ, where Δ/δ is an integer, such that the values for any strategy pair $(c_1(\cdot), c_2(\cdot))$ with $c_i(\cdot) \in \mathcal{C}_i$, and under which the solution to* (2.3) *is well defined,*[6] *can be approximated within ϵ_1 by strategy pairs $c_i^{\mu,\delta,\Delta}(\cdot), i = 1, 2$, of the following form. The realizations of $c_i^{\mu,\delta,\Delta}(\cdot)$ (which depend on the other player's strategy or control) are ordinary controls in $\mathcal{U}_i(\mu, \delta, \Delta)$, and we denote them by $u_i^{\mu,\delta,\Delta}(\cdot)$. For integers n, k, and $k\delta \in [n\Delta, n\Delta + \Delta)$, and α_i taking values in U_i^μ,*

$$\begin{aligned} &P\left\{u_i^{\mu,\delta,\Delta}(k\delta) = \alpha_i \middle| w(s), s \leq k\delta; u_j^{\mu,\delta,\Delta}(l\delta), j = 1, 2, l < k\right\} \\ &= P\left\{u_i^{\mu,\delta,\Delta}(k\delta) = \alpha_i \middle| w(l\Delta), l \leq n; u_j^{\mu,\delta,\Delta}(l\delta), j = 1, 2, l\delta < n\Delta\right\} \\ &= p_{i,k}\left(\alpha_i;\, w(l\Delta), l \leq n; u_j^{\mu,\delta,\Delta}(l\delta), j = 1, 2, l\delta < n\Delta\right), \end{aligned} \tag{4.2}$$

[4]The definition in [6] requires that the controls $r_i(A, \cdot)$ be progressively measurable, and not simply measurable and adapted, for each Borel set A. But due to the approximation results of Theorems 3.1–3.3, this added requirement is unnecessary in our case.

[5]As noted above, one need only restrict attention to strategies that yield control processes that are piecewise-constant, finite-valued, and slightly delayed, or, indeed, to discrete-time systems with such controls. This moderates the assumption considerably. Assumption A4.1 is the weakest possible assumption concerning equilibria. If it does not hold, then there is no numerical problem, since there is no solution. Numerical analysis starts with whatever assumptions are needed to ensure that there is a solution. Criteria for the existence of equilibria are in [2, 4, 11]. See also the seminal work [10] and the review and references for stochastic differential games in general in [20].

[6]One or both of them might be simply fixed relaxed feedback controls.

which defines the functions $p_{i,k}(\cdot)$. For each positive value of μ, δ, Δ, the functions $p_{i,k}(\cdot)$ can be taken to be continuous in the w-arguments for each value of the other arguments.

Suppose that the control process realizations for player i are in $\mathcal{U}_i(\mu, \delta, \Delta)$, but those of the other player are general relaxed controls. Then we interpret (4.2)*, applied to that control, as being based on its discretized approximation, as derived above Theorem* 3.3.

A convenient representation of the values in (4.2). It will be useful for the convergence proofs if the random selections implied by the conditional probabilities in (4.2) are systematized as follows. Let $\{\theta_k\}$ be random variables that are mutually independent and uniformly distributed on $[0,1]$. The $\{\theta_k, k \geq l\}$ will be independent of all system data before time $l\delta$. For each i, n, k, divide $[0,1]$ into (random) subintervals whose lengths are proportional to the conditional probability of the $\alpha_i^{\mu,l}$ as given by (4.2), and select $u_i^{\mu,\delta,\Delta}(k\delta) = \alpha_i^{\mu,l}$ if the random selection of θ_k on $[0,1]$ falls into that subinterval. The same random variables $\{\theta_k\}$ are used for both players, and for all conditional probability rules of the form (4.2). This representation is used only for theoretical purposes.

Proof. By Theorem 3.3, it is sufficient to work with strategies $c_i^{\mu,\delta,\Delta}(\cdot), i = 1, 2$, whose control process realizations are in $\mathcal{U}_i(\mu, \delta, \Delta)$, and that in any interval $[n\Delta, n\Delta + \Delta)$ depend only on the control process values of the other player up to time $n\Delta - \Delta$. So we start with such strategies. However, we need to consider the response of such a strategy if the other player uses controls that are not already discretized. Since, for small μ, δ, Δ, a discretization of the other player's control realizations would have a negligible effect on the costs, uniformly in the control policy realizations of the first player, it is sufficient for the first player to act as though the controls of the other player were already discretized, by computing the discretization.

Continuing, let $u_i^{\mu,\delta,\Delta}(\cdot), i = 1, 2$, denote the policy realizations under the strategies $c_i^{\mu,\delta,\Delta}(\cdot), i = 1, 2$. The probability law of $(u_1^{\mu,\delta,\Delta}(\cdot), u_2^{\mu,\delta,\Delta}(\cdot), w(\cdot))$ determines the law of the corresponding solution to (2.1). The law of evolution of the controls can be written in recursive form, for $i = 1, 2$, and $k\delta \in [n\Delta, n\Delta + \Delta)$,

$$P\left\{u_i^{\mu,\delta,\Delta}(k\delta) = \alpha_i \middle| w(s), s \leq n\Delta; u_j^{\mu,\delta,\Delta}(l\delta), j = 1, 2, l\delta < n\Delta\right\}. \tag{4.3}$$

This yields a "randomized" Elliott–Kalton strategy pair.

Now apply the control rule (4.3) to the piecewise-constant interpolation of the discrete-time system (2.5). The probability law of the solution on $[0, t]$ is determined by the law of $\left(u_1^{\mu,\delta,\Delta}(l\delta), u_2^{\mu,\delta,\Delta}(l\delta), l\delta < t; w(n\Delta), n\Delta \leq t\right)$. Hence, for $k\delta \in [n\Delta, n\Delta + \Delta)$, the probability law of the controls and paths for $x^\Delta(\cdot)$ can be determined from the formula

$$P\left\{u_i^{\mu,\delta,\Delta}(k\delta) = \alpha_i \middle| w(l\Delta), l \leq n; u_j^{\mu,\delta,\Delta}(l\delta), j = 1, 2, l\delta < n\Delta\right\}. \tag{4.4}$$

By Theorem 3.3, for small enough δ, Δ the paths $x^\Delta(\cdot)$ and $x(\cdot)$ and the associated costs are arbitrarily close, uniformly in the controls $u_i^{\mu,\delta,\Delta}(\cdot), i = 1, 2$, where we can suppose (without loss of generality) that the law of evolution of the controls takes the form (4.4). This argument implies that we can restrict the conditioning on $w(\cdot)$ in (4.3) to the samples $w(l\Delta)$.

Now turn to the assertion concerning continuity in the w-arguments. (See also [19, Theorem 10.3.1] on this point.) For $\rho > 0$, consider the smoothed conditional probability defined by

$$p^{\rho}_{i,k}\left(\alpha_i;\, w(l\Delta), l \le n; u^{\mu,\delta,\Delta}_j(l\delta), j = 1,2, l\delta < n\Delta\right)$$
$$= N(\rho)\int e^{-|z-w|^2/2\rho} p_{i,k}\left(\alpha_i; z; u^{\mu,\delta,\Delta}_j(l\delta), j = 1,2, l\delta < n\Delta\right) dz,$$

where $N(\rho)$ is a normalizing constant and $w = \{w(l\Delta), l \le n\}$. The variable z has the same dimension as w. The integral is continuous in the w-variables, uniformly in the others. Also it converges to

$$p_{i,k}\left(\alpha_i;\, w(l\Delta), l \le n; u^{\mu,\delta,\Delta}_j(l\delta), j = 1,2, l\delta < n\Delta\right)$$

for almost all $\{w(l\Delta), l \le n\}$ values as $\rho \to 0$. Hence, by Egoroff's theorem, it converges almost uniformly in any compact set of $\{w(l\Delta), l \le n\}$ values. For almost all $\{w(l\Delta), l \le n\}$ values the smoothed conditional probability will choose the same control values as would the original rule defined by (4.4) with a probability that goes to unity as $\rho \to 0$. Hence, without loss of generality we can suppose that the $p_{i,k}(\cdot)$ are smooth in the w-variables, as asserted. □

5. The Markov chain approximation: Brief review and approximations.

5.1. The Markov chain approximation method. We will start by giving a quick overview of the Markov chain approximation method of [14, 15, 19], starting with some comments for the case where there is only one player. We will then develop some approximation results for the chains that are analogous to those in Theorem 3.3, and which will be crucial for the convergence theorems in section 6. The method consists of two steps. Let $h > 0$ be an approximation parameter. The first step is the determination of a finite-state controlled Markov chain ξ^h_n that has a continuous-time interpolation that is an "approximation" of the process $x(\cdot)$. The second step solves the optimization problem for the chain and a cost function that approximates the one used for $x(\cdot)$. Under a natural "local consistency" condition, the minimal cost function for the chain converges to the minimal cost function for the original problem. In applications, the optimal control for the original problem is also approximated. The approximating chain and local consistency conditions are the same for the game problem. The reference [19] contains a comprehensive discussion of many automatic and simple methods for getting the transition probabilities of the chain. The approximations "stay close" to the physical model and can be adjusted to exploit local features.

The simplest state space for the chain for our model (and the one that we will use for simplicity in the discussion) is based on the regular h-grid S_h in $\mathbb{R}^v$. Define $G_h = S_h \cap G$ and $G^0_h = S_h \cap G^0$. On G^0_h the chain "approximates" the diffusion part of (2.1) or (2.3). Let ∂G_h denote the points in $S_h - G^0_h$ that can be reached in one step from G^0_h under some control. These are the boundary points, and the process stops on first reaching them. It is only the points in $G^0_h \cup \partial G_h$ that are of interest.

Next we define the basic condition of *local consistency.* Let $u^h_n = (u^h_{1,n}, u^h_{2,n})$ denote the controls that are used at step n. Define $\Delta\xi^h_n = \xi^h_{n+1} - \xi^h_n$ and let $E^{h,\alpha}_{x,n}$ denote the expectation given the data to step n (when ξ^h_n has just been computed) with $\xi^h_n = x$ and control value $\alpha = u^h_n$ to be used in the next step. For the game problem, $\alpha = (\alpha_1, \alpha_2)$ with $\alpha_i \in U_i$. Define $a(x) = \sigma(x)\sigma'(x)$. Suppose that there is a

function $\Delta t^h(\cdot)$ (this is obtained automatically when the transition probabilities are calculated; see [19] and the example below) such that (this defines the functions $b^h(\cdot)$ and $a^h(\cdot)$)

$$\begin{aligned} &E^{h,\alpha}_{x,n}\Delta\xi^h_n \equiv b^h(x,\alpha)\Delta t^h(x,\alpha) = b(x,\alpha)\Delta t^h(x,\alpha) + o(\Delta t^h(x,\alpha)),\\ &\operatorname{cov}^{h,\alpha}_{x,n}[\Delta\xi^h_n - E^{h,\alpha}_{x,n}\Delta\xi^h_n] \equiv a^h(x,\alpha)\Delta t^h(x,\alpha) = a(x)\Delta t^h(x,\alpha) + o(\Delta t^h(x,\alpha)),\\ &\lim_{h\to 0}\sup_{x\in G,\alpha\in U}\Delta t^h(x,\alpha) = 0. \end{aligned} \tag{5.1}$$

It can be seen that the chain has the "local properties" (conditional mean change and conditional covariance) of the diffusion process.[7] One can always select the transition probabilities such that the intervals $\Delta t^h(x,\alpha)$ do not depend on the control variable, although the general theory in [19] does not require it. Such a simplification is often done in applications to simplify the coding. Let $p^h(x,y|\alpha)$ denote the probability that the next state is y given that the current state is x and control pair $\alpha = (a_1,\alpha_2)$ is used.

Under our condition that the controls are separated in $b(\cdot)$, in that $b(x,\alpha) = v^1(x,\alpha_1) + b^2(x,\alpha_2)$, if desired one can construct the chain so that the controls are "separated" in which the one-step transition probability has the form

$$p^h(x,y|\alpha) = p^h_1(x,y|\alpha_1) + p^h_2(x,y|\alpha_2). \tag{5.2}$$

A useful representation of the transition probabilities. For the convergence proof, it is useful to have the chains for each h defined on the same probability space, no matter what the controls.[8] This is done as follows. Let $\{\chi_n\}$ be a sequence of mutually independent random variables, uniformly distributed on the interval $[0,1]$ and such that $\{\chi_l, l \geq n\}$ is independent of $\{\xi^h_l, u^h_l, l \leq n\}$. For each value of $x = \xi^h_n, \alpha = u^h_n$, arrange the finite number of possible next states y in some order and divide the interval $[0,1]$ into successive subintervals whose lengths are $p^h(x,y|\alpha)$. Then for $x = \xi^h_n, \alpha = u^h_n$, select the next state according to where the (uniformly distributed) random choice for χ_n falls. The same random variables $\{\chi_n\}$ will be used in all cases, for all controls and values of h. This representation is used only for theoretical purposes.

An example of an approximating chain. The simplest case for illustrative purposes is one-dimensional, and where h is small enough so that $h|b(\alpha,x)| \leq \sigma^2(x)$. Then we can use the transition probabilities and interval, for $x \in G^0_h$ [19, Chapter 5],

$$p^h(x,x\pm h|\alpha) = \frac{\sigma^2(x) \pm hb(x,\alpha)}{2\sigma^2(x)}, \quad \Delta t^h(x,\alpha) = \frac{h^2}{\sigma^2(x)}, \quad \Delta t^h_n = \frac{h^2}{\sigma^2(\xi^h_n)}. \tag{5.3}$$

Admissible controls. Let $\mathcal{F}^h_n$ denote the minimal σ-algebra that measures the control and state data to step n, and let E^h_n denote the expectation conditioned on $\mathcal{F}^h_n$. An admissible control for player i at step n is a U_i-valued random variable that is $\mathcal{F}^h_n$-measurable. Let $\mathcal{U}^h_i$ denote the set of the admissible control processes for player i.

A relaxed control for the chain can be defined as follows. Let $r^h_{i,n}(\cdot)$ be a probability distribution on the Borel sets of U_i such that $r^h_{i,n}(A)$ is $\mathcal{F}^h_n$-measurable for each

[7] Whether the chain is Markovian or not depends on the form of the control that is applied. But the transition probability will always be locally consistent.

[8] This representation of the chain is not needed and was not used for the single player problem or for the zero-sum game. For the nonzero-sum game, it provides a way of dealing with the difficulties in the convergence proof that were described in the introduction.

Borel set $A \in U_i$. Then the $r^h_{i,n}(\cdot)$ are said to be relaxed controls for player i at step n. As for the model (2.3), an ordinary control at step n can be represented by the relaxed control at step n defined by $r^h_{i,n}(A) = I_{\{u^h_{i,n} \in A\}}$ for each Borel set $A \subset U_i$. Define $r^h_n(\cdot)$ by $r^h_n(A_1 \times A_2) = r^h_{1,n}(A_1) r^h_{2,n}(A_2)$, where the A_i are Borel sets in U_i. The associated transition probability is $\int_U p^h(x, y|a) r^h_n(d\alpha)$. If $r^h_{i,n}(A)$ can be written as a measurable function of ξ^h_n for each Borel set A, then the control is said to be relaxed feedback. Under any feedback (or relaxed feedback or randomized feedback) control, the process ξ^h_n is a Markov chain. More general controls, under which there is more "past" dependence and the chain is not Markovian, will be used as well. Let $\mathcal{C}^h_i$ denote the set of control strategies for ξ^h_n.

The cost function. Discretize the costs as follows. The cost functions are the analogs of (2.2) or (2.4). The cost rate for player i is $k_i(x, \alpha_i)\Delta t^h(x, \alpha)$. The stopping costs are $g_i(\cdot)$, and τ^h denotes the first time that the set G^0_h is exited. Let $W^h_i(x, u^h_1, u^h_2)$ denote the expected cost for player i under the control sequences $u^h_i = \{u^h_{i,n}, n \geq 0\}, i = 1, 2$. The numerical problem is to solve the game problem for the approximating chain.

Continuous-time interpolations. The discrete-time chain ξ^h_n is used for the numerical computations. However, for the proofs of convergence, we use a continuous-time interpolation $\xi^h(\cdot)$ of $\{\xi^h_n\}$ that will approximate $x(\cdot)$. This will be a continuous-time process that is constructed as follows. Define $\Delta t^h_n = \Delta t^h(\xi^h_n, u^h_n)$ and $t^h_n = \sum_{i=0}^{n-1} \Delta t^h_i$. Define $\xi^h(t) = \xi^h_n$ on $[t^h_n, t^h_{n+1})$. Define the continuous-time interpolations $u^h_i(\cdot)$ of the control actions for player i by $u^h_i(t) = u^h_{i,n}, t^h_n \leq t < t^h_{n+1}$, and let its (continuous-time) relaxed control representation be denoted by $r^h_i(\cdot)$. Define $r^h(\cdot) = (r^h_1(\cdot), r^h_2(\cdot))$, with time derivative $r^{h,\prime}(\cdot)$. We use $\mathcal{U}^h_i$ for the set of continuous-time interpolations of the control for player i as well.

An alternative interpolation. In [19] an interpolation called $\psi^h(\cdot)$ was used as well, and had some advantages in simplifying the proofs there. We describe it briefly so that the convergence results of [19] can be used where needed. For each h, let $\nu^h_n, n = 0, 1, \ldots,$ be mutually independent and exponentially distributed random variables with unit mean and that are independent of $\{\xi^h_n, u^h_n, n \geq 0\}$. Define $\Delta\tau^h_n = \nu^h_n \Delta t^h_n$, and $\tau^h_n = \sum_{i=0}^{n-1} \Delta\tau^h_i$. Define $\psi^h(t) = \xi^h_n$ and $u^h_\psi(t) = u^h_n$ on $[\tau^h_n, \tau^h_{n+1})$. Now decompose $\psi^h(\cdot)$ in terms of the continuous-time compensator and martingale. Since the intervals between jumps are $\Delta t^h_n \nu^h_n$, where ν^h_n is exponentially distributed and independent of $\mathcal{F}^h_n$, the jump rate of $\psi^h(\cdot)$ when in state x and under control value α is $1/\Delta t^h(x, \alpha)$. Given a jump, the distribution of the next state is given by the $p^h(x, y|\alpha)$, and the conditional mean change is $b^h(x, \alpha)\Delta t^h(x, \alpha)$. So we can write

$$\psi^h(t) = x(0) + \int_0^t b^h(\psi^h(s), u^h_\psi(s))ds + M^h(t), \tag{5.4}$$

where the martingale $M^h(t)$ has quadratic variation process $\int_0^t a^h(\psi^h(s), u^h_\psi(s))ds$. Under any feedback (or randomized feedback) control, the process $\psi^h(\cdot)$ is a continuous-time Markov chain.

It can be shown that [19, sections 5.7.3 and 10.4.1] there is a martingale $\hat{w}^h(\cdot)$ (with respect to the filtration generated by the state and control processes possibly augmented by an "independent" Wiener process) such that

$$M^h(t) = \int_0^t \sigma^h(\psi^h(s), u^h_\psi(s))d\hat{w}^h(s) = \int_0^t \sigma(\psi^h(s))d\hat{w}^h(s) + \epsilon^h(t), \tag{5.5}$$

where $\sigma^h(\cdot)[\sigma^h(\cdot)]' = a^h(\cdot)$ (recall the definition of $a^h(\cdot)$ in (5.1)), and where $\hat{w}^h(\cdot)$ has quadratic variation It and converges weakly to a standard Wiener process. The martingale $\epsilon^h(\cdot)$ is due to the difference between $\sigma(x)$ and $\sigma^h(x)$ (recall the $o(\Delta t^h)$ terms in (5.1)) and

$$\lim_{h\to 0}\sup_{u^h} E\sup_{s\le t}|\epsilon^h(s)|^2 = 0 \tag{5.6}$$

for each t. Thus, where $r^h_\psi(\cdot)$ is the relaxed control representation of $u^h_\psi(\cdot)$,

$$\psi^h(t) = x(0) + \int_0^t\int_U b^h(\psi^h(s),\alpha)r^{h,\prime}_\psi(d\alpha,s)ds + \int_0^t \sigma(\psi^h(s))d\hat{w}^h(s) + \epsilon^h(t). \tag{5.7}$$

The interpolations $\xi^h(\cdot)$ and $\psi^h(\cdot)$ are asymptotically equivalent, as seen in the following theorem, so that any asymptotic results for one are also asymptotic results for the other. We will use $\xi^h(\cdot)$.

THEOREM 5.1. *Assume the local consistency* (5.1). *Then the time scales with intervals Δt^h_n and $\Delta\tau^h_n$ are asymptotically equivalent.*

Proof. Let $f^h(t) = \min\{n : t^h_n \ge t\}$. Write $\Delta\tau^h_n - \Delta t^h_n = (\nu^h_n - 1)\Delta t^h_n$, a martingale difference. By the martingale property we have

$$E\sup_{n<f^h(t)}|t^h_n - \tau^h_n|^2 = E\sup_{n<f^h(t)}\left|\sum_{i=0}^n \Delta t^h_i(v^h_n - 1)\right| \le 4E\sum_{i=0}^{f^h(t)}[\Delta t^h_i]^2 E(v^h_i - 1)^2,$$

which goes to zero as $h\to 0$ by the last line of (5.1). The result is the same if we define $f^h(t) = \min\{n : \tau^h_n \ge t\}$. □

A representation of the approximating chain. By (5.1), we can write

$$\xi^h_{n+1} = \xi^h_n + b^h(\xi^h_n, u^h_n)\Delta t^h_n + \beta^h_n,$$

where β^h_n is a martingale difference with $E^h_n[\beta^h_n][\beta^h_n]' = a^h(\xi^h_n, u^h_n)\Delta t^h_n$. There are martingale differences δw^h_n with conditional (given $\mathcal{F}^h_n$) covariance $\Delta t^h_n I$ such that $\beta^h_n = \sigma^h(\xi^h_n, u^h_n)\delta w^h_n$ [19, section 10.4.1], [14, section 6.6]. Let $w^h(\cdot)$ denote the continuous-time interpolation of $\sum_{i=0}^{n-1}\delta w^h_n$ with intervals Δt^h_n. Then, abusing notation, we can write

$$\begin{aligned}\xi^h(t) &= x(0) + \int_0^t b^h(\xi^h(s), u^h(s))ds + \int_0^t \sigma^h(\xi^h(s))dw^h(s) + \epsilon^h(t),\\ \int_0^t \sigma^h(\xi^h(s), u^h(s))dw^h(s) &= \int_0^t \sigma(\xi^h(s))dw^h(s) + \epsilon^h(t),\end{aligned} \tag{5.8}$$

where $\epsilon^h(\cdot)$ satisfies (5.6) and is due to the $O(\Delta t^h)$ approximation of $a^h(x,\alpha)$ by $\sigma(x)\sigma(x)'$.

Note on convergence. For any subsequence $h\to 0$, there is a further subsequence (also indexed by h for simplicity) such that $(\xi^h(\cdot), r^h_1(\cdot), r^h_2(\cdot),\ w^h(\cdot), \tau^h)$ converges weakly to random processes $(x(\cdot), r_1(\cdot), r_2(\cdot), w(\cdot), \tau)$, where $r_i(\cdot)$ is a relaxed control for player i, $(x(\cdot), r_1(\cdot), r_2(\cdot), w(\cdot), I_{\{\tau^h\le\cdot\}})$ is nonanticipative with respect to the standard vector-valued Wiener process $w(\cdot)$, and, writing $r(\cdot) = (r_1(\cdot), r_2(\cdot))$, the set satisfies

$$x(t) = x(0) + \int_0^t\int_U b(x(s),\alpha)r'(d\alpha,s)ds + \int_0^t \sigma(x(s))dw(s).$$

Also, $W_i^h(x, r_1^h, r_2^h) \to W_i(x, r_1, r_2)$. The proofs of these facts are the same as for the one-player control case in [19, Chapter 10].

On the construction of $\delta w^h(\cdot)$. A special case. Full details for the general method of constructing $w^h(\cdot)$ are in [19, section 10.4.1], [14, section 6.6]. To illustrate the idea, we will consider a very common case that will be needed in Theorems 5.2, 5.3, 5.4, 5.6, and 6.2. Suppose that $\sigma(\cdot) = \sigma$ is a constant. Suppose that the components of x can be partitioned as $x = (x_1, x_2)$, and σ can be partitioned as $\sigma = \left[\begin{smallmatrix} \sigma_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, where the dimension of x_1 is d_1, and σ_1 is a square and invertible matrix of dimension d_1. Partition the $a^h(\cdot)$ in the second line of (5.1) as

$$a^h(x,\alpha) = \begin{bmatrix} a_1^h(x,\alpha) & a_{1,2}^h(x,\alpha) \\ a_{2,1}^h(x,\alpha) & a_2^h(x,\alpha) \end{bmatrix}.$$

As $h \to 0$, $a_1^h(\cdot) \to \sigma_1[\sigma_1]'$ and all other components go to zero, all uniformly in (x, α). Write the analogous partition $w^h(\cdot) = (w_1^h(\cdot), w_2^h(\cdot))$. For any Wiener process $w_2(\cdot)$ that is independent of the other random variables, we can let $w_2^h(\cdot) = w_2(\cdot)$. The only important component of $w^h(\cdot)$ is $w_1^h(\cdot)$, and we can write

$$\begin{aligned} \delta w_{1,n}^h &\equiv w_1^h(t_{n+1}^h) - w_1^h(t_n^h) \\ &= [a_1^h(\xi_n^h, u_n^h)]^{-1/2} \left[\xi_{1,n+1}^h - \xi_{1,n}^h - \int_{t_n^h}^{t_{n+1}^h} \int_U b_1^h(\xi_n^h, \alpha) r^{h,\prime}(s, d\alpha) ds \right] \\ &= [\sigma_1]^{-1} \left[\xi_{1,n+1}^h - \xi_{1,n}^h - \int_{t_n^h}^{t_{n+1}^h} \int_U b_1^h(\xi_n^h, \alpha) r^{h,\prime}(s, d\alpha) ds \right] + \delta\epsilon_n^{1,h}, \end{aligned} \tag{5.9}$$

where $\delta\epsilon_n^{1,h}$ is due to the approximation of $a_1^h(\cdot)$ by $\sigma_1[\sigma_1]'$, and its continuous-time interpolation satisfies (5.6). If an ordinary control is used, then the double integral is just $b_1(\xi_n^h, u_n^h)\Delta t_n^h$.

5.2. First approximations to the chain. Approximation results analogous to those of Theorems 3.1–3.3 can be proved and will be used in the next section. These approximations are of independent interest and should be quite useful for other convergence and approximation analyses for numerical approximations. Theorem 5.2 concerns an approximation to (5.8) that is based on the same $w^h(\cdot)$ process and will be used in Theorem 6.1. The $w^h(\cdot)$ process depends on the control. For the constant σ-case, Theorem 5.3 shows that this control dependence is small and can be factored out, and (uniform in the control) approximations in terms of an independently and identically distributed (i.i.d.) driving sequence are developed. Once this control dependence is factored out, more convenient approximations to the chain can be obtained. This is done in Theorem 5.4, and the results will be used in Theorem 6.2.

Consider the representation (5.8), and for μ, δ, Δ as used in Theorem 3.3 and the $r^h(\cdot) = (r_1^h(\cdot), r_2^h(\cdot))$ in (5.8), define the approximation $u_i^{\mu,\delta,\Delta,h}(\cdot), i = 1, 2$, analogously to what was done above Theorem 3.3. For the process $w^h(\cdot)$ that appears in (5.8) under the original control $r^h(\cdot)$, define the process

$$\xi^{\mu,\delta,\Delta,h}(t) = x(0) + \int_0^t b(\xi^{\mu,\delta,\Delta,h}(s), u^{\mu,\delta,\Delta,h}(s))ds + \int_0^t \sigma(\xi^{\mu,\delta,\Delta,h}(s))dw^h(s). \tag{5.10}$$

Let $r_i^{\mu,\delta,\Delta,h}(\cdot)$ denote the relaxed control representation of $u_i^{\mu,\delta,\Delta,h}(\cdot)$. The process defined by (5.10) is not a Markov chain even if the controls are feedback, since the

$w^h(\cdot)$ is obtained from the process (5.8) under $r^h(\cdot)$ and not under the $r_i^{\mu,\delta,\Delta,h}(\cdot), i = 1,2$. Let $W_i^{\mu,\delta,\Delta,h}(x, r_1^{\mu,\delta,\Delta,h}, r_2^{\mu,\delta,\Delta,h})$ denote the cost for the process (5.10). Define the discrete-time system

(5.11)
$$\tilde{\xi}^{\mu,\delta,\Delta,h}(n\Delta+\Delta) = \tilde{\xi}^{\mu,\delta,\Delta,h}(n\Delta) + \int_{n\Delta}^{n\Delta+\Delta} b(\tilde{\xi}^{\mu,\delta,\Delta,h}(n\Delta), u^{\mu,\delta,\Delta,h}(s))ds + \sigma(\tilde{\xi}^{\mu,\delta,\Delta,h}(n\Delta))[w^h(n\Delta+\Delta) - w^h(n\Delta)],$$

with initial condition $x(0)$ and piecewise-constant continuous-time interpolation denoted by $\tilde{\xi}^{\mu,\delta,\Delta,h}(\cdot)$. Let $\tilde{W}_i^{\mu,\delta,\Delta,h}(x, r_1^{\mu,\delta,\Delta,h}, r_2^{\mu,\delta,\Delta,h})$ denote the associated cost. We have the following analogue of Theorem 3.3.

THEOREM 5.2. *Assume Condition* A2.1. *Given* $(\mu,\delta,\Delta) > 0$, *approximate* $r_i^h(\cdot)$ *as noted above to get* $r_i^{\mu,\delta,\Delta,h}(\cdot)$. *Given* $\epsilon > 0$ *and* $t < \infty$, *there are* $\mu_\epsilon > 0, \delta_\epsilon > 0$, $\Delta_\epsilon > 0$, *and* $\kappa_\epsilon > 0$ *such that, for* $\mu \le \mu_\epsilon, \delta \le \delta_\epsilon, \Delta \le \Delta_\epsilon$ *and* $\delta/\Delta \le \kappa_\epsilon$,

$$\limsup_{h\to 0} \sup_{x, r_1^h, r_2^h} E \sup_{s\le t} \left|\xi^{\mu,\delta,\Delta,h}(s) - \xi^h(s)\right| \le \epsilon, \tag{5.12}$$

and if Condition A2.2 *holds in addition, then*

$$\limsup_{h\to 0} \sup_{x, r_1^h, r_2^h} \left|W_i^{\mu,\delta,\Delta,h}(x, r_1^{\mu,\delta,\Delta,h}, r_2^{\mu,\delta,\Delta,h}) - W_i^h(x, r_1^h, r_2^h)\right| \le \epsilon. \tag{5.13}$$

The expressions (5.12) *and* (5.13) *hold if only one of the controls is approximated and also if* $\xi^{\mu,\delta,\Delta,h}(\cdot)$ *and* $W_i^{\mu,\delta,\Delta,h}(\cdot)$ *are replaced by* $\tilde{\xi}^{\mu,\delta,\Delta,h}(\cdot)$ *and* $\tilde{W}_i^{\mu,\delta,\Delta,h}(\cdot)$, *respectively.*

Comments on the proof. For notational simplicity in the proof, drop the superscripts μ,δ. Define $\delta\xi^{\Delta,h}(t) = \tilde{\xi}^{\Delta,h}(t) - \xi^h(t)$. Then, following the procedure of Theorem 3.1, write

$$\begin{aligned}\delta\xi^{\Delta,h}(t) = &\int_0^t \int_U \left[b(\xi^{\Delta,h}(s),\alpha) - b^h(\xi^h(s),\alpha)\right] r^{h,\prime}(d\alpha, s)ds \\ &+ \int_0^t \left[\sigma(\xi^{\Delta,h}(s)) - \sigma(\xi^h(s))\right] dw^h(s) \\ &+ \int_0^t \int_U b(\xi^{\Delta,h}(s),\alpha)\left[r^{\Delta,h,\prime}(d\alpha, s) - r^{h,\prime}(d\alpha, s)\right] ds + \epsilon_1^h(t).\end{aligned}$$

The $w^h(\cdot), \epsilon_1^h(\cdot)$ are martingales with respect to the filtration induced by the data $(\xi^h(\cdot), r^h(\cdot), w^h(\cdot))$, the martingale $w^h(\cdot)$ has quadratic variation[9] It, and $\epsilon_1^h(\cdot)$ satisfies (5.6). Partition the last integral analogously to what was done in (3.6), with intervals λ. The process $\xi^{\Delta,h}(\cdot)$ satisfies the following version of (3.7):

$$\sup_{\mu,\delta,\Delta} \sup_{r^h} \sup_{l\lambda\le t} E \sup_{s\le\lambda} \left|\xi^{\Delta,h}(l\lambda + s) - \xi^{\Delta,h}(l\lambda)\right|^2 = O(\lambda) + \sup \Delta t^h(x,\alpha).$$

Now, using the martingale property and the Lipschitz condition, one proceeds in the same way that would be used for approximations to (2.3) in Theorem 3.1.

[9] Actually, they are martingales only when evaluated at the time points t_n^h, but the difference is unimportant, since they are constant between such times.

For example, for some constant K (which depends on t), we have the inequality

$$E\sup_{s\le t}\left|\delta\xi^{\Delta,h}(s)\right|^2 \le K\int_0^t E\left|\delta\xi^{\Delta,h}(s)\right|^2 ds + \kappa^{\lambda,h}(t) + O(\lambda)$$
$$+KE\left|\sum_{l=0}^{[t/\lambda]-1}\int_{l\lambda}^{(l+1)\lambda} b(\xi^{\Delta,h}(l\lambda),\alpha)\left[r^{\Delta,h,\prime}(d\alpha,s) - r^{h,\prime}(d\alpha,s)\right]ds\right|^2,$$

where $\kappa^{\lambda,h}(t)$ is due to the $\epsilon_1^h(\cdot)$ and to the use of $\xi^{\Delta,h}(l\lambda)$ in the second line, and $\sup_{s\le t}|\kappa^{\lambda,h}(s)| \to 0$ as $\lambda\to 0, h\to 0$. For each small λ, the last term in the above expression goes to zero uniformly in h as $(\mu,\delta,\Delta)\to 0$, by the method of approximation of the controls. Then (5.12) follows from the resulting inequality and the Bellman–Gronwall lemma. The inequality (5.13) follows from (5.12) and Condition A2.2. □

5.3. Representations and approximations of the chain with control-independent driving noise. The driving noise $w^h(\cdot)$ depends on the path and control. In section 6 it will be useful to have approximations to $\xi^h(\cdot)$ (uniform in the control and initial condition), where the driving noise increments are independent of the path and control. To accomplish this we will need to factor $w^h(\cdot)$ as $w^h(\cdot) = \bar{w}^h(\cdot)+\zeta^h(\cdot)$, where $\bar{w}^h(\cdot)$ does not depend on the control and $\zeta^h(\cdot)$ is "asymptotically negligible." We will work with the model described at the end of subsection 5.1, where $\sigma = \left[\begin{smallmatrix}\sigma_1 & 0\\ 0 & 0\end{smallmatrix}\right]$, the dimension of x_1 is d_1, and σ_1 is a square and invertible matrix of dimension d_1. The approximation and representation results of Theorems 5.3, 5.4, and 5.6 below will hold for such a form. But to simplify the notation and development, we will work with two specific forms, each of which is typical of a large class of models and numerical algorithms. Case 1 below arises when one uses the so-called central-difference approximation to get the transition probabilities. Case 2 arises when one uses a central-difference approximation for the nondegenerate part and a one-sided or "upwind" approximation for the degenerate part [19, Chapter 5]. Both forms are locally consistent. Let $b_i(\cdot)$ denote the ith component of $b(\cdot)$.

Case 1. Suppose that $d_1 = v$, so that σ is invertible. For $a = \sigma\sigma'$, suppose that $a_{i,i} - \sum_{j:j\ne i}|a_{i,j}| \ge 0$. The condition can be weakened if the approximation intervals can depend on the coordinate direction, or if a linear transformation of the state space is used to diagonalize $\sigma\sigma'$ [19, Chapter 5]. Let e_i denote the unit vector in the ith coordinate direction. A central-difference version of the canonical form of the transition probabilities and interpolation interval in [19, equation (3.15), Chapter 5] is[10]

$$p^h(x,x\pm e_ih|\alpha) = \frac{q_{i,i}\pm hb_i(x,\alpha)/2}{Q}, \quad \Delta t^h(x,\alpha) = \Delta t^h = \frac{h^2}{Q},$$
$$p^h(x,x+e_ih+e_jh|\alpha) = p^h(x,x-e_ih-e_jh|\alpha) = \frac{a^+_{i,j}}{2Q},$$
$$p^h(x,x+e_ih-e_jh|\alpha) = p^h(x,x-e_ih+e_jh|\alpha) = \frac{a^-_{i,j}}{2Q}, \tag{5.14}$$
$$Q = \sum_i a_{i,i} - \sum_{i,j:i\ne j}|a_{i,j}|/2, \quad q_{ii} = a_{i,i}/2 - \sum_{j:j\ne i}|a_{i,j}|/2.$$

[10]The form (5.14) and Cases 1 and 2 are selected for specificity in the constructions to follow. Any of the approximations in [19, Chapter 5] could be used, provided that $\sigma(\cdot)$ is constant.

We suppose that $q_{i,i} - h|b_i(x,\alpha)| \geq 0$. A simple computation using (5.14) shows that $b^h(x,\alpha) = b(x,\alpha)$ and $a^h(x,\alpha) = \sigma\sigma' + O(\Delta t^h)$. Also, by (5.14) we can write $\Delta t_n^h = \Delta t^h$. In one dimension, (5.14) reduces to (5.3), where $q_{1,1} = \sigma^2/2$.

Case 2. Suppose that σ can be partitioned as in the last paragraph of subsection 5.1; i.e., $\sigma = \left[\begin{smallmatrix}\sigma_1 & 0\\ 0 & 0\end{smallmatrix}\right]$ where the dimension of x_1 is d_1, and σ_1 is a square and invertible matrix of dimension d_1. The problem concerns the effect of the degenerate part. The following canonical model for such cases is motivated by the general model of [19, Chapter 5]. Define $\bar{b} = \sup_{x,\alpha} \sum_{i=d_1+1}^{v} |b_i(x,\alpha)|$. For this case, redefine $\Delta t^h = \Delta t^h(x,\alpha) = h^2/[Q + h\bar{b}]$. Use the form (5.14) for $i \leq d_1$, with Q replaced by $Q^h = Q + h\bar{b}$. For $i = d_1 + 1, \ldots, v$, use

$$p^h(x, x \pm e_i h|\alpha) = \frac{hb_i^{\pm}(x,\alpha)}{Q^h},$$

and use

$$p^h(x,x|\alpha) = \frac{h\bar{b} - h\sum_{i=d_1+1}^{v} |b_i(x,\alpha)|}{Q^h}.$$

We still have $a^h(x,\alpha) = \sigma\sigma' + O(\Delta t^h)$ and $b^h(x,\alpha) = b(x,\alpha)$. Let E_n^h denote the expectation given all the data up to step n.

THEOREM 5.3. *Use either of the models from Case* 1 *or Case* 2. *Then we can write* $\delta w_n^h = \delta\bar{w}_n^h + \delta\zeta_n^h$, *where the components are martingale differences. The* $\delta\bar{w}_n^h$ *are i.i.d.,* $\{\delta\bar{w}_l^h, l \geq n\}$ *is independent of* $\{\xi_l^h, u_l^h, l \leq n\}$, *and the components have values* $O(h)$. *Also, for either case,* $E_n^h \delta\bar{w}_n^h[\delta\bar{w}_n^h]' = \Delta t^h$, *and* $E_n^h\delta\zeta_n^h[\delta\zeta_n^h]' = O(h\Delta t^h)$, $E_n^h\delta\zeta_n^h[\delta\bar{w}_n^h]' = O(h\Delta t^h)$.

Proof. The proof is a simple construction. The basic approach is to first define δw_n^h as though $b(\cdot) = 0$ and $a^h(x,\alpha) = \sigma\sigma'$. The result will define $\delta\bar{w}_n^h$. Then $\delta\zeta_n^h$ is defined to make up the difference. The facts that the dominant terms in the transition probabilities in (5.14) do not depend on h and that the contributions due to the drift (hence control and state) are proportional to h make this possible. To avoid excessive notation and concentrate on the essential ideas, we start with Case 1 in one dimension. The treatment of the higher-dimensional model follows the same pattern and is illustrated via a two-dimensional case. Then the minor modifications that are required for Case 2 are discussed. The procedure in the general case should be apparent from the three examples.

Case 1, *one dimension.* We can write the double integral term in (5.9) as $b(\xi_n^h, u_n^h)\Delta t^h$, since $b^h(\cdot) = b(\cdot)$. To construct the state transitions, we will use the representation in terms of the random variables χ_n described in the paragraph below (5.2). In one dimension (5.14) is (5.3) and $p^h(x, x \pm h|\alpha) = 0.5 \pm hb(x,\alpha)/[2\sigma^2]$, $\Delta t^h = h^2/\sigma^2$. Now, define $\xi_{n+1}^h - \xi_n^h$ by setting it equal to h if the random sample of χ_n falls in $[0, .5 + hb(\xi_n^h, u_n^h)/2\sigma^2]$, and set it equal to $-h$ otherwise. The "conditional mean" change is $2h[hb(\xi_n^h, u_n^h)/2\sigma^2] = b(\xi_n^h, u_n^h)\Delta t^h$, which is just what is required by the local consistency condition (5.1).

Define the martingale difference term $\delta\bar{w}_n^h$ as follows. Divide $[0,1]$ into the two segments $[0,.5], [.5,1]$. If the random sample of χ_n falls in $[0,.5]$, set $\delta\bar{w}_n^h = h/\sigma$; otherwise set it equal to $-h/\sigma$. It is what δw_n^h would be if $b(\cdot) = 0$ and $a^h(\xi_n^h, u_n^h) = \sigma^2$. Now define $\delta\zeta_n^h$ to make up for the difference. There are two components to $\delta\zeta_n^h$. One component is due to the use of σ in lieu of $[a^h(\xi_n^h, u_n^h)]^{1/2}$ as in the last line of (5.9). Since $a^h(x,\alpha) - \sigma^2 = O(h^2)$, we have $[a^h(x,\alpha)]^{1/2} - \sigma = O(h^2)$, and the corresponding

error in computing the sample values of δw_n^h is $O(h^3)$. The associated interpolated error process clearly satisfies (5.6).

The second component of $\delta\zeta_n^h$ is due to the neglect of the term $b(\cdot)$ in constructing $\delta\bar{w}_n^h$. We handle this as follows. Suppose that $b(\xi_n^h, u_n^h) \geq 0$ (the computation is analogous if $b(\xi_n^h, u_n^h) < 0$). Then, for this second component,[11]

$$\delta\zeta_n^h = (2h - b(\xi_n^h, u_n^h)\Delta t^h)/\sigma \quad \text{if } \chi_n \in [.5, .5 + hb(\xi_n^h, u_n^h)/2\sigma^2],$$

and it equals $-b(\xi_n^h, u_n^h)\Delta t^h/\sigma$ otherwise. The conditional variance of $\delta\zeta_n^h$ is

$$E_n^h \left[2h - b(\xi_n^h, u_n^h)\Delta t^h/\sigma\right]^2 \frac{hb(\xi_n^h, u_n^h)}{2\sigma^2} + E_n^h \left[b(\xi_n^h, u_n^h)\Delta t^h/\sigma\right]^2 \left(1 - \frac{hb(\xi_n^h, u_n^h)}{2\sigma^2}\right) = O(h)\Delta t^h,$$

uniformly in the controls. The $\delta\zeta_n^h$ term depends on the control, but the $\delta\bar{w}_n^h$ term does not. It is simply a Bernoulli sequence, with $\{\delta\bar{w}_l^h, l \geq n\}$ independent of the data up to step n. Also, $E_n^h[\delta\bar{w}_n^h]^2 = \Delta t^h$, $E_n^h \delta\bar{w}_n^h \delta\zeta_n^h = O(h)\Delta t^h$, and $E_n^h[\delta\zeta_n^h]^2 = O(h)\Delta t^h$, uniformly in the controls.

Now, construct the continuous-time martingales $\bar{w}^h(t), \zeta^h(t)$ by interpolating the sums $\sum_{i=0}^{n-1} \delta\bar{w}_i^h$ and $\sum_{i=0}^{n-1} \delta\zeta_i^h$ with intervals Δt^h. Write $w^h(t) = \bar{w}^h(t) + \zeta^h(t)$. The $\bar{w}^h(\cdot)$ does not depend on the control, has quadratic variation It, and $\bar{w}^h(s), s \geq t$, is independent of $\xi^h(s), u^h(s), s \leq t$. The quadratic variation of $\zeta^h(\cdot)$ (and its quadratic covariation with $\bar{w}^h(\cdot)$) is $O(h)$, uniformly in the controls and initial condition.

Comment on the two-dimensional problem for Case 1. The following computation illustrates the procedure in higher dimensions. Now σ is a 2×2 nonsingular matrix. Let $a_{1,2} \geq 0, b_i(\xi_n^h, u_n^h) \geq 0$, for specificity. Divide the unit interval into successive subintervals of lengths $q_{1,1}/Q, q_{1,1}/Q, q_{2,2}/Q, q_{2,2}/Q, a_{1,2}/2Q, a_{1,2}/2Q$. Again, the goal is to reproduce the transition probabilities (5.14). If χ_n falls in $[0, (q_{1,1} + hb_1(\xi_n^h, u_n^h))/Q]$, set $\xi_{1,n+1}^h - \xi_{1,n}^h = h$, and $\xi_{2,n+1}^h - \xi_{2,n}^h = 0$. If χ_n falls in $[(q_{1,1} + hb_1(\xi_n^h, u_n^h))/Q, 2q_{1,1}/Q]$, then set $\xi_{1,n+1}^h - \xi_{1,n}^h = -h$, and $\xi_{2,n+1}^h - \xi_{2,n}^h = 0$. Do the analogous computation for the second component, using the two intervals of length $q_{2,2}/Q$. If χ_n falls in the next to last of the six subintervals, then set $\xi_{n+1}^h - \xi_n^h = (h, h)$, and set it equal to $(-h, -h)$ if χ_n falls in the last of the six subintervals. Define $\delta\bar{w}_n^h$ by repeating the above with $b(x, \alpha) = 0$ and premultiplying by σ^{-1} as follows: For χ_n in the six successive subintervals, define

$$\delta\bar{w}_n^h = \sigma^{-1}\left\{\begin{pmatrix} h \\ 0 \end{pmatrix}, \begin{pmatrix} -h \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ h \end{pmatrix}, \begin{pmatrix} 0 \\ -h \end{pmatrix}, \begin{pmatrix} h \\ h \end{pmatrix}, \begin{pmatrix} -h \\ -h \end{pmatrix}\right\}.$$

The procedure is analogous in any dimension.

Comment on Case 2. For ease of presentation, let us work in two dimensions, where only the first component of $x(\cdot)$ has a Wiener process driving term. Then $\bar{b} = \max_{x,\alpha} |b_2(x, \alpha)|$, $Q = 2q_{1,1} = [\sigma_1]^2$, and $Q^h = Q + h\bar{b}$. Slightly modifying the procedure used for Case 1, divide the unit interval into successive subintervals of lengths

$$\frac{q_{1,1}/2}{Q^h}, \frac{q_{1,1}/2}{Q^h}, \frac{h\bar{b}}{Q^h},$$

[11]The $2h$ in the formula is due to the fact that on the interval $\chi_n \in [.5, .5 + hb(\xi_n^h, u_n^h)/2\sigma^2]$, the difference $\xi_{n+1}^h - \xi_n^h$ was implicitly assigned a value $-h$ when $\delta\bar{w}_n^h$ was constructed, when it should have been assigned the value $+h$.

and divide the last subinterval into two further subintervals of lengths

$$h|b_2(x,\alpha)|/Q^h, \quad h[\bar{b} - |b_2(x,\alpha)|]/Q^h,$$

so that there are now four subintervals. First let us construct $\delta\xi_n^h = \xi_{n+1}^h - \xi_n^h$. Analogously to what was done in the one-dimensional example of Case 1, set $\delta\xi_{2,n}^h = 0$ if $\chi_n \in [0, Q/Q^h]$. If $\chi_n \in (Q/Q^h, 1]$, set $\delta\xi_{1,n}^h = 0$ and $\delta\bar{w}_{1,n}^h = 0$. If the random sample of χ_n falls into the fourth subinterval, set $\delta\xi_{2,n}^h = 0$. If it falls into the third subinterval, set $\delta\xi_{2,n}^h = h\ \text{sign}(b_2(\xi_n^h, u_n^h))$. If $\chi_n \in [0, (q_{11} + hb_1(\xi_n^h, u_n^h))/Q^h]$, set $\delta\xi_{1,n}^h = h$, and set it equal to $-h$ if $\chi_n \in [(q_{11} + hb_1(\xi_n^h, u_n^h))/Q^h, 2q_{11}/Q^h]$.

To complete the construction of $\delta\bar{w}_{1,n}^h$, repeat the procedure with $b(\cdot) = 0$ and divide by σ_1. In particular, $\delta\bar{w}_{1,n}^h = h/\sigma_1$ if $\chi_n \in [0, q_{1,1}/Q^h]$. It is $-h/\sigma_1$ if $\chi_n \in (q_{1,1}/Q^h, 2q_{1,1}/Q^h]$, and it is zero otherwise. The variance is h^2/Q^h, which is the current value of Δt^h. The value of the second component $\delta\bar{w}_{2,n}^h$ is unimportant since it is eventually multiplied by zero. So, let us use an independent Bernoulli sequence with values $\pm h/\sqrt{Q^h}$, each taken with probability 1/2.

These constructions yield (5.14) with $\Delta t^h = h^2/Q^h$. The error terms ζ_n^h for this and the previous example are computed using a procedure that is analogous to that in Case 1. □

In the next theorem, $\sigma(\cdot)$ is just the constant σ. The theorem implies that $\xi^h(\cdot)$ can be written in the form

$$\xi^h(t) = x(0) + \int_0^t \int_U b(\xi^h(s),\alpha) r^{h,\prime}(d\alpha, s)ds + \int_0^t \sigma d\bar{w}^h(s) + \epsilon_2^h(t), \tag{5.15}$$

where $\epsilon_2^h(\cdot)$ equals $\epsilon_1^h(\cdot)$ plus a stochastic integral with respect to $\zeta^h(\cdot)$, and satisfies (5.6). Since the martingale $\bar{w}^h(\cdot)$ does not depend on the control and is essentially the sum of i.i.d. zero mean random variables of size $O(h)$, the form (5.15) can be used to obtain approximation theorems of the type in Theorems 3.1–3.3. The controls can be space and time discretized with arbitrarily small change in the costs, just as in the cited theorems. The quadratic variation process of $\bar{w}^h(\cdot)$ is It.

THEOREM 5.4. *Assume Condition* A2.1 *and the models of Theorem* 5.3. *Define*

$$\bar{\xi}^h(t) = x(0) + \int_0^t \int_U b(\bar{\xi}^h(s),\alpha) r^{h,\prime}(d\alpha, s)ds + \int_0^t \sigma d\bar{w}^h(s). \tag{5.16}$$

Then, for each $t > 0$,

$$\lim_{h\to 0} \sup_{x(0), r^h} E \sup_{s\le t} \left|\xi^h(s) - \bar{\xi}^h(s)\right|^2 = 0. \tag{5.17}$$

If Condition A2.2 *is assumed as well, then the costs for the two processes are arbitrarily close, uniformly in the control and initial condition.*

Now, given $(\mu, \delta, \Delta) > 0$, *let* $u_i^{\mu,\delta,\Delta,h}(\cdot)$ *be the delayed and discretized approximation of* $r_i^h(\cdot)$ *that would be defined by the procedure described above Theorem* 3.3, *with the relaxed control representation of the pair* $(i = 1, 2)$ *of approximations being* $r^{\mu,\delta,\Delta,h}(\cdot)$. *Define the system*

$$\begin{aligned}\bar{\xi}^{\mu,\delta,\Delta,h}(t) = x(0) &+ \int_0^t \int_U b(\bar{\xi}^{\mu,\delta,\Delta,h}(s),\alpha) r^{\mu,\delta,\Delta,h,\prime}(d\alpha, s)ds \\ &+ \int_0^t \sigma d\bar{w}^h(s).\end{aligned} \tag{5.18}$$

Then for $t > 0$ and $\gamma > 0$ there are positive numbers $\mu_\gamma, \delta_\gamma, \Delta_\gamma, h_\gamma, \kappa_\gamma$, such that for $\mu \le \mu_\gamma$, $\delta \le \delta_\gamma$, $\Delta \le \Delta_\gamma$, $h \le h_\gamma$, $\delta/\Delta \le \kappa_\gamma$ we have

$$\sup_{r^h, x(0)} E \sup_{s \le t} \left| \bar{\xi}^{\mu,\delta,\Delta,h}(s) - \bar{\xi}^h(s) \right|^2 \le \gamma. \tag{5.19}$$

If Condition A2.2 *is assumed as well, then for small (μ, δ, Δ, h) the costs are arbitrarily close, uniformly in the control and initial condition.*

Comment on the proof. The proof of the various assertions follows the lines of the arguments used in Theorem 5.2, exploiting the martingale properties and the Lipschitz condition. The details are very similar and are omitted.

The terms $[\bar{w}^h(n\Delta + \Delta) - \bar{w}^h(n\Delta)], n = 0, 1, \ldots$, are i.i.d. and have orthogonal components. The covariance is Δ times the identity matrix, and the processes converge to normally distributed random variables as $h \to 0$. It will be useful to quantify this closeness for use in the next section. This will be done in Theorem 5.6, which requires the following strong approximation theorem for i.i.d. random variables.

LEMMA 5.5 (see [3, Theorem 3]). *Let $\{\phi_n\}$ be a sequence of $\mathbb{R}^d$-valued i.i.d. random variables with zero mean and bounded $(2+\delta)$th moment, where $0 < \delta \le 1$. Suppose that the covariance matrix Γ is nonsingular. Then without changing the distribution, one can redefine the sequence on a richer probability space together with a Wiener process $B(\cdot)$ with covariance matrix Γ such that*

$$\left| \sum_{i \le n} \phi_i - B(n) \right| = o(n^{0.5-c}) \tag{5.20}$$

*w.p.*1 *for large n, for some $0 < c < 0.5$.*

The following theorem asserts that if $\sigma(\cdot)$ is constant, then the process defined by (5.18) can be written essentially as the discrete-time system (2.5), which we now write as $x^{\mu,\delta,\Delta,h}(\cdot)$, since the discretized controls are used. Keep in mind that the controls in (5.18) are obtained from the discretization of the relaxed control representation of the interpolation of $\{u_n^h\}$, the original controls for the chain.

THEOREM 5.6. *Assume Conditions* A2.1 *and* A2.2 *and the models used in Theorem* 5.3. *Then we can define the probability space such that $\bar{w}^h(t) = w(t) + \rho^h(t)$, where $w(\cdot)$ is a vector-valued Wiener process whose covariance matrix is the identity. For each $t > 0$, $E \sup_{s \le t} |\rho^h(s)|^2 \to 0$ as $h \to 0$. Let $x^{\mu,\delta,\Delta,h}(\cdot)$ be the solution to* (2.5) *with the same Wiener process $w(\cdot)$ and with the controls that are used in* (5.18). *Then, for any $t > 0$,*

$$\lim_{h \to 0} \sup_{r^h, x(0)} E \sup_{s \le t} \left| x^{\mu,\delta,\Delta,h}(s) - \bar{\xi}^{\mu,\delta,\Delta,h}(s) \right|^2 = 0. \tag{5.21}$$

Proof. Since we have assumed that the same controls are used for both systems (2.5) and (5.18), some explanation is needed. Consider Case 1 and define random variables ϕ_n by $\delta \bar{w}_n^h = \phi_n \sqrt{h^2/Q}$. (For Case 2, the development is the same, but with the normalization factor Q^h replacing the normalization factor Q.) This can be done since the parameter h is only a linear scale factor in the construction of the $\delta \bar{w}_n^h$. Then $\{\phi_n\}$ satisfies the conditions of Lemma 5.5, and we can suppose that the probability space is such that (5.20) holds for some Wiener process $B(\cdot)$, whose covariance matrix will be the identity. In fact, instead of starting with $\{\phi_n\}$, we can start with the pair $\{\phi_n, \chi_n\}$, with the same law as used originally. Let us do this,

but consider only the approximation of the sums of the ϕ_n by $B(\cdot)$ as in Lemma 5.5. Then, on this probability space define $\delta\bar{w}_n^h$ in terms of the ϕ_n, as above. Without loss of generality, we can suppose that ξ_0^h is also defined on this space. The next step is to define u_0^h, which is just a function of ξ_0^h, on this space. For convenience we use the same notation for the controls and states as on the original space. Continuing, define ξ_1^h on the space by constructing it from ξ_0^h, u_0^h, χ_0^h as done in Theorem 5.3. This procedure can be continued so that all of $\{\xi_n^h, u_n^h, \delta\bar{w}_n^h\}$ is defined on the space, and with the same law as used originally. Now, given the control sequence $\{u_n^h\}$, the continuous-time interpolation can be time and space discretized and delayed as in Theorem 5.4.[12]

From Lemma 5.5, we have, w.p.1 for large n,

$$\left|\sum_{i=0}^{n} h\phi_i - hB(n)\right| = h\,o(n^{0.5-c}). \tag{5.22}$$

The process $w(\cdot) = hB(\cdot/\Delta t^h)/\sqrt{Q}$ is a Wiener process whose covariance is the identity. By the above arguments and (5.22), there is a constant $c > 0$ and a $t_h \to 0$ as $h \to 0$ such that

$$\left|\sum_{i=0}^{t/\Delta t^h} \delta\bar{w}_i^h - w(t)\right| = o(t[\Delta t^h]^c) \tag{5.23}$$

w.p.1 for $t \geq t_h$ and small h. By the above arguments concerning the approximation of $\bar{w}^h(\cdot)$ by $w(\cdot)$, we can write $\bar{w}^h(t) = w(t) + \rho^h(t)$, where the process $\rho^h(\cdot)$ has independent increments, and $\lim_{h\to 0} E\sup_{s\leq t}|\rho^h(s)|^2 = 0$. Rewrite (5.18) as

$$\bar{\xi}^{\mu,\delta,\Delta,h}(t) = x(0) + \int_0^t \int_U b(\bar{\xi}^{\mu,\delta,\Delta,h}(s),\alpha) r^{\mu,\delta,\Delta,h,\prime}(d\alpha, s)ds + \sigma dw(t) + \sigma\rho^h(t),$$

and write (2.5) with the controls $r^{\mu,\delta,\Delta,h}(\cdot)$ as (in interpolated form)

$$x^{\mu,\delta,\Delta,h}(t) = x(0) + \int_0^t \int_U b(x^{\mu,\delta,\Delta,h}(s),\alpha) r^{\mu,\delta,\Delta,h,\prime}(d\alpha, s)ds + \sigma dw(t). \tag{5.24}$$

From this point the proof is standard, using the Lipschitz condition and the martingale properties. □

6. An approximate equilibrium for the diffusion process is an approximate equilibrium for the chain and vice versa.

Representations of the transition probability and controls. In the next two theorems, we will use the representations of the transitions of the Markov chain in terms of the i.i.d. random variables $\{\chi_n\}$ discussed in the paragraph after (5.2), and the similar representation for the realizations of the rule (4.2) in terms of the random variables $\{\theta_l\}$ noted in the discussion just below the statement of Theorem 4.1. This ensures that the sample path of the approximating chain depends only on the selected control values and that the selected control value in (4.2) depends only on the past values of the control and Wiener process.

[12] Actually, it is only required that the controls be approximated and delayed such that the control applied on $[n\Delta, n\Delta+\Delta)$ is $\mathcal{F}^h_{n\Delta-}$-measurable. The other aspects of the discretization are not needed.

THEOREM 6.1. *Assume Conditions* A2.1, A2.2, *and* A4.1. *An* ϵ*-equil- ibrium value for* (2.1) *or* (2.3) *is an* ϵ_1*-equilibrium value for the approximating Markov chain, where* $\epsilon_1 \to 0$ *as* $\epsilon \to 0$.

Proof. Let $\epsilon > 0$ be given. By Condition A4.1, there is an ϵ-equilibrium strategy pair for (2.3) under which the solution to (2.3) is well defined. By Theorem 4.1, without loss of generality, and for small enough μ, δ, and Δ, it can be represented as in (4.2), where we assume that Δ/δ is an integer, and the $p_{i,k}(\cdot)$ are continuous in the w-variables. We can suppose without loss of generality that for each n, k, and i, the rule (4.2) is defined for all possible conditioning u-sequences with values in $U_i^\mu, i = 1, 2$. Let $(\bar{c}_1^\Delta(\cdot), \bar{c}_2^\Delta(\cdot))$ denote this strategy pair. The strategies $\bar{c}_i^\Delta(\cdot)$ depend on μ and δ as well as on Δ, but for simplicity we suppress that dependence in the notation.

Recall that when a strategy that is defined by a rule such as (4.2) is applied to an arbitrary relaxed control, the formula (4.2) is actually applied to the space-time discretization of that relaxed control, as defined above Theorem 3.3. These strategies $\bar{c}_i^\Delta(\cdot)$ will need to be adapted for use on the chain. To do this, simply replace the $w(\cdot)$-samples in (4.2) by samples of the $w^h(\cdot)$ process that was used in (5.8) and whose construction was illustrated in (5.9). Keep in mind that these strategies are used only for theoretical purposes to prove a convergence theorem. They are not for practical implementation. For each integer k, the control value $u_i^{\mu,\delta,\Delta,h}(k\delta)$ that is obtained from the rule (4.2) with $w^h(\cdot)$ used will be applied to the chain for all steps m such that $t_m^h \in [k\delta, k\delta + \delta)$. The resulting strategies for the chain will be denoted by $\bar{c}_i^{\Delta,h}(\cdot)$ and are in $\mathcal{C}_i^h$.

We want to show that for small enough (μ, Δ, δ), there are $\epsilon_0 > 0$ and $h_0 > 0$, where $\epsilon_0 \to 0$ as $\epsilon \to 0$ such that for $h \le h_0$ and any sequence $r_i^h(\cdot)$ of admissible relaxed (or ordinary) controls for the chain,

$$
\begin{aligned}
W_1^h(x, \bar{c}_1^{\Delta,h}, \bar{c}_2^{\Delta,h}) &\ge W_1^h(x, r_1^h, \bar{c}_2^{\Delta,h}) - \epsilon_0, \\
W_2^h(x, \bar{c}_1^{\Delta,h}, \bar{c}_2^{\Delta,h}) &\ge W_2^h(x, \bar{c}_1^{\Delta,h}, r_2^h) - \epsilon_0.
\end{aligned} \tag{6.1}
$$

The notation $W_2^h(x, \bar{c}_1^{\Delta,h}, r_2^h)$ implies that player 1 uses strategy $\bar{c}_1^{\Delta,h}(\cdot)$ and player 2 uses relaxed control $r_2^h(\cdot)$ (in continuous-time interpolation notation) or an ordinary control with this relaxed control representation, with the analogous interpretation when the indices are reversed. The notation $W_1^h(x, \bar{c}_1^{\Delta,h}, \bar{c}_2^{\Delta,h})$ implies that player i uses strategy $\bar{c}_i^{\Delta,h}(\cdot), i = 1, 2$.

Suppose that the pair $\bar{c}_i^{\Delta,h}(\cdot), i = 1, 2$, is used for the chain. Let $\bar{r}_i^{\mu,\delta,\Delta,h}(\cdot), i = 1, 2$, denote the (continuous-time interpolation notation) relaxed control representation of the control actions. Let $\xi^h(\cdot)$, $w^h(\cdot)$, and τ^h denote the corresponding continuous-time interpolation of the chain, the "pre-Wiener" process, and the first exit time, respectively. The sequence $(\xi^h(\cdot), \bar{r}_1^{\mu,\delta,\Delta,h}(\cdot), \bar{r}_2^{\mu,\delta,\Delta,h}(\cdot), w^h(\cdot), \tau^h)$ (parametrized by h for fixed μ, δ, Δ) is tight. Select a weakly convergent subsequence ($h \to 0$) with the limit denoted by $(x(\cdot), r_1(\cdot), r_2(\cdot), w(\cdot), \tau)$. The set $(x(\cdot), r_1(\cdot), r_2(\cdot), I_{\{\tau \le \cdot\}})$ is nonanticipative with respect to the standard vector-valued Wiener process $w(\cdot)$, and the set $(x(\cdot), r_1(\cdot), r_2(\cdot), w(\cdot))$ solves (2.3). The limit τ is the first hitting time of the boundary of G by the limit process $x(\cdot)$. The details concerning the tightness, characterization of the limit processes, and boundary hitting times, and that the limit processes solve (2.3), are the same as for the control problem in [19, Chapters 10, 11].

Henceforth, when weak convergent sequences are dealt with, we will assume (without loss of generality) when needed for simplicity in the argument that the Skorokhod

representation is used so that all processes are defined on the same probability space and the weak convergence is equivalent to convergence w.p.1 in the appropriate topology [7, Theorem 1.8, Chapter 3].

Under the Skorokhod representation, the rule (4.2) with the $w^h(\cdot)$-samples used converges w.p.1 to the same rule with the $w(\cdot)$-samples used, due to the convergence $w^h(\cdot) \to w(\cdot)$ and the continuity of the probabilities in (4.2) in the w-variables. Because of this, the limits $r_i(\cdot), i = 1, 2$, are just realizations of the original ϵ-equilibrium strategies $\bar{c}_i^{\Delta}(\cdot), i = 1, 2$. Since the solution to (2.1) or (2.3) is unique for each admissible pair (control, Wiener process), we can conclude that the probability law of any limit set $(x(\cdot), r_1(\cdot), r_2(\cdot), w(\cdot))$ is the same, no matter what the selected convergent subsequence. Hence the original set of processes (before the subsequence was taken) converges weakly to this (unique in the sense of probability law) limit set, where the control is determined by the rules $\bar{c}_i^{\Delta}(\cdot), i = 1, 2$.

By the weak convergence,

$$W_1(x, \bar{c}_1^{\Delta}, \bar{c}_2^{\Delta}) \leftarrow W_1^h(x, \bar{c}_1^{\Delta,h}, \bar{c}_2^{\Delta,h}) \leq \max_{r_1 \in \mathcal{U}_1^h} W_1^h(x, r_1, \bar{c}_2^{\Delta,h}) = W_1^h(x, \hat{r}_1^h, \bar{c}_2^{\Delta,h}), \tag{6.2}$$

$$W_2(x, \bar{c}_1^{\Delta}, \bar{c}_2^{\Delta}) \leftarrow W_2^h(x, \bar{c}_1^{\Delta,h}, \bar{c}_2^{\Delta,h}) \leq \max_{r_2 \in \mathcal{U}_2^h} W_2^h(x, \bar{c}_1^{\Delta,h}, r_2) = W_1^h(x, \bar{c}_1^{\Delta,h}, \hat{r}_2^h). \tag{6.3}$$

It can be shown by a weak convergence argument working with the chain for any fixed $h > 0$ that the maximizing controls $\hat{r}_i^h(\cdot)$ exist. But we need only work with control processes that approximate the maximum values arbitrarily well, and we assume that the $\hat{r}_i^h(\cdot)$ are such controls.

It will be shown that

$$\limsup_{h \to 0} W_1^h(x, \hat{r}_1^h, \bar{c}_2^{\Delta,h}) \leq W_1(x, \bar{c}_1^{\Delta}, \bar{c}_2^{\Delta}) + \epsilon + \rho(\mu, \delta, \Delta), \tag{6.4}$$

where $\rho(\mu, \delta, \Delta) \to 0$ as $(\mu, \delta, \Delta) \to 0$, with the analogous result for indices $1, 2$ interchanged. Inequalities (6.2), (6.3), and (6.4) imply that if player 2 uses $\bar{c}_2^{\Delta,h}(\cdot)$, then player 1 cannot do better (asymptotically as $h \to 0$ and modulo $\rho(\mu, \delta, \Delta) + \epsilon$) than by using $\bar{c}_1^{\Delta,h}(\cdot)$, with the analogous result holding for the other player. This last fact implies the theorem since (μ, δ, Δ) can be made as small as desired.

Now (6.4) will be shown. Let $\{u_1^{\mu,\delta,\Delta,h}(l\delta)\}$ denote the values that are obtained from $\hat{r}_1^h(\cdot)$ by the space and time discretization given above Theorem 3.3, and which are used by the rule $\bar{c}_2^{\Delta,h}(\cdot)$. Let $\{u_2^{\mu,\delta,\Delta,h}(l\delta)\}$ denote the control choices for player 2, based on the rule $\bar{c}_2^{\Delta,h}(\cdot)$ and the control of player 1. Let $r_i^{\mu,\delta,\Delta,h}(\cdot)$ denote the (continuous-time) relaxed control representation of $\{u_i^{\mu,\delta,\Delta,h}(l\delta)\}$. The processes $\xi^h(\cdot)$ and $w^h(\cdot)$ now denote the interpolation of the chain and the pre-Wiener process, resp., under the strategy $\bar{c}_2^{\Delta,h}(\cdot)$ and control $\hat{r}_1^h(\cdot)$. This $w^h(\cdot)$ process will be fixed for each h and used in the rest of the proof.

Define the process $\xi^{\mu,\delta,\Delta,h}(\cdot)$ by (5.10), driven by $\{u_i^{\mu,\delta,\Delta,h}(l\delta)\}, i = 1, 2$, and $w^h(\cdot)$. Note that $\{u_2^{\mu,\delta,\Delta,h}(l\delta)\}$ is the response of $\bar{c}_2^{\Delta,h}(\cdot)$ to any control of player 1 with discretization $\{u_1^{\mu,\delta,\Delta,h}(l\delta)\}$. By Theorem 5.2, we have, for small h,

$$\left| W_i^h(x, \hat{r}_1^h, \bar{c}_2^{\Delta,h}) - W_i^{\mu,\delta,\Delta,h}(x, r_1^{\mu,\delta,\Delta,h}, r_2^{\mu,\delta,\Delta,h}) \right| \leq \rho_1(\mu, \delta, \Delta), \tag{6.5}$$

where $\rho_1(\mu, \delta, \Delta)$ can be made arbitrarily small, uniformly in $\hat{r}^h(\cdot)$, as $(\mu, \delta, \Delta, h) \to 0$ and as $W_i^{\mu,\delta,\Delta,h}(\cdot)$ was defined above (5.11). Also,

$$W_i^{\mu,\delta,\Delta,h}(x, r_1^{\mu,\delta,\Delta,h}, r_2^{\mu,\delta,\Delta,h}) = W_i^{\mu,\delta,\Delta,h}(x, r_1^{\mu,\delta,\Delta,h}, \bar{c}_2^{\Delta,h}). \tag{6.6}$$

Let $\tau^{\mu,\delta,\Delta,h}$ denote the first hitting time of the boundary for $\xi^{\mu,\delta,\Delta,h}(\cdot)$.

The set $(\xi^{\mu,\delta,\Delta,h}(\cdot), \hat{r}_1^h(\cdot), r_1^{\mu,\delta,\Delta,h}(\cdot), r_2^{\mu,\delta,\Delta,h}(\cdot), w^h(\cdot), \tau^{\mu,\delta,\Delta,h})$ is tight. Extract a weakly convergent subsequence, and index it by h also. Denote the limit of the weakly convergent subsequence by $(x(\cdot), \hat{r}_1(\cdot), r_1^{\mu,\delta,\Delta}(\cdot),\ r_2^{\mu,\delta,\Delta}(\cdot),\ w(\cdot), \tau)$. Then, as was the case in an earlier part of the proof, $(x(\cdot), \hat{r}_1(\cdot),\ r_1^{\mu,\delta,\Delta}(\cdot),\ r_2^{\mu,\delta,\Delta}(\cdot),\ w(\cdot), I_{\{\tau\le\cdot\}})$ is nonanticipative with respect to the standard Wiener process $w(\cdot)$, the set $(x(\cdot), r_1^{\mu,\delta,\Delta}(\cdot), r_2^{\mu,\delta,\Delta}(\cdot), w(\cdot))$ satisfies (2.3), and τ is the first hitting time of the boundary. The $r_i^{\mu,\delta,\Delta}(\cdot)$ is just the relaxed control that is defined by the weak-sense limit $\{u_i^{\mu,\delta,\Delta}(l\delta)\}$ of $\{u_i^{\mu,\delta,\Delta,h}(l\delta)\}$.

We need to show that the limits $u_2^{\mu,\delta,\Delta}(l\delta)$ are chosen by the conditional probability law that defines $\bar{c}_2^{\Delta}(\cdot)$; i.e., that (along the selected subsequence)

$$\begin{aligned} &p_{2,k}\left(\alpha_2;\ w^h(l\Delta), l\le n; u_j^{\mu,\delta,\Delta,h}(l\delta), j=1,2, l\delta < n\Delta\right) \\ &\quad\to p_{2,k}\left(\alpha_2;\ w(l\Delta), l\le n; u_j^{\mu,\delta,\Delta}(l\delta), j=1,2, l\delta < n\Delta\right) \end{aligned} \tag{6.7}$$

for $k\delta \in [n\Delta, n\Delta+\Delta)$. In (6.7), the $w^h(\cdot)$ can be replaced by its limit $w(\cdot)$ due to the continuity in w. Since there are only a finite number of values for the control, for any $t<\infty$ the limit $\{u_1^{\mu,\delta,\Delta}(l\delta), u_2^{\mu,\delta,\Delta}(l\delta), l\delta\le t\}$ will be achieved after a finite number of steps through the convergent subsequence, w.p.1. This implies (6.7). (We will comment further on this point at the end of the proof.) Thus the policy $\bar{c}_2^{\Delta}(\cdot)$ acting on any relaxed control with discretization $\{u_1^{\mu,\delta,\Delta}(l\delta)\}$ will yield the sequence $\{u_2^{\mu,\delta,\Delta}(l\delta)\}$. Thus,

$$W_1^{\mu,\delta,\Delta,h}(x, r_1^{\mu,\delta,\Delta,h}(\cdot), \bar{c}_2^{\Delta,h}) \to W_1(x, r_1^{\mu,\delta,\Delta}(\cdot), \bar{c}_2^{\Delta}),$$

and by (6.5) and (6.6), mod $\rho_1(\mu,\delta,\Delta)$,

$$W_1^h(x, \hat{r}_1^h, \bar{c}_2^{\Delta,h}) \to W_1(x, r_1^{\mu,\delta,\Delta}, \bar{c}_2^{\Delta}).$$

We can conclude that

$$\begin{aligned} \lim_{h\to 0} W_1^h(x, \hat{r}_1^h, c_2^{\Delta,h}) &\le W_1(x, r_1^{\mu,\delta,\Delta}(\cdot), \bar{c}_2^{\Delta}) + \rho_1(\mu,\delta,\Delta) \\ &\le W_1(x, \bar{c}_1^{\Delta}, \bar{c}_2^{\Delta}) + \rho_1(\mu,\delta,\Delta) + \epsilon, \end{aligned} \tag{6.8}$$

where the ϵ is due to the fact that $(\bar{c}_1^{\Delta}(\cdot), \bar{c}_2^{\Delta}(\cdot))$ is an ϵ-equilibrium. The arbitrariness of the subsequence implies (6.4). The same argument is used when the indices $1,2$ are reversed.

Finally, let us comment on (6.7). Recall that the discretizations given above Theorem 3.3 use fixed (and asymptotically unimportant) values on the initial interval $[0,\Delta)$, so let us use $u_1^{\mu,\delta,\Delta,h}(l\delta) = u_1(l\delta)$, $u_2^{\mu,\delta,\Delta,h}(l\delta) = u_2(l\delta)$ for fixed $u_i(l\delta)$ and $l\delta<\Delta$. For $k\delta\in[\Delta, 2\Delta)$, we have the rule

$$p_{2,k}\left(\alpha_2; w^h(\Delta); u_1(l\delta), u_2(l\delta), l\delta<\Delta\right), \tag{6.9}$$

and the probability of selecting any $\alpha_2\in U_2^{\mu}$ converges as $w^h(\Delta)\to w(\Delta)$. Then the limit in (6.9) must be the law of $u_2^{\mu,\delta,\Delta}(l\delta)$ for $\Delta\le l\delta<2\Delta$. Using the method of selecting the control values in terms of the θ_i that was recalled above the theorem statement, we can assume that the convergence $u_2^{\mu,\delta,\Delta,h}(l\delta)\to u_2^{\mu,\delta,\Delta}(l\delta), \Delta\le l\delta<2\Delta$,

occurs in a finite number of steps w.p.1, as $h \to 0$ through the convergent subsequence, with the rule (6.9) used. Next, on $[2\Delta, 2\Delta + \Delta)$, we have the rule

$$p_{2,k}\left(\alpha_2; w^h(l\Delta), l \le 2; u_i(l\delta), l\delta < \Delta; u_i^{\mu,\delta,\Delta,h}(l\delta), \Delta \le l\delta < 2\Delta, i = 1, 2\right).$$

The $u_1^{\mu,\delta,\Delta,h}(l\delta), \Delta \le l\delta < 2\Delta$, can be assumed to converge in a finite number of steps as well, w.p.1, and hence, as above, so do the selected values of $u_2^{\mu,\delta,\Delta,h}(l\delta), \Delta \le l\delta < \Delta + 2\Delta$. Continuing in this way yields the form (6.7). □

The converse result. If the ϵ-equilibrium value for the chain is unique for arbitrarily small ϵ, then the converse result is true, namely, that ϵ-equilibrium values for the chain are ϵ_1-equilibrium values for (2.3), where $\epsilon_1 \to 0$ as $\epsilon \to 0$, and we are done, since Theorem 6.1 then implies that the ϵ-equilibrium values for the diffusion are also unique for small ϵ, and that the numerical solutions will converge to the desired value. If the ϵ-equilibrium value for the chain is not unique for arbitrarily small ϵ, then we will show that this "converse" assertion is true for the models used in Theorem 5.3. We are not able to show the converse result when $\sigma(\cdot)$ depends on x, so $\sigma(\cdot)$ is constant in the next theorem and Theorems 5.3–5.6 are applied. Condition A4.1 is not needed.

THEOREM 6.2. *Assume Conditions* A2.1 *and* A2.2 *and the models used in Theorem* 5.3*, where $\sigma(\cdot)$ is constant. Then for any $\epsilon > 0$ there is $\epsilon_1 > 0$, which goes to zero as $\epsilon \to 0$ such that an ϵ-equilibrium value for the chain ξ_n^h for small h is an ϵ_1-equilibrium value for* (2.3).

Proof. Theorem 5.4 says that the paths and cost functions for (5.15) (which is $\xi^h(\cdot)$ under an arbitrary control), (5.16) (where the control is as in (5.15) but the driving process is $\bar{w}^h(\cdot)$), and (5.18) (which is (5.16) with discretized controls) are arbitrarily close, uniformly in the controls, for small (μ, δ, Δ, h). Theorem 5.6 gives the same result for (5.18) and $x^{\mu,\delta,\Delta,h}(\cdot)$ given by (5.24), which is (2.5) with discretized controls. Theorem 3.3 implies the same thing for $x^{\mu,\delta,\Delta,h}(\cdot)$ and (2.3). These uniform closeness results imply that ϵ-equilibrium values for the chain for small h are ϵ_1-equilibrium values for the diffusion. □

REFERENCES

[1] M. BARDI, M. FALCONE, AND P. SORAVIA, *Numerical methods for pursuit-evasion games via viscosity solutions*, in Stochastic and Differential Games: Theory and Numerical Methods, M. Bardi, T. E. S. Raghavan, and T. Parthasarathy, eds., Birkhäuser Boston, Boston, MA, 1999, pp. 105–175.

[2] A. BENSOUSSAN AND A. FRIEDMAN, *Nonzero-sum stochastic differential games with stopping times and free boundaries*, Trans. Amer. Math Soc., 231 (1977), pp. 275–327.

[3] I. BERKES AND W. PHILIPP, *Approximation theorems for independent and weakly dependent random vectors*, Ann. Probab., 7 (1979), pp. 29–54.

[4] R. BUCKDAHN, P. CARDALIAGUET, AND C. RAINIER, *Nash equilibrium payoffs for nonzero-sum stochastic differential games*, SIAM J. Control Optim., 43 (2004), pp. 624–642.

[5] E. ALTMAN, O. POURTALLIER, A. HAURIE, AND F. MORESINO, *Approximating Nash equilibria in nonzero-sum games*, Int. Game Theory Rev., 2 (2000), pp. 155–172.

[6] R. J. ELLIOTT AND N. J. KALTON, *Existence of Value in Differential Games*, Mem. AMS 126, Amer. Math. Soc., Providence, RI, 1972.

[7] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.

[8] W. F. FLEMING, *Generalized solutions in optimal stochastic control*, in Differential Games and Control Theory: III, P. T. Liu, E. Roxin, and R. Sternberg, eds., Marcel Dekker, New York, 1977, pp. 147–165.

[9] W. H. FLEMING AND P. E. SOUGANIDIS, *On the existence of value functions for two-player zero-sum differential games*, Indiana Univ. Math. J., 38 (1989), pp. 293–314.

[10] A. Friedman, *Stochastic differential games*, J. Differential Equations, 11 (1972), pp. 79–108.
[11] S. Hamadene, *Points d'équilibre dans les jeux stochastiques de somme non nulle*, C. R. Acad. Sci. Paris Sér. I Math., 318 (1994), pp. 251–256.
[12] A. B. Haurie, J. B. Krawczyk, and M. Roche, *Monitoring cooperative equilibria in a stochastic differential game*, J. Optim. Theory Appl., 81 (1994), pp. 73–95.
[13] A. B. Haurie and F. Moreseno, *Computing equilibria in stochastic games of intergenerational equity*, Int. Game Theory Rev., 8 (2006), pp. 273–293.
[14] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.
[15] H. J. Kushner, *Numerical methods for stochastic control problems in continuous time*, SIAM J. Control Optim., 28 (1990), pp. 999–1048.
[16] H. J. Kushner, *Numerical approximations for stochastic differential games*, SIAM J. Control Optim., 41 (2002), pp. 457–486.
[17] H. J. Kushner, *Numerical approximations for stochastic differential games: The ergodic case*, SIAM J. Control Optim., 42 (2004), pp. 1911–1933.
[18] H. J. Kushner, *Numerical methods for stochastic differential games: The ergodic cost criterion*, in Advances in Dynamic Game Theory: Numerical Methods, Algorithms, and Applications to Ecology and Economics, Annals of the International Society of Dynamic Games, S. Jorgensen, M. Quincampoix, and T. Vincent, eds., Birkhäuser Boston, Boston, MA, 2007, pp. 617–638.
[19] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed., Springer-Verlag, Berlin, New York, 2001.
[20] K. M. Ramachandran, *Stochastic differential games and applications*, in Handbook of Stochastic Analysis and Applications, Marcel Dekker, New York, 2002, pp. 473–532.
[21] D. W. Stroock and S. R. S. Varadhan, *On degenerate elliptic and parabolic operators of second order and their associated diffusions*, Comm. Pure Appl. Math., 25 (1972), pp. 651–713.
[22] M. Tidball, *Undiscounted zero sum differential games with stopping times*, in New Trends in Dynamic Games and Applications, G. J. Olsder, ed., Birkhäuser Boston, Boston, MA, 1995, pp. 305–322.
[23] M. Tidball and R. L. V. González, *Zero-sum differential games with stopping times: Some results and about its numerical resolution*, in Advances in Dynamic Games and Applications, T. Basar and A. Haurie, eds., Birkhäuser Boston, Boston, MA, 1994, pp. 106–124.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 1972–1994

ON A MODEL FOR THE EFFICIENT OPERATION OF A BANK OR INSURANCE COMPANY*

JOSEPH G. CONLON† AND HYEKYUNG MIN†

Abstract. In this paper the authors study a model for the optimal operation of a bank or insurance company which was recently introduced by Peura and Keppo. The model generalizes a previous one of Milne and Robertson by allowing the bank to raise capital as well as to pay out dividends. Optimal operation of the bank is determined by solving an optimal control problem. In this paper it is shown that the solution of the optimal control problem proposed by Peura and Keppo exists for all values of the parameters and is unique.

Key words. stochastic control theory, finance

AMS subject classifications. 91B30, 93E20, 60J60

DOI. 10.1137/060653408

1. Introduction. In this paper we study a model for the optimal operation of a bank or insurance company which was introduced by Peura and Keppo [9]. In this model capital is invested in a risky asset whose evolution is described by Brownian motion with drift. Thus if $X(t)$ is the bank's capital at time t, then

$$dX(t) = \mu dt + \sigma dW(t), \tag{1.1}$$

where $W(t)$ is Brownian motion, $\mu > 0$ is the drift, and $\sigma > 0$ the volatility. For an insurance company model, μ represents the expected premium collection rate minus the expected claims payment rate.

In addition to investing in the asset described by (1.1), the bank also pays dividends to its owners and raises capital from them. Dividend payments can be implemented instantaneously, but capital issuance is associated with a delay of length Δ and a fixed cost K. If capital is ordered at time t, it is actually received at time $t+\Delta$. Since there is a fixed cost K associated with ordering the capital, if an amount s of capital is ordered, then $s-K$ of that goes to increasing the total capital of the bank. Furthermore, in this model the owners decide at time $t+\Delta$ on the actual amount of capital to be raised. Hence, while the decision to *raise* capital is based on information up to time t, the decision on the *amount* to be raised is based on information up to time $t+\Delta$. The fixed cost K is also paid at time $t+\Delta$.

The payment of dividends and the raising of capital is controlled by a policy π. For $t > 0$ let $L^\pi(t)$ be defined by

$$L^\pi(t) = \text{ dividends paid out } - \text{ capital raised up to time } t. \tag{1.2}$$

It is assumed that capital is raised at a set of discrete times $t_1^\pi < t_2^\pi < t_3^\pi < \cdots$, where the number $N^\pi(t)$ of times capital is raised up to time t is given by the formula

$$N^\pi(t) = \sup\{i : t_i^\pi \le t\}. \tag{1.3}$$

*Received by the editors March 1, 2006; accepted for publication (in revised form) April 21, 2007; published electronically November 28, 2007. This research was partially supported by NSF under grants DMS-0138519 and DMS-0500608.

http://www.siam.org/journals/sicon/46-6/65340.html

†University of Michigan, Department of Mathematics, Ann Arbor, MI 48109-1109 (conlon@umich.edu, hyekyung@umich.edu).

It is also assumed that the time between raising capital is at least Δ and that no dividends are paid out during the period Δ before a capital raising time. Thus, assuming $L^\pi(t)$ is a right continuous process, $L^\pi(t)$ is constant in the interval $\left(t_i^\pi - \Delta,\ t_i^\pi\right)$. At time t_i^π the bank decides the amount of capital it needs to raise. If its portfolio has performed particularly well in the previous time period of length Δ, then it may not raise capital but actually pay out dividends at time t_i^π. In all cases it has to pay the capital raising cost K. Since dividend payments are nonnegative the process $L^\pi(t)$ is increasing for $t \notin \{t_i^\pi : i = 1, 2, \ldots\}$.

Let $X^\pi(t)$ be the amount of capital the bank has at time t. Then from (1.1), (1.2), (1.3) the evolution of $X^\pi(t)$ is governed by the equation

$$dX^\pi(t) = \mu dt + \sigma dW(t) - dL^\pi(t) - K dN^\pi(t). \tag{1.4}$$

As is usual in control theory it is assumed that the policy π depends only on information up to the present time. Hence one assumes that the process $L^\pi(t)$ is right continuous and measurable with respect to the σ field generated by $X^\pi(s)$, $s < t$.

The main concern of [9] is to determine a policy π which maximizes the expected payments to the owners of the bank over its lifetime. They therefore define a value function $V^\pi(x)$, $x \geq 0$, by

$$V^\pi(x) = E\left[\int_0^{\tau_\pi} e^{-\rho t} dL^\pi(t) \Big| X^\pi(0) = x\right], \tag{1.5}$$

where $\tau_\pi = \sup\{t > 0 : X^\pi(s) > 0,\ 0 < s < t\}$ is the lifetime of the bank, and $\rho > 0$ is a predetermined discount factor. Assigning a value to ρ is probably the most problematic part of matching the model to actual data. In [9] they describe ρ as representing the wedge between debt and equity finance due to capital market frictions such as taxation and agency costs of equity. With ρ given, the optimal value function is then given by

$$V(x) = \sup_\pi\ V^\pi(x), \tag{1.6}$$

where the supremum is taken over all allowable strategies. In [9] an expression for the function V of (1.6) is obtained and a corresponding strategy to realize it is given. The function V is characterized by 2 parameters $u_{1,K}(\Delta),\ u_{2,K}(\Delta)$ satisfying $0 \leq u_{1,K}(\Delta) < u_{2,K}(\Delta) < \infty$. If the capital the bank holds is less than $u_{1,K}(\Delta)$, then a capital raising event is initiated. If the capital exceeds $u_{2,K}(\Delta)$, then dividends are paid out. If the capital the bank holds lies between $u_{1,K}(\Delta)$ and $u_{2,K}(\Delta)$, then it is fully invested in the risky asset described by (1.1). A notable feature of the model is that $\tau_\pi < \infty$ with probability 1 for an optimal strategy. Thus maximally efficient operation of the bank gives rise to zero capital in finite time.

The model of Peura and Keppo generalizes an earlier model of Milne and Robertson [8] which allows dividend payments but not the raising of capital (see also [5, 6]). The Milne–Robertson model can be recaptured from that of Peura and Keppo by simply taking the cost K of raising capital to be sufficient large. In that case $u_{1,K}(\Delta) = 0$ and $u_{2,K}(\Delta) = u_0$, where u_0 is the Milne–Robertson threshold for the payment of dividends. The value function V of (1.6) now satisfies the equation

$$(A - \rho)V(x) = 0, \quad 0 < x < u_0,\ V(0) = 0,\ V'(u_0) = 1, \tag{1.7}$$

where A is the infinitesimal generator for the process (1.1),

$$A = \frac{1}{2}\sigma^2 \frac{d^2}{dx^2} + \mu\ \frac{d}{dx}, \tag{1.8}$$

and $u_0 > 0$ is chosen so that the function V, when continued by a linear function for $x > u_0$, is C^2. The complete value function is then this linearly extended function which satisfies (1.7) for $0 \le x \le u_0$.

When K and Δ are small enough then $u_{1,K}(\Delta) > 0$ and it is determined from the solution of a free boundary problem which is a zero latent heat limit for the Stefan problem (see [4] for a Stefan problem occurring in finance). The approach in [9] to obtaining the solution to (1.6) is to study the Bellman inequalities [2, 3] corresponding to the optimal control problem. These are given as follows:

$$\begin{aligned} &\text{(a)} \quad V(0) = 0, \\ &\text{(b)} \quad V(x) \ge MV(x), \\ &\text{(c)} \quad (A-\rho)V(x) \le 0, \\ &\text{(d)} \quad V'(x) \ge 1, \\ &\text{(e)} \quad [V(x) - MV(x)][(A-\rho)V(x)][V'(x) - 1] = 0, \end{aligned} \tag{1.9}$$

where the operator M in (b) is defined by

$$Mf(x) = E\left[e^{-\rho\Delta} \sup_{s\ge 0}\Big[f\Big(X(\Delta)+s\Big) - s - K\Big] I_{\tau>\Delta} | X(0) = x\right],\ x > 0. \tag{1.10}$$

Here $X(t)$ is the diffusion (1.1) and τ is the first hitting time at 0. In (1.10) note that the supremum is *inside* the expectation value. The parameter s denotes the amount of capital raised, whose value can be determined at time of receipt, which is time Δ after the decision to raise capital has been made. In view of (d) in (1.9) the supremum over $s \ge 0$ when $f = V$ is attained as $s \to \infty$. Thus one has

$$MV(x) = \lim_{s\to\infty} E\left[e^{-\rho\Delta}\Big[V\Big(X(\Delta)+s\Big) - s - K\Big] I_{\tau>\Delta} | X(0) = x\right],\ x > 0. \tag{1.11}$$

Evidently a finite limit as $s \to \infty$ in (1.11) exists only if $\lim_{x\to\infty}[V(x) - x] = \beta + K$ exists. In that case the function MV is given by the formula

$$MV(x) = \beta[1 - p(x,\Delta)] + h(x,\Delta), \tag{1.12}$$

where

$$p(x,\Delta) = E[I_{\tau<\Delta}], \quad h(x,\Delta) = E[X(\Delta) I_{\tau>\Delta}]. \tag{1.13}$$

The determination of the parameter β in (1.12) plays a key role in the analysis of the Peura–Keppo model. If the cost K of raising capital is so large that the model reduces to the Milne–Robertson model, with solution V given in (1.7), then β has the value $\beta = V(u_0) - u_0 - K = \mu/\rho - u_0 - K$, which is independent of Δ. As K decreases there should be some critical value of K, $K_{\text{crit}}(\Delta)$ say, at which it begins to become optimal for the bank to raise capital in certain circumstances. In section 2 we define a function $\beta(\Delta)$ which is related to $K_{\text{crit}}(\Delta)$ by the equation $\beta(\Delta) = \mu/\rho - u_0 - K_{\text{crit}}(\Delta)$. One might reasonably expect $K_{\text{crit}}(\Delta)$ to be a decreasing function of Δ, and hence $\beta(\Delta)$ to be increasing. We show in section 2 that this in fact turns out to be the case.

Our main theorem is that the Peura–Keppo solution to (1.6) is the unique solution to the system of inequalities (1.9).

THEOREM 1.1. *For all $K, \Delta > 0$ there is a unique C^1 solution $V(x)$ to the system of inequalities* (1.9). *Further, for $\varepsilon > 0$ there is a control policy $\pi = \pi_\varepsilon$ such that if V_ε is defined by* (1.5), *then* $\lim_{\varepsilon\to 0} V_\varepsilon(x) = V(x)$, $x \ge 0$.

In [9] a formula for the value function $V(x)$ is given once one knows the two threshold values $u_{1,K}(\Delta)$, $u_{2,K}(\Delta)$. For $0 < x < u_{1,K}(\Delta)$, $V(x)$ is determined from equality in (1.9b). For $u_{1,K}(\Delta) < x < u_{2,K}(\Delta)$, $V(x)$ is determined from equality in (1.9c). For $x > u_{2,K}(\Delta)$, $V(x)$ is determined from equality in (1.9d). The threshold values $u_{1,K}(\Delta)$, $u_{2,K}(\Delta)$ are then determined by the requirement that the function V so constructed is C^1 at $u_{1,K}(\Delta)$and C^2 at $u_{2,K}(\Delta)$. Thus $u_{1,K}(\Delta)$, $u_{2,K}(\Delta)$ are solutions of a nonlinear system of two equations, and it is not clear if a solution exists for all positive Δ and nonnegative K. In [9] it is shown that a solution of these equations exists for certain ranges of the parameters Δ, K and that the constructed value function $V(x)$ satisfies most of the conditions of (1.9). There does not, however, seem to be a proof that (1.9c) holds for $0 < x < u_{1,K}(\Delta)$. In section 2 we construct the solution to (1.9) for all positive Δ and nonnegative K. The main mathematical fact needed is that a solution $u(x,t)$, $t > 0$, of the diffusion equation such that the initial data $u(x,0)$ has just one change of sign has at most one sign change for all $t > 0$. This property of the diffusion equation has been used previously [1, 7], and a proof based on probability has been given. In the appendix we give a proof using the maximum principle [10].

In [9] a limiting formula for the value function as $\Delta \to 0$ is given. This formula has the interesting property that condition (a) of (1.9) no longer holds. It corresponds to the fact that for $\Delta \to 0$ one allows the capital of the bank to drop to an arbitrarily low value, and then immediately raises capital to avoid default. The final section of the paper is devoted to establishing this formula and studying the asymptotic behavior of the thresholds $u_{1,K}(\Delta)$, and $u_{2,K}(\Delta)$ as $\Delta \to 0$. We also show that if $u_{1,K}(\Delta) > 0$ the optimal function $V(x)$, which is C^∞ for $x \neq u_{1,K}(\Delta)$, $u_{2,K}(\Delta)$ and C^2 at $x = u_{2,K}(\Delta)$, is not C^2 at $u_{1,K}(\Delta)$.

In section 3 we construct the policies π_ε and give the proof of uniqueness of V. The policy π_ε consists of paying out an immediate dividend of ε when the capital x of the bank reaches the upper threshold $u_{2,K}(\Delta)$. If $0 < x < u_{1,K}(\Delta)$, sufficient capital is raised (after time Δ) to bring the capital to the upper threshold $u_{2,K}(\Delta)$. If $x > u_{2,K}(\Delta)$, sufficient dividend is immediately paid out to bring the capital down to $u_{2,K}(\Delta) - \varepsilon$. For $u_{1,K}(\Delta) < x < u_{2,K}(\Delta)$, the bank's capital is fully invested in the risky asset described by (1.1). In the proof of uniqueness we need to use the fact that solutions to $u(x,t) = 0$ are nondegenerate.

2. Solution to system of Bellman equations. In this section we construct a solution to the system of inequalities (1.9). To do this we first consider solutions $V(x)$ to the equation

$$(A - \rho)V(x) = 0, \ x \in \mathbf{R}. \tag{2.1}$$

For any $u_0 > 0$, there is a unique solution V_0 to (2.1) with the initial conditions,

$$V_0(u_0) = \mu/\rho, \ V_0'(u_0) = 1. \tag{2.2}$$

Evidently (2.1), (2.2) imply that $V_0''(u_0) = 0$. Since solutions to (2.1) have just one point of inflection it follows that the function V_0 is concave for $x < u_0$ and convex for $x > u_0$. Evidently by translation there is a unique u_0 such that $V_0(0) = 0$. This unique u_0 is given in [9] by the formula

$$u_0 = \frac{1}{r_1 + r_2} \ \ln\left[\frac{\rho + \mu r_2}{\rho - \mu r_1}\right], \tag{2.3}$$

where r_1, r_2 are the characteristic roots for (2.1),

$$r_1 = \frac{-\mu + \sqrt{\mu^2 + 2\sigma^2\rho}}{\sigma^2},\ r_2 = \frac{\mu + \sqrt{\mu^2 + 2\sigma^2\rho}}{\sigma^2}.$$

Note that since V_0 is concave for $x < u_0$ and $V_0'(u_0) = 1$ it follows that u_0 of (2.3) satisfies the inequality

$$0 < u_0 < \mu/\rho. \tag{2.4}$$

Next for the diffusion process with generator A, started at $x > 0$, let τ_x be the first hitting time at 0. We define for $x > 0$, $t > 0$ the function $p(x,t)$ by

$$p(x,t) = P(\tau_x < t). \tag{2.5}$$

Then $p(x,t)$ is a solution to the equation

$$\frac{\partial p}{\partial t} = Ap, \quad x > 0,\ t > 0, \tag{2.6}$$

with boundary and initial conditions

$$p(0,t) = 1,\ t > 0; \quad p(x,0) = 0,\ x > 0. \tag{2.7}$$

It is also evident from the representation (2.5) that $p(x,t)$ satisfies the inequalities

$$\frac{\partial p}{\partial t} \geq 0,\ \frac{\partial p}{\partial x} \leq 0,\ x > 0,\ t > 0. \tag{2.8}$$

It follows from (2.6), (2.8) that for any fixed $t > 0$ the function $p(x,t)$ is a convex function of $x, x > 0$.

Let $h(x,t),\ x > 0, t > 0$, be a solution of the equation

$$\frac{\partial h}{\partial t} = Ah,\ x > 0,\ t > 0,$$

with boundary and initial conditions

$$h(0,t) = 0,\ t > 0;\ h(x,0) = x,\ x > 0.$$

It is easy to see that h and p are the same functions as those defined by (1.13), and that they are related by the formula

$$h(x,t) = x + \mu t - \mu \int_0^t p(x,s)ds. \tag{2.9}$$

LEMMA 2.1. *For fixed $t > 0$ the function $h(x,t)$ is a concave function of $x,\ x > 0$. It also satisfies the inequalities*

$$\mu/\rho - u_0 + h(x,t) \leq e^{\rho t}[x - u_0 + \mu/\rho],\ x \geq u_0, \tag{2.10}$$

$$h(x,t) \leq e^{\rho t} V_0(x),\ x \geq 0. \tag{2.11}$$

Proof. The concavity of h follows from (2.9) and the convexity of p. Inequality (2.10) follows from (2.9). Inequality (2.11) follows from the maximum principle for the diffusion equation since $V_0(x) \geq x,\ x > 0$. □

PROPOSITION 2.1. *Suppose u_0 is given by* (2.3) *and the cost K of capital issuance satisfies $K \geq \mu/\rho - u_0$. Define $V(x)$ by*

$$V(x) = V_0(x),\ 0 \leq x \leq u_0;\ V(x) = x - u_0 + \mu/\rho,\ x > u_0. \tag{2.12}$$

Then $V(x)$ is a C^2 function and satisfies the system of inequalities (1.9).

Proof. It is easy to see that the function $V(x)$ of (2.12) satisfies (1.9a)–(1.9e). To prove (b) we note from (2.12) that the parameter β in (1.12) is given by the formula $\beta = \mu/\rho - u_0 - K$, and hence the function $MV(x)$ is given by the expression

$$MV(x) = e^{-\rho\Delta}\left\{\left[\frac{\mu}{\rho} - u_0 - K\right]\{1 - p(x,\Delta)\} + h(x,\Delta)\right\}. \tag{2.13}$$

Now (b) follows from Lemma 2.1 on noting (2.4). □

From (2.11) we have that

$$\frac{\partial h}{\partial x}(0,t) \leq e^{\rho t}V_0'(0),\ t > 0.$$

Observe also that by the Hopf maximum principle [10] one has $\partial p/\partial x(0,t) < 0$, $t > 0$. Hence one may define a smooth function $\beta(t)$, $t > 0$, by the formula

$$\beta(t) = \left[\frac{\partial h}{\partial x}(0,t) - e^{\rho t}V_0'(0)\right] \Big/ \frac{\partial p}{\partial x}(0,t). \tag{2.14}$$

LEMMA 2.2. *The function $\beta(t)$ of* (2.14) *is strictly monotonic increasing and satisfies*

$$\lim_{t\to 0}\beta(t) = 0,\ \lim_{t\to\infty}\beta(t) = \infty.$$

Proof. Observe by the Hopf maximum principle that $\beta(t) > 0,\ t > 0$. To prove monotonicity let $T > 0$ and consider the function $u(x,t)$ defined by

$$u(x,t) = \beta(T)\{1 - p(x,t)\} + h(x,t) - e^{\rho t}V_0(x).$$

Then $u(x,t)$ is a solution of the diffusion equation (2.6) with initial and boundary conditions

$$u(x,0) = \beta(T) + x - V_0(x),\ u(0,t) = 0.$$

Evidently $u(x,0)$ is a monotonically decreasing function satisfying

$$\lim_{x\to 0}u(x,0) = \beta(T), \quad \lim_{x\to\infty}u(x,0) = -\infty.$$

In particular $u(x,0)$ has exactly one sign change. It follows therefore by Theorem A.1 of the appendix that $u(x,t)$ has at most one sign change for any fixed $t > 0$.

Consider now the function $u(x,T)$. Suppose first that $u(x,T) \leq 0,\ x \geq 0$. Then by the Hopf maximum principle $u(x,t) < 0,\ x > 0$, for any $t > T$ and $\partial u/\partial x(0,t) < 0$. Thus $\beta(t) > \beta(T)$. Alternatively there is an interval $(0,\alpha)$ for which $u(x,T) > 0,\ x \in (0,\alpha)$. By the Hopf principle one must then have $\partial u/\partial x(0,T) > 0$, which contradicts the definition of $\beta(T)$. We have shown that $\beta(t)$ is strictly monotonic increasing.

To find the limit of $\beta(t)$ as $t \to 0$ we compute the limits of the numerator and denominator of (2.14). For the numerator we clearly have that

$$\lim_{t\to 0}\left[\frac{\partial h}{\partial x}(0,t) - e^{\rho t}V_0'(0)\right] = 1 - V_0'(0) < 0. \tag{2.15}$$

To find the limit of the denominator we use the Green's functions $G(x,y,t)$ for the equation (2.6). Thus by the reflection principle we have that

$$G(x,y,t) = \frac{1}{\sqrt{2\pi\sigma^2\, t}}\Bigg\{ \exp\left[-\frac{(x-y+\mu t)^2}{2\sigma^2 t}\right] - \exp\left[-\frac{2\mu x}{\sigma^2} - \frac{(x+y-\mu t)^2}{2\sigma^2 t}\right]\Bigg\}, \quad x,y>0,$$

is the Dirichlet Green's function. Hence $p(x,t)$ is given by the formula

$$p(x,t) = 1 - \int_0^\infty G(x,y,t)dy. \tag{2.16}$$

It is easy to see from this that

$$\lim_{t\to 0}\ \sqrt{2\pi\sigma^2 t}\ \frac{\partial p}{\partial x}(0,t) = -2\ , \tag{2.17}$$

whence (2.15) and (2.17) imply $\lim_{t\to 0}\ \beta(t) = 0$. Similarly, one can easily see that $\lim_{t\to\infty}\ \beta(t) = \infty$. □

We may use the function $\beta(t)$ of (2.14) to improve Proposition 2.1.

PROPOSITION 2.2. *Let $\Delta > 0$ and $\beta(\Delta) > 0$ satisfy the inequality $K \geq \mu/\rho - u_0 - \beta(\Delta)$. Then the function $V(x)$ of* (2.12) *satisfies the system of inequalities* (1.9).

Proof. From Lemma 2.2 the function $MV(x)$ of (2.13) satisfies the inequality $MV(x) \leq V_0(x),\ x \geq 0$. To show that (b) of (1.9) holds, we then need to prove

$$MV(x) \leq x - u_0 + \mu/\rho,\ x > u_0.$$

This follows from Lemma 2.1 since $K \geq 0$. □

Next we consider situations for which K does not satisfy the conditions of Proposition 2.2. To help us understand this we define a function $u_2(\beta,t),\ \beta \geq 0,\ t > 0$, as follows:

$$u_2(\beta,t) = u_0 \text{ if } \beta \leq \beta(t).$$

If $\beta > \beta(t)$, then $u_2(\beta,t)$ is the supremum of all $z \in \mathbf{R}$ such that

$$\beta\{1 - p(x,t)\} + h(x,t) \leq e^{\rho t}\ V_0(x - z + u_0),\ x \geq 0.$$

It is evident that $u_2(\beta,t)$ is a monotonic decreasing function of β.

LEMMA 2.3. *The function $g(\beta) = \beta + u_2(\beta,t),\ \beta > 0$, is strictly monotone increasing.*

Proof. Suppose $\beta > \beta(t)$. Since $V_0'(w) \geq 1,\ w \in \mathbf{R}$, we have that for any $\delta > 0$,

$$e^{\rho t}\ V_0(x - u_2(\beta,t) + \delta + u_0) \geq \delta e^{\rho t} + \beta\{1 - p(x,t)\} + h(x,t),\ x \geq 0.$$

Hence,

$$u_2(\beta + \delta e^{\rho t}, t) \geq u_2(\beta, t) - \delta,$$

whence $g(\beta)$ is strictly monotonic. □

LEMMA 2.4. *For any $\beta > \beta(t)$ there exists a unique point $u_1(\beta, t) > 0$ such that*

$$\beta\{1 - p(x,t)\} + h(x,t) < e^{\rho t}\, V_0(x - u_2(\beta,t) + u_0),\ x \geq 0,\ x \neq u_1(\beta,t),$$

$$\beta\{1 - p(x,t)\} + h(x,t) = e^{\rho t}\, V_0(x - u_2(\beta,t) + u_0),\ x = u_1(\beta,t).$$

Proof. Suppose for a given $T > 0$ we have $\beta > \beta(T)$ and define $u(x,t),\ t > 0, x > 0$, by

$$u(x,t) = \beta\{1 - p(x,t)\} + h(x,t) - e^{\rho t}\, V_0(x - u_2(\beta,T) + u_0).$$

Then $u(x,t)$ is a solution of the diffusion equation (2.6) with initial and boundary conditions

$$u(x,0) = \beta + x - V_0(x - u_2(\beta,T) + u_0),\ u(0,t) = -e^{\rho t}\, V_0(u_0 - u_2(\beta,T)).$$

Evidently $u(x,0)$ is a monotonic decreasing function and $\lim_{x\to\infty} u(x,0) = -\infty$. Hence we must have

$$\beta > V_0(u_0 - u_2(\beta,T)),$$

since otherwise $u(x,T) < 0,\ x \geq 0$, which would contradict the definition of $u_2(\beta,T)$.

We show that for small t the set $\{x > 0 : u(x,t) > 0\}$ is an open interval $\big(a(t), b(t)\big)$ with $0 < a(t) < b(t) < \infty$. To see this, first note from (2.9), (2.17) that since $p(x,t)$ is a convex decreasing function there is the inequality

$$1 \leq \frac{\partial h}{\partial x} \leq 1 + 4\mu\sqrt{t}\,/\,\sqrt{2\pi\sigma^2},\ x > 0. \tag{2.18}$$

We can also see from (2.16) that for any $\delta,\ 0 < \delta < 1$, there exist positive constants $C(\delta)$, $K(\delta)$ such that for $0 < t < 1$,

$$p(x,t) < \delta, \quad x > C(\delta)\,\sqrt{t}, \tag{2.19}$$

$$\frac{\partial p}{\partial x}(x,t) < -\frac{K(\delta)}{\sqrt{t}},\ \ 0 < x < C(\delta)\,\sqrt{t}.$$

Choose now $\delta > 0$ such that

$$\beta(1 - 3\delta) > V_0(u_0 - u_2(\beta,T)).$$

Then from (2.18), (2.19) there exists $\varepsilon > 0$ such that for $0 < t < \varepsilon$ the function $u(x,t)$ is increasing for $0 < x < C(\delta)\,\sqrt{t}$ and $u\big(C(\delta)\sqrt{t},\ t\big) > \beta\delta$.

Next we find a region where $u(x,t)$ is decreasing. To see this observe that we may choose $\varepsilon > 0$ sufficiently small so that for $0 < t < \varepsilon$, there is the inequality

$$\frac{\partial p}{\partial x}(x,t) > -\exp\left[-1/t^{1/6}\right],\ x > t^{1/3}.$$

It follows then from (2.9) that $u(x,t)$ is decreasing for $x > t^{1/3}$, provided $0 < t < \varepsilon$. One also has that

$$\inf\left\{u(x,t) - u(C(\delta)\sqrt{t},\ t) : C(\delta)\sqrt{t} < x < t^{1/3}\right\} \geq -C_1\, t^{1/3}$$

for some constant C_1 which depends on δ. If we choose now ε small enough so that $C_1\varepsilon^{1/3} < \beta\delta$, then it follows that $\{x : u(x,t) > 0\}$ consists of an open interval, provided $0 < t < \varepsilon$.

We now invoke Theorem A.2 of the appendix. By the definition of $u_2(\beta, T)$ one has $u(x,T) \leq 0,\ x \geq 0$, and there is a unique $x = u_1(\beta, T)$ for which $u(x,T) = 0$. □

PROPOSITION 2.3. *Let $\Delta > 0$ and $\beta(\Delta) > 0$ satisfy the inequality $K < \mu/\rho - u_0 - \beta(\Delta)$. Then there is a unique solution $\beta > \beta(\Delta)$ to the equation*

$$\beta + u_2(\beta, \Delta) = \mu/\rho - K. \tag{2.20}$$

For this value of β put $u_2 = u_2(\beta,\Delta)$ and $u_1 = u_1(\beta,\Delta)$. Define the function $V(x)$, $x \geq 0$, by

$$\begin{aligned} &V(x) = x - u_2 + \mu/\rho, \quad x > u_2, \\ &V(x) = V_0(x + u_0 - u_2), \quad u_1 < x \leq u_2, \\ &V(x) = e^{-\rho\Delta}\left\{\beta[1 - p(x,\Delta)] + h(x,\Delta)\right\}, \quad 0 \leq x \leq u_1. \end{aligned} \tag{2.21}$$

Then the function $V(x)$ satisfies the system of inequalities (1.9).

Proof. Since $\beta(\Delta) + u_2\big(\beta(\Delta),\Delta\big) = \beta(\Delta) + u_0 < \mu/\rho - K$ it follows from Lemma 2.3 that there is a unique $\beta > \beta(\Delta)$ satisfying (2.20). We can also see that since $K \geq 0$ there is the inequality $u_1 < u_2$. In fact one has

$$\beta[1 - p(x,\Delta)] + h(x,\Delta) < e^{\rho\Delta}V_0(x + u_0 - u_2),\ x \geq u_2. \tag{2.22}$$

This follows from (2.9) since the left-hand side of (2.22) is strictly less than

$$\mu/\rho - K - u_2 + x + \mu\Delta < e^{\rho\Delta}[\mu/\rho + x - u_2] \leq e^{\rho\Delta}V_0(x + u_0 - u_2),\ x \geq u_2,$$

provided $K \geq 0$.

It is clear now that the function V of (2.21) is a C^1 function and C^2 except possibly at the point $x = u_1$. It is also concave with slope 1 for $x \geq u_2$. Hence (a), (d) of (1.9) hold. Next we prove (b). In view of the concavity of V we have that the function MV of (1.10) is given by the expression

$$MV(x) = e^{-\rho\Delta}\left\{\left[\frac{\mu}{\rho} - u_2 - K\right]\left\{1 - p(x,\Delta)\right\} + h(x,\Delta)\right\}. \tag{2.23}$$

Thus $V(x) = MV(x),\ 0 \leq x \leq u_1$, and by the definition of u_1 one has $MV(x) \leq V(x), u_1 \leq x \leq u_2$. For $x \geq u_2$ we also have $MV(x) \leq V(x)$ by the same argument used to show (2.22). We have proved (b) and also (e).

We are left to prove (c). It is easy to see that $(A - \rho)V(x) \leq 0,\ x > u_1$. We consider then the case $0 < x < u_1$. To see this we observe that the function $u_1(\beta',\Delta),\ \beta' > \beta(\Delta)$, is continuous and satisfies $\lim_{\beta' \to \beta(\Delta)} u_1(\beta',\Delta) = 0$. Hence if $0 < x < u_1$, there exists $\beta',\ 0 < \beta' < \beta$ such that $x = u_1(\beta',\Delta)$. Let $w(z)$ be the function

$$w(z) = \beta'\{1 - p(z,\Delta)\} + h(z,\Delta),\ z > 0.$$

By Lemma 2.4 it follows that $(A-\rho)w(z) \le 0$ at $z = u_1(\beta', \Delta) = x$. Since

$$e^{\rho\Delta}\, MV(z) = (\beta - \beta')\{1 - p(z,\Delta)\} + w(z),$$

and $Ap(z,\Delta) = \partial p/\partial t(z,\Delta) \ge 0$, one has therefore that $(A-\rho)V(x) \le 0$. □

3. Uniqueness of the solution. Here we show that the solution to the system of inequalities (1.9) is unique, provided we make some smoothness assumptions on the function $V(x)$. Our first goal is to show that a limiting set of strategies realizes the function $V(x)$ constructed in Propositions 2.1–2.3. In the following we shall use the convention that if $F(t),\ t > 0$, is a right continuous function of time, then at time τ, $F(\tau^+)$ denotes the limit of $F(t)$ as t converges to τ from above.

We first consider the situation in Propositions 2.1 and 2.2. Let ε satisfy $0 < \varepsilon < u_0$. We define a strategy π_ε for the control process (1.4). Suppose the process begins at x with $0 < x < u_0$. For those paths which exit the interval $[0, u_0]$ through u_0 let τ_1 be the exit time. We set $L(t) = 0,\ t \le \tau_1$, and $L(\tau_1^+) = \varepsilon$. Thus $X(\tau_1^+) = u_0 - \varepsilon$. If the process begins at x with $x \ge u_0$, we set $\tau_1 = 0, L(\tau_1^+) = \varepsilon + x - u_0$, whence again $X(\tau_1^+) = u_0 - \varepsilon$. Next we define $\tau_2 > \tau_1$ as the first time the diffusion process with $X(\tau_1^+) = u_0 - \varepsilon$ hits u_0 for paths which exit the interval $[0, u_0]$ through u_0. We put $L(t) - L(\tau_1^+) = 0$, $\tau_1 < t < \tau_2$, $L(\tau_2^+) - L(\tau_1^+) = \varepsilon$. Thus $X(\tau_2^+) = u_0 - \varepsilon$. We proceed in this manner defining a sequence of stopping times $\tau_1, \tau_2, \ldots$ until the diffusion exits $[0, u_0]$ through 0.

LEMMA 3.1. *Let V_ε be the return function* (1.5) *for the strategy $\pi = \pi_\varepsilon$. Then $\lim_{\varepsilon\to 0} V_\varepsilon(x) = V(x)$, where $V(x)$ is given by* (2.12).

Proof. Evidently we have that

$$V_\varepsilon(x) = x - u_0 + \varepsilon + V_\varepsilon(u_0 - \varepsilon), \quad x \ge u_0. \tag{3.1}$$

For the diffusion process started at $x,\ 0 < x < u_0$, let τ_x be the first exit time from the interval $[0, u_0]$. Then we also have that

$$\begin{aligned} V_\varepsilon(x) &= V_\varepsilon(u_0) E\left[\exp(-\rho\tau_x); X(\tau_x) = u_0\right] \\ &= V_\varepsilon(u_0) w(x), \quad 0 < x < u_0, \end{aligned} \tag{3.2}$$

where the function $w(x)$ satisfies

$$(A-\rho)w(x) = 0,\ 0 < x < u_0, \quad w(0) = 0,\ w(u_0) = 1.$$

It follows that $w(x) = \rho\ V_0(x)/\mu$. Letting $x = u_0 - \varepsilon$ in (3.2) and using (3.1), we conclude that $V_\varepsilon(u_0)$ is given by the formula

$$V_\varepsilon(u_0) = \varepsilon \,/\, [1 - w(u_0 - \varepsilon)].$$

Hence $V_\varepsilon(x),\ 0 < x < u_0$, is given by the formula

$$V_\varepsilon(x) = \left[\frac{V_0(u_0) - V_0(u_0 - \varepsilon)}{\varepsilon}\right]^{-1} V_0(x),\ 0 < x < u_0.$$

Since $V_0'(u_0) = 1$ the result follows. □

Next we consider the situation in Proposition 2.3. We define a strategy π_ε for the control process (1.4) whose limiting return function as $\varepsilon \to 0$ yields the function (2.21). If the process begins at x with $x \ge u_2$, we set $\tau_1 = 0,\ L(\tau_1^+) = \varepsilon + x - u_2$, whence

$X(\tau_1^+) = u_2 - \varepsilon$. We require that $0 < \varepsilon < u_2 - u_1$, whence $u_1 < X(\tau_1^+) < u_2$. If the process begins at x with $u_1 < x < u_2$, we set τ_1 to be the first exit time of the diffusion process from the interval $[u_1, u_2]$. If $X(\tau_1) = u_2$, then we put $L(t) = 0,\ t \le \tau_1$, and $L(\tau_1^+) = \varepsilon$, whence $X(\tau_1^+) = u_2 - \varepsilon$. Suppose now $X(\tau_1) = u_1$. We restrict ourselves to all paths of the diffusion process $X(t),\ \tau_1 \le t \le \tau_1 + \Delta$, which satisfy $X(t) > 0$. For these paths we set $L(t) = 0,\ t \le \tau_1 + \Delta$,

$$L((\tau_1 + \Delta)^+) = X(\tau_1 + \Delta) - u_2 - K. \tag{3.3}$$

Note that if Δ is small, the expression in (3.3) is negative. We finally put $X((\tau_1 + \Delta)^+) = u_2$. For $0 < x \le u_1$ we restrict ourselves to all paths of the diffusion process $X(t), t \le \Delta$, which satisfy $X(t) > 0$. For these paths we set $L(t) = 0,\ t \le \Delta$,

$$L(\Delta^+) = X(\Delta) - u_2 - K. \tag{3.4}$$

Finally we put $X(\Delta^+) = u_2$. The process $L(t), t > \tau_1$, is defined similarly.

LEMMA 3.2. *Let V_ε be the return function* (1.5) *for the strategy $\pi = \pi_\varepsilon$. Then* $\lim_{\varepsilon \to 0} V_\varepsilon(x) = V(x)$, *where $V(x)$ is given by* (2.21).

Proof. Arguing as in Lemma 3.1, we have that

$$V_\varepsilon(x) = x - u_2 + \varepsilon + V_\varepsilon(u_2 - \varepsilon),\ x \ge u_2, \tag{3.5}$$

$$V_\varepsilon(x) = V_\varepsilon(u_2) w_2(x) + V_\varepsilon(u_1) w_1(x),\ u_1 < x < u_2, \tag{3.6}$$

where

$$w_2(x) = E\left[\exp(-\rho\tau_x);\ X(\tau_x) = u_2\right],$$

$$w_1(x) = E\left[\exp(-\rho\tau_x);\ X(\tau_x) = u_1\right],$$

and τ_x is the exit time from the interval $[u_1, u_2]$ for the diffusion process started at x. For $0 < x \le u_1$ we have in addition the identity

$$V_\varepsilon(x) = e^{-\rho\Delta} E\left[X(\Delta) - u_2 - K\ ;\ \tau_x > \Delta\right] + e^{-\rho\Delta} V_\varepsilon(u_2) P(\tau_x > \Delta), \tag{3.7}$$

where τ_x is the first time the diffusion started at x hits 0. From (2.21) we can rewrite (3.7) as

$$V_\varepsilon(x) = V(x) + e^{-\rho\Delta}\left\{V_\varepsilon(u_2) - V(u_2)\right\}[1 - p(x, \Delta)]. \tag{3.8}$$

It is clear that for $u_1 < x < u_2$ the function $V(x)$ may be written as

$$V(x) = V(u_2) w_2(x) + V(u_1) w_1(x),\quad u_1 < x < u_2.$$

Hence if we put $g_\varepsilon(x) = V_\varepsilon(x) - V(x)$, we have from (3.6) that

$$g_\varepsilon(x) = g_\varepsilon(u_2) w_2(x) + g_\varepsilon(u_1) w_1(x),\ u_1 < x < u_2. \tag{3.9}$$

Setting $x = u_1$ in (3.8) we also have that

$$g_\varepsilon(u_1) = e^{-\rho\Delta} g_\varepsilon(u_2)[1 - p(u_1, \Delta)]. \tag{3.10}$$

Putting $x = u_2$ in (3.5) yields the identity

$$g_\varepsilon(u_2 - \varepsilon) = g_\varepsilon(u_2) + [V(u_2) - V(u_2 - \varepsilon) - \varepsilon]. \tag{3.11}$$

If we set $x = u_2 - \varepsilon$ in (3.9) and use (3.10), (3.11), we obtain a formula for $g_\varepsilon(u_2)$,

$$g_\varepsilon(u_2) = -[V(u_2) - V(u_2 - \varepsilon) - \varepsilon] \big/ \left\{1 - w_2(u_2 - \varepsilon) - e^{-\rho\Delta}[1 - p(u_1, \Delta)]w_1(u_2 - \varepsilon)\right\}. \tag{3.12}$$

Since V is concave at u_2 the numerator of (3.12) is negative. The denominator is positive since $w_1(x) + w_2(x) < 1$, $u_1 < x < u_2$. Hence $g_\varepsilon(u_2) < 0$. It follows now from (3.5), (3.8), (3.9) that $V_\varepsilon(x) < V(x)$, $x > 0$. Since the numerator of (3.12) is $O(\varepsilon^2)$ and the denominator is from Hopf's maximum principle bounded below by a positive constant times ε, it follows that $\lim_{\varepsilon\to 0} g_\varepsilon(u_2) = 0$. Hence $\lim_{\varepsilon\to 0} V_\varepsilon(x) = V(x)$, $x \geq 0$. □

We have shown that certain limiting strategies yield the return functions given in Propositions 2.1–2.3. Next let $V(x)$ be a C^1 solution of the system of inequalities (1.9). Since $V'(x) \geq 1$, $x > 0$, the limit

$$\beta + K = \lim_{x\to\infty} [V(x) - x] \text{ exists.} \tag{3.13}$$

This limit must be finite. Otherwise the function $MV(x)$ cannot be finite. Hence from (1.10) we have

$$MV(x) = e^{-\rho\Delta}\left\{\beta[1 - p(x, \Delta)] + h(x, \Delta)\right\}. \tag{3.14}$$

LEMMA 3.3. *There exists u_2, ε with $0 < \varepsilon < u_2$ such that $V(x) = \beta + K + x$ for $x \geq u_2$, and for $u_2 - \varepsilon < x \leq u_2$, $V(x)$ is the solution to the initial value problem*

$$(A - \rho)V(x) = 0, \quad V(u_2) = \mu/\rho, \quad V'(u_2) = 1. \tag{3.15}$$

Further, β and u_2 are related by the identity

$$\beta = \mu/\rho - K - u_2. \tag{3.16}$$

Proof. Suppose $u_2 > 0$ is a point which has the property that for some $\delta > 0$ one has $V'(x) = 1$ for $u_2 \leq x < u_2 + \delta$, and $(A - \rho)V(x) = 0$ for $u_2 - \delta < x < u_2$. Since $(A - \rho)V(x) \leq 0$ for $u_2 \leq x < u_2 + \delta$ it follows that $V(u_2) \geq \mu/\rho$. Using the fact that V is C^1 at u_2 and $(A - \rho)V(x) = 0$, $x < u_2$, we can conclude now that $\lim_{x\to u_2-} V''(x) \geq 0$. If $\lim_{x\to u_2-} V''(x) > 0$, then $V'(x) < 1$ for $x < u_2$ with $u_2 - x$ sufficiently small, in contradiction to (1.9). Hence $\lim_{x\to u_2-} V''(x) = 0$, whence $V(u_2) = \mu/\rho$ since $V'(u_2) = 1$. Thus for $u_2 - \delta < x < u_2$ the function V is the solution to (3.15).

From (2.9), (3.13), (3.14) we see that there exists $u_3 > 0$ such that $MV(x) < V(x)$ for $x \geq u_3$. Hence for each $x > u_3$ the function V must satisfy $(A - \rho)V(x) = 0$ or $V'(x) = 1$. Observe now that (3.13) implies that $\{x : (A - \rho)V(x) = 0\}$ does not include a neighborhood of ∞. Hence by the argument of the previous paragraph $\{x : V'(x) = 1\}$ does include a neighborhood of ∞. We define u_2 by

$$u_2 = \inf\{z : V'(x) = 1, \ x \geq z\}.$$

We show that $MV(u_2) < V(u_2)$. To see this, first note that from (3.13) one has $V(x) = \beta + K + x$, $x > u_2$, and we also have that $V(u_2) \geq \mu/\rho$. Now $V'(x) \geq 1$, $0 < x < u_2$, and $V(0) = 0$. Hence $\beta + K \geq 0$. If $\beta \leq 0$, then (2.9), (3.14) yield the inequality

$$MV(u_2) < e^{-\rho\Delta}(u_2 + \mu\Delta).$$

Since there is also the inequality

$$e^{\rho\Delta}V(u_2) \geq \left[e^{\rho\Delta} - 1\right]\mu/\rho + \beta + K + u_2 \geq \mu\Delta + u_2,$$

it follows that $MV(u_2) < V(u_2)$. If on the other hand $\beta \geq 0$, then

$$MV(u_2) < e^{-\rho\Delta}[\beta + u_2 + \mu\Delta],$$

$$e^{\rho\Delta}V(u_2) \geq \mu\Delta + \beta + K + u_2.$$

Hence again we have $MV(u_2) < V(u_2)$.

The result of the lemma now easily follows since by the previous paragraph there exists $\varepsilon > 0$ such that $(A-\rho)V(x) = 0$ for $u_2 - \varepsilon < x < u_2$. From the first paragraph it follows that V is the solution to (3.15). The identity (3.16) follows from the fact that $V(u_2) = \mu/\rho$. □

Next we define $u_1 < u_2$ by

$$u_1 = \inf\{z > 0 : (A-\rho)V(x) = 0,\ z < x < u_2\}.$$

LEMMA 3.4. *If $u_1 > 0$, then $u_2 < u_0$, $\beta > 0$ and $V(x) = MV(x)$ for $0 < x < u_1$.*

Proof. We proceed as in Lemma 2.4 by considering the function $u(x,t)$ given by

$$u(x,t) = \beta\{1 - p(x,t)\} + h(x,t) - e^{\rho t}V_0(x - u_2 + u_0).$$

Then $u(x,t)$ is a solution of the diffusion equation (2.6) with initial and boundary conditions

$$u(x,0) = \beta + x - V_0(x - u_2 + u_0),\ u(0,t) = -e^{\rho t}V_0(u_0 - u_2).$$

Let us suppose first that $u_2 \geq u_0$. In that case $u(0,t) \geq 0$ and $u(x,0)$ is a monotonic decreasing function with $\lim_{x\to\infty} u(x,0) = -\infty$. It is easy to see from this that for small t the function $u(x,t)$ has at most one sign change. Hence by Theorem A.1, u_1 is the unique solution to the equation $u(x,\Delta) = 0$. From Theorem A.3 it follows that $\partial u/\partial x(x,\Delta) < 0$ at $x = u_1$, but this contradicts the C^1 property of the function $V(x)$ at $x = u_1$. We conclude that $u_2 < u_0$.

Assuming $u_2 < u_0$, then $u(0,t) < 0$. Hence $\beta > V_0(u_0 - u_2) > 0$ since otherwise $u(x,t) < 0,\ x \geq 0,\ t > 0$. By Theorem A.2 the set $\{x > 0 : u(x,\Delta) \geq 0\}$ is a closed interval with u_1 as one of its end points. Evidently u_1 must be the rightmost end point. If the interior of the interval is nonempty, then $\partial u/\partial x(x,\Delta) < 0$ at $x = u_1$ by Theorem A.3 of the appendix. Since this again contradicts the C^1 property of V at u_1 we conclude that $\{x > 0 : u(x,\Delta) \geq 0\} = \{u_1\}$. It is clear now that in the notation of section 2 we have $\beta > \beta(\Delta)$, $u_2 = u_2(\beta,\Delta)$, and $u_1 = u_1(\beta,\Delta)$.

Finally we need to show that $V(x) = MV(x),\ 0 < x < u_1$. Let $u_3 = \inf\{z : 0 < z < u_1,\ V(x) = MV(x) \text{ for } z < x < u_1\}$. Evidently $0 \leq u_3 < u_1$. Suppose now $u_3 > 0$. Since $V(x)$ is concave for $x > u_3$ it follows that $V'(u_3) > 1,\ V(u_3) < \mu/\rho$. Let $u_4 = \inf\{z : 0 < z < u_3,\ (A-\rho)V(x) = 0 \text{ for } z < x < u_3\}$. It is easy to see that $V(x)$ is concave for $u_4 < x < u_2$. Let $w(x) = V(x) - MV(x),\ u_4 < x < u_3$. Then we must have that $w(u_4) = w(u_3) = 0$ and $w(x) \geq 0,\ u_4 < x < u_3$. By the argument of Proposition 2.3 we also have that $(A-\rho)w(x) \geq 0,\ u_4 < x < u_3$. Hence by the maximum principle it follows that $w(x) = 0,\ u_4 < x < u_3$. Since this contradicts the definition of u_3 we conclude that $u_3 = 0$. □

PROPOSITION 3.1. *Let $V(x),\ x \geq 0$, be a C^1 solution to the set of inequalities* (1.9). *Then V is the unique solution given by Propositions* 2.1–2.3.

Proof. This follows from Lemmas 3.3 and 3.4. □

We give an alternative proof of Proposition 3.1 which avoids the use of Theorem A.3. Instead, we shall use the technique of "verification theorem" [2, 3].

LEMMA 3.5. *With u_1, u_2 as defined in Lemma* 3.4 *there is the inequality $u_2 \geq u_2(\beta, \Delta)$, where β is the solution to* (2.20).

Proof. Let π_ε be the strategy defined just before Lemma 3.2, where u_1, u_2 are as in Lemma 3.4. If V_ε is the return function (1.5) corresponding to π_ε, then by the argument of Lemma 3.2 we see that $\lim_{\varepsilon \to 0} V_\varepsilon(x) = V(x)$ for $x \geq u_1$, where $V(x)$ is the function discussed in Lemmas 3.3 and 3.4.

Next let $V_{opt}(x)$ be the solution to the control problem constructed in Propositions 2.1–2.3. We shall show that

$$V_{opt}(x) \geq V_\varepsilon(x), \quad x \geq u_1. \tag{3.17}$$

To see this first let $X(t),\ t > 0$, be the diffusion process with generator A. Assume $X(0) > 0$ and τ is the first hitting time at 0. Since $V_{opt}(x)$ is C^1 for $x \geq 0$ and C^2 for all $x \geq 0$ except possibly $x = u_1(\beta, \Delta)$ with β satisfying (2.20), it follows that

$$e^{-\rho t} V_{opt}\left(X(t \wedge \tau)\right) - \int_0^{t \wedge \tau} e^{-\rho s}(A - \rho) V_{opt}\left(X(s)\right) ds \tag{3.18}$$

is a martingale.

Now let $X(0) = u_2 - \varepsilon$ and let τ_1 be the first exit time from the interval $[u_1, u_2]$. Evidently $\tau_1 < \tau$. Since $(A - \rho)V_{opt}(x) \leq 0,\ x \geq 0$, it follows from (3.18) that

$$V_{opt}(u_2 - \varepsilon) \geq E\left[e^{-\rho \tau_1} V_{opt}\left(X(\tau_1)\right)\right]. \tag{3.19}$$

Since $V'_{opt}(x) \geq 1,\ x \geq 0$, it follows that

$$\begin{aligned} &E\left[e^{-\rho \tau_1} V_{opt}\left(X(\tau_1)\right); X(\tau_1) = u_2\right] \\ &\quad \geq E\left[e^{-\rho \tau_1} V_{opt}(u_2 - \varepsilon); X(\tau_1) = u_2\right] + E\left[e^{-\rho \tau_1} \varepsilon; X(\tau_1) = u_2\right]. \end{aligned} \tag{3.20}$$

On using the fact that $MV_{opt}(x) \leq V_{opt}(x),\ x \geq 0$, we also have that

$$\begin{aligned} &E\left[e^{-\rho \tau_1} V_{opt}\left(X(\tau_1)\right); X(\tau_1) = u_1\right] \\ &\quad \geq E\left[e^{-\rho(\tau_1 + \Delta)} V_{opt}(u_2); X(\tau_1) = u_1,\ \tau_1 + \Delta < \tau\right] \\ &\qquad + E\left[e^{-\rho(\tau_1 + \Delta)}\left[X(\tau_1 + \Delta) - u_2 - K\right]; X(\tau_1) = u_1, \tau_1 + \Delta < \tau\right]. \end{aligned}$$

Hence there is the inequality

$$\begin{aligned} &E\left[e^{-\rho \tau_1} V_{opt}\left(X(\tau_1)\right); X(\tau_1) = u_1\right] \\ &\quad \geq E\left[e^{-\rho(\tau_1 + \Delta)} V_{opt}(u_2 - \varepsilon); X(\tau_1) = u_1,\ \tau_1 + \Delta < \tau\right] \\ &\qquad + E\left[e^{-\rho(\tau_1 + \Delta)} \varepsilon; X(\tau_1) = u_1,\ \tau_1 + \Delta < \tau\right] \\ &\qquad + E\left[e^{-\rho(\tau_1 + \Delta)}\left[X(\tau_1 + \Delta) - u_2 - K\right]; X(\tau_1) = u_1, \tau_1 + \Delta < \tau\right]. \end{aligned} \tag{3.21}$$

If we now define τ_1^* as $\tau_1^* = \tau_1$ if $X(\tau_1) = u_2$, $\tau_1^* = \tau_1 + \Delta$ if $X(\tau_1) = u_1$, we have from (3.19), (3.20), (3.21) the inequality

$$V_{opt}(u_2 - \varepsilon) \geq E\left[e^{-\rho\tau_1^*} V_{opt}(u_2 - \varepsilon); \tau_1^* < \tau\right] + E\left[\int_0^{\tau_1^*} e^{-\rho t} dL(t); \tau_1^* < \tau\right], \tag{3.22}$$

where $L(t)$ is the return function associated with the strategy π_ε. Evidently if we iterate the inequality (3.22), we obtain (3.17) for $x = u_2 - \varepsilon$. The inequality for all $x \geq u_1$ follows in a similar way.

If we let $\varepsilon \to 0$ in (3.17), we obtain the inequality $V_{opt}(x) \geq V(x)$, $x \geq u_1$, which implies the result. □

LEMMA 3.6. *With u_1, u_2 as defined in Lemma* 3.4 *there is the inequality $u_2 \leq u_2(\beta, \Delta)$, where β is the solution to* (2.20).

Proof. Let $\pi_{opt,\varepsilon}$ be the strategy of Lemma 3.2 and $V_{opt,\varepsilon}$ the corresponding return function. Then by Lemma 3.2 we have that $\lim_{\varepsilon \to 0} V_{opt,\varepsilon}(x) = V_{opt}(x)$, where V_{opt} is the function given by (2.21). We shall show that for V, the function discussed in Lemmas 3.3 and 3.4, there is the inequality

$$V(x) \geq V_{opt,\varepsilon}(x), \ x \geq u_1(\beta, \Delta), \tag{3.23}$$

where β is the solution to (2.20). In fact the proof of (3.23) is identical to the proof of (3.17) since V satisfies the variational inequalities (1.9). The result follows by letting $\varepsilon \to 0$. □

Proof of Proposition 3.1. We have shown in Lemmas 3.5 and 3.6 that $u_2 = u_2(\beta, \Delta)$ with β satisfying (2.20). Since MV is given by (3.14) and $MV(u_1) = V(u_1)$ we must have $u_1 = u_1(\beta, \Delta)$. The fact that $V(x) = MV(x)$ for $0 < x < u_1$ follows by the argument at the end of Lemma 3.4. □

4. Properties of the thresholds u_1, u_2. In this section we shall study the properties of u_1, u_2 as defined in Proposition 2.3. Evidently u_1, u_2 are functions of $K \geq 0$ and $\Delta > 0$. If $K \geq \mu/\rho - u_0$, then $u_2 = u_0, u_1 = 0$. We shall therefore be interested in the situation where $0 \leq K < \mu/\rho - u_0$.

LEMMA 4.1. *Suppose $0 \leq K < \mu/\rho - u_0$ and for $\Delta > 0$ let $u_{1,K}(\Delta)$, $u_{2,K}(\Delta)$ be the values of u_1, u_2 determined by K, Δ. Then $u_{1,K}, u_{2,K}$ are continuous functions satisfying*

$$\lim_{\Delta \to 0} u_{1,K}(\Delta) = 0, \quad \lim_{\Delta \to 0} u_{2,K}(\Delta) = u_0 - \hat{u}_K, \tag{4.1}$$

where $z = \hat{u}_K$ is the unique solution to the equation

$$V_0(z) = z + \mu/\rho - K - u_0. \tag{4.2}$$

$u_{1,K}(\Delta) = 0$, $u_{2,K}(\Delta) = u_0$, provided $\Delta \geq \Delta_0$, where $\Delta = \Delta_0$ is the unique solution to the equation

$$\beta(\Delta) = \mu/\rho - K - u_0.$$

Proof. It is easy to see that $u_{1,K}(\Delta)$ and $u_{2,K}(\Delta)$, $\Delta > 0$, are continuous functions. Evidently Proposition 2.2 implies that $u_{1,K}(\Delta) = 0$, $u_{2,K}(\Delta) = u_0$ if

$\Delta \geq \Delta_0$. We consider the case $\Delta \to 0$. Then the function $u_2(\beta, \Delta)$ defined just before Lemma 2.3 satisfies

$$\lim_{\Delta \to 0} u_2(\beta, \Delta) = u_0 - z_\beta,$$

where z_β is the unique solution to the equation $V_0(z_\beta) = \beta$. From Proposition 2.3 it follows therefore that $\hat{u}_K = z_\beta$, where β satisfies $\beta + u_0 - z_\beta = \mu/\rho - K$. This equation is evidently the same as (4.2). Note that $0 < \hat{u}_K \leq u_0$ since $V_0(0) = 0$, $V_0(u_0) = \mu/\rho$ and $V_0'(z) > 1$, $z > 0$. □

Next we obtain the first order behavior of $\beta(\Delta), u_{1,K}(\Delta), u_{2,K}(\Delta)$ as $\Delta \to 0$. The first order behavior of $\beta(\Delta)$ can easily be obtained from (2.14), (2.17), (2.18). Thus we have

$$\lim_{\Delta \to 0} \beta(\Delta)/\sqrt{\Delta} = [V_0'(0) - 1]\sqrt{\pi\sigma^2/2} > 0.$$

To obtain the first order behavior of $u_{1,K}(\Delta), u_{2,k}(\Delta)$ as $\Delta \to 0$ we consider the behavior of the functions $u_1(\beta, \Delta), u_2(\beta, \Delta)$ as $\Delta \to 0$, where $\beta > 0$.

LEMMA 4.2. *Let $u_1(\beta, \Delta), u_2(\beta, \Delta)$ be the functions defined in Lemmas* 2.3 *and* 2.4. *Then if $\beta > 0$ there are the limits*

$$\lim_{\Delta \to 0} u_1(\beta, \Delta)/\sigma\sqrt{\Delta}|\ln \Delta|^{1/2} = 1,$$

$$\lim_{\Delta \to 0} \{u_2(\beta, \Delta) - [u_0 - \hat{u}_\beta]\} \Big/ \sigma\sqrt{\Delta}|\ln \Delta|^{1/2} = 1 - 1/V_0'(\hat{u}_\beta),$$

where $\hat{u}_\beta$ is the unique solution $z = \hat{u}_\beta$ to the equation $V_0(z) = \beta$.

Proof. Just as in the proof of Lemma 4.1 we see that $u_1(\beta, \Delta), u_2(\beta, \Delta)$ satisfy

$$\lim_{\Delta \to 0} u_1(\beta, \Delta) = 0, \lim_{\Delta \to 0} u_2(\beta, \Delta) = u_0 - \hat{u}_\beta. \tag{4.3}$$

For $\Delta > 0$ it follows from Lemma 2.4 that $u_1 = u_1(\beta, \Delta)$ and $u_2 = u_2(\beta, \Delta)$ are the unique positive solutions to the system of equations

$$\beta\{1 - p(u_1, \Delta)\} + h(u_1, \Delta) = e^{\rho\Delta} V_0(u_1 - u_2 + u_0), \tag{4.4}$$

$$-\beta \frac{\partial p}{\partial x}(u_1, \Delta) + \frac{\partial h}{\partial x}(u_1, \Delta) = e^{\rho\Delta} V_0'(u_1 - u_2 + u_0). \tag{4.5}$$

We shall look for solutions to (4.4), (4.5) which satisfy (4.3). To do this we first note from (2.16) that $\partial p/\partial x$ is given by the formula

$$-\frac{\partial p}{\partial x}(x, t) = \frac{2}{\sqrt{2\pi\sigma^2 t}} \left\{ \exp\left[-\frac{(x + \mu t)^2}{2\sigma^2 t}\right] + \frac{\mu}{\sigma^2} \exp\left[-\frac{2\mu x}{\sigma^2}\right] \int_0^\infty \exp\left[-\frac{(x + y - \mu t)^2}{2\sigma^2 t}\right] dy \right\}. \tag{4.6}$$

Let $g(z), z > 0$, be the function

$$g(z) = \sqrt{2\pi\sigma^2\Delta} \left\{ -\beta \frac{\partial p}{\partial x}\left(z\sqrt{\Delta}, \Delta\right) + \frac{\partial h}{\partial x}\left(z\sqrt{\Delta}, \Delta\right) - e^{\rho\Delta} V_0'\left(z\sqrt{\Delta} - u_2 + u_0\right) \right\}, \tag{4.7}$$

where u_2 is a fixed parameter restricted to lie in the region

$$\hat{u}_\beta/2 < u_0 - u_2 < (u_0 + \hat{u}_\beta)/2. \tag{4.8}$$

Observe now that in view of (4.8) there are constants $\Delta_0, K_0 > 0$ such that if $0 < \Delta < \Delta_0$ and $a(\Delta), b(\Delta) > 0$ are defined by the identities

$$a(\Delta)^2/2\sigma^2 = -\frac{1}{2}\ln\Delta - K_0, \tag{4.9}$$

$$b(\Delta)^2/2\sigma^2 = -\frac{1}{2}\ln\Delta + K_0,$$

then the function g is strictly monotonic decreasing in the interval $[a(\Delta), b(\Delta)]$ with $g(a(\Delta)) > 0,\ g(b(\Delta)) < 0$. It follows that there is a unique solution $z = z_\Delta(u_2)$ of the equation $g(z) = 0$ in the interval $(a(\Delta), b(\Delta))$. We have shown then that (4.5) gives $u_1 = \sqrt{\Delta}\ z_\Delta(u_2)$ as a unique function of u_2, provided $0 < \Delta < \Delta_0$ and u_2 satisfies (4.8).

Next we wish to estimate the left-hand side of (4.4) when $u_1 = \sqrt{\Delta}\ z$ and $z \in [a(\Delta), b(\Delta)]$. To do this we write

$$p\big(\sqrt{\Delta}\ z, \Delta\big) = -\int_{\sqrt{\Delta}\ z}^{\infty} \frac{\partial p}{\partial x}\ (x, \Delta)dx$$

and use the formula (4.6). Observe now that

$$\begin{aligned}
&\frac{2}{\sqrt{2\pi\sigma^2\Delta}} \int_{\sqrt{\Delta}\ z}^{\infty} \exp\left[-\frac{(x+\mu\Delta)^2}{2\sigma^2\Delta}\right] dx \\
&\quad = \exp\left[-\frac{1}{2\sigma^2}\left(z+\mu\sqrt{\Delta}\right)^2\right] \frac{2}{\sqrt{2\pi\sigma^2}} \int_0^\infty d\xi\ \exp\left[-\frac{\xi(z+\mu\sqrt{\Delta})}{\sigma^2} - \frac{\xi^2}{2\sigma^2}\right] \\
&\quad \le \exp\left[-\frac{1}{2\sigma^2}\left(z+\mu\sqrt{\Delta}\right)^2\right] \frac{2\sigma}{\sqrt{2\pi}(z+\mu\sqrt{\Delta})}.
\end{aligned}$$

We can similarly estimate the contribution to $p(\sqrt{\Delta}\ z, \Delta)$ from the second term in (4.6). Thus we have

$$\begin{aligned}
&\frac{2}{\sqrt{2\pi\sigma^2\Delta}} \int_0^\infty \exp\left[-\frac{(x+y-\mu\Delta)^2}{2\sigma^2\Delta}\right] dy \\
&\qquad \le \frac{2\sigma\sqrt{\Delta}}{\sqrt{2\pi}(x-\mu\Delta)}\ \exp\left[-\frac{1}{2\sigma^2\Delta}(x-\mu\Delta)^2\right], \\
&\int_{\sqrt{\Delta}\ z}^\infty dx\ \frac{2\sigma\sqrt{\Delta}}{\sqrt{2\pi}(x-\mu\Delta)} \frac{\mu}{\sigma^2} \exp\left[-\frac{2\mu x}{\sigma^2}\right] \exp\left[-\frac{1}{2\sigma^2\Delta}(x-\mu\Delta)^2\right] \\
&\qquad \le \frac{2\sigma\sqrt{\Delta}}{\sqrt{2\pi}(z+\mu\sqrt{\Delta})} \exp\left[-\frac{1}{2\sigma^2}(z+\mu\sqrt{\Delta})^2\right],
\end{aligned}$$

provided $z \ge 2\mu\max(1, \sqrt{\Delta_0})$. We conclude then that for Δ_0 sufficiently small there is a constant C such that

$$0 < p\Big(\sqrt{\Delta}\ z, \Delta\Big) \le C\sqrt{\Delta}/|\ln\ \Delta|^{1/2}, \tag{4.10}$$

provided $0 < \Delta \leq \Delta_0,\ z \in [a(\Delta), b(\Delta)]$, and u_2 satisfies (4.8).

Consider now the function $F(u_2)$ defined by

$$F(u_2) = e^{\rho\Delta}\, V_0(u_1 - u_2 + u_0) - h(u_1, \Delta) - \beta\{1 - p(u_1, \Delta)\},\ 0 < \Delta < \Delta_0,$$

for u_2 in the region (4.8) and $u_1 = \sqrt{\Delta}\, z_\Delta(u_2)$ the unique solution of (4.5), $z_\Delta(u_2) \in (a(\Delta), b(\Delta))$. In view of (4.10) it is clear that

$$F(u_0 - \hat{u}_\beta) > 0,\ F\left(u_0 - \hat{u}_\beta + \sqrt{\Delta}\, b(\Delta)\right) < 0,$$

whence there is a solution u_2 to the equation $F(u_2) = 0$ in the region $u_0 - \hat{u}_\beta < u_2 < u_0 - \hat{u}_\beta + \sqrt{\Delta}\ b(\Delta)$. We have therefore shown the existence of a solution (u_1, u_2) to the set of equations (4.4), (4.5), provided $0 < \Delta < \Delta_0$. By Lemma 2.4 the solution is unique. One can now easily derive the asymptotics of $u_1(\beta, \Delta), u_2(\beta, \Delta)$ as $\Delta \to 0$. In fact the asymptotics of $u_1(\beta, \Delta)$ are already a consequence of the fact that $z_\Delta(u_2) \in (a(\Delta), b(\Delta))$. To obtain the asymptotics of $u_2(\beta, \Delta)$ we do a Taylor expansion of $F(u_2)$ about $u_2 = u_0 - \hat{u}_\beta$. Thus we have

$$F(u_2) = e^{\rho\Delta}\left\{V_0(\hat{u}_\beta) + [u_1 - u_2 + u_0 - \hat{u}_\beta]V_0'(\hat{u}_\beta) + O\left([u_1 - u_2 + u_0 - \hat{u}_\beta]^2\right)\right\}$$

$$- u_1 - \beta + O\left(\sqrt{\Delta}/|\ln\ \Delta|^{1/2}\right).$$

Hence the asymptotic form of $u_2 = u_2(\beta, \Delta)$ is obtained from the equation

$$[u_1 - u_2 + u_0 - \hat{u}_\beta]V_0'(\hat{u}_\beta) - u_1 = 0,$$

where $u_1 = u_1(\beta, \Delta)$ has the asymptotic form $u_1(\beta, \Delta) = \sigma\sqrt{\Delta}|\ln\ \Delta|^{1/2}$. □

PROPOSITION 4.1. *Let $u_{1,K}(\Delta), u_{2,K}(\Delta)$ be the functions of Δ defined in Lemma* 4.1. *Then there are the limits*

$$\lim_{\Delta\to 0} u_{1,K}(\Delta)/\sigma\sqrt{\Delta}\ |\ln\ \Delta|^{1/2} = 1, \tag{4.11}$$

$$\lim_{\Delta\to 0} \{u_{2,K}(\Delta) - [u_0 - \hat{u}_K]\}/\sigma\sqrt{\Delta}\ |\ln\ \Delta|^{1/2} = 1, \tag{4.12}$$

where $\hat{u}_K$ is as in Lemma 4.1.

Proof. We use Lemma 4.2 and (2.20). Evidently (4.11) follows directly from Lemma 4.2. To get (4.12) we substitute from Lemma 4.2 the formula for $u_2(\beta, \Delta)$ and solve approximately for $\hat{u}_\beta$. Thus on writing $\hat{u}_\beta = \hat{u}_K + \delta$ we have to highest order

$$V_0(\hat{u}_K + \delta) + u_0 - \hat{u}_K - \delta$$

$$+ [1 - 1/V_0'(\hat{u}_K)]\sigma\sqrt{\Delta}\ |\ln\ \Delta|^{1/2} = \mu/\rho - K.$$

Taylor expanding this last identity about $\delta = 0$ and solving for δ yields

$$\delta = -\sigma\sqrt{\Delta}\ |\ln\ \Delta|^{1/2}/V_0'(\hat{u}_K).$$

Hence we have to highest order

$$\begin{aligned} u_{2,K}(\Delta) &= [u_0 - \hat{u}_\beta] + \sigma\sqrt{\Delta}\ |\ln\ \Delta|^{1/2}\{1 - 1/V_0'(\hat{u}_\beta)\} \\ &= [u_0 - \hat{u}_K - \delta] + \sigma\sqrt{\Delta}\ |\ln\ \Delta|^{1/2}\{1 - 1/V_0'(\hat{u}_K)\} \\ &= u_0 - \hat{u}_K + \sigma\sqrt{\Delta}\ |\ln\ \Delta|^{1/2}. \end{aligned}$$

□

Finally we wish to show that the function $V(x), x \geq 0$, of Proposition 2.3, which is a C^1 function, fails to be twice differentiable at $x = u_1$. To do this let u_2 satisfy $0 < u_2 < u_0$ and $\beta > V_0(u_0 - u_2)$. We consider the function $u(x,t)$ defined by

$$u(x,t) = \beta\{1 - p(x,t)\} + h(x,t) - e^{\rho t} V_0(x - u_2 + u_0). \tag{4.13}$$

Evidently $u(x,0) = \beta + x - V_0(x - u_2 + u_0)$ and $u(0,t) = -e^{\rho t}\ V_0(u_0 - u_2)$, whence $u(0,t) < 0,\ t > 0$.

LEMMA 4.3. *Let $u(x,t)$ be the function* (4.13). *Then there exist $\varepsilon > 0$ and $x(t) > 0$, $0 < t < \varepsilon$, such that $\partial u/\partial x(x,t) > 0$ for $0 \leq x < x(t)$ and $\partial u/\partial x(x,t) < 0$ for $x > x(t)$.*

Proof. We proceed as in Lemma 4.2 observing that the function g of (4.7) is given by

$$g(z) = \sqrt{2\pi\sigma^2\Delta}\ \partial u/\partial x\Big(z\sqrt{\Delta},\ \Delta\Big). \tag{4.14}$$

With $a(\Delta), b(\Delta)$ defined by (4.9) we have seen that, provided $0 < \Delta < \Delta_0$, then $g(z) > 0$ for $0 < z \leq a(\Delta)$, $g(z) < 0$ for $z \geq b(\Delta)$ and g is strictly monotonic decreasing in the interval $[a(\Delta), b(\Delta)]$. □

LEMMA 4.4. *Let $u(x,t)$ be the function* (4.13). *Then there exist $\varepsilon > 0$ and $a(t), b(t) > 0,\ 0 < t < \varepsilon$, such that $\{x > 0 : \partial^2 u/\partial x^2(x,t) > 0\} = (a(t), b(t))$.*

Proof. From (2.9), (4.13) we have that

$$\partial^2 u/\partial x^2(x,t) = -\beta\partial^2 p/\partial x^2(x,t) - \mu \int_0^t \partial^2 p/\partial x^2(x,s)ds - e^{\rho t}\ V_0''(x - u_2 + u_0).$$

Since $p(x,t)$ is a convex function of x it follows that $\partial^2 u/\partial x^2(x,t) < 0$ for $x > u_2$. Since $V_0''(u_0) = 0$, $V_0'''(u_0) = 2\rho/\sigma^2$ we conclude that there exist $\Delta_0 > 0$ and $b(t) > 0,\ 0 < t < \Delta_0$, with the property that $\lim_{t\to 0} b(t) = u_2$, and $\partial^2 u/\partial x^2(x,t) < 0$ if $x > b(t)$ and $\partial^2 u/\partial x^2(x,t) > 0$ if $t^{1/3} < x < b(t)$.

We consider now the interval $0 < x < t^{1/3}$. From (4.14) we may consider the function $g'(z)$ in the region $0 < z < \Delta^{-\,1/6}$ instead of $\partial^2 u/\partial x^2(x,t), 0 < x < t^{1/3}$. We have that

$$g'(z) = \frac{-2\beta(z + \mu\sqrt{\Delta})}{\sigma^2} \exp\left[-\frac{(z + \mu\sqrt{\Delta})^2}{2\sigma^2}\right] \tag{4.15}$$

$$+ G(z,\Delta) - \sqrt{2\pi\sigma^2}\ e^{\rho\Delta}\ \Delta\ V_0''(z\sqrt{\Delta} - u_2 + u_0).$$

We can estimate the function $G(z,\Delta)$ in the same way we obtained the estimate (4.10). In fact differentiating (4.6) and estimating as before, we have the inequality

$$2\left[\frac{x}{\sigma^2 t} + \frac{2\mu}{\sigma^2}\right] \exp\left[-\frac{(x + \mu t)^2}{2\sigma^2 t}\right] \leq \sqrt{2\pi\sigma^2 t}\ \frac{\partial^2 p}{\partial x^2}\ (x,t) \tag{4.16}$$

$$\leq 2\left[\frac{x}{\sigma^2 t} + \frac{4\mu}{\sigma^2}\right] \exp\left[-\frac{(x + \mu t)^2}{2\sigma^2 t}\right],\quad x \geq 2\mu t.$$

From (4.16) we can also estimate $\partial^2 h/\partial x^2$ using the identity

$$-\frac{\partial^2 h}{\partial x^2}\ (x,t) = \mu \int_0^t \frac{\partial^2 p}{\partial x^2}\ (x,s)ds. \tag{4.17}$$

We can estimate the right-hand side of (4.17) by substituting the right-hand side of (4.16) and making the change of variable $w = (x + \mu s)/\sqrt{s}$. Thus we have

$$-\frac{\partial^2 h}{\partial x^2}(x,t) \le \frac{32\mu}{\sigma^2\sqrt{2\pi\sigma^2}} \int_{(x+\mu\sqrt{t})/t}^{\infty} \exp\left[-\frac{w^2}{2\sigma^2}\right] dw \tag{4.18}$$

$$\le \frac{11}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x+\mu t)^2}{2\sigma^2 t}\right],\ x \ge 2\mu t.$$

The function $G(z,\Delta)$ is therefore bounded from the estimates (4.16), (4.18) as

$$0 \le -G(z,\Delta) \le \left(\frac{6\beta\mu}{\sigma^2} + 11\right)\sqrt{\Delta}\ \exp\left[-\frac{(z+\mu\sqrt{\Delta})^2}{2\sigma^2}\right],\ z \ge 2\mu\sqrt{\Delta}. \tag{4.19}$$

From (4.15), (4.19) we see that there exists $\Delta_0 > 0$ such that for $0 < \Delta < \Delta_0$ one has $g'(z) < 0$ for $0 < z < \sigma\sqrt{2|\ln\ \Delta|}$ and $g'(z) > 0$ for $\sigma\sqrt{3|\ln\ \Delta|} < z < \Delta^{-1/6}$. Here there is a solution $z = a(\Delta)/\sqrt{\Delta}$ of the equation $g'(z) = 0$ in the interval $\sigma\sqrt{2\ln\ \Delta} < z < \sigma\sqrt{3\ln\ \Delta}$. Evidently then $\lim_{\Delta\to 0} a(\Delta) = 0$.

We complete the proof by showing that $g''(z) > 0$ for $\sigma\sqrt{2|\ln\ \Delta|} < z < \sigma\sqrt{3|\ln\ \Delta|}$, provided $0 < \Delta < \Delta_0$. This is accomplished by estimating $\partial^3 p/\partial x^3$ in a similar way to how we estimated $\partial^2 p/\partial x^2$. □

LEMMA 4.5. *Let $u(x,t)$ be the function* (4.13), *and define $T > 0$ as $T = \sup\{t > 0 : \sup_{x>0} u(x,t) > u(0,t)\}$. Then for $0 < t < T$ there exists unique $x(t) > 0$ satisfying $\lim_{t\to 0} x(t) = 0$, with the property that $\partial u/\partial x(x,t) > 0$ for $0 \le x < x(t)$ and $\partial u/\partial x(x,t) < 0$ for $x > x(t)$. Furthermore $\partial^2 u/\partial x^2(x(t),t) < 0$.*

Proof. By the definition of T one must have $\partial u/\partial x(x,t) > 0$ for some x when $0 < t < T$. Since $\partial u/\partial x(x,t) < 0$ for x large it follows from Theorem A.1 that a unique $x(t)$ exists for $0 < t < T$. Now we apply the argument in Theorem A.3, using Lemma 4.4 to conclude that $\partial^2 u/\partial x^2(x(t),t) < 0$. □

PROPOSITION 4.2. *The function $V(x)$ defined by* (2.21) *is C^1 but not C^2 at $x = u_1$.*

Proof. Let β be as in (2.21) and let $u(x,t)$ be the function (4.13). Then $u(u_1,\Delta) = \partial u/\partial x(u_1,\Delta) = 0$, $u(0,\Delta) < 0, u(x,\Delta) \to -\infty$ as $x \to \infty$. Then by Lemma 4.5 we have $\partial^2 u/\partial x^2(u_1,\Delta) < 0$, whence V is not C^2. □

Appendix. Some consequences of the maximum principle. Here we prove some general results for the heat equation which are used in earlier sections. Let $a(x,t), b(x,t)$ be uniformly bounded smooth functions in (x,t), $x \in \mathbf{R}$, $t \ge 0$, with the property that a is also uniformly bounded from below by a positive constant. We define the operator L on C^2 functions $u(x,t)$ by

$$Lu(x,t) = a(x,t)\,\frac{\partial^2 u}{\partial x^2} + b(x,t)\frac{\partial u}{\partial x} - \frac{\partial u}{\partial t}.$$

THEOREM A.1. *Suppose $u(x,t)$ is a C^2 function in $\{(x,t) \in \mathbf{R}^2 : x \ge 0,\ t \ge 0\}$ satisfying $Lu \equiv 0$. Suppose further that $u(0,t) = 0$, $t \ge 0$, and the set $\{x > 0 : u(x,0) > 0\}$ is a semi-infinite interval. Then for any $t > 0$ there is at most one point $x(t) > 0$ satisfying $u(x(t),\ t) = 0$.*

Proof. Observe that since the function $u(x,0)$, $x \ge 0$, can be negative only on a bounded set it follows that $u(x,t)$, $x \ge 0$, is uniformly bounded below for all $t \ge 0$. Further, one has $\liminf_{x\to\infty} u(x,t) \ge 0$ for all $t \ge 0$. For some $T > 0$

suppose there is an $x_0 > 0$ with $u(x_0, T) < 0$. Let D be the maximal domain containing (x_0, T) such that $u(x,t) < 0$ for $(x,t) \in D$. Evidently $u(x,t) = 0$ for $(x,t) \in \partial D \cap \mathbf{R} \times (0, \infty)$. Hence by the maximum principle the minimum of u occurs at an interior point unless there exists $(x_1, 0) \in \partial D$ with $u(x_1, 0) < 0$. We conclude that there is a path $\gamma(s),\ 0 \le s \le T$, with $\gamma(T) = x_0$ and $u(\gamma(s), s) < 0,\ 0 \le s \le T$.

Now we define Ω to be the domain $\Omega = \{(x,s) : 0 < x < \gamma(s),\ 0 < s < T\}$. Then $u \le 0$ on $\partial\Omega \cap \mathbf{R} \times [0, T)$ and strictly negative on part of this boundary. The maximum principle therefore implies that u is strictly negative on $\partial\Omega \cap \mathbf{R}^+ \times \{T\}$. Thus u is strictly negative on the interval $(0, x_0]$. We define now $x(t)$ as $x(t) = 0$ if $u(x,t) \ge 0,\ x > 0;\ x(t) = \infty$ if $u(x,t) < 0, x > 0;\ x(t) = \limsup\{x > 0 : u(x,t) < 0\}$ otherwise. If $0 < x(t) < \infty$, then the maximum principle implies that $u(x,t) > 0$ for $x > x(t)$. □

THEOREM A.2. *Suppose $u(x,t)$ is a C^2 function in $\{(x,t) \in \mathbf{R}^2 : x \ge 0,\ t \ge 0\}$ satisfying $Lu \equiv 0$. Suppose further that $u(0,t) < 0,\ t \ge 0$, and that the set $\{x > 0 : u(x,0) > 0\}$ is an open interval. Then for any $t > 0$ the set $\{x > 0 : u(x,t) \ge 0\}$ is either empty, a single point, or a closed interval. If the set is a closed interval, then u is strictly positive on its interior.*

Proof. By Theorem A.1 we may assume that the set $\{x > 0 : u(x,0) > 0\}$ is a finite interval. Hence for all $t \ge 0$, $\limsup_{x\to\infty} u(x,t) \le 0$. Now let $D = \{(x,t) : x > 0,\ t > 0,\ u(x,t) > 0\}$. Since u is uniformly bounded above it follows by the maximum principle that D is connected and that $\{(x,t) : x > 0,\ u(x,0) > 0\} \subset \partial D$. We show that the intersection of D with any line $t =$ constant is either empty or an open interval. To do this let $a(t), b(t)$ be defined by

$$a(t) = \inf\{x > 0 : (x,t) \in D\},$$

$$b(t) = \sup\{x > 0 : (x,t) \in D\}.$$

We define a domain $\Omega = \{(x,t) : t > 0,\ a(t) < x < b(t)\}$, whence $D \subset \Omega$. For $T > 0$ let D_T and Ω_T be defined by $D_T = D \cap [\mathbf{R} \times (0,T)]$, $\Omega_T = \Omega \cap [\mathbf{R} \times (0,T)]$. Suppose for some T one has $D_T \ne \Omega_T$. By the maximum principle if the minimum of u on Ω_T is negative, then it must be at a point (x_0, T) with $a(T) < x_0 < b(T)$. This contradicts Lemma 3 of Chapter 3, section 2 of [10]. Hence $u \ge 0$ on Ω_T and consequently by the maximum principle again $u > 0$ on Ω_T. Applying the maximum principle to the complement of Ω_T in $\mathbf{R} \times (0,T)$ we conclude that the set $\{x \in \mathbf{R} : u(x,t) \ge 0\} = [a(t), b(t)]$.

Finally suppose there is a minimum T such that $\Omega = \Omega_T$. As before the maximum principle implies that u is strictly negative on the complement of $\bar{\Omega}_T$ in $\mathbf{R} \times (0, \infty)$. This completes the proof. □

Next we wish to show that the solution $x(t)$ in Theorem A.1 of the equation $u(x,t) = 0$ is nondegenerate. As a consequence it follows that $x(t)$ is a smooth function of t for $t > 0$.

THEOREM A.3. *Suppose $u(x,t),\ x \ge 0,\ t \ge 0$, satisfies the assumptions of Theorem* A.1. *Let $W_0 = \{x > 0 : \partial u/\partial x(x,0) > 0\}$ and assume that the boundary ∂W_0 of W_0 has no finite limit points. Then there is the inequality $\partial u/\partial x(x(t),t) > 0,\ t > 0$.*

Proof. For $t > 0$ let $W_t = \{x > 0 : \partial u/\partial x(x,t) > 0\}$. Since $\partial u/\partial x(x,t)$ also satisfies a diffusion equation it follows that ∂W_t has no finite limit points. Furthermore, for t small and $x \notin \partial W_t$ the derivative $\partial u/\partial x(x,t)$ is nonzero.

We first show that $\partial u/\partial x(x(t),t) > 0$ for $t > 0$ sufficiently small. We argue by contradiction. Suppose $t_n,\ n = 1, 2, \ldots$, is a positive sequence with $\lim_{n\to\infty} t_n = 0$

such that $x(t_n) \in \partial W_{t_n}$, $n = 1, 2, \ldots$. We may assume without loss of generality that $\lim_{n\to\infty} x(t_n) = x_\infty$. There is then a possibly nontrivial closed interval $[a, b]$ such that $x_\infty \in [a, b]$, $\partial u/\partial x(x, 0) = 0$, $x \in [a, b]$, and $\partial u/\partial x(x, 0) \neq 0$, $x \notin [a, b]$, but sufficiently close to a or b. Suppose now that $\partial u/\partial x(x, 0) > 0$ for x close to a or close to b. Then by the maximum principle it follows that $\partial u/\partial x(x, t) > 0$ for x in a neighborhood of $[a, b]$ when t is sufficiently small. In particular $\partial u/\partial x(x(t_n), t_n) > 0$ for n large, which is a contradiction. Alternatively we can have $\partial u/\partial x(x, 0) > 0$ for x close to a and $\partial u/\partial x(x, 0) < 0$ for x close to b. In that case the maximum principle implies $u(x, t) < 0$ for x in a neighborhood of $[a, b]$ when t is small. This again contradicts the fact that $u(x(t_n), t_n) = 0$ as $n \to \infty$. We conclude therefore that $\partial u/\partial x(x(t), t) > 0$ for $t > 0$ sufficiently small.

Finally we show that $\partial u/\partial x(x(t), t) > 0$ for all t. To see this consider $T > 0$ and suppose that $\partial u/\partial x(x(t), t) > 0$ for all $0 < t < T$, $\lim_{t\to T} x(t) = x_T$. If $x_T = \infty$, then $u(x, T) \leq 0, x \geq 0$. If $x_T = 0$, then $u(x, T) \geq 0, x \geq 0$. Suppose now $0 < x_T < \infty$, whence $u(x, T)$, $x > 0$, takes on both positive and negative values. For $0 < t < T$ there exists $a(t), b(t)$ with $0 \leq a(t) < x(t) < b(t) \leq \infty$ such that $\partial u/\partial x(x, t) > 0$ for $a(t) < x < b(t)$ and $a(t), b(t) \in \partial W_t$ if $a(t) > 0$ and $b(t) < \infty$. Since $u(0, t) = 0$ it follows that $\partial u/\partial x(x, t) < 0$ for $x > 0$ small. Hence $a(t) > 0$.

We show that x_T cannot be the limit as $t \to T$ of any points in $\partial W_t \cap [0, x(t)]$. Suppose first that $\partial W_t \cap [0, x(t)] = \{a(t)\}$. Then $\partial u/\partial x(x, t)$, $0 < x < x(t)$, has just one sign change, whence $\partial u/\partial x(x, T)$, $0 < x < x_T$, has also at most one sign change. Since $u(x_T, T) = u(0, T) = 0$ it follows that $\lim_{t\to T} a(t) < x_T$. Alternatively let $a^*(t) \in \partial W_t \cap [0, x(t)]$ have the property that

$$u(a^*(t), t) = \sup\{u(x, t) : x \in \partial W_t \cap [0, x(t)]\}.$$

Evidently $u(a^*(t), t) < 0$. By the maximum principle it follows that $u(a^*(t), t)$ is a decreasing function of t. Hence x_T cannot be the limit as $t \to T$ of any point in $\partial W_t \cap [0, x(t)]$.

Since we can argue similarly that x_T is not the limit as $t \to T$ of any point in $\partial W_t \cap [x(t), \infty)$ we conclude that $\partial u/\partial x(x(T), T) > 0$. We have shown that the set $\{t > 0 : \partial u/\partial x(x(t), t) > 0\}$ is both open and closed, whence it must be the interval $(0, \infty)$. □

Acknowledgment. The authors would like to thank Jeffrey Rauch and Peter Bates for helpful conversations.

REFERENCES

[1] M. Bramson, *Convergence of solutions of the Kolmogorov equation to travelling waves*, Mem. Amer. Math. Soc., 44 (1983), p. 285.

[2] W. Fleming and R. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[3] W. Fleming and H. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[4] A. Friedman and W. Shen, *A variational inequality approach to financial valuation of retirement benefits based on salary*, Finance Stoch., 6 (2002), pp. 273–302.

[5] B. Hojgaard and M. Taksar, *Controlling risk exposure and dividends payout schemes: Insurance company example*, Math. Finance, 9 (1999), pp. 153–182.

[6] B. Hojgaard and M. Taksar, *Optimal risk control for a large corporation in the presence of returns on investments*, Finance Stoch., 5 (2001), pp. 527–547.

[7] H. McKean, *Application of Brownian motion to the equation of Kolmogorov-Petrowskii-Piskounov*, Commun. Pure Appl. Math., 28 (1975), pp. 323–331.

[8] A. Milne and D. Robertson, *Firm behavior under the threat of liquidation*, J. Econom. Dynam. Control, 20 (1996), pp. 1427–1449.
[9] S. Peura and J. Keppo, *Optimal bank capital with costly recapitalization*, J. Business, 79 (2006), pp. 2163–2201.
[10] M. Protter and H. Weinberger, *Maximum Principle in Differential Equations*, Springer-Verlag, New York, 1984 (corrected reprint of the 1967 original).

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 1995–2012

STABILITY OF NONLINEAR FEEDBACK SYSTEMS: A NEW SMALL-GAIN THEOREM*

ANNA L. CHEN†, GUI-QIANG CHEN‡, AND RANDY A. FREEMAN†

Abstract. For the feedback interconnection of general nonlinear systems, the classical small-gain condition is sufficient but not necessary for robust stability. We introduce a weaker notion of gain which yields a small-gain condition that is both necessary and sufficient for robust stability. We also discuss conditions under which the two notions coincide, and we further provide results for dissipation performance measures that are more general than the classical gain measures.

Key words. small-gain theorem, conditional gain, dissipativity, performance analysis, nonlinear systems, feedback connection

AMS subject classifications. 93D09, 93D05, 93D25, 93B52, 93C25, 34D23

DOI. 10.1137/S0363012904440812

1. Introduction.

The classical small-gain theorem [28, 4] provides a sufficient condition for the stability of the feedback interconnection of nonlinear systems, and it has been studied extensively by many researchers. In [28], Zames derived a sufficient condition for the small-gain theorem for general nonlinear systems by using the concepts of loop gain and positivity. Willems [24] systematically studied the passivity theorem, and Anderson proved in [1] the equivalence between the classical small-gain theorem and the passivity theorem for general feedback systems. In [25], Willems established a theorem for dissipative systems in the context of finite-dimensional stationary linear systems with quadratic supply rates, deriving a necessary and sufficient condition in the frequency domain for dissipativity. Hill and Moylan in [10, 17] established conditions for stability and instability for interconnected systems in terms of the properties of dissipative subsystems. In [2], Chen and Desoer derived a necessary and sufficient condition for the robust stability of linear distributed feedback systems. Extensions of these classical results to nonlinear notions of gain appear in [15, 13, 12].

Of course, one may have stability even if the small-gain condition is violated, but it is of interest to know whether the small-gain condition becomes necessary when we treat one of the systems as uncertainty and consider the *robust* stabilization problem. Positive results are available for the case in which the nominal system is linear [5, 3, 20]. More recently, Shamma [19] demonstrated the necessity of the small-gain condition for the robust stability of a class of nonlinear systems with *fading memory*. Shamma and Zhao further developed this approach in [21] and gave a necessary condition for uniform robust invertibility, whereas Gonçalves and Dahleh provided another necessary condition in [7] for nonlinear systems with fading memory.

Unfortunately, the small-gain condition is not necessary for robust stability for

*Received by the editors February 11, 2004; accepted for publication (in revised form) May 22, 2007; published electronically November 30, 2007. This work was supported in part by the National Science Foundation under grant ECS-0115317.

http://www.siam.org/journals/sicon/46-6/44081.html

†Department of Electrical Engineering and Computer Science, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208-3118 (anna@ece.northwestern.edu, freeman@ece.northwestern.edu).

‡Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208-2730 (gqchen@math.northwestern.edu). This author was supported in part by an Alexander von Humboldt Foundation Fellowship.

more general classes of nonlinear systems (even time-invariant, first-order systems), as demonstrated by an example in [6]. In this paper, we show how to weaken the standard notion of gain so that we preserve the sufficiency of the small-gain condition while recovering its necessity for the robust stability of general nonlinear systems, even those without fading memory. We call the weakened property *conditional gain*,[1] as it is related to the conditional integral quadratic constraints proposed in [16] (in fact, for the $\mathcal{L}_2$ case, one can develop frequency-domain conditions for conditional gain using such integral quadratic constraints, but we will not pursue this here). Our sufficiency result improves upon the classical small-gain theorem by replacing the gain assumption on one of the subsystems in the loop with the weaker conditional gain assumption. In our necessity result, we show that it suffices to consider only time-invariant, second-order uncertainties, even when the nominal system itself is a distributed and/or time-varying system.

Throughout this paper, we use the behavioral theory of dynamical systems developed in [27, 18], modified to account for the fact that different trajectories may be defined on different time intervals (e.g., for systems admitting finite escape times). We found this behavioral approach to be ideal for our analysis, allowing us to simplify proofs and avoid technicalities associated with other classes of dynamical system models (such as input/output operator-theoretic models or differential-equation state models). Other authors have also found this framework useful in the study of dissipative systems [8, 22].

The organization of this paper is as follows. In section 2, we develop the notions of dissipativity in the context of behavioral dynamical systems. We revisit classical $\mathcal{L}_p$-gain in this context in section 3, where we also present our definition of conditional gain along with the main results of this paper. Moreover, in this section, we present a sufficient condition under which conditional and classical $\mathcal{L}_p$-gains are equivalent notions; this sufficient condition is in the form of a resetting property related to the ones used in [19, 21] to prove the necessity of the small-gain theorem for the robust stability of nonlinear systems with fading memory. This sufficient condition also implies that the conditional and classical notions of $\mathcal{L}_p$-gain are equivalent for linear time-invariant systems. Finally, we extend our results to more general dissipation performance measures in section 4.

2. Dynamical systems, dissipativity, and storage functions. In this section we discuss dissipativity in the context of behavioral dynamical systems [27, 18]. We restrict our attention to continuous-time systems for simplicity. A *trajectory* on a nonempty set $\mathbb{V}$ is a mapping $v : I \to \mathbb{V}$ on a nonempty interval domain $\mathrm{Dom}(v) = I \subset \mathbb{R}$. We use the following notation from [27]: given two trajectories v_1, v_2 on $\mathbb{V}$ and a time $t \in \mathrm{Dom}(v_1) \cap \mathrm{Dom}(v_2)$, we define the *concatenation* of v_1 and v_2 at time t to be the trajectory $v_1 \wedge^t v_2$ on $\mathbb{V}$ given by

$$(v_1 \wedge^t v_2)(\tau) = \begin{cases} v_1(\tau) & \text{when } \tau < t, \\ v_2(\tau) & \text{when } \tau \geqslant t, \end{cases} \tag{2.1}$$

with interval domain $\mathrm{Dom}(v_1 \wedge^t v_2) = \big[(-\infty, t] \cap \mathrm{Dom}(v_1)\big] \cup \big[[t, \infty) \cap \mathrm{Dom}(v_2)\big]$.

A *trajectory space* on $\mathbb{V}$ is a nonempty collection $\mathfrak{T}(\mathbb{V})$ of trajectories on $\mathbb{V}$ which is closed under concatenations: for any $v_1, v_2 \in \mathfrak{T}(\mathbb{V})$ with $t \in \mathrm{Dom}(v_1) \cap \mathrm{Dom}(v_2)$,

[1] The term "conditional gain" was also used in [23] to describe the ultimate virtual-dissipativity performance measure discussed in [9]. In general this system property neither implies nor is implied by our conditional gain property.

we have $v_1 \wedge^t v_2 \in \mathfrak{T}(\mathbb{V})$. A trajectory space $\mathfrak{T}(\mathbb{V})$ together with a subset $\mathcal{B} \subset \mathfrak{T}(\mathbb{V})$ is called a *dynamical system* $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B})$, and the set $\mathcal{B}$ is called the *behavior* of the system H (cf. [27, 18]). When the underlying trajectory space $\mathfrak{T}(\mathbb{V})$ is apparent from context, we will use the terms "system H" and "behavior $\mathcal{B}$" interchangeably. Note that the trajectory space used in [27] is the set of all mappings from a fixed time interval $\mathbb{T}$ into $\mathbb{V}$; here we instead allow different trajectories of a system to be defined on different time intervals. We also allow the behavior $\mathcal{B}$ to be the empty set $\mathcal{B} = \varnothing$ and refer to empty and nonempty systems H accordingly.

Given a trajectory $v \in \mathfrak{T}(\mathbb{V})$ and a time $t \in \mathbb{R}$, we let $\sigma^t v$ denote the time-shifted trajectory $(\sigma^t v)(\cdot) = v(\cdot + t)$ with $\mathrm{Dom}(\sigma^t v) = [\mathrm{Dom}(v) - t]$. We say that a system is *time-invariant* when its behavior $\mathcal{B}$ satisfies $\sigma^t \mathcal{B} = \mathcal{B}$ for all $t \in \mathbb{R}$.

To develop the notion of a state-space system, we introduce a *state space* $\mathbb{X}$ and a mapping $\xi : \mathbb{V} \to \mathbb{X}$ which assigns a state value $\xi q \triangleq \xi(q) \in \mathbb{X}$ to each signal value $q \in \mathbb{V}$. We recall the following from [27].

Definition 2.1. *A system $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B})$ satisfies the* axiom of state *with respect to a mapping $\xi : \mathbb{V} \to \mathbb{X}$ when the following holds: whenever a pair of trajectories $v_1, v_2 \in \mathcal{B}$ and a time $t \in \mathrm{Dom}(v_1) \cap \mathrm{Dom}(v_2)$ are such that $\xi v_1(t) = \xi v_2(t)$, then the concatenation $v_1 \wedge^t v_2$ also belongs to $\mathcal{B}$.*

A system H which satisfies the axiom of state is called a *state-space system*, and we write H as the triple $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B}, \xi)$. Note that the development in [18] requires the axiom of state to hold only at points of continuity in the state trajectory,[2] but we adopt the original definition in [27], which does not require the state space to be endowed with a topology.

Given a trajectory $v \in \mathfrak{T}(\mathbb{V})$ and a state mapping $\xi : \mathbb{V} \to \mathbb{X}$, we let $\mathfrak{G}(v) \subset \mathbb{R} \times \mathbb{X}$ denote the graph of the state trajectory $\xi v(\cdot)$, namely, the set of all pairs $(t, \xi v(t))$ as t varies over $\mathrm{Dom}(v)$. Likewise, given a behavior $\mathcal{B} \subset \mathfrak{T}(\mathbb{V})$, we let $\mathfrak{G}(\mathcal{B})$ denote the union of all graphs $\mathfrak{G}(v)$ as v varies over $\mathcal{B}$. In other words, $(t_0, x_0) \in \mathfrak{G}(\mathcal{B})$ if and only if there is some trajectory $v \in \mathcal{B}$ such that $\xi v(t_0) = x_0$.

One can characterize state-space systems using differential-algebraic equations. For example, let $\mathbb{X} = \mathbb{R}^n$, let $\mathbb{V} = \mathbb{W} \times X$ for some auxiliary signal space $\mathbb{W}$, let ξ be the natural projection of $\mathbb{V}$ onto $\mathbb{X}$, and let $\mathfrak{T}(\mathbb{V})$ be a trajectory space on $\mathbb{V}$ whose members all have domains with nonempty interiors. Given a mapping $F : \mathbb{R} \times \mathbb{V} \times \mathbb{R}^n \to \mathbb{R}^m$, let $\mathcal{B}$ denote the set of trajectories $(w, x) \in \mathfrak{T}(\mathbb{V})$ such that $x(\cdot)$ is absolutely continuous on $\mathrm{Dom}(w, x)$ and

$$F\big(t,\, w(t),\, x(t),\, \dot{x}(t)\big) = 0 \tag{2.2}$$

for almost all $t \in \mathrm{Dom}(w, x)$. Then the system $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B}, \xi)$ is a state-space system. Here the auxiliary signal $w(t)$ can represent a combination of inputs, outputs, or other signals, although at this point there is no need to label them as such. Also, there is no need to assume the existence or uniqueness of solutions to (2.2); whatever solutions do exist belong to the behavior $\mathcal{B}$. This class of systems (2.2) encompasses many of the nonlinear state-space systems studied in standard texts [14].

We next introduce the concepts of abstract energy, storage, and dissipation. Roughly speaking, if a system is dissipative, then it will absorb more energy from the external world than it supplies. To make this idea precise, we will adopt various definitions from [25, 26, 9, 10, 8], modifying them as appropriate to fit our particular behavioral context.

[2] This weaker version of the axiom of state is called the *property of state* in [18].

DEFINITION 2.2. *A* supply rate $\mathbf{S}$ *for a trajectory space* $\mathfrak{T}(\mathbb{V})$ *is a function* $\mathbf{S} : \mathbb{R} \times \mathbb{V} \to \mathbb{R}$ *such that, for every* $v \in \mathfrak{T}(\mathbb{V})$*, the mapping* $t \mapsto \mathbf{S}(t, v(t))$ *is locally integrable on* $\mathrm{Dom}(v)$.

DEFINITION 2.3. *Let* $\mathbf{S}$ *be a supply rate for a trajectory space* $\mathfrak{T}(\mathbb{V})$*, and let* $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B}, \xi)$ *be a state-space system. Given* $(t_0, x_0) \in \mathfrak{G}(\mathcal{B})$*, we define the* available storage $\phi_a(t_0, x_0)$ *for* H *(with respect to* $\mathbf{S}$*) to be the quantity*

$$\phi_a(t_0, x_0) \triangleq - \inf_{\substack{v \in \mathcal{B} \\ t \geqslant t_0}} \int_{t_0}^{t} \mathbf{S}(\tau, v(\tau))\, d\tau, \tag{2.3}$$

where the infimum is taken over all trajectories $v \in \mathcal{B}$ *and all times* $t \geqslant t_0$ *such that* $t_0, t \in \mathrm{Dom}(v)$ *and* $\xi v(t_0) = x_0$*. We say that* H *is* dissipative *with respect to* $\mathbf{S}$ *when* $\phi_a(t_0, x_0) < \infty$ *for every* $(t_0, x_0) \in \mathfrak{G}(\mathcal{B})$.

Note that $\mathfrak{G}(\mathcal{B}) = \varnothing$ if and only if H is empty, in which case H is trivially dissipative. Hence, from this point on we will assume that H is nonempty. Also, because we can take $t = t_0$ to get zero as a possible value of the integral in (2.3), we see that $\phi_a(t_0, x_0) \geqslant 0$ for all $(t_0, x_0) \in \mathfrak{G}(\mathcal{B})$.

DEFINITION 2.4. *Let* $\mathbf{S}$ *be a supply rate for a trajectory space* $\mathfrak{T}(\mathbb{V})$*, and let* $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B}, \xi)$ *be a state-space system. A* virtual storage function *for* H *(with respect to* $\mathbf{S}$*) is a function* $U : \mathfrak{G}(\mathcal{B}) \to \mathbb{R}$ *such that the* dissipation inequality

$$U(t, \xi v(t)) \leqslant U(t_0, \xi v(t_0)) + \int_{t_0}^{t} \mathbf{S}(\tau, v(\tau))\, d\tau \tag{2.4}$$

holds for any $v \in \mathcal{B}$ *and any* $t_0, t \in \mathrm{Dom}(v)$ *with* $t_0 \leqslant t$*. A* storage function *is a virtual storage function which is nonnegative everywhere.*

We now give a behavioral version of the classical result equating dissipativity to the existence of a storage function [25, 26, 9, 8] (e.g., Theorem IV-2 in [26] or Theorem 1 in [8]). Note that the only aspect of the system dynamics needed here is the axiom of state; there is no need to distinguish between inputs and outputs or assume any additional structure of these dynamics.

PROPOSITION 2.5. *A state-space system* $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B}, \xi)$ *is dissipative with respect to a supply rate* $\mathbf{S}$ *if and only if it admits a storage function. Furthermore, if* H *is dissipative, then* ϕ_a *is a storage function and any other storage function* U *satisfies* $U(t_0, x_0) \geqslant \phi_a(t_0, x_0)$ *for all* $(t_0, x_0) \in \mathfrak{G}(\mathcal{B})$.

Proof. First suppose that U is a storage function. Then, since $U \geqslant 0$, we have

$$0 \leqslant U(t_0, x_0) + \int_{t_0}^{t} \mathbf{S}(\tau, v(\tau))\, d\tau \tag{2.5}$$

along all trajectories $v \in \mathcal{B}$ satisfying the initial condition $\xi v(t_0) = x_0$. It follows that $\phi_a(t_0, x_0)$ in (2.3) is bounded from above by the finite value $U(t_0, x_0)$.

Conversely, suppose that H is dissipative and that there exist $v_1 \in \mathcal{B}$ and $t_0, t_1 \in \mathrm{Dom}(v_1)$ with $t_0 \leqslant t_1$ such that

$$\phi_a(t_1, \xi v_1(t_1)) > \phi_a(t_0, \xi v_1(t_0)) + \int_{t_0}^{t_1} \mathbf{S}(\tau, v_1(\tau))\, d\tau. \tag{2.6}$$

Then, from (2.3), there exist $v_2 \in \mathcal{B}$ and $t_2 \geqslant t_1$ such that $t_1, t_2 \in \mathrm{Dom}(v_2)$, $\xi v_2(t_1) = \xi v_1(t_1)$, and

$$-\int_{t_1}^{t_2} \mathbf{S}(\tau, v_2(\tau))\, d\tau > \phi_a(t_0, \xi v_1(t_0)) + \int_{t_0}^{t_1} \mathbf{S}(\tau, v_1(\tau))\, d\tau. \tag{2.7}$$

It follows from the axiom of state that the trajectory $v \triangleq v_1 \wedge^{t_1} v_2$ also belongs to $\mathcal{B}$. Thus, from (2.7), we have

$$-\int_{t_0}^{t_2} \mathbf{S}(\tau, v(\tau))\, d\tau > \phi_a(t_0, \xi v(t_0)), \tag{2.8}$$

which contradicts (2.3). □

Note that if U is a storage function for a system H, then so is $U - \inf U$ where the infimum is taken over $\mathfrak{G}(\mathcal{B})$. Hence, from Proposition 2.5, we have $U - \inf U \geqslant \phi_a$ on $\mathfrak{G}(\mathcal{B})$, which immediately gives us the following.

PROPOSITION 2.6. *If $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B}, \xi)$ is dissipative, then*

$$\inf\big\{\phi_a(t_0, x_0) \;:\; (t_0, x_0) \in \mathfrak{G}(\mathcal{B})\big\} = 0. \tag{2.9}$$

We may interpret Proposition 2.6 as a statement about "bias" in a dissipative system H: for any $\varepsilon > 0$, there exists an initial condition (t_0, x_0) such that $\phi_a(t_0, x_0) \leqslant \varepsilon$, which implies from (2.3) that

$$\int_{t_0}^{t} \mathbf{S}(\tau, v(\tau))\, d\tau \geqslant -\varepsilon \tag{2.10}$$

along every trajectory $v \in \mathcal{B}$ with $\xi v(t_0) = x_0$. In other words, the "bias" ε can be made arbitrarily small by choice of initial condition. If the infimum in (2.9) is actually a minimum, then we can choose $\varepsilon = 0$ and achieve zero bias from some nonempty set of initial conditions.

3. Dissipativity, $\mathcal{L}_p$-gain, and robust stability. Throughout this section, we will consider the $\mathcal{L}_p$ spaces for $1 \leqslant p < \infty$; the case $p = \infty$ is also of interest but requires a somewhat different formulation [11]. We say that a trajectory y on a normed linear space $(\mathbb{Y}, |\cdot|)$ belongs to $\mathcal{L}_p$ when the integral in the norm definition

$$\|y\|_p \triangleq \Big(\int_{\mathrm{Dom}(y)} |y(\tau)|^p\, d\tau\Big)^{1/p} \tag{3.1}$$

exists and is finite. If $t_0, t \in \mathrm{Dom}(y)$ with $t_0 \leqslant t$, then we let $y_{[t_0,t]}$ denote the restriction of the trajectory y to the domain $[t_0, t]$, whereas we let $y_{[t_0]}$ denote the restriction of the trajectory y to the domain $[t_0, \infty) \cap \mathrm{Dom}(y)$. We say that y belongs to the extended space $\mathcal{L}_{pe}$ when $y_{[t_0,t]} \in \mathcal{L}_p$ for every $t_0, t \in \mathrm{Dom}(y)$ with $t_0 \leqslant t$. Likewise, we write $y \in \mathcal{L}_p^+$ when $y_{[t_0]} \in \mathcal{L}_p$ for every $t_0 \in \mathrm{Dom}(y)$. Note that if $y \in \mathcal{L}_{pe}$, then $y \in \mathcal{L}_p^+$ if and only if $y_{[t_0]} \in \mathcal{L}_p$ for *some* $t_0 \in \mathrm{Dom}(y)$.

In this section we assume that the signal space $\mathbb{V}$ is the product $\mathbb{V} = \mathbb{E} \times \mathbb{Y} \times \mathbb{X}$ of a normed linear input space $(\mathbb{E}, |\cdot|)$, a normed linear output space $(\mathbb{Y}, |\cdot|)$, and a state space $\mathbb{X}$. We also assume that the trajectory space $\mathfrak{T}(\mathbb{V})$ is the set of all trajectories (e, y, x) such that $e, y \in \mathcal{L}_{pe}$. Given a nonnegative gain parameter γ, we say that a state-space system $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B}, \xi)$ (with ξ being the natural projection of $\mathbb{V}$ onto $\mathbb{X}$) has $\mathcal{L}_p$-gain $\leqslant \gamma$ when there exists a mapping $\beta : \mathfrak{G}(\mathcal{B}) \to \mathbb{R}$ such that, for all trajectories $(e, y, x) \in \mathcal{B}$ and all $t, t_0 \in \mathrm{Dom}(e, y, x)$ with $t \geqslant t_0$, we have

$$\|y_{[t_0,t]}\|_p^p \leqslant \gamma^p \|e_{[t_0,t]}\|_p^p + \beta\big(t_0, x(t_0)\big). \tag{3.2}$$

Here $\beta(\cdot)$ represents a finite bias which depends only on the initial conditions. Note that we have not defined $\mathcal{L}_p$-gain per se, but only the property of having $\mathcal{L}_p$-gain $\leqslant \gamma$

(this will be sufficient for our purposes). We will also say that H has $\mathcal{L}_p$-gain $< \gamma$ whenever it has $\mathcal{L}_p$-gain $\leqslant \gamma_0$ for some $\gamma_0 < \gamma$.

The classical $\mathcal{L}_p$-gain supply rate for $\mathfrak{T}(\mathbb{V})$ is

$$\mathbf{S}(e, y) = \gamma^p |e|^p - |y|^p, \tag{3.3}$$

and it follows from Definition 2.3 that H is dissipative with respect to this supply rate if and only if, for every initial condition $(t_0, x_0) \in \mathfrak{G}(\mathcal{B})$,

$$\inf_{\substack{e,y,x \\ t \geqslant t_0}} \int_{t_0}^{t} \left(\gamma^p |e(\tau)|^p - |y(\tau)|^p\right) d\tau \triangleq -\phi_a(t_0, x_0) > -\infty, \tag{3.4}$$

where the infimum is taken over all trajectories $(e, y, x) \in B$ and all $t \geqslant t_0$ such that $t, t_0 \in \mathrm{Dom}(e, y, x)$ and $\xi(e, y, x)(t_0) = x(t_0) = x_0$. Thus it is clear that H is dissipative with respect to (3.3) if and only if H has $\mathcal{L}_p$-gain $\leqslant \gamma$. Moreover, we can choose the bias as $\beta \equiv \phi_a$, and it follows from Proposition 2.6 that this bias can always be made arbitrarily small by choosing appropriate initial conditions.

We will consider these classical notions of $\mathcal{L}_p$-gain and dissipation throughout this section. To further motivate our analysis, we first recall the example in [6] which demonstrates that the small-gain condition is not necessary for the robust stability of general nonlinear feedback interconnections.

3.1. A motivating example. This example is a single-input-single-output, time-invariant, first-order nonlinear system H which does *not* have an $\mathcal{L}_2$-gain $\leqslant 1$ but for which nevertheless the system obtained by connecting H in feedback with an uncertain system Δ as shown in Figure 3.1 is stable for any Δ having an $\mathcal{L}_2$-gain < 1.

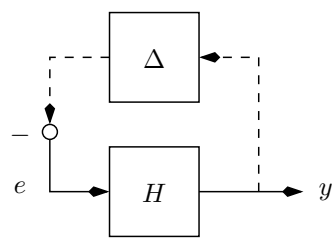


FIG. 3.1. *Performance versus robust stability.*

The system H is described by the following state-space equations:

$$H: \quad \begin{cases} \dot{x} = -\phi'(x) + 2e, & x \in \mathbb{R}, \\ y = \phi'(x) - e, \end{cases} \tag{3.5}$$

where $\phi'(x)$ denotes the derivative of some smooth function $\phi : \mathbb{R} \to \mathbb{R}$. From (3.5), we find that $e^2 - y^2 = \phi'(x)\dot{x}$ and, upon integrating both sides,

$$\phi(x(t)) = \phi(x(t_0)) + \int_{t_0}^{t} \left(e^2(\tau) - y^2(\tau)\right) d\tau, \tag{3.6}$$

which holds along any trajectory of the system H. It follows from Definition 2.4, Proposition 2.5, and the controllability of H that H has $\mathcal{L}_2$-gain $\leqslant 1$ if and only if the function $\phi(\cdot)$ is bounded from below over $\mathbb{R}$, in which case

$$U(x) = \phi(x) - \inf_{x \in \mathbb{R}} \phi(x) \tag{3.7}$$

is a corresponding storage function. Consider now the feedback connection of Figure 3.1, where the (possibly nonlinear, time-varying) uncertain system Δ has an $\mathcal{L}_2$-gain $\leqslant \gamma$ for some $\gamma \in [0,1]$. Let $U_\Delta(t, x_\Delta)$ denote a storage function for Δ which satisfies the corresponding dissipation inequality,

$$U_\Delta(t, x_\Delta(t)) \leqslant U_\Delta(t_0, x_\Delta(t_0)) + \int_{t_0}^{t} \left(\gamma^2 y(\tau)^2 - e(\tau)^2\right) d\tau, \tag{3.8}$$

along trajectories of Δ, where x_Δ denotes the internal state of Δ. We now define a quantity $W(t, x, x_\Delta)$:

$$W(t, x, x_\Delta) = \frac{1}{2}\left(1+\gamma^2\right)\phi(x) + U_\Delta(t, x_\Delta). \tag{3.9}$$

Using (3.6) and (3.8), we see that trajectories of the closed-loop system are such that

$$W(t, x(t), x_\Delta(t)) \leqslant W(t_0, x(t_0), x_\Delta(t_0)) - \frac{1}{2}\left(1-\gamma^2\right) \int_{t_0}^{t} \left(e(\tau)^2 + y(\tau)^2\right) d\tau. \tag{3.10}$$

Because U_Δ is a nonnegative function, it follows from (3.9) and (3.10) that

$$\begin{aligned} \phi(x(t)) &\leqslant \frac{2}{1+\gamma^2}\, W(t_0, x(t_0), x_\Delta(t_0)) - \frac{1-\gamma^2}{1+\gamma^2} \int_{t_0}^{t} \left(e(\tau)^2 + y(\tau)^2\right) d\tau \\ &\leqslant \frac{2}{1+\gamma^2}\, W(t_0, x(t_0), x_\Delta(t_0)) \end{aligned} \tag{3.11}$$

along trajectories of the closed-loop system. In particular, we see from (3.11) that $\phi(x(t))$ is bounded from above in forward time. Now suppose that ϕ satisfies the following:

(i) $\inf \phi = -\infty$ (i.e., ϕ is not bounded from below over $\mathbb{R}$);
(ii) for all $a \in \mathbb{R}$, every connected component of the set $\{\phi \leqslant a\}$ is compact;
(iii) ϕ is positive definite in a neighborhood of zero.

An example of such a function is $\phi(x) = x^2 \cos(x)$. Because the state trajectory $x(t)$ is a continuous function of time, we can conclude from (3.11) and (ii) that $|x(t)|$ and thus also $|\phi(x(t))|$ are bounded in forward time from any initial state of the interconnected system. Furthermore, from (iii), we observe that we have Lyapunov-like stability of the point $x = 0$, namely, for any neighborhood of zero in $\mathbb{R}$, the state $x(t)$ does not leave this neighborhood in forward time provided that the initial state $x(t_0)$ and initial energy $U_\Delta(t_0, x_\Delta(t_0))$ stored in system Δ are sufficiently small. If in addition the gain parameter γ of system Δ is strictly less than one, then we can conclude further from (3.11) that the signals e and y belong to $\mathcal{L}_2^+$. Finally, we know from (i) that the system H does not have an $\mathcal{L}_2$-gain $\leqslant 1$. To summarize, the closed-loop system of Figure 3.1 is robustly stable even though the conditions of the classical small-gain theorem are not satisfied. More precisely, even though H does not have $\mathcal{L}_2$-gain $\leqslant 1$, there does not exist an uncertainty Δ with $\mathcal{L}_2$-gain < 1 which causes the interconnection to exhibit unstable behavior. Even if we allow $\gamma = 1$, we cannot cause the internal state of H to become unbounded or lose the Lyapunov-like stability of its zero equilibrium value. Notice that this does not contradict the known results for linear systems: the Jacobi linearization of H about $x = 0$ indeed has $\mathcal{L}_2$-gain $\leqslant 1$; hence the small-gain condition is satisfied in a neighborhood of the point $x = 0$.

This example demonstrates that for general nonlinear systems the classical small-gain condition is not necessary for robust stability. Our goal is to weaken the definition of gain so that we preserve the sufficiency of the small-gain condition while at the same time recovering its necessity for robust stability.

3.2. Conditional $\mathcal{L}_p$-gain. We have seen that a system has $\mathcal{L}_p$-gain $\leqslant \gamma$ if and only if inequality (3.4) holds for all system trajectories satisfying the initial conditions. To weaken this definition of gain, we require this inequality to hold not for all such system trajectories but only for those that satisfy an auxiliary inequality.

DEFINITION 3.1. *A state-space system $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B}, \xi)$ has* conditional $\mathcal{L}_p$*-gain* $\leqslant \gamma$ *when, for every initial condition $(t_0, x_0) \in \mathfrak{G}(\mathcal{B})$, every $C > 0$, and every $\overline{\gamma} > \gamma$,*

$$\inf_{\substack{e,y,x \\ t \geqslant t_0}} \int_{t_0}^{t} \left(\gamma^p |e(\tau)|^p - |y(\tau)|^p\right) d\tau > -\infty, \tag{3.12}$$

where the infimum is taken over all trajectories $(e, y, x) \in \mathcal{B}$ and all $t \geqslant t_0$ such that $t, t_0 \in \mathrm{Dom}(e, y, x)$, $\xi(e, y, x)(t_0) = x(t_0) = x_0$, and furthermore

$$\sup_{\substack{t \in \mathrm{Dom}(e,y,x) \\ t \geqslant t_0}} \int_{t_0}^{t} \left(\overline{\gamma}^p |e(\tau)|^p - |y(\tau)|^p\right) d\tau < C. \tag{3.13}$$

In other words, a system has conditional $\mathcal{L}_p$-gain when it is not possible to extract arbitrarily large amounts of (abstract) energy from it without pumping arbitrarily large amounts of energy into it in the process. It is clear that any system having $\mathcal{L}_p$-gain $\leqslant \gamma$ also has conditional $\mathcal{L}_p$-gain $\leqslant \gamma$, but the converse need not hold. For example, the system H in (3.5) with $\phi(x) = x^2 \cos(x)$ has conditional $\mathcal{L}_2$-gain $\leqslant 1$ even though the same is not true for the classical, unconditional $\mathcal{L}_2$-gain.

If a system H has $\mathcal{L}_p$-gain $\leqslant \gamma$ for some finite γ, then it follows from (3.2) that H is also $\mathcal{L}_p$-BIBO (bounded-input/bounded-output) stable, namely, that inputs e belonging to $\mathcal{L}_p^+$ generate outputs y which also belong to $\mathcal{L}_p^+$. Systems having only the weaker conditional $\mathcal{L}_p$-gain share the following stability property.

PROPOSITION 3.2. *If a state-space system $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B}, \xi)$ has conditional $\mathcal{L}_p$-gain $\leqslant \gamma$ for some finite γ, then H is $\mathcal{L}_p$-BIBO stable.*

Proof. Fix a trajectory $(e, y, x) \in \mathcal{B}$ such that the input e belongs to $\mathcal{L}_p^+$, choose any $t_0 \in \mathrm{Dom}(e, y, x)$, and choose any $\overline{\gamma} > \gamma$. Then there exists $C > 0$ such that

$$\sup_{\substack{t \in \mathrm{Dom}(e,y,x) \\ t \geqslant t_0}} \int_{t_0}^{t} \overline{\gamma}^p |e(\tau)|^p \, d\tau < C. \tag{3.14}$$

It follows from Definition 3.1 that

$$C - \sup_{\substack{t \in \mathrm{Dom}(e,y,x) \\ t \geqslant t_0}} \int_{t_0}^{t} |y(\tau)|^p \, d\tau > \inf_{\substack{t \in \mathrm{Dom}(e,y,x) \\ t \geqslant t_0}} \int_{t_0}^{t} \left(\gamma^p |e(\tau)|^p - |y(\tau)|^p\right) d\tau > -\infty,$$

from which we conclude that $y_{[t_0]}$ belongs to $\mathcal{L}_p$. ∎

It is of interest to identify classes of systems for which our definitions of $\mathcal{L}_p$-gain and conditional $\mathcal{L}_p$-gain coincide. One such class is characterized by the following result, which employs a resetting property related to the ones used in [19, 21] to prove the necessity of the small-gain theorem for the robust stability of nonlinear systems with fading memory.

THEOREM 3.3. *Let $H = (\mathfrak{T}(\mathbb{V}), \mathcal{B}, \xi)$ be a time-invariant state-space system. Let $Q \subset \mathbb{X}$ be such that H is controllable from Q in the following manner: for every $x_0 \in \mathbb{X}$, there exists $\alpha > 0$ such that, for every $q \in Q$, there exist a trajectory $(e, y, x) \in \mathcal{B}$ and a time interval $[0, t] \subset \mathrm{Dom}(e, y, x)$ such that $x(0) = q$, $x(t) = x_0$,*

and $\|e_{[0,t]}\|_p \leqslant \alpha$. *Suppose also that* H *has the following resetting property with respect to* Q: *there exists* $\beta > 0$ *such that, for all* $x_0 \in \mathbb{X}$, *there exist a trajectory* $(e, y, x) \in \mathcal{B}$ *and a time interval* $[0, t] \subset \mathrm{Dom}(e, y, x)$ *such that* $x(0) = x_0$, $x(t) \in Q$, *and* $\|e_{[0,t]}\|_p \leqslant \beta$. *Then* H *has* $\mathcal{L}_p$*-gain* $\leqslant \gamma$ *if and only if it has conditional* $\mathcal{L}_p$*-gain* $\leqslant \gamma$.

Proof. Suppose that H does not have $\mathcal{L}_p$-gain $\leqslant \gamma$; then there exists $x_0 \in \mathbb{X}$ such that, for any $M > 0$, there exist a trajectory $(e_1, y_1, x_1) \in \mathcal{B}$ and a time interval $[0, t_1] \subset \mathrm{Dom}(e_1, y_1, x_1)$ such that $x_1(0) = x_0$ and

$$\int_0^{t_1} \left(\gamma^p |e_1(\tau)|^p - |y_1(\tau)|^p\right) d\tau \leqslant -M. \tag{3.15}$$

Given this state x_0, we choose α as in the theorem statement and pick $M = \gamma^p \alpha^p + \gamma^p \beta^p + 1$. From the resetting property, there exist a trajectory $(e_2, y_2, x_2) \in \mathcal{B}$ and a time interval $[t_1, t_2] \subset \mathrm{Dom}(e_2, y_2, x_2)$ such that $x_2(t_1) = x_1(t_1)$, $x_2(t_2) \in Q$, and $\|e_{2[t_1,t_2]}\|_p \leqslant \beta$. From the controllability property, there exist a trajectory $(e_3, y_3, x_3) \in \mathcal{B}$ and a time interval $[t_2, t_3] \subset \mathrm{Dom}(e_3, y_3, x_3)$ such that $x_3(t_2) = x_2(t_2)$, $x_3(t_3) = x_0$, and $\|e_{3[t_2,t_3]}\|_p \leqslant \alpha$. We now concatenate these three trajectories using the axiom of state to create a periodic trajectory $(e, y, x) \in \mathcal{B}$ with period t_3 which matches (e_1, y_1, x_1) on the interval $[0, t_1]$, matches (e_2, y_2, x_2) on the interval $[t_1, t_2]$, and matches (e_3, y_3, x_3) on the interval $[t_2, t_3]$. If we now integrate the supply rate over one period, we obtain the bound

$$\int_0^{t_3} \left(\gamma^p |e(\tau)|^p - |y(\tau)|^p\right) d\tau \leqslant -M + \gamma^p \beta^p + \gamma^p \alpha^p = -1. \tag{3.16}$$

If we choose $\overline{\gamma} > \gamma$ such that

$$\left(\overline{\gamma}^p - \gamma^p\right) \int_0^{t_3} |e(\tau)|^p \, d\tau < 1, \tag{3.17}$$

then upon adding inequalities (3.16) and (3.17) we obtain

$$\int_0^{t_3} \left(\overline{\gamma}^p |e(\tau)|^p - |y(\tau)|^p\right) d\tau < 0. \tag{3.18}$$

With C defined as

$$C = 1 + \sup_{0 \leqslant t \leqslant t_3} \int_0^t \left(\overline{\gamma}^p |e(\tau)|^p - |y(\tau)|^p\right) d\tau, \tag{3.19}$$

it follows from (3.18) and (3.19) that

$$\sup_{t \geqslant 0} \int_0^t \left(\overline{\gamma}^p |e(\tau)|^p - |y(\tau)|^p\right) d\tau = \sup_{0 \leqslant t \leqslant t_3} \int_0^t \left(\overline{\gamma}^p |e(\tau)|^p - |y(\tau)|^p\right) d\tau < C, \tag{3.20}$$

but, from (3.16), we have

$$\inf_{t \geqslant 0} \int_0^t \left(\gamma^p |e(\tau)|^p - |y(\tau)|^p\right) d\tau = -\infty. \tag{3.21}$$

We conclude from (3.20), (3.21), and Definition 3.1 that H does not have conditional $\mathcal{L}_p$-gain $\leqslant \gamma$. ∎

We can use Proposition 3.2 and Theorem 3.3 to show that, for linear, time-invariant, finite-dimensional, controllable, and observable state-space systems, the notions of $\mathcal{L}_p$-gain and conditional $\mathcal{L}_p$-gain coincide. Indeed, if such a system H has conditional $\mathcal{L}_p$-gain $\leqslant \gamma$, then we conclude from Proposition 3.2 and the observability of H that all internal modes of H have strictly negative real parts. Hence, with $Q = \{0\}$, the internal stability of H guarantees that the resetting property of Theorem 3.3 holds.

3.3. A conditional small-gain theorem. The behavioral approach allows us to consider feedback interconnections like those of Figure 3.2 without having to impose (or even define) well-posedness. Indeed, let $\mathbb{V}_1 = \mathbb{E} \times \mathbb{Y} \times \mathbb{X}_1$ and $\mathbb{V}_2 = \mathbb{Y} \times \mathbb{E} \times \mathbb{X}_2$, let $\mathfrak{T}(\mathbb{V}_1)$ be the set of all trajectories (e_1, y_1, x_1) such that $e_1, y_1 \in \mathcal{L}_{pe}$, let $\mathfrak{T}(\mathbb{V}_2)$ be the set of all trajectories (e_2, y_2, x_2) such that $e_2, y_2 \in \mathcal{L}_{pe}$, and let ξ_1 and ξ_2 be the natural projections of $\mathbb{V}_1$ and $\mathbb{V}_2$ onto $\mathbb{X}_1$ and $\mathbb{X}_2$, respectively. Now given two state-space systems $H_1 = (\mathfrak{T}(\mathbb{V}_1), \mathcal{B}_1, \xi_1)$ and $H_2 = (\mathfrak{T}(\mathbb{V}_2), \mathcal{B}_2, \xi_2)$, we can define the behavior of their interconnection in Figure 3.2 to be simply the collection of all trajectories $(u_1, u_2, e_1, e_2, y_1, y_2, x_1, x_2)$ satisfying $(e_1, y_1, x_1) \in \mathcal{B}_1$, $(e_2, y_2, x_2) \in \mathcal{B}_2$, and the interconnection constraints $e_1 \equiv u_1 - y_2$ and $e_2 \equiv u_2 + y_1$.

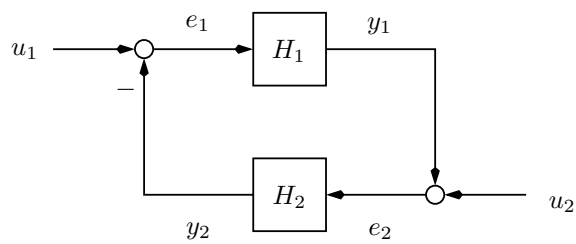


FIG. 3.2. *Canonical feedback interconnection.*

We say that the feedback connection of Figure 3.2 is $\mathcal{L}_p$-stable when, for every initial state condition (t_0, x_{10}, x_{20}) and every input pair $u_1, u_2 \in \mathcal{L}_p^+$, there exists a uniform upper bound on $\|y_{1[t_0]}\|_p$ and $\|y_{2[t_0]}\|_p$ over all trajectories of the interconnection which satisfy the initial conditions $x_1(t_0) = x_{10}$ and $x_2(t_0) = x_{20}$ (in particular, the trajectories e_1, e_2, y_1, and y_2 all belong to $\mathcal{L}_p^+$). Likewise, we say that the feedback connection is $\mathcal{L}_p$-unstable when it is not $\mathcal{L}_p$-stable. In section 4, we will extend our results to more general stability measures.

We will present our new small-gain theorem in two parts, the first one strengthening the classical sufficiency result (see [4], for example) by weakening the assumption on the system H_1, and the second one demonstrating the necessity of conditional gain for robust stability (treating the system H_2 as uncertainty).

THEOREM 3.4 (sufficiency). *If H_1 has conditional $\mathcal{L}_p$-gain $\leqslant \gamma_1$ and H_2 has (unconditional) $\mathcal{L}_p$-gain $\leqslant \gamma_2$, and if $\gamma_1\gamma_2 < 1$, then the feedback connection of Figure* 3.2 *is $\mathcal{L}_p$-stable.*

THEOREM 3.5 (necessity). *If H_1 does not have conditional $\mathcal{L}_p$-gain $\leqslant \gamma_1$, then there exist $\gamma_2 < 1/\gamma_1$ and a time-invariant state-space system H_2 with (unconditional) $\mathcal{L}_p$-gain $\leqslant \gamma_2$ such that the feedback connection of Figure* 3.2 *is $\mathcal{L}_p$-unstable.*

It is significant that the destabilizing system H_2 in Theorem 3.5 can always be chosen as a time-invariant system, even if H_1 is not time-invariant. Moreover, in the proof of this theorem in section 3.5 below, H_2 is constructed as a second-order system, namely, its state space is $\mathbb{X}_2 = \mathbb{R}^2$. Thus, to show that a system H_1 has conditional

gain, it suffices to prove the $\mathcal{L}_p$-stability of the feedback connection of Figure 3.2 for all second-order, time-invariant systems H_2. There is an alternative construction of H_2 which is first-order but time-varying.

3.4. Proof of Theorem 3.4. Fix an initial condition (t_0, x_{10}, x_{20}) and an input pair $u_1, u_2 \in \mathcal{L}_p^+$, and suppose $v = (u_1, u_2, e_1, e_2, y_1, y_2, x_1, x_2)$ is a trajectory of the interconnection. Now H_2 has $\mathcal{L}_p$-gain $\leqslant \gamma_2$, which implies that, for any $t \in \mathrm{Dom}(v)$ with $t \geqslant t_0$, we have

$$\int_{t_0}^{t} |y_2(\tau)|^p \, d\tau \leqslant \gamma_2^p \int_{t_0}^{t} |e_2(\tau)|^p \, d\tau + \beta_2, \tag{3.22}$$

where $\beta_2 = \beta_2(t_0, x_{20})$ is a finite bias. By taking the pth root of both sides of (3.22), we obtain the estimate

$$\|y_{2[t_0,t]}\|_p \leqslant \gamma_2 \|e_{2[t_0,t]}\|_p + \beta_2^{1/p}. \tag{3.23}$$

Because $\gamma_1\gamma_2 < 1$, we can fix $\overline{\gamma}_1 > \gamma_1$ such that $\overline{\gamma}_1\gamma_2 < 1$. We next obtain the following bounds:

$$\begin{aligned} \overline{\gamma}_1^p \|e_{1[t_0,t]}\|_p^p - \|y_{1[t_0,t]}\|_p^p &\leqslant \overline{\gamma}_1^p \big(\|u_{1[t_0,t]}\|_p + \|y_{2[t_0,t]}\|\big)^p - \|(e_2 - u_2)_{[t_0,t]}\|_p^p \\ &\leqslant \overline{\gamma}_1^p (\|u_{1[t_0,t]}\|_p + \gamma_2 \|e_{2[t_0,t]}\|_p + \beta_2^{1/p})^p \\ &\qquad - \|(e_2 - u_2)_{[t_0,t]}\|_p^p. \end{aligned} \tag{3.24}$$

For the case $\|e_{2[t_0,t]}\|_p \leqslant \|u_{2[t_0]}\|_p$, we see from (3.24) that

$$\overline{\gamma}_1^p \|e_{1[t_0,t]}\|_p^p - \|y_{1[t_0,t]}\|_p^p \leqslant \overline{\gamma}_1^p (\|u_{1[t_0]}\|_p + \gamma_2 \|u_{2[t_0]}\|_p + \beta_2^{1/p})^p + 2^p \|u_{2[t_0]}\|_p^p. \tag{3.25}$$

For the case $\|e_{2[t_0,t]}\|_p > \|u_{2[t_0]}\|_p$, it follows from (3.24) that

$$\begin{aligned} \overline{\gamma}_1^p \|e_{1[t_0,t]}\|_p^p - \|y_{1[t_0,t]}\|_p^p &\leqslant \big(\overline{\gamma}_1 \|u_{1[t_0]}\|_p + \overline{\gamma}_1\gamma_2 \|e_{2[t_0,t]}\|_p + \overline{\gamma}_1 \beta_2^{1/p}\big)^p \\ &\qquad - \big(\|e_{2[t_0,t]}\|_p - \|u_{2[t_0]}\|_p\big)^p. \end{aligned} \tag{3.26}$$

Since $\overline{\gamma}_1\gamma_2 < 1$, the right-hand side of (3.26) is bounded from above by a constant which is independent of e_2 or t, but depends only on the fixed quantities t_0, $\|u_{i[t_0]}\|_p$, p, $\overline{\gamma}_1$, γ_2, and x_{20}. The same is true for the right-hand side of (3.25), and we conclude that there exists a constant $C > 0$ depending only on these fixed quantities such that

$$\sup_{\substack{t \in \mathrm{Dom}(v) \\ t \geqslant t_0}} \int_{t_0}^{t} (\overline{\gamma}_1^p |e_1(\tau)|^p - |y_1(\tau)|^p) \, d\tau < C. \tag{3.27}$$

It now follows from Definition 3.1 that there exists a constant $\delta > 0$, depending only on the fixed quantities t_0, $\|u_{i[t_0]}\|_p$, p, $\overline{\gamma}_1$, γ_2, and x_{i0}, such that

$$\|y_{1[t_0,t]}\|_p \leqslant \gamma_1 \|e_{1[t_0,t]}\|_p + \delta^{1/p}. \tag{3.28}$$

Proceeding from (3.23) and (3.28), the rest of the proof is virtually identical to the proof of the classical small-gain theorem as presented in [4, p. 41]. For example, we obtain a bound on $\|y_{1[t_0]}\|_p$ as follows:

$$\begin{aligned} \|y_{1[t_0,t]}\|_p &\leqslant \gamma_1 \|u_{1[t_0,t]}\|_p + \gamma_1 \|y_{2[t_0,t]}\|_p + \delta^{1/p} \\ &\leqslant \gamma_1 \|u_{1[t_0,t]}\|_p + \gamma_1\gamma_2 \|e_{2[t_0,t]}\|_p + \gamma_1 \beta_2^{1/p} + \delta^{1/p} \\ &\leqslant \gamma_1 \|u_{1[t_0]}\|_p + \gamma_1\gamma_2 \|u_{2[t_0]}\|_p + \gamma_1\gamma_2 \|y_{1[t_0,t]}\|_p + \gamma_1 \beta_2^{1/p} + \delta^{1/p}. \end{aligned} \tag{3.29}$$

Upon collecting the $\|y_{1[t_0,t]}\|_p$ terms on the right-hand side, we obtain the bound

$$\|y_{1[t_0,t]}\|_p \leqslant \frac{1}{1-\gamma_1\gamma_2}\Big(\gamma_1\|u_{1[t_0]}\|_p + \gamma_1\gamma_2\|u_{2[t_0]}\|_p + \gamma_1\beta_2^{1/p} + \delta^{1/p}\Big). \tag{3.30}$$

The left-hand side of (3.30) is independent of t, so this provides a bound on $\|y_{1[t_0]}\|_p$ as desired. The calculation for $\|y_{2[t_0]}\|_p$ is analogous.

3.5. Proof of Theorem 3.5. If $H_1 = (\mathfrak{T}(\mathbb{V}_1), \mathcal{B}_1, \xi_1)$ does not have conditional $\mathcal{L}_p$-gain $\leqslant \gamma_1$, then there exist an initial condition $(t_0, x_{10}) \in \mathfrak{G}(\mathcal{B}_1)$, constants $C_1 > 0$ and $\overline{\gamma}_1 > \gamma_1$, a sequence of trajectories $(e_{1n}, y_{1n}, x_{1n}) \in \mathcal{B}_1$ with $t_0 \in \mathrm{Dom}(e_{1n}, y_{1n}, x_{1n})$ and $x_{1n}(t_0) = x_{10}$ for each $n \geqslant 1$, and a sequence of times $t_n \in \mathrm{Dom}(e_{1n}, y_{1n}, x_{1n})$ with $t_n \geqslant t_0$ such that

$$\int_{t_0}^{t_n} \left(\gamma_1^p|e_{1n}(\tau)|^p - |y_{1n}(\tau)|^p\right) d\tau < -n \tag{3.31}$$

and

$$\sup_{\substack{t\in\mathrm{Dom}(e_{1n},y_{1n},x_{1n})\\ t\geqslant t_0}} \int_{t_0}^{t} \left(\overline{\gamma}_1^p|e_{1n}(\tau)|^p - |y_{1n}(\tau)|^p\right) d\tau < C_1 \tag{3.32}$$

for each $n \geqslant 1$. From the axiom of state, the concatenations

$$(e_{11}, y_{11}, x_{11}) \wedge^{t_0} (e_{1n}, y_{1n}, x_{1n})$$

also belong to $\mathcal{B}_1$ and clearly satisfy (3.31)–(3.32), so we may assume without loss of generality that these trajectories are identical before time t_0, namely, that

$$\mathrm{Dom}(e_{1n}, y_{1n}, x_{1n}) \cap (-\infty, t_0) = \mathrm{Dom}(e_{11}, y_{11}, x_{11}) \cap (-\infty, t_0)$$

and

$$(e_{1n}, y_{1n}, x_{1n})(t) = (e_{11}, y_{11}, x_{11})(t)$$

for all $n \geqslant 1$ and all $t \in \mathrm{Dom}(e_{11}, y_{11}, x_{11}) \cap (-\infty, t_0)$. For each $n \geqslant 1$, define

$$e_{2n} = y_{1n}, \qquad y_{2n} = -e_{1n},$$

and

$$x_{2n}(t) = \left(\tanh(t-t_0),\ \frac{1}{n}\tanh(t-t_0)1(t-t_0)\right) \in \mathbb{R}^2 \tag{3.33}$$

for $t \in \mathrm{Dom}(e_{1n}, y_{1n}, x_{1n})$, where $1(s)$ is the unit step function which is 1 when $s \geqslant 0$ and 0 when $s < 0$. By construction, the trajectories (e_{2n}, y_{2n}, x_{2n}) for $n \geqslant 1$ are identical before time t_0, so the concatenation of two of these trajectories at a time before t_0 does not yield a new trajectory. We let $H_2 = (\mathfrak{T}(\mathbb{V}_2), \mathcal{B}_2, \xi_2)$, where $\mathcal{B}_2$ is the collection of all time-shifted versions of these trajectories:

$$\mathcal{B}_2 = \big\{\sigma^t(e_{2n}, y_{2n}, x_{2n}) \ :\ t \in \mathbb{R},\, n \geqslant 1\big\}. \tag{3.34}$$

Clearly, H_2 is time-invariant. To verify the axiom of state, suppose that there exist $t_1, t_2 \in \mathbb{R}$ and $n, m \geqslant 1$ such that $(\sigma^{t_1}x_{2n})(t) = (\sigma^{t_2}x_{2m})(t)$ for some $t \in \mathbb{R}$. By

matching the first state coordinates, we obtain $\tanh(t+t_1-t_0)=\tanh(t+t_2-t_0)$, which implies $t_1=t_2$. By matching the second state coordinates, we then obtain $\frac{1}{n}\tanh(t+t_1-t_0)1(t+t_1-t_0)=\frac{1}{m}\tanh(t+t_1-t_0)1(t+t_1-t_0)$, which implies either $t+t_1\leqslant t_0$ or $n=m$. In either case, it is clear that the concatenation $\sigma^{t_1}(e_{2n},y_{2n},x_{2n})\wedge^t\sigma^{t_1}(e_{2m},y_{2m},x_{2m})$ also belongs to B_2.

We next verify that H_2 has $\mathcal{L}_p$-gain less than or equal to $\gamma_2\triangleq 1/\overline{\gamma}_1$ (which clearly satisfies $\gamma_1\gamma_2<1$). Fix $x_{20}=(a_0,b_0)\in\mathbb{R}^2$ such that $(\sigma^t x_{2n})(0)=x_{20}$ for some $t\in\mathbb{R}$ and some $n\geqslant 1$. Then we have $t=t_0+\tanh^{-1}(a_0)$ and, if $t>t_0$, then also $n=a_0/b_0$. Fix $s\geqslant 0$ and suppose first that $s+t\leqslant t_0$. Because the trajectories (e_{2n},y_{2n},x_{2n}) for $n\geqslant 1$ are identical before time t_0, we have

(3.35)
$$\begin{aligned}\int_0^s\left(\gamma_2^p|e_{2n}(\tau+t)|^p-|y_{2n}(\tau+t)|^p\right)d\tau&=\int_t^{s+t}\left(\gamma_2^p|e_{2n}(\tau)|^p-|y_{2n}(\tau)|^p\right)d\tau\\&\geqslant\min_{t\leqslant r\leqslant t_0}\int_t^r\left(\gamma_2^p|e_{21}(\tau)|^p-|y_{21}(\tau)|^p\right)d\tau,\end{aligned}$$

where the minimum exists because $t,t_0\in\mathrm{Dom}(e_{21},y_{21},x_{21})$, and the input and output trajectories belong to $\mathcal{L}_{pe}$ on their domains. Suppose next that $t\leqslant t_0<s+t$: in this case, we use (3.32) to obtain

(3.36)
$$\begin{aligned}\int_0^s\left(\gamma_2^p|e_{2n}(\tau+t)|^p-|y_{2n}(\tau+t)|^p\right)d\tau&=\int_t^{t_0}\left(\gamma_2^p|e_{2n}(\tau)|^p-|y_{2n}(\tau)|^p\right)d\tau\\&\quad+\int_{t_0}^{s+t}\left(\gamma_2^p|e_{2n}(\tau)|^p-|y_{2n}(\tau)|^p\right)d\tau\\&=\int_t^{t_0}\left(\gamma_2^p|e_{2n}(\tau)|^p-|y_{2n}(\tau)|^p\right)d\tau\\&\quad-\gamma_2^p\int_{t_0}^{s+t}\left(\bar{\gamma}_1^p|e_{1n}(\tau)|^p-|y_{1n}(\tau)|^p\right)d\tau\\&\geqslant\int_t^{t_0}\left(\gamma_2^p|e_{21}(\tau)|^p-|y_{21}(\tau)|^p\right)d\tau-\gamma_2^pC_1.\end{aligned}$$

Finally, if $t>t_0$, then $n=a_0/b_0$ and we obtain again from (3.32) that

(3.37)
$$\begin{aligned}\int_0^s\left(\gamma_2^p|e_{2n}(\tau+t)|^p-|y_{2n}(\tau+t)|^p\right)d\tau&=-\int_{t_0}^{t}\left(\gamma_2^p|e_{2n}(\tau)|^p-|y_{2n}(\tau)|^p\right)d\tau\\&\quad+\int_{t_0}^{s+t}\left(\gamma_2^p|e_{2n}(\tau)|^p-|y_{2n}(\tau)|^p\right)d\tau\\&\geqslant-\int_{t_0}^{t}\left(\gamma_2^p|e_{2n}(\tau)|^p-|y_{2n}(\tau)|^p\right)d\tau-\gamma_2^pC_1.\end{aligned}$$

In all three cases, we obtain finite lower bounds (3.35), (3.36), and (3.37) on the integral of the $\mathcal{L}_p$-gain supply rate which are independent of both n and the upper integration limit s (the lower bound (3.37) appears to depend on n but, in this third case, n is fixed at $n=a_0/b_0$). We conclude that H_2 has $\mathcal{L}_p$-gain $\leqslant\gamma_2$.

Finally, we need to show that the feedback connection of Figure 3.2 is $\mathcal{L}_p$-unstable. Choose inputs $u_1 \equiv 0$ and $u_2 \equiv 0$; then, by construction, for each $n \geqslant 1$, the trajectory $(u_1, u_2, e_{1n}, e_{2n}, y_{1n}, y_{2n}, x_{1n}, x_{2n})$ belongs to the behavior of the interconnection and satisfies the initial condition $(x_{1n}(t_0), x_{2n}(t_0)) = (x_{10}, 0)$. It follows from (3.31) that, for this initial condition, there is no upper bound on the $\mathcal{L}_p$-norms $\|y_{1n[t_0]}\|_p$.

4. Conditional dissipativity. In this section we extend our results to general performance measures characterized by dissipation inequalities, $\mathcal{L}_p$-gain being but one such measure. We will consider the general interconnection G of systems G_1 and G_2 as illustrated in Figure 4.1. Here the interconnection signal w might contain components labeled as inputs and outputs of G_1 and G_2; however, in this behavioral context, there is no need to label them as such. To simplify the sufficiency part of our result, we will assume that the interconnection G is not driven by exogenous signals such as the inputs u_1 and u_2 of Figure 3.2. To simplify the necessity part, we will assume throughout this section that all supply rates are time-invariant, namely, that they do not depend explicitly on the time variable.

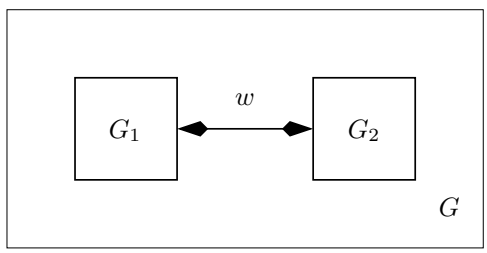


FIG. 4.1. *Behavioral interconnection.*

We first make precise our notion of interconnection in Figure 4.1. Let $\mathbb{W}$ be a nonempty signal set, define $\mathbb{V}_1 \triangleq \mathbb{W} \times \mathbb{X}_1$ and $\mathbb{V}_2 \triangleq \mathbb{W} \times \mathbb{X}_2$ for some nonempty state spaces $\mathbb{X}_1$ and $\mathbb{X}_2$, and let ξ_1 and ξ_2 represent the respective projections of $\mathbb{V}_1$ and $\mathbb{V}_2$ onto $\mathbb{X}_1$ and $\mathbb{X}_2$. Given trajectory spaces $\mathfrak{T}(\mathbb{V}_1)$ and $\mathfrak{T}(\mathbb{V}_2)$, we define $\mathbb{V} = \mathbb{W} \times \mathbb{X}_1 \times \mathbb{X}_2$, we let $\mathfrak{T}(\mathbb{V})$ denote the collection of trajectories (w, x_1, x_2) such that $(w, x_1) \in \mathfrak{T}(\mathbb{V}_1)$ and $(w, x_2) \in \mathfrak{T}(\mathbb{V}_2)$, and we let ξ denote the projection of $\mathbb{V}$ onto $\mathbb{X}_1 \times \mathbb{X}_2$. Finally, we let $\mathfrak{T}(\mathbb{W})$ denote the appropriate projection of $\mathfrak{T}(\mathbb{V})$, namely, $w_1 \in \mathfrak{T}(\mathbb{W})$ whenever there exists $(w, x_1, x_2) \in \mathfrak{T}(\mathbb{V})$ such that $w \equiv w_1$. Note that if $\mathbf{S}$ is a supply rate for $\mathfrak{T}(\mathbb{W})$, then, by simple extension, we may also consider $\mathbf{S}$ to be a supply rate for $\mathfrak{T}(\mathbb{V}_1)$, $\mathfrak{T}(\mathbb{V}_2)$, or $\mathfrak{T}(\mathbb{V})$. Now, given state-space systems $G_1 = (\mathfrak{T}(\mathbb{V}_1), \mathcal{B}_1, \xi_1)$ and $G_2 = (\mathfrak{T}(\mathbb{V}_2), \mathcal{B}_2, \xi_2)$, we define their intersection $G = G_1 \cap G_2$ to be the system $G = (\mathfrak{T}(\mathbb{V}), \mathcal{B}, \xi)$ with behavior

$$\mathcal{B} = \big\{(w, x_1, x_2) \in \mathfrak{T}(\mathbb{V}) \ : \ (w, x_1) \in \mathcal{B}_1 \text{ and } (w, x_2) \in \mathcal{B}_2\big\}. \tag{4.1}$$

It is straightforward to verify that G satisfies the axiom of state. We will describe the desired performance of the interconnection G by means of a supply rate N for $\mathfrak{T}(\mathbb{W})$. For example, if $\mathbb{W}$ is a normed linear space and we are interested in the $\mathcal{L}_p$-stability, we would choose $N(w) = -|w|^p$ (so that G is dissipative with respect to N if and only if $\|w_{[t_0]}\|_p$ is bounded by a constant depending only on t_0 and the initial states).

DEFINITION 4.1. *Given a supply rate N for $\mathfrak{T}(\mathbb{W})$, we say that supply rates $\mathbf{S}_1$ and $\mathbf{S}_2$ for $\mathfrak{T}(\mathbb{W})$ are* N-complementary *when there exist strictly positive constants a,*

b, c, and d such that

$$a\mathbf{S}_1(w) + b\mathbf{S}_2(w) \leqslant N(w) \leqslant c\mathbf{S}_1(w) - d\mathbf{S}_2(w) \tag{4.2}$$

for all $w \in \mathbb{W}$.

For the $\mathcal{L}_p$-gain scenario considered in section 3, we would assume $\mathbb{W} = \mathbb{E} \times \mathbb{Y}$ for normed linear spaces $\mathbb{E}$ and $\mathbb{Y}$, and we would choose the following:

$$N(w) = -|e|^p - |y|^p, \tag{4.3}$$

$$\mathbf{S}_1(w) = \gamma^p|e|^p - |y|^p, \tag{4.4}$$

$$\mathbf{S}_2(w) = -\overline{\gamma}^p|e|^p + |y|^p \tag{4.5}$$

for nonnegative constants γ and $\overline{\gamma}$. In this case, it is straightforward to show that if $\gamma < \overline{\gamma}$, then $\mathbf{S}_1$ and $\mathbf{S}_2$ are N-complementary with $c = d = 1/2$ and

$$a = \frac{\overline{\gamma}^p + 1}{\overline{\gamma}^p - \gamma^p}, \qquad b = \frac{\gamma^p + 1}{\overline{\gamma}^p - \gamma^p}. \tag{4.6}$$

At this point the classical sufficiency of the small-gain condition can be phrased as follows: if G_1 is dissipative with respect to $\mathbf{S}_1$ and G_2 is dissipative with respect to $\mathbf{S}_2$, then if $\gamma < \overline{\gamma}$, the interconnection G is dissipative with respect to N. We will use such phrasing in our general sufficiency result below.

Another common scenario involves *passivity*: for example, suppose $\mathbb{W} = \mathbb{E} \times \mathbb{Y}$ for some real inner product space $\mathbb{E} = \mathbb{Y}$, and consider the following:

$$N(w) = -\langle e, e\rangle - \langle y, y\rangle, \tag{4.7}$$

$$\mathbf{S}_1(w) = \langle e, y\rangle - \varepsilon\langle e, e\rangle, \tag{4.8}$$

$$\mathbf{S}_2(w) = -\langle e, y\rangle - \delta\langle y, y\rangle \tag{4.9}$$

for constants ε and δ. If ε and δ are both strictly positive, then $\mathbf{S}_1$ and $\mathbf{S}_2$ are N-complementary with

$$a = b = \frac{1}{\varepsilon} + \frac{1}{\delta}, \qquad c = d = \frac{1}{1+\varepsilon}. \tag{4.10}$$

Here the classical result states that the feedback connection of two input (or output) strictly passive systems is $\mathcal{L}_2$-stable [14, Lemma 6.8]. In other words, if G_1 is dissipative with respect to $\mathbf{S}_1$ and G_2 is dissipative with respect to $\mathbf{S}_2$, then the interconnection G is dissipative with respect to N.

To strengthen these classical sufficiency results, we introduce the following conditional version of dissipativity.

DEFINITION 4.2. *A state-space system* $G_1 = (\mathfrak{T}(\mathbb{V}_1), \mathcal{B}_1, \xi_1)$ *is* N-conditionally dissipative *with respect to a supply rate* $\mathbf{S}_1$ *when, for every initial* $(t_0, x_{10}) \in \mathfrak{G}(\mathcal{B}_1)$, *every* $C > 0$, *and every supply rate* $\mathbf{S}_2$ *which is* N*-complementary to* $\mathbf{S}_1$,

$$\inf_{\substack{w,x_1 \\ t \geqslant t_0}} \int_{t_0}^{t} \mathbf{S}_1(w(\tau))\, d\tau > -\infty, \tag{4.11}$$

where the infimum is taken over all trajectories $(w, x_1) \in \mathcal{B}_1$ *and all* $t \geqslant t_0$ *such that* $t, t_0 \in \mathrm{Dom}(w, x_1)$, $x_1(t_0) = x_{10}$, *and furthermore*

$$\inf_{\substack{t \in \mathrm{Dom}(w,x_1) \\ t \geqslant t_0}} \int_{t_0}^{t} \mathbf{S}_2(w(\tau))\, d\tau > -C. \tag{4.12}$$

THEOREM 4.3 (sufficiency). *If G_1 is N-conditionally dissipative with respect to $\mathbf{S}_1$, if G_2 is dissipative with respect to $\mathbf{S}_2$, and if $\mathbf{S}_1$ and $\mathbf{S}_2$ are N-complementary, then the intersection $G = G_1 \cap G_2$ is dissipative with respect to N.*

Proof. Fix $(t_0, x_{10}, x_{20}) \in \mathfrak{G}(\mathcal{B})$, and let $(w, x_1, x_2) \in \mathcal{B}$ be such that $x_1(t_0) = x_{10}$ and $x_2(t_0) = x_{20}$. Because G_2 is dissipative with respect to $\mathbf{S}_2$, there exists $C_2 > 0$, depending only on the initial condition (t_0, x_{20}), such that

$$\inf_{\substack{t \in \mathrm{Dom}(w) \\ t \geqslant t_0}} \int_{t_0}^{t} \mathbf{S}_2(w(\tau))\, d\tau > -C_2. \tag{4.13}$$

By the N-conditional dissipativity of G_1, there exists $C_1 > 0$, depending only on the initial condition (t_0, x_{10}, x_{20}), such that

$$\inf_{\substack{t \in \mathrm{Dom}(w) \\ t \geqslant t_0}} \int_{t_0}^{t} \mathbf{S}_1(w(\tau))\, d\tau > -C_1. \tag{4.14}$$

It now follows from (4.2), (4.13), and (4.14) that

$$\inf_{\substack{t \in \mathrm{Dom}(w) \\ t \geqslant t_0}} \int_{t_0}^{t} N(w(\tau))\, d\tau > -aC_1 - bC_2, \tag{4.15}$$

and we conclude that G is dissipative with respect to N. ∎

As before, the use of conditional dissipativity also allows us to recover the corresponding necessity result.

THEOREM 4.4 (necessity). *If G_1 is not N-conditionally dissipative with respect to $\mathbf{S}_1$, then there exists a supply rate $\mathbf{S}_2$ which is N-complementary to $\mathbf{S}_1$ and a time-invariant state-space system G_2 which is dissipative with respect to $\mathbf{S}_2$ such that the intersection $G = G_1 \cap G_2$ is not dissipative with respect to N.*

Proof. Negating Definition 4.2, there exist an initial condition $(t_0, x_{10}) \in \mathfrak{G}(\mathcal{B}_1)$, a constant $C_1 > 0$, a supply rate $\mathbf{S}_2$ which is N-complementary to $\mathbf{S}_1$, a sequence of trajectories $(w_n, x_{1n}) \in \mathcal{B}_1$ with $t_0 \in \mathrm{Dom}(w_n, x_{1n})$ and $x_{1n}(t_0) = x_{10}$ for each $n \geqslant 1$, and a sequence of times $t_n \in \mathrm{Dom}(w_n, x_{1n})$ with $t_n \geqslant t_0$ such that

$$\int_{t_0}^{t_n} \mathbf{S}_1(w_n(\tau))\, d\tau < -n \tag{4.16}$$

and

$$\inf_{\substack{t \in \mathrm{Dom}(w_n, x_{1n}) \\ t \geqslant t_0}} \int_{t_0}^{t} \mathbf{S}_2(w_n(\tau))\, d\tau > -C_1 \tag{4.17}$$

for each $n \geqslant 1$. As argued in section 3.5, we may assume without loss of generality that these trajectories are identical before time t_0. To construct G_2, we define a sequence of state trajectories x_{2n} as in (3.33), and we let the behavior $\mathcal{B}_2$ of G_2 be

$$\mathcal{B}_2 = \big\{\sigma^t(w_n, x_{2n}) \;:\; t \in \mathbb{R},\, n \geqslant 1\big\}. \tag{4.18}$$

Using arguments similar to those in section 3.5, we can show that G_2 is time-invariant, satisfies the axiom of state, and is (unconditionally) dissipative with respect to the supply rate $\mathbf{S}_2$. Furthermore, we conclude from (4.2), (4.16), and (4.17) that

$$\int_{t_0}^{t_n} N(w_n(\tau))\, d\tau \leqslant c \int_{t_0}^{t_n} \mathbf{S}_1(w_n(\tau))\, d\tau - d \int_{t_0}^{t_n} \mathbf{S}_2(w_n(\tau))\, d\tau \leqslant -cn + dC_1 \tag{4.19}$$

for each $n \geqslant 1$. Because the trajectories (w_n, x_{1n}, x_{2n}) belong to the behavior (4.1) of the intersection G, we conclude that G is not dissipative with respect to N. □

An example application of Theorem 4.4 is the following. Suppose we can prove that the feedback connection of a system G_1 with any input and output strictly passive system G_2 is $\mathcal{L}_2$-stable. Then we cannot conclude in general that G_1 is a passive system, but Theorem 4.4 guarantees that G_1 is conditionally passive.

5. Conclusion. It was shown in [6] that the classical definition of $\mathcal{L}_p$-gain is too strong to provide an equivalence between performance and robust stability. We have presented a weaker notion of gain which recovers the necessity while preserving the sufficiency of the small-gain condition. In the necessity proof, the uncertainty can be chosen to be second-order and time-invariant, even when the plant is time-varying and/or infinite-dimensional. We have also shown that these two notions of gain are equivalent for systems having a resetting property (such as fading-memory systems or linear time-invariant systems). Finally, we extended our results to more general dissipation performance measures.

REFERENCES

[1] B. D. O. Anderson, *The small-gain theorem, the passivity theorem, and their equivalence*, J. Franklin Inst., 298 (1972), pp. 105–115.

[2] M. J. Chen and C. A. Desoer, *Necessary and sufficient condition for robust stability of linear distributed feedback systems*, Internat. J. Control, 35 (1982), pp. 255–267.

[3] M. A. Dahleh and Y. Ohta, *A necessary and sufficient condition for robust BIBO stability*, Systems Control Lett., 11 (1988), pp. 271–275.

[4] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

[5] J. C. Doyle and G. Stein, *Multivariable feedback design: Concepts for a classical/modern synthesis*, IEEE Trans. Automat. Control, 26 (1981), pp. 4–16.

[6] R. A. Freeman, *On the necessity of the small-gain theorem in the performance analysis of nonlinear systems*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 51–56.

[7] J. M. Gonçalves and M. A. Dahleh, *Necessary condition for robust stability of a class of nonlinear systems*, Automatica J. IFAC, 34 (1998), pp. 705–714.

[8] D. J. Hill, *Dissipative nonlinear systems: Basic properties and stability analysis*, in Proceedings of the 31st IEEE Conference on Decision and Control, Tucson, AZ, 1992, pp. 3259–3264.

[9] D. J. Hill and P. J. Moylan, *Dissipative dynamical systems: Basic input-output and state properties*, J. Franklin Inst., 309 (1980), pp. 327–357.

[10] D. J. Hill and P. J. Moylan, *General instability results for interconnected systems*, SIAM J. Control Optim., 21 (1983), pp. 256–279.

[11] S. Huang and M. R. James, *ℓ_∞-bounded robustness for nonlinear systems: Analysis and synthesis*, IEEE Trans. Automat. Control, 48 (2003), pp. 1875–1891.

[12] Z.-P. Jiang, I. Mareels, and Y. Wang, *A Lyapunov formulation of nonlinear small gain theorem for interconnected systems*, in Proceedings of the IFAC Nonlinear Control Systems Design Symposium, Tahoe City, CA, 1995, pp. 666–671.

[13] Z.-P. Jiang, A. R. Teel, and L. Praly, *Small-gain theorem for ISS systems and applications*, Math. Control Signals Systems, 7 (1994), pp. 95–120.

[14] H. K. Khalil, *Nonlinear Systems*, 3rd ed., Prentice–Hall, Upper Saddle River, NJ, 2002.

[15] I. M. Y. Mareels and D. J. Hill, *Monotone stability of nonlinear feedback systems*, J. Math. Systems Estim. Control, 2 (1992), pp. 275–291.

[16] A. Megretski and A. Rantzer, *System analysis via integral quadratic constraints*, IEEE Trans. Automat. Control, 42 (1997), pp. 819–830.

[17] P. J. Moylan and D. J. Hill, *Tests for stability and instability of interconnected system*, IEEE Trans. Automat. Control, 24 (1979), pp. 574–575.

[18] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, Springer-Verlag, New York, 1998.

[19] J. S. Shamma, *The necessary of the small-gain theorem for time-varying and nonlinear systems*,

IEEE Trans. Automat. Control, 36 (1991), pp. 1138–1147.

[20] J. S. Shamma and M. A. Dahleh, *Time-varying versus time-invariant compensation for rejection of persistent bounded disturbances and robust stabilization*, IEEE Trans. Automat. Control, 36 (1991), pp. 838–847.

[21] J. S. Shamma and R. Zhao, *Fading-memory feedback systems and robust stability*, Automatica J. IFAC, 29 (1993), pp. 191–200.

[22] H. L. Tentelman and J. C. Willems, *Every storage function is a stable function*, Systems Control Lett., 32 (1997), pp. 249–259.

[23] M. Vidyasagar, *Input-Output Analysis of Large-Scale Interconnected Systems*, Lecture Notes in Control and Inform. Sci. 29, Springer-Verlag, Berlin, 1981.

[24] J. C. Willems, *Stability, instability, invertibility, and causality*, SIAM J. Control, 7 (1969), pp. 645–671.

[25] J. C. Willems, *Dissipative dynamical systems* I: *General theory*, Arch. Rational Mech. Anal., 45 (1972), pp. 321–351.

[26] J. C. Willems, *Mechanisms for the stability and instability in feedback systems*, Proc. IEEE, 64 (1976), pp. 24–35.

[27] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

[28] G. Zames, *On the input-output stability of time-varying nonlinear feedback systems* I: *Conditions using concepts of loop gain, conicity, and positivity*, IEEE Trans. Automat. Control, 11 (1966), pp. 228–238.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2013–2030

ON WORST-CASE PORTFOLIO OPTIMIZATION*

RALF KORN† AND MOGENS STEFFENSEN‡

Abstract. We formulate a worst-case portfolio optimization problem that technically appears as a game where the investor chooses a portfolio and his opponent, the market, chooses some market crashes. The asymmetry of the opponents' decision processes leads to a new and delicate generalization of the classical Hamilton–Jacobi–Bellman equation in stochastic control. We characterize the optimal controls in general and specify them further in the cases of Hara, logarithmic, and exponential utilities of the investor.

Key words. Hamilton–Jacobi–Bellman equation, stochastic differential games, market crash, utility optimization

AMS subject classifications. 93E20, 91B28, 60H30

DOI. 10.1137/060657145

1. Introduction. The problem of finding an optimal investment strategy for an investor with a given utility function and a fixed initial endowment—the so-called portfolio optimization problem—is one of the classical problems in financial mathematics and its applications in insurance mathematics. The corresponding modern continuous-time approach was pioneered by Merton [8, 9], who applied classical stochastic control methods to reduce the portfolio problem to a matter of solving a Hamilton–Jacobi–Bellman partial differential equation (HJB equation).

Since Merton's pioneering work, many attempts have been made to solve the portfolio optimization problem in a framework that allows for more realistic models of stock prices, in particular for models that can explain large price movements. Examples where portfolio optimization problems are treated in more general settings are portfolio optimization in jump-diffusion models (see, e.g., Aase [1]) or in a general semimartingale framework (see, e.g., Kramkov and Schachermayer [7]).

A different portfolio problem that includes dramatic negative changes of the stock prices (so-called crashes) has been introduced by Korn and Wilmott [6]. Their main idea consists of two aspects: the separation between normal times where the stock price behaves as a geometric Brownian motion and crash times where it jumps downwards and the introduction of a worst-case functional that resembles the form of a game-theoretic max-min approach. While in Korn and Wilmott [6] the problem is solved only for the choice of the logarithmic utility function, a more general problem of the worst-case form is treated in Korn and Menkens [5]. At first sight, their approach seems to be an approach similar to the HJB-equation approach of stochastic control. However, their arguments are based on equilibrium and indifference considerations, and they derive differential equations for the optimal portfolio processes and not for the value function. Even more, they could prove optimality of their proposed portfolio processes only within the class of (piecewise) deterministic control strategies.

*Received by the editors April 13, 2006; accepted for publication (in revised form) June 29, 2007; published electronically November 30, 2007.
http://www.siam.org/journals/sicon/46-6/65714.html
†Fachbereich Mathematik, Universität Kaiserslautern and Fraunhofer Institut ITWM, 67653 Kaiserslautern, Germany (korn@mathematik.uni-kl.de).
‡Laboratory of Actuarial Mathematics, Department of Applied Mathematics and Statistics, University of Copenhagen, Denmark (mogens@math.ku.dk).

The main purpose of this paper is to put the worst-case portfolio optimization in the HJB-equation framework and thus to connect it with the mainstream of stochastic control theory. This leads us to a type of continuous-time game problem that, to our knowledge, is new in control theory. It is the asymmetry of the opponents' decision processes that makes the game so interesting and challenging from a control theoretical point of view: The investor decides the portfolio process, whereas the opponent, the market, decides when the stock market crashes. One could fear that this very asymmetry prevents solutions to the game problem, but we show that the problem, indeed, has a solution and that the solution can be characterized by a generalized HJB equation.

A related financial game problem is approached by Talay and Zheng [10]. They solve a problem where the opponent of the investor, the market, decides the parameters in a diffusion model. Thus, the idea of seeing the market as an opponent is exactly the same as ours. But since their price processes are continuous, the decisions of both the investor and the market affect only the coefficients of the continuous portfolio process. Therefore, from a financial modeling point of view their problem is completely different from ours.

We do not necessarily believe that the investment decision is a part of a game really played in the sense that the market really tries to hurt the investor. But, on the other hand, even a rational investor could choose to invest as if this were really the case. This is exactly what comes out of basing decision on a worst-case scenario. Thus, when we use the word game throughout, this is mainly because the technical formulation of the worst-case investment problem conforms with the general idea of a stochastic differential game.

Our main result is a verification theorem asserting that a system consisting of an HJB-type inequality, a relation between value functions before and after an action of the market, final conditions, and a complementarity condition determine the value function. This result and its consequences are highlighted by some explicit examples. Note in particular that we do not have to restrict ourselves to (piecewise) deterministic controls. Therefore, the existing work on worst-case portfolio optimization is substantially generalized.

Even more, the explicit form of the optimal portfolio strategy that results from our worst-case approach closely resembles the form of strategies that—in real life applications—are suggested for use in pension savings plans. Thus, our model can also be seen as a theoretical justification of a market practice.

One can imagine other areas where the structure of the problem and its solution can find applications. For example, an insurance company decides on reinsurance against large claims, e.g., triggered by a storm or earthquakes. In its battle against the merciless Mother Earth, the company could adopt a worst-case basis for making certain decisions, e.g., the extent of reinsurance protection. The structure of the HJB equation for such a problem is similar, and the relative explicit characterization of the optimal decision obtained here holds out every promise of success in other areas.

2. The model and the preferences. Take as given a probability space $(\Omega, \mathcal{F}, P)$. Let W be a standard Brownian motion defined on this probability space. Let us consider an agent over a fixed time interval $[0, T]$. At time 0 the agent is endowed with initial wealth x_0, and his problem is to allocate investments over the given time horizon. We assume that the agent's investment opportunities are given by the

following financial market:

$$\begin{aligned} dB(t) &= rB(t)\,dt, \\ B(0) &= 1, \\ dS(t) &= S(t-)\left(\alpha dt + \sigma dW(t) - \beta dN(t)\right), \\ S(0) &= s_0, \end{aligned}$$

where r, α, σ, and β are constants. For $N(t) = 0$, this market is a classical Black–Scholes market. We introduce, however, jumps in the Black–Scholes market and let N be a counting process counting the number of jumps such that

$$N(t) = \#\{0 < s \leq t : S(t) \neq S(t-)\}.$$

Remark 1. (a) Referring to Korn and Wilmott [6] our model above can also be seen as a worst-case approach when downward jumps of arbitrary size from the interval $[0, \beta]$ are possible. Since it is shown there that only the extreme jump size of β enters the following considerations, we have chosen to work only with this jump size above. Note also that we consider relative jump sizes which means that the jump size actually depends on the past performance of our stock. However, we have not yet modeled external market influences on the jump size which would be a possible extension of our model.

(b) We choose to work with constant coefficients here, but generalizations to, e.g., N-dependent coefficients should be straightforward. We do believe that generalizations to markets where the coefficients r, α, σ react on crashes, henceforth called "crashed coefficients," are important and that our results carry over to this situation. Crashed coefficients are considered in Korn [4] and in Korn and Menkens [5], where, however, weaker optimality results than ours below are given. Here we focus on the qualitative form of the new HJB-system characterization. With crashed coefficients the qualitative form of this system will not change, but in the examples and illustrations one would have to distinguish between several different subcases.

In usual jump-diffusion models the counting process is now assumed to follow some probability law on $(\Omega, \mathcal{F}, P)$. One could, e.g., let N be a Poisson process or a Cox process. Here, however, we take the counting process to be chosen by the market, which from the point of view of the agent is considered as an opponent. In other words, the market decides when the stock jumps, and N is to be considered as a decision process held by the market. We speak of jumps in N as interventions.

We assume that the market is able to decide on only a limited number of interventions and denote the maximal number of interventions over $[0, T]$ by n_0. Figure 1 illustrates the process of interventions. The market can, however, also choose not to exercise its intervention options, so at time T the process of interventions can be in any of the states in Figure 1. Further, to avoid technical complications, we assume that the market never intervenes more than once in one time instant (a multiple intervention of the market in one time instant is also not reasonable from an intuitive point of view).

The investment behavior of the agent is modeled by a predictable portfolio strategy π denoting the proportion of wealth invested in S; i.e., π is a decision variable held by the agent. Restricting ourselves to self-financing portfolio strategies, the wealth process follows the differential equation

$$\begin{aligned} X(0) &= x_0, \\ dX(t) &= X(t-)\left(rdt + \pi(t)\left((\alpha - r)\,dt + \sigma dW(t) - \beta dN(t)\right)\right). \end{aligned}$$

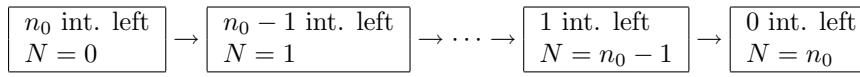


FIG. 1. *Process of interventions.*

The differential equation can be considered as a controlled differential equation with a pair of controls being a pair of portfolio strategies and interventions, i.e., (π, N). The agent is allowed to choose $\pi \in \mathcal{A}$, and the market is allowed to choose an intervention $N \in \mathcal{B}$ such that the pair of controls (π, N) leads to a well-posed optimization problem below. Even more, we consider $\mathcal{A}$ to be the set of all predictable processes (with respect to the σ-algebra generated by the stock price process and the counting process and which in particular carries the information about how many jumps can still at most appear) such that we have

$$E\int_0^T |\pi(t)|^m \, ds < \infty \text{ for } m = 1, 2, \ldots,$$
$$\pi(t)\,\beta < 1 \text{ for all } t \in [0, T].$$

These requirements in particular ensure that the wealth process stays nonnegative and has finite moments of all order. Note that, in a model with jumps, predictability of strategies becomes very important (in contrast to diffusion models, where predictability and adaptedness collapse), since otherwise the investor could choose his position during a crash after the crash has been observed.

Then, given a pair of controls (π, N), the controlled differential equation describing the wealth is given by

$$\begin{aligned} X^{\pi N}(0) &= x_0, \\ dX^{\pi N}(t) &= X^{\pi N}(t-)\left(r\,dt + \pi(t)\left((\alpha - r)\,dt + \sigma\,dW(t) - \beta\,dN(t)\right)\right). \end{aligned} \tag{1}$$

For a fixed time s and given $X^{\pi,N}(s) = x_s$, if τ is the first intervention time after s, we can obviously write

$$\begin{aligned} X^{\pi N}(s) &= x_s, \\ dX^{\pi N}(t) &= X^{\pi N}(t)\left(\left(r + \pi(t)(\alpha - r)\right)dt + \pi(t)\,\sigma\,dW(t)\right), s < t < \tau, \\ X^{\pi N}(\tau) &= X^{\pi N}(\tau-)\left(1 - \pi(\tau)\,\beta\right). \end{aligned}$$

If we only need to consider X until the first intervention time, we can just as well denote the argument N by τ, and we do so below.

We assume that the investor chooses a portfolio process to maximize worst-case expected utility of terminal wealth in the sense of the following optimization problem:

$$\sup_{\pi \in \mathcal{A}} \inf_{N \in \mathcal{B}} E\left[U\left(X^{\pi N}(T)\right)\right].$$

For each function $v \in C^{1,2}$ we define the differential operator $\mathcal{L}^\pi v$ by

$$\mathcal{L}^\pi v(t,x) = v_t(t,x) + v_x(t,x)\left(r + \pi(\alpha - r)\right)x + \frac{1}{2}v_{xx}(t,x)\,\pi^2\sigma^2x^2.$$

3. The Bellman system and the verification theorem. In this section we present and prove the Bellman system connected with the control problem described in the previous section.

We define the value function $\mathcal{J}^n(t,x,\pi)$ by

$$\mathcal{J}^n(t,x,\pi) = E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right],$$

where $E_{t,x,n}$ denotes conditional expectation given that $X(t) = x$ and that there are at most n possible jumps left. We define the optimal value function $V^n(t,x)$ by

$$V^n(t,x) = \sup_{\pi\in\mathcal{A}} \inf_{N\in\mathcal{B}} \mathcal{J}^n(t,x,\pi,\tau).$$

We can now present a Bellman system in a verification theorem, the proof of which can be found in the appendix.

THEOREM 2 (verification theorem). 1. Assume that $v^0(t,x)$ is a classical solution of

$$\begin{aligned} 0 &= \sup_{\pi\in\mathcal{A}}\left[\mathcal{L}^\pi v^0(t,x)\right], \\ v^0(T,x) &= U(x), \end{aligned}$$

which is polynomially bounded, and that

$$p^0(t,x) = \arg\sup_{\pi\in\mathcal{A}}\left[\mathcal{L}^\pi v^0(t,x)\right]$$

is an admissible control function. Then we have

$$V^0(t,x) = v^0(t,x),$$

and the optimal control function exists and is given by

$$\pi^{0*}(t,x) = p^0(t,x).$$

2. For $n\in\mathbf{N}$ and every function $v^n\in C^{1,2}$, define the sets $\mathcal{A}'_n(t,x)$ and $\mathcal{A}''_n(t,x)$ by

$$\begin{aligned} \mathcal{A}'_n(t,x) &= \left\{\pi : \pi\in\mathcal{A}, 0\le \mathcal{L}^\pi v^n(t,x)\right\}, \\ \mathcal{A}''_n(t,x) &= \left\{\pi : \pi\in\mathcal{A}, 0\le v^{n-1}(t,x(1-\beta\pi)) - v^n(t,x)\right\}, \end{aligned}$$

respectively. Assume that there exists a polynomially bounded $C^{1,2}$-solution of

$$\begin{aligned} 0 &\le \sup_{\pi\in\mathcal{A}''_n(t,x)}\left[\mathcal{L}^\pi v^n(t,x)\right], \\ 0 &\le \sup_{\pi\in\mathcal{A}'_n(t,x)}\left[v^{n-1}(t,x(1-\beta\pi)) - v^n(t,x)\right], \\ 0 &= \sup_{\pi\in\mathcal{A}''_n(t,x)}\left[\mathcal{L}^\pi v^n(t,x)\right] \sup_{\pi\in\mathcal{A}'_n(t,x)}\left[v^{n-1}(t,x(1-\beta\pi)) - v^n(t,x)\right], \\ v^n(T,x) &= U(x) \end{aligned}$$

and that

$$\begin{aligned} p^n(t,x) &= \arg\sup_{\pi\in\mathcal{A}''_n(t,x)}\left[\mathcal{L}^\pi v^n(t,x)\right], \\ \theta^n(t,x) &= \inf_{s:s\ge t}\left[v^{n-1}\left(s,X^{\pi N}(s)(1-\beta\pi)\right) - v^n\left(s,X^{\pi N}(s)\right)\le 0\right], \end{aligned}$$

where $X^{\pi N}(t) = x$ and s is a stopping time, is a pair of admissible control functions. Then

$$V^n(t,x) = v^n(t,x),$$

and the optimal control functions exist and are given by

$$\pi^{n*}(t,x) = p^n(t,x),$$
$$\tau^{n*}(t,x) = \theta^n(t,x).$$

In the proof of the verification theorem, the following lemma, which is also proved in the appendix, is used.

LEMMA 3. *The value function can be represented in the following ways:*

$$\begin{aligned} V^n(t,x) &= \sup_{\pi}\inf_{N} E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right] \\ &= \inf_{N}\sup_{\pi} E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right] \\ &= \sup_{\pi}\inf_{\tau} E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi\tau}(\tau-)\left(1-\beta\pi(\tau)\right)\right)\right] \\ &= \inf_{\tau}\sup_{\pi} E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi\tau}(\tau-)\left(1-\beta\pi(\tau)\right)\right)\right]. \end{aligned}$$

A careful inspection of the proof of Theorem 2 shows that the expectation requirements on the admissible controls and the polynomial growth condition for the value function are indeed only needed for the expectation of the stochastic integrals to vanish just before relations (16) and (21). Of course, these requirements are only sufficient for the proof to go through. The assumption of polynomial growth of the value function is not satisfied for our examples of the logarithmic utility and the exponential utility function below. However, one can directly check that the above proof still goes through for those special choices of the utility functions as it can be verified that the expectations of the two mentioned stochastic integrals vanish.

4. Characterization of the solution. To apply the verification theorem we are now going to construct (in a heuristic way) general candidates for the value functions V^n and the optimal controls along the lines of the theorem. In the following sections it will be shown that for special choices of the utility function U these heuristically derived candidates indeed satisfy all of the requirements of the verification theorem and are thus solutions of the control problem. Let us start by considering the inequality

$$0 \le \sup_{\pi\in\mathcal{A}'_n(t,x)}\left[V^{n-1}\left(t, x\left(1-\beta\pi\right)\right) - V^n(t,x)\right]. \tag{2}$$

Since for $\beta > 0$ and a (strictly) increasing utility function U we have that $V^{n-1}(t, x(1-\beta\pi))$ is a decreasing function of π, the supremum in (2) is obtained for the smallest π with

$$V^n_t(t,x) \ge -V^n_x(t,x)\left(r+\pi\left(\alpha-r\right)\right)x - \frac{1}{2}V^n_{xx}(t,x)\,\pi^2\sigma^2x^2. \tag{3}$$

Under the assumption of a concave V^n we have that the supremum in (2) is attained for the smallest value of π for which (3) holds as an equality. We can now consider

the obvious choice for the separation of the (t,x)-space into the set $\mathcal{M}$, where the right-hand side of the inequality (2) is strictly positive, and its complement, i.e.,

$$\mathcal{M}=\left\{(t,x):\sup_{\pi\in\mathcal{A}_n'(t,x)}\left[V^{n-1}\left(t,x\left(1-\beta\pi\right)\right)-V^n\left(t,x\right)\right]>0\right\}.$$

Outside $\mathcal{M}$, π and V are determined by the set of equations

$$\begin{aligned}V^n\left(t,x\right)&=V^{n-1}\left(t,x\left(1-\beta\pi\right)\right),\\ V_t^n\left(t,x\right)&=-V_x^n\left(t,x\right)\left(r+\pi\left(\alpha-r\right)\right)x-\frac{1}{2}V_{xx}^n\left(t,x\right)\pi^2\sigma^2x^2.\end{aligned}\tag{4}$$

Note that the first equation has to hold by the complementarity condition in the verification theorem. The second equation is argued for above.

We now argue that, even at some points (t,x) inside $\mathcal{M}$, the set of equations (4) determine π and V, which finally leads to an alternative and more relevant separation of the (t,x)-space. Intuitively, this is due to the constraint on π in $\mathcal{A}''$. Inside $\mathcal{M}$ we must have $\sup_{\pi\in\mathcal{A}_n''(s)}\left[\mathcal{L}^\pi v^n\left(s,X^{\pi,\theta}\left(s\right)\right)\right]=0$, again by complementarity. Ignoring the constraint $\pi\in\mathcal{A}_n''\left(t,x\right)$ we can compute the usual candidate for an optimal portfolio process by the first order conditions as:

$$\pi=-\frac{V_x^n\left(t,x\right)}{V_{xx}^n\left(t,x\right)x}\frac{\alpha-r}{\sigma^2}.\tag{5}$$

If, for the strategy (5), we have that

$$V^n\left(t,x\right)\leq V^{n-1}\left(t,x\left(1-\beta\pi\right)\right),$$

then (5) indeed satisfies the constraint $\pi\in\mathcal{A}_n''\left(t,x\right)$ and can be considered as the candidate optimal portfolio. If, however, for π as given in (5), we have that

$$V^n\left(t,x\right)>V^{n-1}\left(t,x\left(1-\beta\pi\right)\right),$$

then again (under suitable assumptions on U and β as mentioned above) we know that $V^{n-1}\left(t,x\left(1-\beta\pi\right)\right)$ decreases as a function of π. We further assume that

$$V^{n-1}\left(t,x\left(1-\beta\pi\right)\right)\to\infty$$

for $x\to\infty$ and $\beta\pi<1$ (this always has to be checked for concrete choices of the utility function U when even more explicit computations are performed in later sections). Since

$$V_x^n\left(t,x\right)\left(r+\pi\left(\alpha-r\right)\right)x+\frac{1}{2}V_{xx}^n\left(t,x\right)\pi^2\sigma^2x^2$$

is increasing for $\pi<-\frac{V_x^n(t,x)}{V_{xx}^n(t,x)x}\frac{\alpha-r}{\sigma^2}$, then $\sup_{\pi\in\mathcal{A}_n''(s)}\left[\mathcal{L}^\pi v^1\left(s,X^{\pi,\theta}\left(s\right)\right)\right]$ is obtained for the π for which

$$V^n\left(t,x\right)=V^{n-1}\left(t,x\left(1-\beta\pi\right)\right)$$

holds, and consequently π and V are determined by the set of equations

$$\begin{aligned}V^n\left(t,x\right)&=V^{n-1}\left(t,x\left(1-\beta\pi\right)\right),\\ V_t^n\left(t,x\right)&=-V_x^n\left(t,x\right)\left(r+\pi\left(\alpha-r\right)\right)x-\frac{1}{2}V_{xx}^n\left(t,x\right)\pi^2\sigma^2x^2.\end{aligned}$$

Since this is the same case as outside $\mathcal{M}$ (see (4)), we realize that $\mathcal{M}$ is not the relevant set that decomposes the state space in an appropriate way. Instead, we consider now a set $\mathcal{N}$ for which it is really (5) that determines the optimal portfolio. Thus, we separate the (t,x)-space into

$$\begin{aligned}(6)\qquad \pi(t,x) &= -\frac{V_x^n(t,x)}{V_{xx}^n(t,x)\,x}\frac{\alpha-r}{\sigma^2},\\ V_t^n(t,x) &= -V_x^n(t,x)\,(r+\pi(\alpha-r))\,x-\frac{1}{2}V_{xx}^n(t,x)\,\pi^2\sigma^2x^2,\end{aligned}$$

and its complement characterized by (4), i.e.,

$$\begin{aligned}(7)\qquad V^n(t,x) &= V^{n-1}(t,x\,(1-\beta\pi)),\\ V_t^n(t,x) &= -V_x^n(t,x)\,(r+\pi(\alpha-r))\,x-\frac{1}{2}V_{xx}^n(t,x)\,\pi^2\sigma^2x^2.\end{aligned}$$

Formally,

$$\begin{aligned}\mathcal{N}=\Bigg\{&(t,x):\pi(t,x)=-\frac{V_x^n(t,x)}{V_{xx}^n(t,x)\,x}\frac{\alpha-r}{\sigma^2},\\ &V_t^n(t,x)=-V_x^n(t,x)\,(r+\pi(\alpha-r))\,x-\frac{1}{2}V_{xx}^n(t,x)\,\pi^2\sigma^2x^2\Bigg\}.\end{aligned}$$

Note in particular that for $n=0$ we have that $\mathcal{N}$ typically equals the whole possible (t,x)-space, while for $n>1$ it might be possible that $\mathcal{N}$ is empty as we show for the specific choices of the utility functions below. But there are also examples where $\mathcal{N}$ will not be empty for $n>1$. In our examples below, this is so for $\beta<1$, where, however, the complement of $\mathcal{N}$ is empty. Examples where neither $\mathcal{N}$ nor its complement are empty for a given n may require a generalized model, such as, e.g., the case of "crashed coefficients" where the diffusion coefficients react on crashes.

5. Power utility. In this section we consider the case of power utility

$$U(x)=\frac{1}{\gamma}x^\gamma,\quad \gamma<1,\quad \gamma\neq 0,$$

and assume $\alpha>r$. Inspired by the solution of the usual portfolio problem we try a solution of the form

$$V^n(t,x)=\frac{1}{\gamma}f^n(t)\left(\frac{x}{f^n(t)}\right)^\gamma$$

leading to

$$\begin{aligned}V_t^n(t,x) &= \frac{1-\gamma}{\gamma}f_t^n(t)\left(\frac{x}{f^n(t)}\right)^\gamma,\\ V_x^n(t,x) &= \left(\frac{x}{f^n(t)}\right)^{\gamma-1},\\ V_{xx}^n(t,x) &= -(1-\gamma)\frac{1}{f^n(t)}\left(\frac{x}{f^n(t)}\right)^{\gamma-2}.\end{aligned}$$

With these relations we obtain the optimal portfolio

$$\pi^{*n}=\begin{cases}\frac{1}{1-\gamma}\frac{\alpha-r}{\sigma^2}, & (t,x,n)\in\mathcal{N},\\ \frac{1}{\beta}\left(1-\left(\frac{f^n(t)}{f^{n-1}(t)}\right)^{\frac{1-\gamma}{\gamma}}\right), & (t,x,n)\notin\mathcal{N}\end{cases}$$

(here we have again implicitly assumed strict concavity of V^n!). The usual solution in the case of $n = 0$ is well-known and given by

$$\pi^{*0}(t,x) = \frac{1}{1-\gamma}\frac{\alpha - r}{\sigma^2},$$
$$f^0(t) = e^{-\left(r+\frac{1}{2}\frac{\alpha-r}{\sigma^2}\frac{1}{1-\gamma}\right)\frac{(T-t)(1-\gamma)}{\gamma}}.$$

As the investor has a strictly increasing utility function, he would like to choose a portfolio process as close as possible to π^{*0} to obtain a high final expected utility. On the other hand, if he chooses $\pi^{*n}(t) = \pi^{*0}$, then the market could do him most possible harm by choosing an immediate crash under the assumption of $\beta > 0$. In the case of $\beta > 0$, we have thus argued for that $\mathcal{N}$ is indeed empty for $n > 0$; i.e., for $n > 0$ we assume that we are always in the situation of

$$\pi^{*n}(t,x) = \frac{1}{\beta}\left(1 - \left(\frac{f^n(t)}{f^{n-1}(t)}\right)^{\frac{1-\gamma}{\gamma}}\right),$$
$$\mathcal{L}^{\pi^{*n}} v^n(t,x) = 0.$$

If we now plug in our guess for $\pi^{*n}(t,x)$ and for $v^n(t,x)$ into the second condition above and use the final condition of

$$v^n(T,x) = v^{n-1}(T,x) = \cdots = v^0(T,x) = \frac{1}{\gamma}x^\gamma$$

implying

$$\pi^{*n}(T,x) = 0,$$

we arrive at the following ordinary differential equation for f (skipping the t-argument in ordinary differential equations below and throughout):

$$f_t^n = f^n\left(-\frac{\gamma}{1-\gamma}\left(r + (\alpha - r)\pi^{*n}\right) + \gamma\frac{1}{2}\left(\pi^{*n}\right)^2\sigma^2\right), f^n(T) = 1,$$

for $n = 1, 2, \ldots$. With the help of this equation and the definition of π^{*n} we can derive an ordinary differential equation for $\pi^{*n}(t)$ which holds for $(t,x,n) \notin \mathcal{N}$:

$$\pi_t^{*n} = \frac{1}{\beta}\left(1 - \pi^{*n}\beta\right)\left((\alpha - r)\left(\pi^{*n} - \pi^{*n-1}\right) - \frac{1}{2}(1-\gamma)\sigma^2\left(\left(\pi^{*n}\right)^2 - \left(\pi^{*n-1}\right)^2\right)\right).$$

Using this differential equation and its final condition $\pi^{*n}(T) = 0$, one can show via induction that its solution satisfies

$$0 \le \pi^{*n}(t) \le \pi^{*n-1}(t) \le \cdots \le \pi^{*0},$$

is unique, and for $n = 1$ can be explicitly given as the solution of a nonlinear equation (see Korn and Wilmott [6] and Korn and Menkens [5]). A consequence of this is in particular that the solution of the differential equation for $f^n(t)$ is always positive, which implies that V^n of the above form is a concave function in x, as desired. Thus, all assumptions of the verification theorem are satisfied, and we have indeed computed the optimal portfolio.

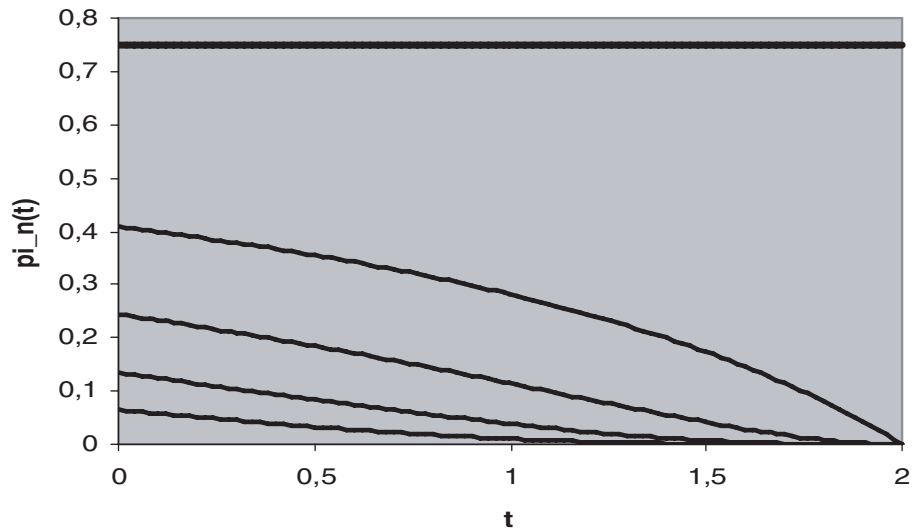


FIG. 2. $\pi^{*n}(t)$ *for* $n = 0, 1, 2, 3, 4$ *(from top to bottom), parameters:* $\beta = 0.05,\ \alpha = 0.11,\ r = 0.05,\ \sigma = 0.4,\ T = 2,\ \gamma = 0.5$.

The form of the optimal portfolio processes is illustrated by Figure 2, where the maximum number n of crashes that can still occur determines which of the five lines is relevant for the optimal portfolio process. If there are still four possible crashes, the investor chooses the portfolio process given by the lowest line in Figure 2. Note that this is a nonconstant process which decreases with time. After a crash (or if, for some other reason, the investor now assumes that there are only three possible crashes left), the investor immediately shifts his portfolio process up to the next line, stays there until the next crash has happened, then jumps up again, etc. Unless the last jump possible has occurred, the investor follows the decreasing line until maturity, and by then he has reduced his fraction of risky investments to zero. This form is very reasonable, because as long as there is a possibility for a crash, the investor is hit by it harder and harder as time goes by since he loses his possibility of compensating by posterior risky investments. It is common advice to pension savers to reduce their risky investments as time to retirement decreases. Thus, our new worst-case approach can be seen as a theoretical framework supporting this market practice.

In the case of $\beta < 0$ it can easily be verified that it is never optimal for the market to intervene if the investor uses the portfolio process $\pi^{*0} = \frac{1}{1-\gamma}\frac{\alpha-r}{\sigma^2}$. Consequently, in this setting we have

$$v^n(t,x) = v^{n-1}(t,x) = \cdots = v^0(t,x),$$

and the optimality criteria of the verification theorem can be satisfied only for the strategy that consists of no intervention before time T at all. This is also intuitively clear, because this portfolio process leads to the highest expected utility in the standard market setting on one hand, and on the other hand a jump of positive size (which is the case for $\beta < 0$) would even make the situation of the investor better. Hence, the infimum over the intervention strategies is attained for the above-mentioned no-jump strategy.

6. Log utility. The situation in the case of the logarithmic utility function is very similar to that of the Hara utility. In fact, it can essentially be solved by using the results of the foregoing section and setting $\gamma = 0$. We therefore shorten its presentation. Consider

$$U(x) = \log x,$$

and assume $\alpha > r$. The main difference to the Hara case is our guess on the form of the value functions (again inspired by the case $n = 0$):

$$\begin{aligned} V^n(t,x) &= \log x + f^n(t), \\ V^n_t(t,x) &= f^n_t(t), \\ V^n_x(t,x) &= \frac{1}{x}, \\ V^n_{xx}(t,x) &= -\frac{1}{x^2}. \end{aligned}$$

Inside $\mathcal{N}$ we obtain the form of π^{*n} as in the case of $n = 0$, while outside $\mathcal{N}$ the (candidate for the) optimal portfolio process is determined by the indifference requirement $V^n(t,x) = V^{n-1}(t, x(1-\beta\pi))$. This leads to

$$\pi^* = \begin{cases} \frac{\alpha - r}{\sigma^2}, & (t,x,n) \in \mathcal{N}, \\ \frac{1 - e^{f^n(t) - f^{n+1}(t)}}{\beta}, & (t,x,n) \notin \mathcal{N}. \end{cases}$$

Again, the case of $n = 0$ is well-known and given by (see, e.g., Korn [3])

$$\begin{aligned} \pi^{*0}(t,x) &= \frac{\alpha - r}{\sigma^2}, \\ f^0(t) &= \left(r + \frac{1}{2} \frac{\alpha - r}{\sigma^2} \right) (T - t). \end{aligned}$$

With the same argument as in the Hara case above, for $\beta > 0$ we conclude that $\mathcal{N}$ is empty for $n > 0$. Then, inserting our resulting guess for the form of the optimal portfolio into (6) leads to a differential equation for f:

$$f^n_t(t) = -\left(r + \pi^{*n}(\alpha - r)\right) + \frac{1}{2}\left(\pi^{*n}\right)^2 \sigma^2, \quad f(T) = 0,$$

which again leads to an ordinary differential equation for π which holds for $(t,x,n) \notin \mathcal{N}$:

$$\begin{aligned} \pi^{*n}_t &= \frac{1}{\beta}(1 - \pi^{*n}\beta)\left(f^{n+1}_t - f^n_t\right) \\ &= \frac{1}{\beta}(1 - \pi^{*n}\beta)\left((\alpha - r)\left(\pi^{*n} - \pi^{*n+1}\right) - \frac{1}{2}\sigma^2\left(\left(\pi^{*n}\right)^2 - \left(\pi^{*n+1}\right)^2\right)\right). \end{aligned}$$

As shown in Korn and Menkens [5] it has a unique solution which is bounded by 0 from below and by $\pi^{*(n-1)}_t$ from above for $n \geq 1$. Also, for numerical examples which are similar to the one given in the Hara utility section, we refer to Korn and Menkens [5].

Further, it is obvious that in the case of $\beta < 0$ the optimal intervention strategy consists of never doing a jump at all.

7. Exponential utility. In this section we consider the case of exponential utility, i.e.,

$$U(x) = -e^{-\theta x},$$

for some $\theta > 0$. Compared to the foregoing examples of the log-utility and the Hara case, the situation for the exponential is fundamentally different with respect to two aspects. First of all, a separation of the t- and the x-variables in the HJB equation is not possible, a property that is essentially due to the fact that the derivative of the exponential function is itself the exponential function. It is well known from standard portfolio optimization (see, e.g., Browne [2]) that it is therefore more suitable to consider the amount of money invested in the risky stock at a time, in our notation $\pi(t) X(t)$, as control variables as opposed to the portfolio process itself. As the second difference, compared to the two utility functions considered above, note that the exponential utility function has a finite slope in $x = 0$, which results in the fact that the (unconstrained) optimal wealth process can attain negative values. Again, this is well known (compare again with Browne [2]). To apply our main verification it would therefore be necessary to refine our definition of an admissible control. This can be done along the lines of Browne [2], but details are left to the reader. Note that in contrast to Korn [4] we do not have to restrict to deterministic strategies. Keeping all of these considerations in mind, we guess the following form of the value function, inspired by the case $n = 0$:

$$V^n(t,x) = -e^{-\theta f(t) x - g^n(t)},$$

leading to

$$\begin{aligned}
V_t^n(t,x) &= -e^{-\theta f(t) x - g^n(t)} \left(-\theta f_t(t) x - g_t^n(t)\right), \\
V_x^n(t,x) &= \theta f(t) e^{-\theta f(t) x - g^n(t)}, \\
V_{xx}^n(t,x) &= -\theta^2 f(t)^2 e^{-\theta f(t) x - g^n(t)}.
\end{aligned}$$

Assuming strict concavity of V^n in x—which is given if f is nonvanishing—(suitable) application of the verification theorem yields the following candidate for the optimal amount of money invested in the stock:

$$\pi^*(t,x) x = \begin{cases} \frac{1}{\theta f(t)} \frac{\alpha - r}{\sigma^2}, & (t,x,n) \in \mathcal{N}, \\ \frac{1}{\theta f(t)} \frac{g^{n+1}(t) - g^n(t)}{\beta}, & (t,x,n) \notin \mathcal{N}. \end{cases}$$

In the case of $n = 0$ it is well known that we have

$$\begin{aligned}
f(t) &= \exp\left(r(T-t)\right), \\
g^0(t) &= \frac{1}{2} \left(\frac{\alpha - r}{\sigma}\right)^2 (T-t), \\
\pi^{*0}(t,x) x &= \frac{1}{\theta} \frac{\alpha - r}{\sigma^2} e^{-r(T-t)}.
\end{aligned}$$

Now since $\mathcal{N}$ is empty for $n > 0$ (the argument for this is similar to the preceding two examples), we use the above derived form of our candidate optimal strategy $\pi^* x$ and insert our guesses in (6) to obtain a differential equation for g^n:

$$g_t^n = -\frac{g^{n+1} - g^n}{\beta} (\alpha - r) + \frac{1}{2} \frac{\left(g^{n+1} - g^n\right)^2}{\beta^2} \sigma^2.$$

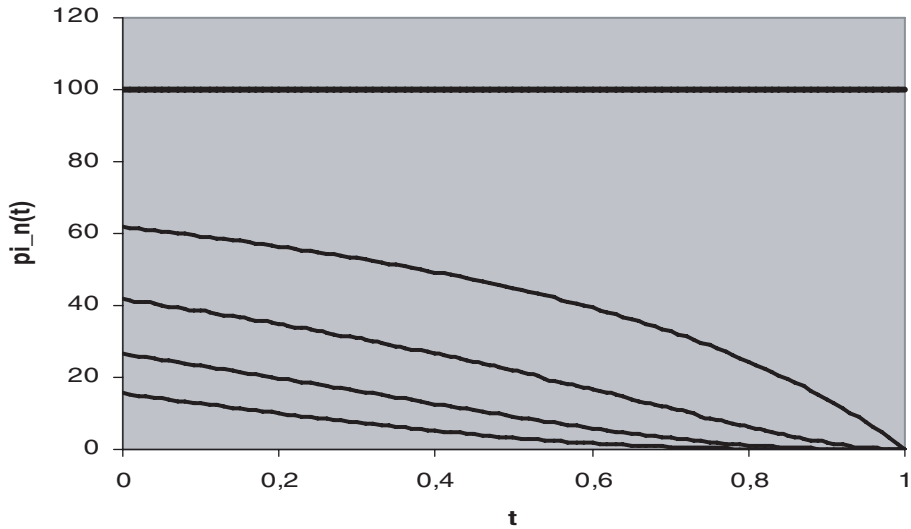


FIG. 3. $\pi^{*n}(t)\,x$ *for* $n = 0, 1, 2, 3, 4$ *(from top to bottom), parameters:* $\beta = 0.05,\ \alpha = 0.16,\ r = 0,\ \sigma = 0.4,\ T = 1,\ \theta = 0.01$.

This results in an ordinary differential equation for $\pi^* x$ for $(t, x, n) \notin \mathcal{M}$:

$$\begin{aligned} \pi_t^{*n} x &= r\pi^{*n} x - \left(\pi^{*n+1} x - \pi^{*n} x\right) \frac{\alpha - r}{\beta} \\ &\quad + \left(\left(\pi^{*n+1} x\right)^2 - \left(\pi^{*n} x\right)^2\right) \frac{\frac{1}{2}\theta f \sigma^2}{\beta}, \\ \pi^{*n}(T)\, x &= 0, \end{aligned}$$

for which we can show with standard arguments that a unique bounded and non-negative solution exists. Before we illustrate the form of the optimal strategy, let us remark that, in the case of $r = 0$, an explicit solution for $n = 1$ exists which has the form (see Korn [4] for a different derivation)

$$\pi^{*1}(t, x)\, x = \frac{\alpha}{\theta\sigma^2} + \frac{2\beta}{\theta\sigma^2[(T-t) - \frac{2\beta}{\alpha}]}.$$

The form of the optimal trading strategies are illustrated in Figure 3. They look very similar to the optimal portfolio processes of Figure 2, and of course the comments for their use depending on the maximum number n of crashes remain valid. However, note that, if we would plot the portfolio processes, we would have very irregular curves as they are inversely proportional to the actual wealth process curve, and for small values of the wealth process the portfolio process can grow above all limits (but the amount of money invested in the stock stays bounded).

8. Conclusion and further aspects. In this paper we have put the worst-case approach to portfolio optimization as developed by Korn and Wilmott [6] into a generalized HJB-equation framework. This has the particular advantage that the restriction to deterministic control processes is no longer required. This framework can be used for a worst-case approach in other areas than finance. The ideas are generally applicable to situations, e.g., in insurance or engineering, where risk is managed by

combining an expectation functional on "normal risks" and a worst-case functional on "extreme risks." But even within the portfolio application, there remain many open problems and generalizations for future research such as

- explicit solution of problems with many stocks (in contrast to the Korn and Wilmott [6] approach, this should be possible in a more explicit way using our new approach),
- explicit solution of problems with nonconstant β (this should again be possible in an easier way in our HJB-equation framework), and
- weakening the regularity assumptions of the verification theorem (maybe via the use of viscosity solution techniques).

Appendix.

Proof of Lemma 3. Let $\varepsilon > 0$. Then, for a given first intervention time τ, we can choose a portfolio strategy π^* which is $\varepsilon/4$-optimal until time τ and a portfolio strategy π^{**} which is arbitrary until time τ and $\varepsilon/4$-optimal after time τ in the sense that the following two inequalities hold (note that we cannot yet assume that V^n is indeed the value function):

$$\begin{aligned}(8)\qquad &\sup_{\pi} E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi\tau}\left(\tau-\right)\left(1-\beta\pi\left(\tau\right)\right)\right)\right]\\ &\leq E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi^*\tau}\left(\tau-\right)\left(1-\beta\pi^*\left(\tau\right)\right)\right)\right]+\varepsilon/4,\\ (9)\qquad &\sup_{\pi}\inf_{N} E_{\tau,x,n}\left[U\left(X^{\pi N}\left(T\right)\right)\right]\\ &\leq \inf_{N} E_{\tau,X^{\pi^{**}N}(\tau),n-1}\left[U\left(X^{\pi^{**}N}\left(T\right)\right)\right]+\varepsilon/4.\end{aligned}$$

Further, for a given portfolio strategy, introduce an $\varepsilon/4$-optimal strategy for the first intervention τ^* and, given an arbitrary first intervention time τ, an $\varepsilon/4$-optimal intervention strategy N^* after time τ, again in the sense that the following two inequalities are valid:

$$\begin{aligned}(10)\qquad &\inf_{\tau} E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi\tau}\left(\tau-\right)\left(1-\beta\pi\left(\tau\right)\right)\right)\right]\\ &\geq E_{t,x,n}\left[V^{n-1}\left(\tau^*\left(\pi\right), X^{\pi\tau^*}\left(\tau^*\left(\pi\right)-\right)\left(1-\beta\pi\right)\right)\right]-\varepsilon/4,\\ (11)\qquad &\inf_{N} E_{\tau,X^{\pi}(\tau),n-1}\left[U\left(X^{\pi N}\left(T\right)\right)\right]\\ &\geq E_{\tau,X^{\pi N^*}(\tau),n-1}\left[U\left(X^{\pi N^*}\left(T\right)\right)\right]-\varepsilon/4.\end{aligned}$$

Then we have the following list of inequalities (explanation follows after):

$$\begin{aligned}(12)\qquad &\sup_{\pi}\inf_{N} E_{t,x,n}\left[U\left(X^{\pi N}\left(T\right)\right)\right]\\ &\geq \inf_{N} E_{t,x,n}\left[E_{\tau,X^{\pi^{**}N}(\tau),n-1}\left[U\left(X^{\pi^{**}N}\left(T\right)\right)\right]\right]\\ &\geq \inf_{\tau} E_{t,x,n}\left[\inf_{N} E_{\tau,X^{\pi^{**}N}(\tau),n-1}\left[U\left(X^{\pi^{**}N}\left(T\right)\right)\right]\right]\\ &\geq \inf_{\tau} E_{t,x,n}\left[\sup_{\pi}\inf_{N} E_{\tau,X^{\pi N}(\tau),n-1}\left[U\left(X^{\pi N}\left(T\right)\right)\right]\right]-\varepsilon/4\\ &\geq \inf_{\tau} E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi\tau}\left(\tau-\right)\left(1-\beta\pi\left(\tau\right)\right)\right)\right]-\varepsilon/4\\ &\geq E_{t,x,n}\left[V^{n-1}\left(\tau^*, X^{\pi\tau^*}\left(\tau^*-\right)\left(1-\beta\pi\left(\tau^*\right)\right)\right)\right]-\varepsilon/2.\end{aligned}$$

The first inequality follows from plugging in the portfolio strategy π^{**} and the tower property. The second inequality follows from interchanging the first expectation and the infimum over intervention strategies after the first intervention. The third inequality follows from (9). The fourth inequality follows from the definition of V. The fifth inequality follows from (10). Taking supremum on both sides gives

$$\begin{aligned}(13)\qquad &\sup_{\pi}\inf_{N} E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right]\\ &\geq \sup_{\pi} E_{t,x,n}\left[V^{n-1}\left(\tau^*, X^{\pi\tau^*}(\tau^*-)\left(1-\beta\pi(\tau^*)\right)\right)\right]-\varepsilon/2.\end{aligned}$$

We also have the following list of inequalities (explanation follows after):

$$\begin{aligned}&\inf_{N}\sup_{\pi} E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right]\\ &\leq \sup_{\pi} E_{t,x,n}\left[E_{\tau,X^{\pi N^*}(\tau),n-1}\left[U\left(X^{\pi N^*}(T)\right)\right]\right]\\ &\leq \sup_{\pi} E_{t,x,n}\left[\inf_{N} E_{\tau,X^{\pi N}(\tau),n-1}\left[U\left(X^{\pi N}(T)\right)\right]\right]+\varepsilon/4\\ &\leq \sup_{\pi} E_{t,x,n}\left[\sup_{\pi}\inf_{N} E_{\tau,X^{\pi N}(\tau),n-1}\left[U\left(X^{\pi N}(T)\right)\right]\right]+\varepsilon/4\\ &\leq \sup_{\pi} E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi N}(\tau-)\left(1-\beta\pi(\tau)\right)\right)\right]+\varepsilon/4.\end{aligned}$$

The first inequality follows from plugging in the intervention strategy N^* and the tower property. The second inequality follows from (11). The third inequality is obvious. The fourth inequality follows from the definition of V. Taking infimum on both sides results in

$$\begin{aligned}(14)\qquad &\inf_{N}\sup_{\pi} E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right]\\ &\leq \inf_{\tau}\sup_{\pi} E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi\tau}(\tau-)\left(1-\beta\pi(\tau)\right)\right)\right]+\varepsilon/4.\end{aligned}$$

Finally we can gather the inequalities (explanation follows after):

$$\begin{aligned}&\sup_{\pi}\inf_{N} E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right]\\ &\geq \sup_{\pi} E_{t,x,n}\left[V^{n-1}\left(\tau^*, X^{\pi\tau^*}(\tau^*-)\left(1-\beta\pi(\tau^*)\right)\right)\right]-\varepsilon/2\\ &\geq \inf_{\tau}\sup_{\pi} E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi\tau}(\tau-)\left(1-\beta\pi(\tau)\right)\right)\right]-\varepsilon/2\\ &\geq \inf_{\tau} E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi^*\tau}(\tau-)\left(1-\beta\pi^*(\tau)\right)\right)\right]-\varepsilon/2\\ &\geq \inf_{\tau}\sup_{\pi} E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi\tau}(\tau-)\left(1-\beta\pi(\tau)\right)\right)\right]-3\varepsilon/4\\ &\geq \inf_{N}\sup_{\pi} E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right]-\varepsilon\\ &\geq \sup_{\pi}\inf_{N} E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right]-\varepsilon.\end{aligned}$$

The first inequality is just (13). The second inequality follows from giving up the specification of the first intervention. The third inequality follows from plugging in the portfolio strategy π^*. The fourth inequality follows from (8). The fifth inequality follows from (14). The sixth inequality is the usual $\sup\inf \leq \inf\sup$ relation.

The reversed line of arguments (left to the reader) gives that

$$\begin{aligned}
&\inf_N \sup_\pi E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right] \\
&\leq \sup_\pi \inf_\tau E_{t,x,n}\left[V^{n-1}\left(\tau, X^{\pi N}(\tau-)(1-\beta\pi)\right)\right] + \varepsilon/2 \\
&\leq \sup_\pi \inf_N E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right] + \varepsilon \\
&\leq \inf_N \sup_\pi E_{t,x,n}\left[U\left(X^{\pi N}(T)\right)\right] + \varepsilon.
\end{aligned}$$

Since all inequalities above hold for any ε, they must hold as equalities, and consequently the theorem is proved. □

Proof of Theorem 2 (verification theorem). The assertions in the first part of the verification (corresponding to 0 interventions left) are classical and have a proof which can be found in any textbook on dynamic portfolio optimization, e.g., Korn [3].

The second part is proved by induction. First, we prove the verification theorem for $n = 1$. Here the control N and the control τ are equivalent, and we can denote $X^{\pi,N}$ by $X^{\pi,\tau}$.

Choose an arbitrary control (π, τ), and fix a point (t, x). Let X follow the dynamics given in (1) with the time point 0 replaced by the time point t. Inserting $X^{\pi,\tau}$ in v^1 and using Ito's formula we obtain

$$\begin{aligned}
&v^1\left(t, X^{\pi,\tau}(t)\right) = v^1(t,x), \\
&dv^1\left(s, X^{\pi,\tau}(s)\right) = \mathcal{L}^\pi v^1\left(s, X^{\pi,\tau}(s)\right) ds + v_x^1\left(s, X^{\pi,\tau}(s)\right) \sigma X^{\pi,\tau}(s)\, dW(s), t < s < \tau, \\
&dv^1\left(\tau, X^{\pi,\tau}(\tau)\right) = v^1\left(\tau, X^{\pi,\tau}(\tau-)(1-\pi(\tau)\beta)\right) - v^1\left(\tau-, X^{\pi,\tau}(\tau-)\right)
\end{aligned}$$

such that

$$\begin{aligned}
v^1\left(\tau-, X^{\pi,\tau}(\tau-)\right) - v^1(t,x) = &\int_t^\tau \mathcal{L}^\pi v^1\left(s, X^{\pi,\tau}(s)\right) ds \\
&+ \int_t^\tau v_x^1\left(s, X^{\pi,\tau}(s)\right) \sigma X^{\pi,\tau}(s)\, dW(s).
\end{aligned} \tag{15}$$

Now fix the investment strategy $\pi(s) = p(s)$, $t \leq s \leq \tau$, where we use the letter p, sloppily but intuitively clear, for the process $p\left(s, X^{\pi,\tau}(s-)\right)$ coming from applying the function $p(s,x)$ on the process $X^{\pi,\tau}$such that the strategy is predictable as it should be. Then we know from the Bellman system that, for $t \leq s \leq \tau$,

$$0 \leq \mathcal{L}^p v^1\left(s, X^{p,\tau}(s)\right),$$

and, since $p(s) \in \mathcal{A}_1''(s)$, also

$$0 \leq v^0\left(s, X^{p,\tau}(s)(1-\beta p(s))\right) - v^1\left(s, X^{p,\tau}(s)\right).$$

But this means that inserting p in (15) gives the inequality

$$v^1(t,x) \leq v^0\left(\tau, X^{p,\tau}(\tau-)(1-\beta p(\tau))\right) - \int_t^\tau v_x^1\left(s, X^{p,\tau}(s)\right) \sigma X^{p,\tau}(s)\, dW(s).$$

Due to our requirements on the admissible controls and on the value function, the stochastic integral vanishes when taking expectation, leaving us with the inequality

$$v^1(t,x) \leq E_{t,x}\left[v^0\left(\tau, X^{p,\tau}(\tau-)(1-\beta p(\tau))\right)\right]. \tag{16}$$

Then, on the one hand, we can immediately conclude that

$$v^1(t,x) \leq \sup_{\pi} E_{t,x}\left[v^0\left(\tau, X^{\pi,\tau}(\tau-)\left(1-\beta\pi(\tau)\right)\right)\right]$$

such that taking infimum over τ on both sides gives

$$v^1(t,x) \leq \inf_{\tau}\sup_{\pi} E_{t,x}\left[v^0\left(\tau, X^{\pi,\tau}(\tau-)\left(1-\beta\pi(\tau)\right)\right)\right]. \tag{17}$$

On the other hand, taking infimum over τ on both sides of (16) gives

$$v^1(t,x) \leq \inf_{\tau} E_{t,x}\left[v^0\left(\tau, X^{p,\tau}(\tau-)\left(1-\beta p(\tau)\right)\right)\right],$$

and then we can conclude that also

$$v^1(t,x) \leq \sup_{\pi}\inf_{\tau} E_{t,x}\left[v^0\left(\tau, X^{\pi,\tau}(\tau-)\left(1-\beta\pi(\tau)\right)\right)\right]. \tag{18}$$

Consider again (15). Now fix the time $\tau=\theta$. Then we know that

$$v^0\left(s, X^{\pi,\theta}(s-)\left(1-\beta\pi(s)\right)\right) - v^1\left(s-, X^{\pi,\theta}(s-)\right) > 0, t \leq s < \theta, \tag{19}$$

$$v^0\left(\theta, X^{\pi,\theta}(\theta-)\left(1-\beta\pi(\theta)\right)\right) - v^1\left(\theta, X^{\pi,\theta}(\theta-)\right) \leq 0. \tag{20}$$

Now, either $0 > \mathcal{L}^{\pi}v^1(s, X^{\pi,\tau}(s))$ or $0 \leq \mathcal{L}^{\pi}v^1(s, X^{\pi,\tau}(s))$. But if $0 \leq \mathcal{L}^{\pi}v^1(s, X^{\pi,\tau}(s))$, then $\pi \in \mathcal{A}_1'$, and then (19) gives us that

$$\sup_{\pi\in\mathcal{A}_1'(s)}\left[v^0\left(s, X^{\pi,\theta}(s-)\left(1-\beta\pi(s)\right)\right) - v^1\left(s-, X^{\pi,\theta}(s-)\right)\right] > 0.$$

By complementarity, we then know that

$$\sup_{\pi\in\mathcal{A}_1''(s)}\left[\mathcal{L}^{\pi}v^1\left(s, X^{\pi,\theta}(s)\right)\right] = 0.$$

But since $\pi(s) \in \mathcal{A}_1''(s)$ by (19), we then know that

$$\mathcal{L}^{\pi}v^1\left(s, X^{\pi,\theta}(s)\right) \leq 0.$$

So, in any case, $\mathcal{L}^{\pi}v^1\left(s, X^{\pi,\theta}(s)\right) \leq 0$, $s < \theta$. But this means that inserting θ in (15) and using (20) gives the inequality

$$v^1(t,x) \geq -\int_t^{\theta} v_x\left(s, X^{\pi,\theta}(s)\right)\sigma X^{\pi,\theta}(s)\,dW(s) + v^0\left(\theta, X^{\pi,\theta}(\theta-)\left(1-\beta\pi(\theta)\right)\right).$$

Due to our requirements on the admissible controls and on the value function, the stochastic integral vanishes when taking expectation, leaving us with the inequality

$$v^1(t,x) \geq E_{t,x}\left[v^0\left(\theta, X^{\pi,\theta}(\theta-)\left(1-\beta\pi(\theta)\right)\right)\right]. \tag{21}$$

Then, on the one hand, we can conclude that

$$v^1(t,x) \geq \inf_{\tau} E_{t,x}\left[v^0\left(\tau, X^{\pi,\tau}(\tau-)\left(1-\beta\pi(\tau)\right)\right)\right]$$

such that taking supremum over π on both sides gives

$$v^1(t,x) \geq \sup_{\pi}\inf_{\tau} E_{t,x}\left[v^0\left(\tau, X^{\pi,\tau}(\tau-)\left(1-\beta\pi(\tau)\right)\right)\right]. \tag{22}$$

On the other hand, taking supremum over π on both sides of (21) gives

$$v^1(t,x) \geq \sup_\pi E_{t,x}\left[v^0\left(\theta, X^{\pi,\theta}(\theta-)(1-\beta\pi(\theta))\right)\right]$$

such that we can conclude that

$$v^1(t,x) \geq \inf_\tau \sup_\pi E_{t,x}\left[v^0\left(\tau, X^{\pi,\tau}(\tau-)(1-\beta\pi(\tau))\right)\right]. \tag{23}$$

From (17), (18), (22), and (23), we conclude that

$$\begin{aligned} v^1(t,x) &= \inf_\tau \sup_\pi E_{t,x}\left[v^0\left(\tau, X^{\pi}(\tau-)(1-\beta\pi(\tau))\right)\right] \\ &= \sup_\pi \inf_\tau E_{t,x}\left[v^0\left(\tau, X^{\pi}(\tau-)(1-\beta\pi(\tau))\right)\right], \end{aligned}$$

and it only remains to be realized that V is characterized by this equation. □

Acknowledgments. We thank two anonymous referees for their comments, which greatly helped to improve the paper. The work of Ralf Korn was supported by the Rheinland-Pfalz cluster of excellence "Dependable adaptive systems and mathematical models."

REFERENCES

[1] K. K. Aase, *Optimum portfolio diversification in a general continuous-time model*, Stochastic Process. Appl., 18 (1984), pp. 81–98.
[2] S. Browne, *Optimal investment policies for a firm with a random risk process: Exponential utility and minimizing the probability of ruin*, Math. Oper. Res., 20 (1995), pp. 937–957.
[3] R. Korn, *Optimal Portfolios*, World Scientific, Singapore, 1997.
[4] R. Korn, *Worst-case scenario investment for insurers*, Insurance Math. Econom., 36 (2005), pp. 1–11.
[5] R. Korn and O. Menkens, *Worst-case scenario portfolio optimization: A new stochastic control approach*, Math. Methods Oper. Res., 62 (2005), pp. 123–140.
[6] R. Korn and P. Wilmott, *Optimal portfolios under the threat of a crash*, Int. J. Theor. Appl. Finance, 5 (2002), pp. 171–187.
[7] D. Kramkov and W. Schachermayer, *The asymptotic elasticity of utility functions and optimal investment in incomplete mrkets*, Ann. Appl. Probab., 9 (1999), pp. 904–950.
[8] R. C. Merton, *Lifetime portfolio selection under uncertainty: The continuous time case*, Rev. Econom. Stat., 51 (1969), pp. 247–257.
[9] R. C. Merton, *Optimum consumption and portfolio rules in a continuous time model*, J. Econom. Theory, 3 (1971), pp. 373–413; Erratum 6 (1973), pp. 213–214.
[10] D. Talay and Z. Zheng, *Worst case model risk management*, Finance Stoch., 6 (2002), pp. 517–537.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2031–2051

PRIMAL-DUAL SYMMETRIC INTRINSIC METHODS FOR FINDING ANTIDERIVATIVES OF CYCLICALLY MONOTONE OPERATORS*

HEINZ H. BAUSCHKE†, YVES LUCET‡, AND XIANFU WANG†

Abstract. A fundamental result due to Rockafellar states that every cyclically monotone operator A admits an antiderivative f in the sense that the graph of A is contained in the graph of the subdifferential operator ∂f. Given a method $\mathfrak{m}$ that assigns every finite cyclically monotone operator A some antiderivative $\mathfrak{m}_A$, we say that the method is *primal-dual symmetric* if $\mathfrak{m}$ applied to the inverse of A produces the Fenchel conjugate of $\mathfrak{m}_A$. Rockafellar's antiderivatives do not possess this property. Utilizing Fitzpatrick functions and the proximal average, we present novel primal-dual symmetric intrinsic methods. The antiderivatives produced by these methods provide a solution to a problem posed by Rockafellar in 2005. The results leading to this solution are illustrated by various examples.

Key words. antiderivative, convex function, cyclically monotone operator, Fenchel conjugate, Fitzpatrick function, maximal monotone operator, n-cyclically monotone operator, proximal average, Rockafellar's antiderivative, Rockafellar function, subdifferential operator

AMS subject classifications. Primary 47H05; Secondary 26B25, 52A41, 90C25

DOI. 10.1137/060675794

1. Introduction.

Suppose that X is a real Banach space with continuous dual X^*, dual pairing $\langle \cdot, \cdot \rangle$, and norm $\|\cdot\|$. We start by recalling some known notions and results concerning (cyclically) monotone operators. These operators play a fundamental role in modern optimization as well as convex and variational analysis; see, e.g., [11, 12, 18, 20, 21, 22, 24] for further information and notation not explicitly defined here. Let A be a *set-valued operator* from X to X^*, i.e., $(\forall x \in X)\ Ax \subseteq X^*$; thus, A is a mapping from X to the power set of X^*. We use the notation $A\colon X \rightrightarrows X^*$ and remark that A can be identified with its *graph* $\operatorname{gra} A := \big\{(x, x^*) \in X \times X^* \mid x^* \in Ax\big\}$. The *domain* of A is $\operatorname{dom} A := \big\{x \in X \mid Ax \neq \varnothing\big\}$ and the *range* of A is $\operatorname{ran} A := A(X) = \bigcup_{x \in X} Ax$. The *inverse* of A is the operator $A^{-1}\colon X^* \rightrightarrows X$, defined by $x \in A^{-1}x^* \Leftrightarrow x^* \in Ax$. Furthermore, let $n \in \{2, 3, \dots\}$. Then A is *n-cyclically monotone* [1, 2, 3, 7, 10, 23] if the implication

$$\left.\begin{array}{c}(a_1, a_1^*) \in \operatorname{gra} A,\\ \vdots\\ (a_n, a_n^*) \in \operatorname{gra} A\\ a_{n+1} := a_1\end{array}\right\} \quad \Rightarrow \quad \sum_{i=1}^{n} \langle a_{i+1} - a_i, a_i^* \rangle \leq 0 \tag{1}$$

*Received by the editors November 23, 2006; accepted for publication (in revised form) August 16, 2007; published electronically November 30, 2007.
http://www.siam.org/journals/sicon/46-6/67579.html

†Mathematics, Irving K. Barber School, University of British Columbia Okanagan, Kelowna, BC V1V 1V7, Canada (heinz.bauschke@ubc.ca, shawn.wang@ubc.ca). The first author was partially supported by the Natural Sciences and Engineering Research Council of Canada and by the Canada Research Chair Program. The third author was partially supported by the Natural Sciences and Engineering Research Council of Canada.

‡Computer Science, Irving K. Barber School, University of British Columbia Okanagan, Kelowna, BC V1V 1V7, Canada (yves.lucet@ubc.ca). This author was partially supported by the Natural Sciences and Engineering Research Council of Canada.

holds. Note that 2-monotonicity simplifies to

$$\big(\forall (x,x^*)\in \operatorname{gra} A\big)\big(\forall (y,y^*)\in \operatorname{gra} A\big)\quad \langle x-y, x^*-y^*\rangle \geq 0, \tag{2}$$

i.e., to ordinary *monotonicity*. *Cyclic monotonicity* describes the situation when A is m-cyclically monotone for every $m \in \{2,3,\dots\}$. The operator A is *maximal n-cyclically monotone* if A is n-cyclically monotone and no proper extension (in the sense of inclusion of graphs) of A is n-cyclically monotone. Zorn's lemma guarantees that every n-cyclically monotone operator admits a maximal n-cyclically monotone extension. At one end of the spectrum of maximal n-cyclically monotone operators are the maximal 2-monotone, i.e., the *maximal monotone* operators. At the other end are the *maximal cyclically monotone operators*, which Rockafellar in a groundbreaking paper [19] (see Fact 3.4 below) revealed to be precisely the subdifferential operators of functions that are convex, lower semicontinuous, and proper.

This paper is motivated by the following question posed by Rockafellar in 2005 during open-problem sessions at conferences in Borovets (Bulgaria) and Banff (Canada).

> *Given a cyclically monotone operator A with a finite graph, find a method that produces an antiderivative of A that preserves the natural symmetry induced by convex duality.*

One motivation for the above question that we feel will become particularly relevant for applications as numerical convex analysis matures is the efficient storage and representation of convex functions. This is a surprisingly difficult problem. The perhaps most natural approach of storing grid points (x_i, y_i) causes significant problems because Lagrangian interpolation can fail to recover a convex function [13]. We now describe three other possible approaches. First, one could solve for subgradients x_i^* at each point x_i, or store such data in the first place, and then recover the function via $f(x) = \max_i \big(\langle x - x_i, x_i^*\rangle + y_i\big)$. The resulting function is piecewise linear with a full domain; thus, its conjugate has a bounded domain. Second, one could restrict the model of the function to the convex hull of the points x_i and set the function equal to $+\infty$ outside. Third, one could store the points and subgradients (x_i, x_i^*) along with a scalar y_0 and then recover a function f that satisfies $x_i^* \in \partial f(x_i)$ and $f(x_0) = y_0$. However, the existing representations in the literature [9, 19] are based on piecewise linear functions; so, in the finite graph case, one has to unavoidably privilege either the primal or the dual space in the very model used to recover the function.

In this paper, we provide constructive answers to Rockafellar's question. In fact, we shall exhibit methods for constructing antiderivatives that we call *primal-dual symmetric*. These methods have the property that, when they are applied to A^{-1} instead of A, the Fenchel conjugate of the antiderivative of A is obtained. The mere existence of such methods struck us initially as quite remarkable since antiderivatives are at best unique up to additive constants. These methods also allow for the design of models of convex functions that inherit the symmetry induced by convex duality in the given discrete data. Our constructions are based on Rockafellar's classical construction of an antiderivative as well as on recent work on Fitzpatrick functions and the proximal average operator, which has a close connection to fundamental objects of optimization such as Moreau envelopes and proximal mappings [4, 6]. Another pleasant consequence of primal-dual symmetric methods is their "slope 1" property—we believe that this will aid in efforts to represent convex functions in a numerically stable way (the "slope 1" property guarantees that the derivatives outside the domain of interest have slopes that are neither too small nor too large in magnitude). This is an area of active research that lies beyond the scope of this paper; see [15] for a

one-dimensional framework that is capable to express such antiderivatives and that serves as a starting point for further research.

The remainder of this paper can be summarized as follows. In section 2, we introduce the common ancestor and Fitzpatrick functions [2]. These functions have turned out to be immensely useful in the study of—and they are intimately tied to—n-cyclic monotonicity. We provide a recursion formula for the Fitzpatrick functions (Proposition 2.13) and show that they stabilize when applied to cyclically monotone operators with a finite graph (Theorem 2.16). In section 3, we revisit Rockafellar's classical antiderivative result (Fact 3.4) in the context of Fitzpatrick functions. In fact, his antiderivative satisfies a certain minimality property (Theorem 3.5), it is related to the common ancestor function (Corollary 3.11), and a closed form can be found for some finite-graph operators on $\mathbb{R}$ (Theorem 3.14)—parts of these results, of which we were unaware during the preparation of the originally submitted version of this paper, were previously obtained by Lambert et al. in their interesting work [14] in which they focus on finding upper and lower bounds for antiderivatives using a linear programming formulation. The supremum of all Rockafellar antiderivatives is expressible in terms of a Fitzpatrick function (Theorem 3.15 and Corollary 3.16). Section 4 introduces the notion of a *primal-dual symmetric* method for antiderivatives (Definition 4.6). Such methods provide antiderivatives that depend only on the graph—which makes them *intrinsic*—and that return the Fenchel conjugate of the antiderivative when applied to the inverse operator. Neither Rockafellar's classical antiderivatives nor simple symmetrizations of them have this property (Proposition 4.7). Based on recent work on the proximal average operator, we proceed to present our main result which provides a general construction of primal-dual symmetric methods (Theorem 4.13). Concrete instances are proximal-average-based symmetrizations of the maximum and of the average of Rockafellar's antiderivatives (Examples 4.19 and 4.20). We then present a result (Corollary 4.23) that leads to a resolution of Rockafellar's problem (Corollary 4.26 and Remark 4.27). We conclude the paper with a numerical example (Example 4.28).

Our notation is standard. The subdifferential operator of a convex function f is denoted by ∂f, its Fenchel conjugate by f^*, and its domain by $\operatorname{dom} f$. For a set S, we use $\operatorname{conv} S$, $\overline{\operatorname{conv}}\, S$, $\operatorname{int} S$, and $\overline{S}$ to denote its convex hull, its closed convex hull, its interior, and its closure, respectively. For a nonempty convex subset C of X and a point $x \in C$, the tangent and the normal cone of C at x are denoted by $T_C(x)$ and by $N_C(x)$, respectively. Finally, the set of all functions from X to $]-\infty,+\infty]$ that are convex, lower semicontinuous, and proper is denoted by $\Gamma(X)$ or simply by Γ.

2. The common ancestor and Fitzpatrick functions.

DEFINITION 2.1 (see [2, Definition 2.1]). *Let* $A\colon X \rightrightarrows X^*$, *and let* $(a_1, a_1^*) \in \operatorname{gra} A$. *The* common ancestor functions *are defined by*

$$C_{A,2,(a_1,a_1^*)}\colon X \times X^* \to\,]-\infty,+\infty] : (x,x^*) \mapsto \langle x, a_1^*\rangle + \langle a_1, x^*\rangle - \langle a_1, a_1^*\rangle, \tag{3}$$

and, for every $n \in \{3,4,\dots\}$, *by*

$$\begin{aligned} C_{A,n,(a_1,a_1^*)}\colon X \times X^* &\to\,]-\infty,+\infty] \\ (x,x^*) &\mapsto \sup_{\substack{(a_2,a_2^*)\in\operatorname{gra} A,\\ \vdots \\ (a_{n-1},a_{n-1}^*)\in\operatorname{gra} A}} \left(\sum_{i=1}^{n-2}\langle a_{i+1}-a_i, a_i^*\rangle\right) + \langle x-a_{n-1}, a_{n-1}^*\rangle + \langle a_1, x^*\rangle. \end{aligned} \tag{4}$$

We also set

$$C_{A,\infty,(a_1,a_1^*)} = \sup_{n\in\{2,3,\dots\}} C_{A,n,(a_1,a_1^*)}. \tag{5}$$

It is clear that

$$\big(\forall n \in \{2,3,\dots\}\big) \quad C_{A,n,(a_1,a_1^*)} \text{ is convex and lower semicontinuous,} \tag{6}$$

that the sequence

$$\big(C_{A,n,(a_1,a_1^*)}\big)_{n\in\{2,3,\dots\}} \text{ is increasing and pointwise convergent to } C_{A,\infty,(a_1,a_1^*)}, \tag{7}$$

and that $C_{A,\infty,(a_1,a_1^*)}$ is convex and lower semicontinuous. Moreover,

$$\text{gra } A \text{ finite} \quad \Rightarrow \quad \big(\forall n \in \{2,3,\dots\}\big)\ C_{A,n,(a_1,a_1^*)} \text{ is polyhedral and continuous.} \tag{8}$$

The next result shows that common ancestor functions are closely related to n-cyclic monotonicity. The proof is straightforward and thus omitted.

PROPOSITION 2.2. *Let $A\colon X \rightrightarrows X^*$, and let $n \in \{2,3,\dots\}$. Then A is n-cyclically monotone if and only if*

$$\big(\forall (a,a^*) \in \text{gra } A\big)\big(\forall (b,b^*) \in \text{gra } A\big) \quad C_{A,n,(a,a^*)}(b,b^*) \le \langle b, b^*\rangle. \tag{9}$$

Computationally convenient is the following recursive formula.

PROPOSITION 2.3 (recursion). *Let $A\colon X \rightrightarrows X^*$, let $(a_1,a_1^*) \in \text{gra } A$, let $n \in \{2,3,\dots\}$, and let $(x,x^*) \in X \times X^*$. Then*

$$C_{A,n+1,(a_1,a_1^*)}(x,x^*) = \sup_{(a,a^*)\in \text{gra } A} C_{A,n,(a_1,a_1^*)}(a,x^*) + \langle x-a, a^*\rangle. \tag{10}$$

Proof. By definition, $C_{A,n+1,(a_1,a_1^*)}(x,x^*)$ is the supremum of the terms

$$\begin{aligned}\left(\sum_{i=1}^{n-1}\langle a_{i+1}-a_i, a_i^*\rangle\right) + \langle x-a_n, a_n^*\rangle + \langle a_1, x^*\rangle \\ = \left(\sum_{i=1}^{n-2}\langle a_{i+1}-a_i, a_i^*\rangle\right) + \langle a_n - a_{n-1}, a_{n-1}^*\rangle + \langle a_1, x^*\rangle + \langle x-a_n, a_n^*\rangle,\end{aligned} \tag{11}$$

where $(a_2,a_2^*),\dots,(a_n,a_n^*)$ in gra A. Supremizing first over $(a_2,a_2^*),\dots,(a_{n-1},a_{n-1}^*)$, followed by supremizing over (a_n,a_n^*), we obtain the conclusion. ∎

Due to their implementability, operators with finite graphs are of particular interest. The next result demonstrates that, if a sufficiently high order of cyclic monotonicity is achieved, the common ancestor functions stabilize.

THEOREM 2.4. *Let $A\colon X \rightrightarrows X^*$, let $(a_1,a_1^*) \in \text{gra } A$, and let $n \in \{2,3,\dots\}$. Suppose that A is n-cyclically monotone and that* gra A *has at most n points. Then $C_{A,\infty,(a_1,a_1^*)} = C_{A,n+1,(a_1,a_1^*)}$.*

Proof. Take $(x,x^*) \in X \times X^*$, and take $m \in \{n+2, n+3,\dots\}$. It suffices to show that

$$C_{A,m,(a_1,a_1^*)}(x,x^*) \le C_{A,m-1,(a_1,a_1^*)}(x,x^*), \tag{12}$$

since this and (7) then imply that $C_{A,n+1,(a_1,a_1^*)} = C_{A,n+2,(a_1,a_1^*)} = \cdots = C_{A,\infty,(a_1,a_1^*)}$. Take $(a_2,a_2^*),\dots,(a_{m-1},a_{m-1}^*)$ in gra A. Since gra A contains at most n points and

since $m-1 \geq n+1$, there exist integers k and l such that $1 \leq k < l \leq m-1$ and $a_k = a_l$. Hence

$$\begin{aligned}(13)\quad &\sum_{i=1}^{m-2} \langle a_{i+1} - a_i, a_i^* \rangle + \langle x - a_{m-1}, a_{m-1}^* \rangle + \langle a_1, x^* \rangle \\ &= \sum_{i=1}^{k-1} \langle a_{i+1} - a_i, a_i^* \rangle + \sigma + \sum_{i=l}^{m-2} \langle a_{i+1} - a_i, a_i^* \rangle + \langle x - a_{m-1}, a_{m-1}^* \rangle + \langle a_1, x^* \rangle,\end{aligned}$$

where

$$(14)\qquad \sigma = \sum_{i=k}^{l-1} \langle a_{i+1} - a_i, a_i^* \rangle.$$

We claim that

$$(15)\qquad \sigma \leq 0.$$

Note that σ contains $l-k$ terms. If $l-k \leq n$, then the n-cyclic monotonicity of A implies that $\sigma \leq 0$. Otherwise, $l-k > n$, and we may analogously and recursively split up σ until it is a finite sum of negative terms. This verifies (15). Now (13) implies (12). □

Example 2.5. Suppose that X is a Hilbert space. Let $e \in X$ be such that $\|e\| = 1$, and define A via $\operatorname{gra} A := \{(-e, -e), (e, e)\}$. Then A is (2-cyclically) monotone, and for every $(x, x^*) \in X \times X$ we have

$$(16)\qquad C_{A,2,(-e,-e)}(x, x^*) = -\langle x + x^*, e \rangle - 1,$$

$$(17)\qquad C_{A,2,(e,e)}(x, x^*) = \langle x + x^*, e \rangle - 1,$$

$$(18)\qquad C_{A,3,(-e,-e)}(x, x^*) = \max\big\{ -\langle x + x^*, e \rangle - 1, \langle x - x^*, e \rangle - 3 \big\},$$

$$(19)\qquad C_{A,3,(e,e)}(x, x^*) = \max\big\{ \langle x^* - x, e \rangle - 3, \langle x + x^*, e \rangle - 1 \big\}.$$

THEOREM 2.6. *Let $A\colon X \rightrightarrows X^*$, and let $(a_1, a_1^*) \in \operatorname{gra} A$. Suppose that A is not cyclically monotone. Then $C_{A,\infty,(a_1,a_1^*)} \equiv +\infty$.*

Proof. There exist n points $(a_2, a_2^*), \ldots, (a_{n+1}, a_{n+1}^*)$ in $\operatorname{gra} A$, where $n \in \{2, 3, \ldots\}$, such that

$$(20)\qquad \sigma := \sum_{i=2}^{n+1} \langle a_{i+1} - a_i, a_i^* \rangle > 0, \quad \text{where } a_{n+2} := a_2.$$

Take $(x, x^*) \in X \times X^*$. Take $k \in \{2, 3, \ldots\}$, and define

$$(21)\qquad a_{kn+2} := a_{(k-1)n+2} := \cdots := a_2,$$

$$(22)\qquad a_{kn+1} := a_{(k-1)n+1} := \cdots := a_{n+1},$$

$$(23)\qquad a_{kn} := a_{(k-1)n} := \cdots := a_n,$$

$$(24)\qquad \vdots$$

$$(25)\qquad a_{(k-1)n+3} := a_{(k-2)n+3} := \cdots := a_3,$$

and analogously for $a_{n+2}^*, \ldots, a_{kn+2}^*$. Then

$$C_{A,kn+3,(a_1,a_1^*)}(x,x^*) \geq \sum_{i=1}^{kn+1} \langle a_{i+1}-a_i, a_i^*\rangle + \langle x-a_{kn+2}, a_{kn+2}^*\rangle + \langle a_1, x^*\rangle \tag{26}$$

$$= \langle a_2-a_1, a_1^*\rangle + k\sigma + \langle x-a_2, a_2^*\rangle + \langle a_1, x^*\rangle \tag{27}$$

$$\to +\infty \quad \text{as } k \to +\infty. \tag{28}$$

Therefore, $\lim_{k\to+\infty} C_{A,kn+3,(a_1,a_1^*)}(x,x^*) = +\infty$, and the result now follows from (7). □

Example 2.7. Suppose that X is a Hilbert space. Let $e \in X$ be such that $\|e\| = 1$, and define A via $\operatorname{gra} A := \{(-e,e),(e,-e)\}$. Then A is not monotone, and for every $k \in \{2,3,\ldots\}$ and $(x,x^*) \in X \times X$ we have

$$C_{A,2,(-e,e)}(x,x^*) = \langle x-x^*, e\rangle + 1, \tag{29}$$

$$C_{A,2,(e,-e)}(x,x^*) = \langle x^*-x, e\rangle + 1, \tag{30}$$

$$C_{A,2k-1,(-e,e)}(x,x^*) = 4(k-1) - 2 + \max\big\{\langle x-x^*, e\rangle - 1, -\langle x+x^*, e\rangle + 1\big\}, \tag{31}$$

$$C_{A,2k-1,(e,-e)}(x,x^*) = 4(k-1) - 2 + \max\big\{\langle x+x^*, e\rangle + 1, \langle x^*-x, e\rangle - 1\big\}, \tag{32}$$

$$C_{A,2k,(-e,e)}(x,x^*) = 4(k-1) + \max\big\{\langle x-x^*, e\rangle + 1, -\langle x+x^*, e\rangle - 1\big\}, \tag{33}$$

$$C_{A,2k,(e,-e)}(x,x^*) = 4(k-1) + \max\big\{\langle x+x^*, e\rangle - 1, \langle x^*-x, e\rangle + 1\big\}, \tag{34}$$

$$C_{A,\infty,(-e,e)}(x,x^*) = +\infty, \tag{35}$$

$$C_{A,\infty,(e,-e)}(x,x^*) = +\infty. \tag{36}$$

We now turn to Fitzpatrick functions.

Definition 2.8 (Fitzpatrick functions [2, Definition 2.2]). *Let $A\colon X \rightrightarrows X^*$. For every $n \in \{2,3,\ldots\}$, the* Fitzpatrick function of A of order n *is*

$$F_{A,n} := \sup_{(a,a^*)\in\operatorname{gra} A} C_{A,n,(a,a^*)}. \tag{37}$$

The Fitzpatrick function of A of infinite order *is*

$$F_{A,\infty} := \sup_{n\in\{2,3,\ldots\}} F_{A,n} = \sup_{(a,a^*)\in\operatorname{gra} A} C_{A,\infty,(a,a^*)}. \tag{38}$$

It is clear that each $F_{A,n}$ is convex and lower semicontinuous; moreover, if $\operatorname{gra} A$ is finite, then each $F_{A,n}$ is polyhedral and continuous. The sequence $(F_{A,n})_{n\in\{2,3,\ldots\}}$ is increasing and pointwise convergent to $F_{A,\infty}$, which is convex and lower semicontinuous. An immediate consequence of Definition 2.8 is the following result.

Proposition 2.9 ([2, Proposition 2.3]). *Let $A\colon X \rightrightarrows X^*$, and let $n \in \{2,3,\ldots\}$. Then $F_{A,n}\colon X \times X^* \to [-\infty,+\infty]$ is convex and lower semicontinuous. At $(x,x^*) \in X \times X^*$, the value of $F_{A,n}$ is given by*

$$\sup_{\substack{(a_1,a_1^*)\in\operatorname{gra} A,\\ \vdots \\ (a_{n-1},a_{n-1}^*)\in\operatorname{gra} A}} \left(\sum_{i=1}^{n-2}\langle a_{i+1}-a_i, a_i^*\rangle\right) + \langle x-a_{n-1}, a_{n-1}^*\rangle + \langle a_1, x^*\rangle. \tag{39}$$

Moreover,

$$F_{A,n} \geq \langle\cdot,\cdot\rangle \text{ on } \operatorname{gra} A. \tag{40}$$

PROPOSITION 2.10. *Let $A\colon X \rightrightarrows X^*$, let $n \in \{2,3,\dots\}$, and let $(x,x^*) \in X \times X^*$. Then $F_{A^{-1},n}(x^*,x) = F_{A,n}(x,x^*)$ and $F_{A^{-1},\infty}(x^*,x) = F_{A,\infty}(x,x^*)$.*

Proof. Take $(b_1^*,b_1),\dots,(b_{n-1}^*,b_{n-1})$ in $\operatorname{gra} A^{-1}$ and set

$$\big(\forall i \in \{1,\dots,n-1\}\big) \quad (a_i,a_i^*) := (b_{n-i},b_{n-i}^*) \in \operatorname{gra} A. \tag{41}$$

Then

$$\begin{aligned}\sum_{i=1}^{n-2}\langle b_i, b_{i+1}^* - b_i^*\rangle + \langle b_{n-1}, x^* - b_{n-1}^*\rangle + \langle x, b_1^*\rangle \\ &= \sum_{i=1}^{n-2}\langle b_i, b_{i+1}^*\rangle - \sum_{i=1}^{n-1}\langle b_i, b_i^*\rangle + \langle b_{n-1}, x^*\rangle + \langle x, b_1^*\rangle \\ &= \sum_{i=1}^{n-2}\langle a_{i+1}, a_i^*\rangle - \sum_{i=1}^{n-1}\langle a_i, a_i^*\rangle + \langle a_1, x^*\rangle + \langle x, a_{n-1}^*\rangle \\ &= \sum_{i=1}^{n-2}\langle a_{i+1}-a_i, a_i^*\rangle + \langle x - a_{n-1}, a_{n-1}^*\rangle + \langle a_1, x^*\rangle.\end{aligned} \tag{42}$$

The result follows by supremizing. ∎

FACT 2.11 ([2, Proposition 2.4 and Corollary 2.5]). *Let $A\colon X \rightrightarrows X^*$, and let $n \in \{2,3,\dots\}$. Then*

$$A \text{ is } n\text{-cyclically monotone} \Leftrightarrow F_{A,n} \le \langle\cdot,\cdot\rangle \text{ on } \operatorname{gra} A \Leftrightarrow F_{A,n} = \langle\cdot,\cdot\rangle \text{ on } \operatorname{gra} A, \tag{43}$$

and

$$A \text{ is cyclically monotone} \Leftrightarrow F_{A,\infty} \le \langle\cdot,\cdot\rangle \text{ on } \operatorname{gra} A \Leftrightarrow F_{A,\infty} = \langle\cdot,\cdot\rangle \text{ on } \operatorname{gra} A. \tag{44}$$

COROLLARY 2.12. *Let $A\colon X \rightrightarrows X^*$, and let $n \in \{2,3,\dots\}$. Then A is n-cyclically monotone if and only if A^{-1} is.*

The recursion formula for Fitzpatrick functions that we present next is an immediate consequence of Proposition 2.3. (A special case of it was utilized in [3].)

PROPOSITION 2.13 (recursion). *Let $A\colon X \rightrightarrows X^*$, let $n \in \{2,3,\dots\}$, and let $(x,x^*) \in X \times X^*$. Then*

$$F_{A,n+1}(x,x^*) = \sup_{(a,a^*)\in\operatorname{gra} A} F_{A,n}(a,x^*) + \langle x, a^*\rangle - \langle a, a^*\rangle. \tag{45}$$

Combining Fact 2.11 and Proposition 2.13, we obtain the following result which underlines the importance of the values of the Fitzpatrick function on $\operatorname{dom} A \times \operatorname{ran} A$.

COROLLARY 2.14. *Let $A\colon X \rightrightarrows X^*$, and let $n \in \{3,4,\dots\}$. Then A is n-cyclically monotone if and only if*

$$\big(\forall (a,a^*) \in \operatorname{gra} A\big)\big(\forall (b,b^*) \in \operatorname{gra} A\big) \quad F_{A,n-1}(a,b^*) - \langle a, b^*\rangle \le \langle a-b, a^*-b^*\rangle. \tag{46}$$

Example 2.15. Let $A\colon X \rightrightarrows X^*$ be monotone such that its graph contains two points, and let $n \in \{2,3,\dots\}$. Then $F_{A,n} = \langle\cdot,\cdot\rangle$ on $\operatorname{dom} A \times \operatorname{ran} A$; consequently, A is cyclically monotone.

Proof. The fact that $F_{A,n} = \langle\cdot,\cdot\rangle$ is proved readily by induction. The cyclic monotonicity of A now follows from Corollary 2.14 and from the monotonicity of A. (Alternatively, use Corollary 2.18 below.) ∎

THEOREM 2.16. *Let $A\colon X \rightrightarrows X^*$ be such that* gra A *contains at most n points, where $n \in \{2,3,\ldots\}$. Suppose that A is n-cyclically monotone. Then A is $(n+1)$-cyclically monotone and*

$$F_{A,n+1} = F_{A,n+2} = \cdots = F_{A,\infty}. \tag{47}$$

Proof. Take

$$\big\{(b_1,b_1^*),\ldots,(b_{n+1},b_{n+1}^*)\big\} \subseteq \operatorname{gra} A. \tag{48}$$

We must show that

$$\sigma := \sum_{i=1}^{n+1}\langle b_{i+1}-b_i, b_i^*\rangle \leq 0, \quad \text{where } b_{n+2} := b_1. \tag{49}$$

Since gra A contains no more than n points, there exist integers k and l such that

$$b_k = b_l \quad \text{and} \quad 1 \leq k < l \leq n+1. \tag{50}$$

Then

$$\sigma = \sigma_1 + \sigma_2, \tag{51}$$

where

$$\sigma_1 := \sum_{i=k}^{l-1}\langle b_{i+1}-b_i, b_i^*\rangle \quad \text{and} \quad \sigma_2 := \sum_{i=l}^{n+1}\langle b_{i+1}-b_i, b_i^*\rangle + \sum_{i=1}^{k-1}\langle b_{i+1}-b_i, b_i^*\rangle \tag{52}$$

are two cyclic sums, each of which contains at least one term and hence at most n terms. Since A is n-cyclically monotone, we see that $\sigma_1 \leq 0$ and that $\sigma_2 \leq 0$. Therefore, $\sigma = \sigma_1 + \sigma_2 \leq 0$. The statement concerning the Fitzpatrick functions follows from Theorem 2.4 and Definition 2.8. ∎

Example 2.17. Let $A\colon X \rightrightarrows X^*$ be such that $\operatorname{gra} A = \{(a,a^*)\}$ for some $(a,a^*) \in X \times X^*$. Then A is cyclically monotone, and for every $(x,x^*) \in X \times X^*$ we have

$$F_{A,2}(x,x^*) = F_{A,3}(x,x^*) = \cdots = F_{A,\infty}(x,x^*) = \langle a,x^*\rangle + \langle x,a^*\rangle - \langle a,a^*\rangle. \tag{53}$$

COROLLARY 2.18. *Let $A\colon X \rightrightarrows X^*$ be such that* gra A *contains at most n points, where $n \in \{2,3,\ldots\}$. Suppose that A is n-cyclically monotone. Then A is cyclically monotone.*

Proof. By Theorem 2.16, A is $(n+1)$-cyclically monotone and

$$F_{A,n+1} = F_{A,\infty}. \tag{54}$$

On the other hand, Fact 2.11 yields

$$F_{A,n+1} = \langle\cdot,\cdot\rangle \text{ on } \operatorname{gra} A. \tag{55}$$

The result follows by combining (54), (55), and Fact 2.11. ∎

COROLLARY 2.19. *Let $A\colon X \rightrightarrows X^*$ be such that* gra A *contains at most n points, where $n \in \{2,3,\ldots\}$. Then A is cyclically monotone if and only if*

$$\big(\forall (a,a^*) \in \operatorname{gra} A\big) \quad F_{A,n}(a,a^*) = \langle a,a^*\rangle. \tag{56}$$

Proof. "⇒": On $\operatorname{gra} A$, we always have $\langle \cdot, \cdot \rangle \leq F_{A,n} \leq F_{A,\infty}$. Since A is cyclically monotone, $F_{A,\infty} = \langle \cdot, \cdot \rangle$ on $\operatorname{gra} A$ and hence $F_{A,n} = \langle \cdot, \cdot \rangle$ on $\operatorname{gra} A$. "⇐": By Fact 2.11, A is n-cyclically monotone. The result now follows from Corollary 2.18. ∎

Example 2.20. Suppose that X is a Hilbert space. Let $e \in X$ such that $\|e\| = 1$, and define A via $\operatorname{gra} A = \{(-e, -e), (e, e)\}$. Then A is cyclically monotone but $F_{A,2} \neq F_{A,3}$; in fact, for every $(x, x^*) \in X \times X$, we have

$$F_{A,2}(x, x^*) = \max\big\{ -1 \pm \langle x + x^*, e \rangle \big\} \tag{57}$$

and

$$F_{A,3}(x, x^*) = \cdots = F_{A,\infty}(x, x^*) = \max\big\{ -1 \pm \langle x + x^*, e \rangle, -3 \pm \langle x - x^*, e \rangle \big\}. \tag{58}$$

Proof. The operator A is cyclically monotone since $\operatorname{gra} A \subset \operatorname{gra} \operatorname{Id} = \operatorname{gra} \partial \frac{1}{2} \| \cdot \|^2$. The formulas for $F_{A,2}$ and $F_{A,3}$ follow from Example 2.5. Theorem 2.16 shows that $F_{A,3} = \cdots = F_{A,\infty}$. Finally, we note that $F_{A,2}(2e, -2e) = -1$, whereas $F_{A,3}(2e, -2e) = 1$. ∎

The next example, which is an immediate consequence of Example 2.7 and Definition 2.8, illustrates the nonmonotone case.

Example 2.21. Suppose that X is a Hilbert space. Let $e \in X$ be such that $\|e\| = 1$, and define A via $\operatorname{gra} A := \{(-e, e), (e, -e)\}$. Then A is not monotone, and for every $k \in \{2, 3, \dots\}$ and $(x, x^*) \in X \times X$ we have

$$F_{A,2}(x, x^*) = 1 + \max\big\{ \pm \langle x - x^*, e \rangle \big\}, \tag{59}$$

$$F_{A,2k-1}(x, x^*) = 4(k-1) - 2 + \max\big\{ -1 \pm \langle x - x^*, e \rangle, 1 \pm \langle x + x^*, e \rangle \big\}, \tag{60}$$

$$F_{A,2k}(x, x^*) = 4(k-1) + \max\big\{ 1 \pm \langle x - x^*, e \rangle, -1 \pm \langle x + x^*, e \rangle \big\}, \tag{61}$$

$$F_{A,\infty}(x, x^*) = +\infty. \tag{62}$$

3. Rockafellar functions.

Definition 3.1. *Let $A\colon X \rightrightarrows X^*$, and let $f \in \Gamma$. Then f is an* antiderivative *of A if*

$$\operatorname{gra} A \subseteq \operatorname{gra} \partial f. \tag{63}$$

The following result will turn out to be useful.

Proposition 3.2. *Suppose that X is reflexive. Let $A\colon X \rightrightarrows X^*$, let $f \in \Gamma$, and suppose that f is an antiderivative of A such that $\overline{\operatorname{ran} \partial f} \subseteq \operatorname{conv} \operatorname{ran} A$. Then $\operatorname{dom} f^* = \operatorname{conv} \operatorname{ran} A$.*

Proof. On the one hand, since f is an antiderivative of A, we deduce that $\operatorname{gra} A \subseteq \operatorname{gra} \partial f \Leftrightarrow \operatorname{gra} A^{-1} \subseteq \operatorname{gra}(\partial f)^{-1} = \operatorname{gra} \partial f^* \Rightarrow \operatorname{ran} A = \operatorname{dom} A^{-1} \subseteq \operatorname{dom} \partial f^* \subseteq \operatorname{dom} f^* \Rightarrow \operatorname{conv} \operatorname{ran} A \subseteq \operatorname{conv} \operatorname{dom} f^* = \operatorname{dom} f^*$. Because $\overline{\operatorname{ran} \partial f} \subseteq \operatorname{conv} \operatorname{ran} A$, we see that $\operatorname{dom} f^* \subseteq \overline{\operatorname{dom} f^*} = \overline{\operatorname{dom} \partial f^*} = \overline{\operatorname{dom} (\partial f)^{-1}} = \overline{\operatorname{ran} \partial f} \subseteq \operatorname{conv} \operatorname{ran} A$. Altogether, $\operatorname{dom} f^* = \operatorname{conv} \operatorname{ran} A$. ∎

Definition 3.3 (Rockafellar function). *Let $A\colon X \rightrightarrows X^*$. Then the* Rockafellar functions *are defined by*

$$\big(\forall (a, a^*) \in \operatorname{gra} A\big) \quad R_{A,(a,a^*)}\colon X \to \left]-\infty, +\infty\right] : x \mapsto \sup_{n \in \{2,3,\dots\}} C_{A,n,(a,a^*)}(x, 0). \tag{64}$$

The importance of the Rockafellar functions stems from a fundamental result due to Rockafellar (see [19] or [24, Proposition 2.4.3, Theorem 3.2.8, and Corollary 3.2.11]),

which states that maximal cyclically monotone operators are precisely the subdifferential operators of convex, lower semicontinuous, and proper functions. The following part of Rockafellar's result will be utilized later.

FACT 3.4 (Rockafellar [19] or [24, Proposition 2.4.3 and Corollary 3.2.11]). *Let $A\colon X \rightrightarrows X^*$ be cyclically monotone, and let $(a, a^*) \in \operatorname{gra} A$. Then the following hold:*

(i) *$R_{A,(a,a^*)}$ is convex, lower semicontinuous, and proper, $R_{A,(a,a^*)}(a) = 0$, and $R_{A,(a,a^*)}$ is an antiderivative of A.*

(ii) *If A is maximal cyclically monotone, then any two antiderivatives of A differ only by a constant.*

Among all antiderivatives, Rockafellar functions have a special status due to the following minimality property, which was first observed in [14, Theorem 3.4] for cyclically monotone operators with a finite graph.

THEOREM 3.5. *Let $A\colon X \rightrightarrows X^*$ be cyclically monotone, and let $a \in \operatorname{dom} A$. Then*

(65)
$$(\forall a^* \in Aa)\quad R_{A,(a,a^*)} = \min\big\{f \in \Gamma(X) \mid f \text{ is an antiderivative of } A \text{ with } f(a) \geq 0\big\}$$

(66)
$$= \min\big\{f \in \Gamma(X) \mid f \text{ is an antiderivative of } A \text{ with } f(a) = 0\big\}.$$

Proof. Suppose that $f \in \Gamma$ is an antiderivative of A with $f(a) \geq 0$, and take $a^* \in Aa$ and $x \in X$. Then, for every $x \in X$, $n \in \{1, 2, \dots\}$, and $(a_1, a_1^*), \dots, (a_n, a_n^*)$ belonging to $\operatorname{gra} A$, we have

$$\begin{aligned} f(x) &\geq f(x) - f(a_n) + \left(\sum_{i=1}^{n-1} f(a_{i+1}) - f(a_i)\right) + f(a_1) - f(a) \\ &\geq \langle x - a_n, a_n^*\rangle + \left(\sum_{i=1}^{n-1} \langle a_{i+1} - a_i, a_i^*\rangle\right) + \langle a_1 - a, a^*\rangle. \end{aligned} \tag{67}$$

This implies

$$f \geq R_{A,(a,a^*)}. \tag{68}$$

In view of Fact 3.4(i), the proof is complete. □

COROLLARY 3.6. *Let $A\colon X \rightrightarrows X^*$ be cyclically monotone, let $a \in \operatorname{dom} A$, let $a_1^* \in Aa$, and let $a_2^* \in Aa$. Then $R_{A,(a,a_1^*)} = R_{A,(a,a_2^*)}$.*

Corollary 3.6 and Theorem 2.6 make the following definition well-defined.

DEFINITION 3.7. *Let $A\colon X \rightrightarrows X^*$, and let $a \in \operatorname{dom} A$. Then we set*

$$R_{A,a} := R_{A,(a,a^*)}, \tag{69}$$

where a^* is an arbitrary point in Aa.

COROLLARY 3.8. *Let $A\colon X \rightrightarrows X^*$ be cyclically monotone, and let $a \in \operatorname{dom} A$. Set $B\colon X \rightrightarrows X^*\colon x \mapsto \operatorname{conv}(Ax)$. Then B is cyclically monotone, and $R_{B,a} = R_{A,a}$.*

Proof. It is readily verified that B is cyclically monotone. Hence $R_{B,a}$ is an antiderivative of B and of A such that $R_{B,a}(a) = 0$. By Theorem 3.5, $R_{B,a} \geq R_{A,a}$. On the other hand, $R_{A,a}$ is also an antiderivative of B; thus, again by Theorem 3.5, $R_{A,a} \geq R_{B,a}$. Altogether, $R_{B,a} = R_{A,a}$. □

COROLLARY 3.9. *Suppose that X is reflexive. Let $A\colon X \rightrightarrows X^*$ be cyclically monotone, and let $(a, a^*) \in \operatorname{gra} A$. Then*

$$R_{A,a}^* = \max\big\{g \in \Gamma(X^*) \mid g \text{ is an antiderivative of } A^{-1} \text{ and } g(a^*) = \langle a, a^*\rangle\big\}. \tag{70}$$

Proof. Take $g \in \Gamma(X^*)$ such that g is an antiderivative of A^{-1} and $g(a^*) = \langle a, a^* \rangle$. Then $g^*(a) = 0$, and g^* is an antiderivative of A. By Theorem 3.5, $g^* \geq R_{A,a}$, and therefore $g^{**} = g \leq R^*_{A,a}$. ∎

Corollary 3.9 results in the following interesting counterpart to Theorem 3.5; see also [14, Proposition 4.2].

COROLLARY 3.10. *Suppose that X is reflexive. Let $A\colon X \rightrightarrows X^*$ be cyclically monotone, and let $(a, a^*) \in \operatorname{gra} A$. Then*

(71)
$$R^*_{A^{-1},a^*} - \langle a, a^* \rangle = \max\big\{f \in \Gamma(X) \mid f \text{ is an antiderivative of } A \text{ and } f(a) = 0\big\}.$$

The next result will be used later.

COROLLARY 3.11. *Let $A\colon X \rightrightarrows X^*$, let $(a, a^*) \in \operatorname{gra} A$, and let $n \in \{2, 3, \dots\}$. Suppose that $\operatorname{gra} A$ contains at most n points and that A is n-cyclically monotone. Then A is cyclically monotone, and for every $x \in X$ we have*

$$R_{A,a}(x) = C_{A,n+1,(a,a^*)}(x, 0) \tag{72}$$

$$= \max_{\substack{(a_2,a_2^*)\in\operatorname{gra} A,\\ \vdots \\ (a_n,a_n^*)\in\operatorname{gra} A}} \langle x - a_n, a_n^* \rangle + \langle a_n - a_{n-1}, a_{n-1}^* \rangle + \cdots + \langle a_2 - a, a^* \rangle. \tag{73}$$

Consequently, $R_{A,a}$ is a polyhedral and continuous antiderivative of A with $\operatorname{ran} \partial R_{A,a} \subset \operatorname{conv} \operatorname{ran} A$.

Proof. This follows from Corollary 2.18, Theorem 2.4, (8), Fact 3.4(i), and the Ioffe–Tikhomirov theorem (see, e.g., [24, Theorem 2.4.18]). ∎

Fact 3.4(ii) implies that, if A is maximal cyclically monotone, the Rockafellar functions $\{R_{A,a}\}_{a\in\operatorname{dom} A}$ differ only by constants. For finite-graph operators, this is no longer true as the following consequence of Example 2.5 and Definition 3.3 shows.

Example 3.12. Suppose that X is a Hilbert space. Let $e \in X$ be such that $\|e\| = 1$, and define A via $\operatorname{gra} A := \{(-e, -e), (e, e)\}$. Then for every $x \in X$ we have

$$R_{A,-e}(x) = \max\big\{-\langle x, e\rangle - 1, \langle x, e\rangle - 3\big\} = -2 + |\langle x, e\rangle - 1| \tag{74}$$

and

$$R_{A,e}(x) = \max\big\{\langle x, e\rangle - 1, -\langle x, e\rangle - 3\big\} = -2 + |\langle x, e\rangle + 1|. \tag{75}$$

Consequently, $R_{A,e} \not\geq R_{A,-e}$ and $R_{A,e} \not\leq R_{A,-e}$.

Remark 3.13. Let $A\colon X \rightrightarrows X^*$, and suppose that A is not cyclically monotone. Then Theorem 2.6 and Definition 3.3 imply that the Rockafellar functions $\{R_{A,a}\}_{a\in\operatorname{dom} A}$ are all identically equal to $+\infty$. For a concrete example, see Example 2.7.

Turning momentarily to the case when $X = \mathbb{R}$, we now present not only a considerable generalization of Example 3.12 but also an explicit formula for any Rockafellar function and its subdifferential operator of a cyclically monotone operator with a finite graph. See also [14, section 7].

THEOREM 3.14. *Let $A\colon \mathbb{R} \rightrightarrows \mathbb{R}$ have a finite graph, and suppose that the graph of $B\colon \mathbb{R} \rightrightarrows \mathbb{R}\colon x \mapsto \operatorname{conv}(Ax)$ is*

$$\bigcup_{i=1}^{n} \big(\{a_i\} \times [b_i^-, b_i^+]\big), \tag{76}$$

where $n \in \{1,2,\ldots\}$, $a_1 < a_2 < \cdots < a_n$, *and* $b_1^- \leq b_1^+ \leq b_2^- \leq \cdots \leq b_n^- \leq b_n^+$. *Set* $a_0 := -\infty$ *and* $a_{n+1} := +\infty$. *Suppose that* $k \in \{1,\ldots,n\}$. *Then* R_{A,a_k} *is given by*

$$\mathbb{R} \to \mathbb{R}\colon x \mapsto \begin{cases} (x-a_i)b_i^- + \sum_{j=i+1}^{k} (a_{j-1}-a_j)b_j^- & \text{if } a_{i-1} < x \leq a_i \leq a_k, \\ (x-a_i)b_i^+ + \sum_{j=k}^{i-1} (a_{j+1}-a_j)b_j^+ & \text{if } a_k \leq a_i \leq x < a_{i+1}, \end{cases} \tag{77}$$

and $\partial R_{A,a_k}$ *is given by*

$$\mathbb{R} \rightrightarrows \mathbb{R}\colon x \mapsto \begin{cases} \{b_i^-\} & \text{if } a_{i-1} < x < a_i \leq a_k, \\ [b_i^-, b_{i+1}^-] & \text{if } x = a_i < a_k, \\ [b_k^-, b_k^+], & \text{if } x = a_k, \\ [b_{i-1}^+, b_i^+] & \text{if } a_k < x = a_i, \\ \{b_i^+\} & \text{if } a_k \leq a_i < x < a_{i+1}. \end{cases} \tag{78}$$

Proof. Clearly, A and B are cyclically monotone. Denote the function described in (77) by R, and observe that R is piecewise linear, continuous everywhere, and well-defined at a_k, with

$$R(a_k) = 0. \tag{79}$$

Moreover, (77) implies that ∂R is given by (78), which is clearly monotone. Thus

$$R \in \Gamma(\mathbb{R}). \tag{80}$$

Take $i \in \{1,2,\ldots,n\}$. If $i < k$, then $Aa_i \subseteq Ba_i = [b_i^-, b_i^+] \subseteq [b_i^-, b_{i+1}^-] = \partial R(a_i)$. If $i = k$, then $Aa_i = Aa_k \subseteq Ba_k = [b_k^-, b_k^+] = \partial R(a_k)$. If $k < i$, then $Aa_i \subseteq Ba_i = [b_i^-, b_i^+] \subseteq [b_{i-1}^+, b_i^+] = \partial R(a_i)$. Thus,

$$R \text{ is an antiderivative of } A. \tag{81}$$

Since $\bigcup_{i=1}^n \{(a_i, b_i^-), (a_i, b_i^+)\} \subseteq \operatorname{gra} A$, we deduce from (73) and (77) that

$$R_{A,a_k} \geq R. \tag{82}$$

Hence (79), (80), (81), (82), and Theorem 3.5 imply that $R = R_{A,a_k}$. ∎

The next result links Rockafellar functions to Fitzpatrick functions.

THEOREM 3.15. *Let* $A\colon X \rightrightarrows X^*$. *Then*

$$\big(\forall (x,x^*) \in X \times X^*\big) \quad F_{A,\infty}(x,x^*) = \sup_{a \in \operatorname{dom} A} \langle a, x^* \rangle + R_{A,a}(x). \tag{83}$$

Proof. (See also the proof of [2, Theorem 3.5] for a variant.) Take $(x,x^*) \in X \times X^*$. Using Definitions 2.8, 2.1, 3.3, and 3.7, we see that

$$F_{A,\infty}(x,x^*) = \sup_{n \in \{2,3,\ldots\}} F_{A,n}(x,x^*) = \sup_{n \in \{2,3,\ldots\}} \sup_{(a,a^*) \in \operatorname{gra} A} C_{A,n,(a,a^*)}(x,x^*) \tag{84}$$

$$= \sup_{(a,a^*) \in \operatorname{gra} A} \sup_{n \in \{2,3,\ldots\}} C_{A,n,(a,a^*)}(x,0) + \langle a, x^* \rangle \tag{85}$$

$$= \sup_{(a,a^*) \in \operatorname{gra} A} \langle a, x^* \rangle + R_{A,(a,a^*)}(x) \tag{86}$$

$$= \sup_{a \in \operatorname{dom} A} \langle a, x^* \rangle + R_{A,a}(x), \tag{87}$$

as required. ∎

We deduce that the Fitzpatrick function of infinite order with the second variable set to zero is exactly the supremum of all Rockafellar functions.

COROLLARY 3.16. *Let* $A\colon X \rightrightarrows X^*$. *Then*

$$F_{A,\infty}(\cdot,0) = \sup_{a\in\operatorname{dom} A} R_{A,a}. \tag{88}$$

Remark 3.17. Let $A\colon X \rightrightarrows X^*$ be maximal cyclically monotone. Rockafellar [19] (see Fact 3.4) proved that $A = \partial f$, where $f \in \Gamma$ is uniquely determined up to additive constants. By [2, Theorem 3.5], $F_{A,\infty} = f \oplus f^*$. Thus Corollary 3.16 implies that

$$\sup_{a\in\operatorname{dom} A} R_{A,a} \equiv +\infty \;\Leftrightarrow\; 0 \notin \operatorname{dom} f^* \;\Leftrightarrow\; \inf f(X) = -\infty. \tag{89}$$

COROLLARY 3.18. *Let* $A\colon X \rightrightarrows X^*$ *be cyclically monotone with a finite graph, let* $x^* \in X^*$, *and set* $f := F_{A,\infty}(\,\cdot\,, x^*)$. *Then* f *is polyhedral, continuous, with a full domain,* $\operatorname{gra} A \subset \operatorname{gra} \partial f$, *and* $\overline{\operatorname{ran}\,\partial f} \subseteq \operatorname{conv}\operatorname{ran} A$.

Proof. Since $\operatorname{gra} A$ is finite, Theorem 3.15 yields

$$f = \max_{a\in\operatorname{dom} A} \langle a, x^*\rangle + R_{A,a}. \tag{90}$$

The function f is continuous, polyhedral, with a full domain, as it is the finite maximum of such functions (see Corollary 3.11). Fix $x \in X$, and set $D_x := \big\{a \in \operatorname{dom} A \mid f(x) = \langle a, x^*\rangle + R_{A,a}(x)\big\}$. On the one hand, using the Ioffe–Tikhomirov theorem (see, e.g., [24, Theorem 2.4.18]) and Corollary 3.11, we have

$$\partial f(x) = \overline{\operatorname{conv}}^*\bigcup_{a\in D_x} \partial R_{A,a}(x) \subseteq \overline{\operatorname{conv}}^*\bigcup_{a\in D_x} \operatorname{conv}\operatorname{ran} A = \operatorname{conv}\operatorname{ran} A, \tag{91}$$

where $\overline{\operatorname{conv}}^*$ denotes the weak* closed convex hull operator. Hence $\operatorname{ran}\partial f \subseteq \operatorname{conv}\operatorname{ran} A$, and thus $\overline{\operatorname{ran}\,\partial f} \subseteq \operatorname{conv}\operatorname{ran} A$, since $\operatorname{conv}\operatorname{ran} A$ is compact as a convex hull of finitely many points. On the other hand, Fact 3.4(i) implies that

$$\big(\forall a \in \operatorname{dom} A\big) \quad \operatorname{gra} A \subset \operatorname{gra}\partial R_{A,a}. \tag{92}$$

Combining (91) and (92), we conclude altogether that $\operatorname{gra} A \subset \operatorname{gra}\partial f$. ∎

Example 3.19. Suppose that X is a Hilbert space. Let $e \in X$ such that $\|e\| = 1$, and define A via $\operatorname{gra} A = \{(-e,-e),(e,e)\}$. Then A is cyclically monotone, and for every $x \in X$ we have

$$F_{A,\infty}(x,0) = \max\big\{R_{A,-e}(x), R_{A,e}(x)\big\} = \max\big\{-1 \pm \langle x, e\rangle\big\} = -1 + |\langle x, e\rangle|. \tag{93}$$

Proof. Combine Example 3.12 and Corollary 3.16. ∎

We conclude this section with a result which illustrates how Fitzpatrick functions give rise to the smallest nonnegative antiderivative.

COROLLARY 3.20. *Let* $A : X \rightrightarrows X^*$ *be cyclically monotone with a finite graph. Then*

$$F_{A,\infty}(\cdot,0) = \min\big\{f \in \Gamma(X) \mid f \text{ is an antiderivative of } A \text{ such that } f \geq 0 \text{ on } \operatorname{dom} A\big\}. \tag{94}$$

Proof. Take $f \in \Gamma(X)$ such that f is an antiderivative of A and $f \geq 0$ on $\operatorname{dom} A$. Then Theorem 3.5 implies that $(\forall a \in \operatorname{dom} A)\ f \geq R_{A,a}$; hence, by Corollary 3.16,

$$f \geq \max_{a\in\operatorname{dom} A} R_{A,a} = F_{A,\infty}(\cdot,0). \tag{95}$$

In view of Corollary 3.18 and Fact 3.4(i), $F_{A,\infty}(\cdot,0)$ is an antiderivative of A that is nonnegative on $\operatorname{dom} A$. ∎

4. Intrinsic and primal-dual symmetric methods. From now on,

(96) $\mathcal{A}$ is the set of all cyclically monotone operators on X with finite nonempty graphs.

DEFINITION 4.1. *An* intrinsic method for finding antiderivatives*—or simply an* intrinsic method*—is a mapping* $m\colon \mathcal{A} \to \Gamma\colon A \mapsto m_A$ *such that, for every* $A \in \mathcal{A}$, m_A *is an antiderivative of* A.

Example 4.2. Let $A\colon X \rightrightarrows X^*$ be cyclically monotone, and let $(a, a^*) \in \operatorname{gra} A$. Then the Rockafellar function $R_{A,(a,a^*)} = R_{A,a}$ is an antiderivative (see Fact 3.4) but—due to the dependency on a and the resulting nonuniqueness of Rockafellar functions—there is no corresponding intrinsic method m that produces Rockafellar functions. See Example 3.12 for a concrete example.

Remark 4.3. Given $A \in \mathcal{A}$, an intrinsic method m provides an antiderivatives m_A as a mapping depending only on A or, equivalently, only on the (unordered) graph of A. This key property of intrinsic methods explains why the process of providing Rockafellar functions considered in Example 4.2 is not intrinsic. Similarly, if a method computes antiderivatives by using an enumeration of the graph of A and if a different enumeration may result in a different antiderivative, then such a method cannot be intrinsic.

We now provide two intrinsic methods.

Example 4.4. Let $m\colon \mathcal{A} \to \Gamma\colon A \mapsto F_{A,\infty}(\cdot, 0) = \max_{(a,a^*)\in\operatorname{gra} A} R_{A,(a,a^*)}$. Corollaries 3.16 and 3.18 imply that m is an intrinsic method. Moreover, for every $A \in \mathcal{A}$, the antiderivative m_A has a full domain and $\overline{\operatorname{ran} \partial m_A} \subseteq \operatorname{conv} \operatorname{ran} A$.

Example 4.5. Let $A\colon X \rightrightarrows X^*$ be cyclically monotone such that $\operatorname{gra} A$ contains exactly n points, where $n \in \{1, 2, \dots\}$, and set

$$m_A := \sum_{(a,a^*)\in\operatorname{gra} A} \frac{1}{n} R_{A,(a,a^*)}. \tag{97}$$

Then m_A is an antiderivative of A that is polyhedral and continuous with a full domain and $\overline{\operatorname{ran} \partial m_A} \subseteq \operatorname{conv} \operatorname{ran} A$. Furthermore, the corresponding method $m\colon \mathcal{A} \to \Gamma\colon A \mapsto m_A$ is intrinsic.

Proof. Note that m_A is continuous and polyhedral with a full domain, as a finite sum of such functions. The sum rule (see, e.g., [24, Theorem 2.8.7(iii)]) and Fact 3.4 imply that, for every $(x, x^*) \in \operatorname{gra} A$, we have $x^* \in Ax \subseteq \sum_{(a,a^*)\in\operatorname{gra} A} \frac{1}{n} Ax \subseteq \sum_{(a,a^*)\in\operatorname{gra} A} \frac{1}{n} \partial R_{A,(a,a^*)}(x) = \partial m_A(x)$. Corollary 3.11 shows that, for every $x \in X$, we have $\partial m_A(x) = \sum_{(a,a^*)\in\operatorname{gra} A} \frac{1}{n} \partial R_{A,(a,a^*)}(x) \subseteq \sum_{(a,a^*)\in\operatorname{gra} A} \frac{1}{n} \operatorname{conv} \operatorname{ran} A = \operatorname{conv} \operatorname{ran} A$. Consequently, $\operatorname{ran} \partial m_A \subseteq \operatorname{conv} \operatorname{ran} A$, and hence $\overline{\operatorname{ran} \partial m_A} \subseteq \overline{\operatorname{conv} \operatorname{ran} A} = \operatorname{conv} \operatorname{ran} A$. It is clear that m is intrinsic. ∎

We assume from now on that

$$X \text{ is a Hilbert space.} \tag{98}$$

DEFINITION 4.6. *An intrinsic method* $\mathfrak{m}\colon \mathcal{A} \to \Gamma\colon A \mapsto \mathfrak{m}_A$ *is* primal-dual symmetric *if*

$$\big(\forall A \in \mathcal{A}\big) \quad \mathfrak{m}_{A^{-1}} = \mathfrak{m}_A^*. \tag{99}$$

PROPOSITION 4.7. *While intrinsic, neither*

$$\mathcal{A} \to \Gamma\colon A \mapsto F_{A,\infty}(\cdot, 0) = \max_{(a,a^*)\in\operatorname{gra} A} R_{A,(a,a^*)} \tag{100}$$

nor

(101)
$$\mathcal{A} \to \Gamma\colon A \mapsto \sum_{(a,a^*)\in \operatorname{gra} A} \frac{1}{n_A} R_{A,(a,a^*)}, \quad \text{where } n_A \text{ is the number of points in } \operatorname{gra} A,$$

is primal-dual symmetric.

Proof. On the one hand, both methods produce polyhedral continuous functions with a *full domain*. On the other hand, the Fenchel conjugates of such functions have a *bounded domain*. ∎

Since antiderivatives are only (and at best; see Example 3.12) unique up to a constant, it is perhaps surprising that primal-dual symmetric methods even exist. The remainder of this section is devoted to the derivation of such methods. We shall require several known notions, which we review now.

Let $A\colon X \rightrightarrows X$ be a monotone operator. The *resolvent* of A is (the single-valued, firmly nonexpansive operator) $J_A := (\operatorname{Id}+A)^{-1}$, where Id denotes the identity operator. A classical result due to Minty [16] asserts that J_A has a full domain if and only if A is maximal monotone. The proof of the following result is straightforward and hence omitted.

PROPOSITION 4.8. *Let $A\colon X \rightrightarrows X$ and $B\colon X \rightrightarrows X$ be monotone operators. Then the following are equivalent:*

(i) $\operatorname{gra} A \subseteq \operatorname{gra} B$.
(ii) $\operatorname{gra} A^{-1} \subseteq \operatorname{gra} B^{-1}$.
(iii) *J_B is an extension of J_A; i.e., $J_B = J_A$ on $\operatorname{dom} J_A = \operatorname{ran}(\operatorname{Id}+A)$.*

We further recall that, given $f \in \Gamma$, the *proximal mapping* [17] of f is $\operatorname{Prox}(f) := J_{\partial f}$. It is clear from the definition that, for two points x and x^* in X, one has

$$x^* \in \partial f(x) \quad \Leftrightarrow \quad x = \operatorname{Prox}(f)(x + x^*). \tag{102}$$

PROPOSITION 4.9. *Let $f \in \Gamma$, let $(a, a^*) \in \operatorname{gra} \partial f$, and suppose that $y \in N_{\operatorname{dom} f}(a)$. Then $a = \operatorname{Prox}(f)(2y + a + a^*)$.*

Proof. Since $N_{\operatorname{dom} f}(a)$ is a cone, we have $2y \in N_{\operatorname{dom} f}(a)$. Hence $2y + a + a^* \in a + a^* + \partial\iota_{\operatorname{dom} f}(a) \subseteq a + \partial f(a) + \partial\iota_{\operatorname{dom} f}(a) \subseteq a + \partial(f + \iota_{\operatorname{dom} f})(a) = a + \partial f(a) = (\operatorname{Id}+\partial f)(a)$, and thus $a = \operatorname{Prox}(f)(2y + a + a^*)$. ∎

We need one more notion.

DEFINITION 4.10. *Let $f_0 \in \Gamma$, and let $f_1 \in \Gamma$. The* proximal midpoint average *of f_0 and f_1 is the function*

$$\mathcal{P}(f_0, f_1) := \Big(\tfrac{1}{2}\big(f_0 + \tfrac{1}{2}\|\cdot\|^2\big)^* + \tfrac{1}{2}\big(f_1 + \tfrac{1}{2}\|\cdot\|^2\big)^*\Big)^* - \tfrac{1}{2}\|\cdot\|^2. \tag{103}$$

The proximal average, which is a generalization of the proximal midpoint average with a parameter $\lambda \in [0, 1]$ (the choice $\lambda = \frac{1}{2}$ yields the proximal midpoint average), was introduced in [6] and further studied in [4, 5, 8].

We require the following properties.

FACT 4.11. *Let $f_0 \in \Gamma$, and let $f_1 \in \Gamma$. Then the following hold:*

(i) $\mathcal{P}(f_0, f_1) = \mathcal{P}(f_1, f_0)$.
(ii) $(\mathcal{P}(f_0, f_1))^* = \mathcal{P}(f_0^*, f_1^*)$.
(iii) $\mathcal{P}(f_0, f_1) \in \Gamma$.
(iv) $\operatorname{Prox}(\mathcal{P}(f_0, f_1)) = \frac{1}{2}\operatorname{Prox}(f_0) + \frac{1}{2}\operatorname{Prox}(f_1)$.
(v) *If $f_0 \le f_1$, then $f_0 \le \mathcal{P}(f_0, f_1) \le f_1$.*
(vi) $(\forall \gamma \in \mathbb{R})\ \mathcal{P}(f_0, f_0 + \gamma) = f_0 + \frac{1}{2}\gamma$.

Proof. (i): This is clear from the definition. (ii): See [6, Theorem 6.1]. (iii): This follows from (ii). (iv): See [6, Theorem 6.1]. (v) and (vi) follow readily from the definition. (See also [5, Remark 4.15 and Example 7.1].) ∎

COROLLARY 4.12. *Let $A\colon X \rightrightarrows X$ be cyclically monotone, and let f_0 and f_1 be antiderivatives of A. Then $\mathcal{P}(f_0, f_1)$ is also an antiderivative of A.*

Proof. By assumption, $\operatorname{gra} A \subseteq \operatorname{gra} \partial f_0$ and $\operatorname{gra} A \subseteq \operatorname{gra} \partial f_1$. Using Proposition 4.8, we see that both $\operatorname{Prox}(f_0)$ and $\operatorname{Prox}(f_1)$ extend J_A. Thus, by Fact 4.11(iv), $\operatorname{Prox}(\mathcal{P}(f_0, f_1))$ also extends J_A. Utilizing Proposition 4.8 once more, we deduce that $\mathcal{P}(f_0, f_1)$ is an antiderivative of A. ∎

We are now ready for our main result.

THEOREM 4.13 (symmetrization). *Let $m\colon \mathcal{A} \to \Gamma\colon A \mapsto m_A$ be an intrinsic method. Set*

$$\mathfrak{m}\colon \mathcal{A} \to \Gamma\colon A \mapsto \mathcal{P}\big(m_A, m^*_{A^{-1}}\big). \tag{104}$$

Then $\mathfrak{m}$ is a primal-dual symmetric intrinsic method.

Proof. Fix $A \in \mathcal{A}$. Observe that Corollaries 2.12 and Corollary 4.12 imply that $\mathfrak{m}_A$ is an antiderivative of A; thus, $\mathfrak{m}$ is an intrinsic method. On the one hand, the definitions and Fact 4.11(i) yield

$$\mathfrak{m}_{A^{-1}} = \mathcal{P}\big(m_{A^{-1}}, m^*_{(A^{-1})^{-1}}\big) = \mathcal{P}\big(m_{A^{-1}}, m^*_A\big) = \mathcal{P}\big(m^*_A, m_{A^{-1}}\big). \tag{105}$$

On the other hand, Fact 4.11(ii) implies

$$\mathfrak{m}^*_A = \Big(\mathcal{P}\big(m_A, m^*_{A^{-1}}\big)\Big)^* = \mathcal{P}\big(m^*_A, m^{**}_{A^{-1}}\big) = \mathcal{P}\big(m^*_A, m_{A^{-1}}\big). \tag{106}$$

Altogether, we obtain that $\mathfrak{m}_{A^{-1}} = \mathfrak{m}^*_A$. Therefore, $\mathfrak{m}$ is primal-dual symmetric. ∎

Example 4.14. Let $m\colon \mathcal{A} \to \Gamma\colon A \mapsto m_A$ be intrinsic, and set $\mathfrak{m}\colon \mathcal{A} \to \Gamma\colon A \mapsto \mathcal{P}(m_A, m^*_{A^{-1}})$. Let $A \in \mathcal{A}$ be such that $\operatorname{gra} A \subset \operatorname{gra} \operatorname{Id}$. Then $\mathfrak{m}_A = \frac{1}{2}\|\cdot\|^2$.

Proof. Since $A = A^{-1}$, we have $\mathfrak{m}^*_A = (\mathcal{P}(m_A, m^*_{A^{-1}}))^* = (\mathcal{P}(m_A, m^*_A))^* = \mathcal{P}(m^*_A, m^{**}_A) = \mathcal{P}(m^*_A, m_A) = \mathcal{P}(m_A, m^*_A) = \mathcal{P}(m_A, m^*_{A^{-1}}) = \mathfrak{m}_A$, and the result follows. ∎

Before we present further applications of Theorem 4.13, let us discuss a non-intrinsic variant based on the original Rockafellar function.

THEOREM 4.15. *Let $A\colon X \rightrightarrows X^*$ be cyclically monotone, and let $(a, a^*) \in \operatorname{gra} A$. Set*

$$f_{A,(a,a^*)} := \mathcal{P}\big(R_{A,(a,a^*)}, R^*_{A^{-1},(a^*,a)}\big). \tag{107}$$

Then $f^*_{A,(a,a^*)} = f_{A^{-1},(a^*,a)} := \mathcal{P}\big(R_{A^{-1},(a^*,a)}, R^*_{A,(a,a^*)}\big)$. Moreover,

$$A \text{ is maximal cyclically monotone } \Rightarrow\ f_{A,(a,a^*)} = R_{A,(a,a^*)} + \tfrac{1}{2}\langle a, a^*\rangle. \tag{108}$$

Proof. The proof of $f^*_{A,(a,a^*)} = f_{A^{-1},(a^*,a)}$ is analogous to the one of Theorem 4.13. Now assume that A is maximal cyclically monotone. Since $R_{A,(a,a^*)}$ is an antiderivative of A and $R_{A^{-1},(a^*,a)}$ is an antiderivative of A^{-1}, there exists $\gamma \in \mathbb{R}$ such that

$$R^*_{A^{-1},(a^*,a)} = R_{A,(a,a^*)} + \gamma. \tag{109}$$

Conjugating (109) followed by evaluating at a^* yields $0 = R^*_{A,(a,a^*)}(a^*) - \gamma = \langle a, a^*\rangle - R_{A,(a,a^*)}(a) - \gamma = \langle a, a^*\rangle - \gamma$. Hence

$$R^*_{A^{-1},(a^*,a)} = R_{A,(a,a^*)} + \langle a, a^*\rangle, \tag{110}$$

and this readily implies that $f_{A,(a,a^*)} = \mathcal{P}(R_{A,(a,a^*)}, R_{A,(a,a^*)} + \langle a, a^*\rangle) = R_{A,(a,a^*)} + \frac{1}{2}\langle a, a^*\rangle$. ∎

If the intrinsic method m in Theorem 4.13 produces "nice" antiderivatives, then so does sometimes the symmetrized method $\mathfrak{m}$. Before we state the corresponding result more precisely, we recall the required properties of the proximal midpoint average. Since these properties were stated in finite-dimensional Hilbert spaces, we assume from now on that

$$X \text{ is a finite-dimensional Hilbert space.} \tag{111}$$

Recall that $f \in \Gamma$ is *piecewise linear-quadratic* if $\operatorname{dom} f$ can be written as a finite union of polyhedral sets on which f is of the form $\langle x, Ax\rangle + \langle x, b\rangle + \gamma$, where $A\colon X \to X$ is linear, $b \in X$, and $\gamma \in \mathbb{R}$. The piecewise linear-quadratic functions on X have many nice properties; see [21, sections 10.E and 11.D].

FACT 4.16. *Let $f_0 \in \Gamma$, and let $f_1 \in \Gamma$ be such that f_0 and f_1^* have a full domain. Then the following hold:*

(i) *$\mathcal{P}(f_0, f_1)$ and $\mathcal{P}(f_0^*, f_1^*)$ have a full domain.*
(ii) *If f_0 and f_1 are piecewise linear-quadratic, then so is $\mathcal{P}(f_0, f_1)$.*
(iii) *If f_0 is differentiable and f_1 is strictly convex, then both $\mathcal{P}(f_0, f_1)$ and its conjugate are differentiable and strictly convex.*

Proof. (i): See [5, Theorem 6.2.(i)]. (ii): The functions f_0, f_1, and $\frac{1}{2}\|\cdot\|^2$ are piecewise linear-quadratic. The operations employed to create $\mathcal{P}(f_0, f_1)$ do not lead outside the class of piecewise linear-quadratic functions; consequently, $\mathcal{P}(f_0, f_1)$ is piecewise linear-quadratic as well. (See also [15, Corollary 5.3].) (iii): This follows from (i) and [5, Theorems 6.2.(ii) and 6.2.(iii)]. ∎

COROLLARY 4.17. *Let $m\colon \mathcal{A} \to \Gamma\colon A \mapsto m_A$ be an intrinsic method that produces antiderivatives with a full domain. Set*

$$\mathfrak{m}\colon \mathcal{A} \to \Gamma\colon A \mapsto \mathcal{P}\big(m_A, m_{A^{-1}}^*\big). \tag{112}$$

Then $\mathfrak{m}$ is a primal-dual symmetric intrinsic method, and the following hold:

(i) *$(\forall A \in \mathcal{A})$ $\mathfrak{m}_A$ and $\mathfrak{m}_A^*$ have a full domain.*
(ii) *If $(\forall A \in \mathcal{A})$ m_A is piecewise linear-quadratic, then $(\forall A \in \mathcal{A})$ $\mathfrak{m}_A$ and $\mathfrak{m}_A^*$ are both piecewise linear-quadratic.*

Proof. Theorem 4.13 states that $\mathfrak{m}$ is primal-dual symmetric. Now fix $A \in \mathcal{A}$, set $f_0 := m_A$, and set $f_1 := m_{A^{-1}}^*$. Then $\mathfrak{m}_A = \mathcal{P}(f_0, f_1)$ and, by Fact 4.11(i) and (ii), $\mathfrak{m}_A^* = \mathcal{P}(f_0^*, f_1^*) = \mathcal{P}(f_1^*, f_0^*)$. (i): Since f_0 and f_1^* have a full domain, Fact 4.16(i) (applied to f_0 and f_1 and to f_1^* and f_0^*) implies that $\mathfrak{m}_A$ and $\mathfrak{m}_A^*$ have a full domain. (ii): This is clear from Fact 4.16(ii). ∎

Remark 4.18. Consider Corollary 4.17. In general, antiderivatives are neither differentiable nor strictly convex. However, if for a particularly nice instance $A \in \mathcal{A}$ both m_A and $m_{A^{-1}}$ are differentiable, then we deduce from Fact 4.16(iii) that $\mathfrak{m}_A$ and $\mathfrak{m}_A^*$ are both differentiable and strictly convex. Analogous comments can be made for other symmetrizations of (not necessarily intrinsic) methods based on the proximal midpoint average.

We are now able to provide two examples of primal-dual symmetric intrinsic methods with very nice properties. These are in striking contrast to Proposition 4.7.

Example 4.19. Set $m\colon \mathcal{A} \to \Gamma\colon A \mapsto \max_{(a,a^*)\in \operatorname{gra} A} R_{A,(a,a^*)}$ and $\mathfrak{m}\colon \mathcal{A} \to \Gamma\colon A \mapsto \mathcal{P}(m_A, m_{A^{-1}}^*)$. Then $\mathfrak{m}$ is a primal-dual symmetric intrinsic method, and for every $A \in \mathcal{A}$ both $\mathfrak{m}_A$ and $\mathfrak{m}_A^*$ have a full domain and are piecewise linear-quadratic antiderivatives of A and A^{-1}, respectively.

Proof. For every $A \in \mathcal{A}$, m_A is a convex polyhedral (hence piecewise linear-quadratic) antiderivative of A with full domain (Example 4.4). The conclusion is now a consequence of Corollary 4.17. □

Example 4.20. Set $m\colon \mathcal{A} \to \Gamma\colon A \mapsto \frac{1}{n_A}\sum_{(a,a^*)\in \operatorname{gra} A} R_{A,(a,a^*)}$, where n_A is the number of points in $\operatorname{gra} A$, and $\mathfrak{m}\colon \mathcal{A} \to \Gamma\colon A \mapsto \mathcal{P}(m_A, m^*_{A^{-1}})$. Then $\mathfrak{m}$ is a primal-dual symmetric intrinsic method, and for every $A \in \mathfrak{m}$ both $\mathfrak{m}_A$ and $\mathfrak{m}^*_A$ have a full domain and are piecewise linear-quadratic antiderivatives of A and A^{-1}, respectively.

Proof. For every $A \in \mathcal{A}$, m_A is a convex polyhedral (hence piecewise linear-quadratic) antiderivative of A with a full domain (Example 4.5). The conclusion follows from Corollary 4.17. □

In the remainder of this section, we aim to extract further nice properties enjoyed by these two methods. We require the following results on the proximal midpoint average.

PROPOSITION 4.21. *Let $f_0 \in \Gamma$, let $f_1 \in \Gamma$, and set $f := \mathcal{P}(f_0, f_1)$. Suppose that $a^* \in \partial f_0(a) \cap \partial f_1(a)$. Then*

$$\big(\forall x \in N_{\operatorname{dom} f_0}(a) \cap N_{\operatorname{dom} f_1^*}(a^*)\big) \quad x + a^* \in \partial f(x + a). \tag{113}$$

Proof. Take $x \in N_{\operatorname{dom} f_0}(a) \cap N_{\operatorname{dom} f_1^*}(a^*)$. Proposition 4.9 yields

$$a = \operatorname{Prox}(f_0)(2x + a + a^*). \tag{114}$$

The same result (applied to f_1^*) shows that $a^* = \operatorname{Prox}(f_1^*)(2x + a^* + a)$, which is equivalent to $a^* = (\operatorname{Id} - \operatorname{Prox}(f_1))(2x + a + a^*)$, i.e., to

$$2x + a = \operatorname{Prox}(f_1)(2x + a + a^*). \tag{115}$$

Add (114) and (115), divide the result by 2, and recall Fact 4.11(iv) to deduce that

$$x + a = \operatorname{Prox}(f)\big((x + a) + (x + a^*)\big). \tag{116}$$

The conclusion now follows from (102). □

THEOREM 4.22. *Let $f_0 \in \Gamma$, let $f_1 \in \Gamma$, and set $f := \mathcal{P}(f_0, f_1)$. Suppose that $a^* \in \partial f_0(a) \cap \partial f_1(a)$, and set $N := (N_{\operatorname{dom} f_0}(a) \cap N_{\operatorname{dom} f_1^*}(a^*)) \cup (N_{\operatorname{dom} f_1}(a) \cap N_{\operatorname{dom} f_0^*}(a^*))$. Then the following hold:*

(i) $(\forall y \in a + N)\ y + a^* - a \in \partial f(y)$.

(ii) *f is differentiable at every point $y \in a + \operatorname{int} N$, with $\nabla f(y) = y + a^* - a$.*

Proof. Proposition 4.21 implies (i). On $a + \operatorname{int} N$, we note that $y \mapsto y + a^* - a$ is a continuous selection of ∂f; therefore, $\nabla f(y) = y + a^* - a$ by [18, Proposition 2.8], and (ii) holds. □

COROLLARY 4.23. *Let $m\colon \mathcal{A} \to \Gamma\colon A \mapsto m_A$ be an instrinsic method such that, for every $A \in \mathcal{A}$, $\overline{\operatorname{ran} \partial m_A} \subseteq \operatorname{conv} \operatorname{ran} A$. Set*

$$\mathfrak{m}\colon \mathcal{A} \to \Gamma\colon A \mapsto \mathcal{P}\big(m_A, m^*_{A^{-1}}\big), \tag{117}$$

take $A \in \mathcal{A}$, take $(a, a^*) \in \operatorname{gra} A$, and set $N := N_{\operatorname{conv} \operatorname{dom} A}(a) \cap N_{\operatorname{conv} \operatorname{ran} A}(a^*)$. Then the following hold:

(i) $(\forall y \in a + N)\ y + a^* - a \in \partial \mathfrak{m}_A(y)$.

(ii) $(\forall y \in a + \operatorname{int} N)\ \mathfrak{m}_A$ *is differentiable at* y, *with* $\nabla \mathfrak{m}_A(y) = y + a^* - a$.

Proof. Set $f_0 := m_A$, and set $f_1 := m^*_{A^{-1}}$ so that $\mathfrak{m}_A = \mathcal{P}(f_0, f_1)$. Proposition 3.2 implies that $\operatorname{dom} f_0^* = \operatorname{dom} m_A^* = \operatorname{conv} \operatorname{ran} A$ and that $\operatorname{dom} f_1 = \operatorname{dom} m^*_{A^{-1}} = \operatorname{conv} \operatorname{ran} A^{-1} = \operatorname{conv} \operatorname{dom} A$. The conclusion is therefore a consequence of Theorem 4.22. ∎

Remark 4.24. In view of Examples 4.4 and 4.5, we observe that Corollary 4.23 is applicable when m is either as in Example 4.19 or as in Example 4.20.

Example 4.25. Let m and $\mathfrak{m}$ be as in Corollary 4.23, let $(a, a^*) \in X \times X$, and suppose that $A\colon X \rightrightarrows X$ is given by $\operatorname{gra} A = \{(a, a^*)\}$. Then there exists $\gamma \in \mathbb{R}$ such that $\mathfrak{m}_A = \frac{1}{2}\|\cdot\|^2 + \langle \cdot, a^* - a\rangle + \gamma$.

Proof. Indeed, the set N in Corollary 4.23 is the entire space X, and hence item (ii) of that result implies that $\nabla \mathfrak{m}_A\colon X \to X\colon x \mapsto x + a^* - a$. ∎

We observe next that, on the real line, the subdifferential extending A is actually single-valued—i.e., it corresponds to a gradient—with slope one outside the box $(\operatorname{conv} \operatorname{dom} A) \times (\operatorname{conv} \operatorname{ran} A)$.

COROLLARY 4.26. *Suppose that* $X = \mathbb{R}$, *let* m *and* $\mathfrak{m}$ *be as in Corollary* 4.23, *and let* $A\colon \mathbb{R} \rightrightarrows \mathbb{R}$ *have finite graph* $\{(a_1, a_1^*), \ldots, (a_n, a_n^*)\}$, *where* $a_1 \le a_2 \le \cdots \le a_n$ *and* $a_1^* \le a_2^* \le \cdots \le a_n^*$. *Then*

$$(118)\quad (\forall x < a_1)\quad \mathfrak{m}'_A(x) = x - a_1 + a_1^* \qquad \text{and} \qquad (\forall x > a_n)\quad \mathfrak{m}'_A(x) = x - a_n + a_n^*.$$

Proof. Since $\operatorname{conv} \operatorname{dom} A = [a_1, a_n]$ and $\operatorname{conv} \operatorname{ran} A = [a_1^*, a_n^*]$, the result follows from Corollary 4.23(ii) (applied at (a_1, a_1^*) and at (a_n, a_n^*)). ∎

Remark 4.27. Primal-dual symmetry and the "slope one" property of the extension of the cyclically monotone operator A in Corollary 4.26 were properties deemed desirable by Rockafellar. In view of Remark 4.24, there exist two explicit methods that generate antiderivatives with these desirable properties. Although not the product of an intrinsic method, the function $f_{A,(a,a^*)}$ of Theorem 4.15 has the same properties.

We conclude this paper by numerically illustrating an antiderivative produced by the primal-dual symmetric intrinsic method of Example 4.19.

Example 4.28. Define $A\colon \mathbb{R} \rightrightarrows \mathbb{R}$ via $\operatorname{gra} A := \{(a, \exp(a)) \mid a \in \{0, \pm\frac{1}{2}, \pm 1\}\}$. Because $\operatorname{gra} A \subset \operatorname{gra}(\exp)$, the operator A is cyclically monotone. We interpret A as a 5-point sample of the gradient of the exponential function. Let m and $\mathfrak{m}$ be as in Example 4.19. Figure 1 visualizes the exponential function, the antiderivative m_A, the antiderivative $m^*_{A^{-1}}$, and the antiderivative $\mathfrak{m}_A$ produced by the primal-dual symmetric intrinsic method $\mathfrak{m}$. As predicted by Corollary 3.20, the function m_A is nonnegative on $\operatorname{dom} A$. In Figure 2, we visualized the derivative of the exponential function, its 5-point sample corresponding to A, and the maximal cyclically monotone extension $\partial \mathfrak{m}_A$. Note that by Theorem 3.14 the Rockafellar functions are piecewise linear, and hence their subdifferential operators have a "staircase" graph. On the other hand, $\mathfrak{m}_A$ is piecewise linear-quadratic, and its subdifferential operator displays the "slope one" property guaranteed by Corollary 4.26 outside the rectangle $\operatorname{dom} A \times \operatorname{ran} A$. Both plots were generated in Scilab utilizing software packages discussed in [15]; further details on the numerical implementation will appear elsewhere.

Acknowledgments. The authors thank Jean-Baptiste Hiriart-Urruty for making them aware of [14] and two referees for their pertinent comments.

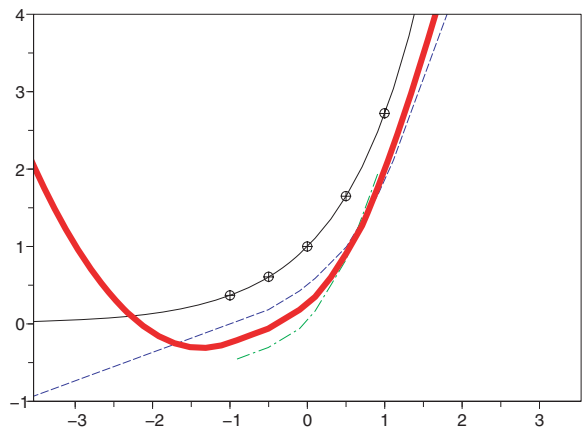


FIG. 1. *The graph of the exponential function (thin black curve) and the* 5 *points (circled) on its graph that led to the operator* A*, the antiderivative* m_A *(dashed blue curve), the antiderivative* $m^*_{A^{-1}}$ *(dashed-dotted green curve), and the proximal-average-based antiderivative* $\mathfrak{m}_A$ *(thick red curve) are shown. Note that* $m_A \geq 0$ *on* $\operatorname{dom} A$*, in accordance with Corollary* 3.20.

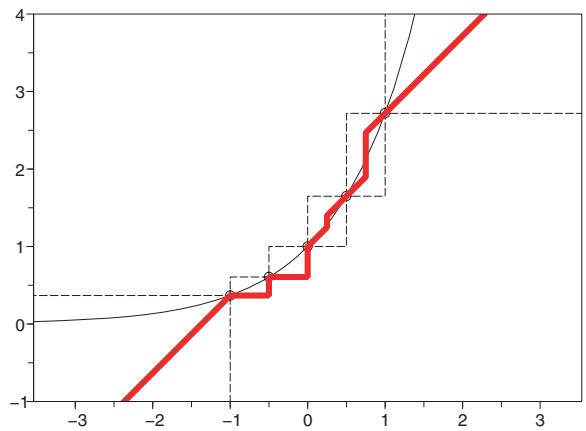


FIG. 2. *The finite graph operator* A *is shown as points (circled) on the graph of the exponential function (thin black curve), which is the same as its derivative. The reconstructed subdifferential operator* $\partial\mathfrak{m}_A$ *(thick red curve) stays inside the rectangles (dashed line) imposed on any monotone extension of* A*. Note the "slope one" property of* $\partial\mathfrak{m}_A$ *outside the rectangle* $\operatorname{conv}\operatorname{dom} A \times \operatorname{conv}\operatorname{ran} A$*, as guaranteed by Corollary* 4.26.

REFERENCES

[1] E. ASPLUND, *A Monotone Convergence Theorem for Sequences of Nonlinear Mappings*, Nonlinear Functional Analysis, in Proceedings of Symposia in Pure Mathematics XVIII Part 1, American Mathematical Society, Chicago, 1970, pp. 1–9.

[2] S. BARTZ, H. H. BAUSCHKE, J. M. BORWEIN, S. REICH, AND X. WANG, *Fitzpatrick functions, cyclic monotonicity, and Rockafellar's antiderivative*, Nonlinear Anal., 66 (2007), pp. 1198–1223.

[3] H. H. BAUSCHKE, J. M. BORWEIN, AND X. WANG, *Fitzpatrick functions and continuous linear monotone operators*, SIAM J. Optim., 18 (2007), pp. 789–809.

[4] H. H. BAUSCHKE, R. GOEBEL, Y. LUCET, AND X. WANG, *The proximal average: Basic theory*, SIOPT, submitted.

[5] H. H. Bauschke, Y. Lucet, and M. Trienis, *How to transform one convex function continuously into another*, SIAM Rev., to appear.
[6] H. H. Bauschke, E. Matoušková, and S. Reich, *Projection and proximal point methods: Convergence results and counterexamples*, Nonlinear Anal., 56 (2004), pp. 715–738.
[7] H. H. Bauschke and X. Wang, *A convex-analytical approach to extension results for n-cyclically monotone operators*, Set-Valued Anal., 15 (2007), pp. 297–306.
[8] H. H. Bauschke and X. Wang, *The Kernel Average for Two Convex Functions and Its Application to the Extension and Representation of Monotone Operators*, preprint available at http://www.optimization-online.org/DB_HTML/2007/05/1658.html
[9] J. Benoist and A. Daniilidis, *Subdifferential representation of convex functions: Refinements and applications*, J. Convex Anal., 12 (2005), pp. 255–265.
[10] J. M. Borwein, *Maximal monotonicity via convex analysis*, J. Convex Anal., 13 (2006), pp. 561–586.
[11] J. M. Borwein and Q. J. Zhu, *Techniques of Variational Analysis*, Springer-Verlag, Berlin, 2005.
[12] H. Brézis, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North-Holland, Amsterdam, 1973.
[13] P. Choné and H. V. J. Le Meur, *Non-convergence result for conformal approximation of variational problems subject to a convexity constraint*, Numer. Funct. Anal. Optim., 22 (2001), pp. 529–547.
[14] D. Lambert, J.-P. Crouzeix, V. H. Nguyen, and J.-J. Strodiot, *Finite convex integration*, J. Convex Anal., 11 (2004), pp. 131–146.
[15] Y. Lucet, H. H. Bauschke, and M. Trienis, *The piecewise linear-quadratic model for computational convex analysis*, Comput. Optim. Appl., to appear.
[16] G. J. Minty, *On the maximal domain of a monotone function*, Michigan Math. J., 8 (1961), pp. 135–137.
[17] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[18] R. R. Phelps, *Convex Functions, Monotone Operators, and Differentiability*, second ed., Lecture Notes in Math. 1364, Springer-Verlag, Berlin, 1993.
[19] R. T. Rockafellar, *On the maximal monotonicity of subdifferential mappings*, Pacific J. Math., 33 (1970), pp. 209–216.
[20] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[21] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[22] S. Simons, *Minimax and Monotonicity*, Lecture Notes in Math. 1693, Springer-Verlag, Berlin, 1998.
[23] M. D. Voisei, *Extension theorems for k-monotone operators*, Stud. Cercet. Stiinţ. Ser. Mat. Univ. Bacău, 9 (1999), pp. 235–242.
[24] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2052–2070

COMPUTING THE L_1-NORM OF CONTINUOUS-TIME LINEAR SYSTEMS*

ARNO LINNEMANN†

Abstract. An algorithm is presented for the computation of the L_1-norm of a linear time-invariant continuous-time system. It is based on the numerical integration of the impulse response and is shown to be quadratically convergent.

Key words. system norm, L_1-norm, computation, impulse response, matrix exponential, sensitivity, integration, trapezoidal rule, nondifferentiable integrand, truncation error, control system

AMS subject classifications. 15A23, 65D30, 65D32, 65F30, 93B17, 93B35, 93B40, 93C05, 93C15, 93C35, 93C73, 93D09

DOI. 10.1137/040613482

1. Introduction.

System norms are an established tool for measuring performance and robustness of linear systems. Most important are the H_∞-, H_2-, and L_1-norms. Therefore, efficient and reliable algorithms for the computation of these norms are required. The H_2-norm is easiest to compute using appropriate Lyapunov equations. The H_∞-norm can be determined by analyzing the eigenvalues of an associated Hamiltonian matrix, and a careful implementation of corresponding algorithms works fine in most cases. See [2, 4, 6, 7, 9, 10, 14] and the references given therein. In contrast to these results, almost nothing is known regarding the computation of the L_1-norm. For stable single-input single-output continuous-time linear systems

$$\Sigma: \quad \dot{x} = Ax + Bu,\ y = Cx \tag{1}$$

the L_1-norm is given by

$$||\Sigma|| := \int_0^\infty |Ce^{At}B|\, dt. \tag{2}$$

To compute an estimate of the L_1-norm, it is generally advised (see, e.g., [3]) to numerically integrate the function $|Ce^{At}B|$ over a finite interval, where the discrete function values are obtained by simulating the differential equation.

A straightforward implementation of this approach faces two main problems. First, the integrand is nondifferentiable, so that rather slow convergence of standard numerical integration algorithms is expected [12]. Second, it is not clear how the truncation error (obtained by integrating over a finite interval only) behaves.

This paper presents an efficient and reliable algorithm to compute the L_1-norm using the above "simulate and integrate" approach. More specifically, it is shown how the standard trapezoidal rule of numerical integration can be modified to efficiently handle the nondifferentiable integrand. It is shown that the new numerical integration procedure converges quadratically. Moreover, the truncation error is estimated by a computable L_2-norm using the Hölder inequality. Finally, and most importantly for

*Received by the editors August 16, 2004; accepted for publication (in revised form) May 29, 2007; published electronically December 5, 2007.
http://www.siam.org/journals/sicon/46-6/61348.html
†Control and Automation, Department of Electrical Engineering and Computer Science, University of Kassel, D-34109 Kassel, Germany (linnemann@uni-kassel.de).

quick convergence in applications, very efficient bounds for the impulse response are derived. These bounds can be employed to integrate without any error in many time intervals, and with relatively small errors in other intervals.

The paper is organized as follows. In section 2, a basic algorithm is described to compute an estimate of the L_1-norm for given discretization points $t_1, t_2, \ldots, t_N$. The details of the algorithm and the underlying theory are developed in sections 3 to 5. Section 6 shows how to adaptively add discretization points to the list $t_1, t_2, \ldots, t_N$ in order to systematically improve the norm estimate. This results in a quadratically convergent algorithm, which computes the L_1-norm to any given accuracy in finitely many steps. Section 7 provides an estimate of the computational cost, and section 8 presents some examples. Sections 2 to 8 are devoted to the single-input single-output case. Extensions to multivariable systems with feedthrough are presented in section 9. The proofs are collected in Appendices A to E. Appendix F contains a closed formula for the L_1-norm of second order systems.

2. The basic algorithm. In this section, an algorithm is described to compute the L_1-norm of a linear system Σ along with a guaranteed error bound. This means that an estimate η for the norm and an error bound ε are computed such that

$$\big|\|\Sigma\| - \eta\,\big| \le \varepsilon. \tag{3}$$

The algorithm is based on the numerical integration of the absolute value of the impulse response

$$g(t) := Ce^{At}B,$$

defined by the matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$, and $C \in \mathbb{R}^{1\times n}$ in (1). It is assumed that the discretization points $t_1 = 0 < t_2 < t_3 < \cdots < t_N$ are given and fixed. The multivariable case and an adaptive version of the algorithm, where the discretization grid is automatically refined and enlarged, are presented later in this paper.

The basic algorithm consists of N parts, namely, to find $\eta_1, \ldots, \eta_{N-1}$ and $\varepsilon_1, \ldots, \varepsilon_{N-1}$ such that

$$\left|\ \int_{t_i}^{t_{i+1}} |g(t)|\,dt - \eta_i\ \right| \le \varepsilon_i,$$

$i = 1, 2, \ldots, N-1$, and to find ε_N such that

$$\int_{t_N}^{\infty} |g(t)|\,dt \le \varepsilon_N.$$

The norm estimate and the associated error in (3) are then given by $\eta = \sum_{i=1}^{N-1} \eta_i$ and $\varepsilon = \sum_{i=1}^{N} \varepsilon_i$, respectively. The numbers $\varepsilon_1, \ldots, \varepsilon_{N-1}$ and the number ε_N are bounds for the so-called approximation and truncation errors, respectively.

The intervals $[t_i, t_{i+1}]$, $i = 1, \ldots, N-1$, can be divided into two categories: The first category consists of those intervals in which it can be guaranteed that $g(t)$ does not change sign. In these intervals, the integral can be exactly computed:

$$\eta_i = \left|CA^{-1}\left(e^{At_{i+1}}B - e^{At_i}B\right)\right|; \quad \varepsilon_i = 0. \tag{4}$$

The second category consists of those intervals which involve a possible sign change in $g(t)$. In these intervals, the estimate η_i can be computed using numerical integration and ε_i is an associated error bound.

To numerically exclude sign changes, lower and upper bounds $\underline{G}_i$ and $\overline{G}_i$ are computed such that

$$\underline{G}_i \le g(t) \le \overline{G}_i, \quad t \in [t_i, t_{i+1}]. \tag{5}$$

The tighter these bounds are, the more intervals exist where the integral is computed using (4) without approximation error. Therefore, the efficiency of the algorithm depends crucially on these bounds. A large part of this paper (section 3 and Appendices A–C) is devoted to deriving new and efficient bounds $\underline{G}_i$ and $\overline{G}_i$.

In the second category, the standard Newton–Cotes formulas of numerical integration cannot be applied, because the associated error bounds assume differentiability of the integrand [12]. Therefore, in section 4 of this paper, the trapezoidal rule and the associated error bounds are modified to be applicable to the possible nondifferentiable integrand $|g(t)|$. They require bounds on $\ddot{g}(t) = CA^2e^{At}B$, which can again be computed using the results of section 3. The resulting bounds $\varepsilon_1, \ldots, \varepsilon_{N-1}$ for the approximation error are new and computable. They are very efficient, since they are either zero or proportional to $(t_{i+1} - t_i)^3$.

The bound ε_N for the truncation error is determined in section 5 by relating the L_1-norm to the L_2-norm using the Hölder inequality and exponential weighting [5].

3. Bounds for the impulse response. The performance of the algorithm presented in section 2 depends crucially on bounds $\underline{G}_i$ and $\overline{G}_i$ such that (5) is satisfied. In this section, these bounds are derived by taking ideas from [11, 13]. These references present various bounds for the sensitivity of the matrix exponential $e^{A+\Delta}$, where Δ is an unstructured perturbation matrix of the same size as A. They show that the tightest bounds are generally provided by the diagonal form, provided the condition number $||S|| \cdot ||S^{-1}||$ of the eigenvector matrix of A is not too large. For an ill-conditioned eigenvector matrix or for multiple eigenvalues, the Schur form provides tighter bounds; see [11, 13]. The advantages of both approaches (diagonal form and Schur form) can be combined by considering a block-diagonal ordered upper-triangular (BDOUT) form, which can be computed using an algorithm suggested in [1]. More specifically, a matrix $A \in \mathbb{C}^{n\times n}$ is said to be in BDOUT form if $A = D + N$, where D is diagonal with eigenvalues ordered according to decreasing real parts and N is strictly upper triangular and block diagonal. More specifically, for stable matrices A,

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}, \quad \lambda_i = \alpha_i + j\omega_i,\ \omega_i \in \mathbb{R},\ \alpha_n \le \alpha_{n-1} \le \cdots \le \alpha_1 < 0,$$

$$N = \begin{bmatrix} N_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & N_k \end{bmatrix}, \ N_j = \begin{bmatrix} 0 & a_{j12} & \cdots & a_{j1n_j} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{j,n_j-1,n_j} \\ 0 & \cdots & \cdots & 0 \end{bmatrix} \in \mathbb{C}^{n_j\times n_j}.$$

Note that the diagonal form, the Jordan form, and the Schur form are special cases of the BDOUT form. Reference [1] describes an algorithm to put a matrix A into BDOUT form by similarity transformations, where the dimensions n_j of the blocks are as small as possible while avoiding ill-conditioned similarity transformations. Note

that similarity transformations SAS^{-1}, SB, CS^{-1} do not affect $g(t)$, so that, without loss of generality, it will be assumed that A is in BDOUT form.

The bounds $\underline{G}_i$ and $\overline{G}_i$ refer to a "perturbation" of the time t in the matrix exponential e^{At}. Compared to the perturbations Δ in [11, 13] this means that Δ has the very special structure $\Delta = A\tau$, where τ is a scalar perturbation. It is clear that much stronger results can be expected under this restricted perturbation class. This is the topic of the present section.

For a (possibly complex) matrix X, let $|X|$ be the real matrix of the same size as X, which is obtained by taking the absolute value componentwise. Moreover, for a (possibly complex) matrix A and a positive scalar T, let

$$\delta(A,T) := \max\{|e^{At} - I|;\ 0 \le t \le T\},$$

where the maximum is taken componentwise. Thus, $\delta(A,T)$ is a matrix of the same size as A.

3.1. Using perturbation results. Taking $T := t_{i+1} - t_i$, the following bounds are obvious:

$$\underline{G}_i := g(t_i) - |C| \cdot \delta(A,T) \cdot |e^{At_i}B|; \quad \overline{G}_i := g(t_i) + |C| \cdot \delta(A,T) \cdot |e^{At_i}B|.$$

"Pulling out" C and $e^{At_i}B$ componentwise using $|\cdot|$, rather than in terms of norms, generally results in tighter bounds, which can be seen from very simple examples like $C = [1\ 1]$, $e^{At_i}B = [1;\ 1]$, $\delta(A,T) = [1\ 1;\ 1\ 0]$. Since $g(t_i)$ and $e^{At_i}B$ are available from the simulation, it remains to derive bounds for $\delta(A,T)$.

If A is diagonal, then $\delta(A,T)$ is also diagonal, with ith entry given by $\delta(\lambda_i,T)$. If λ_i is real, then $\delta(\lambda_i,T)$ can be efficiently bounded by $1 - e^{\lambda_i T}$. The following lemma handles complex λ_i's. Its proof is given in Appendix A.

LEMMA 1. *Let $\lambda = \alpha + j\omega$, $\alpha < 0$, $\omega \neq 0$, and $T > 0$ be given. Let η be an arbitrary number such that $0 \le \eta \le 1$. Then $\delta(\lambda,T) \le \hat{\delta}(\lambda,T,\eta)$, where*

$$\hat{\delta}(\lambda,T,\eta) := \begin{cases} e^{|\lambda|T} - 1 & \text{if} \quad T \le t_1, \\ \max\{\eta,\ (1 + e^{2\alpha t_1} - 2e^{\alpha T}\cos(\omega T))^{\frac{1}{2}}\} & \text{if} \quad t_1 \le T \le t_2, \\ (1 + e^{2\alpha t_1} - 2e^{\alpha t_2}\cos(\omega t_2))^{\frac{1}{2}} & \text{if} \quad T \ge t_2, \end{cases}$$

and

$$t_1 := \frac{\ln(1+\eta)}{|\lambda|}, \quad t_2 := \frac{1}{|\omega|}\left(\pi + \arctan\left(\frac{\alpha}{|\omega|}\right)\right).$$

Given λ, T, and η, the bound $\hat{\delta}(\lambda,T,\eta)$ is quickly computed. The best bound is obtained by minimizing $\hat{\delta}(\lambda,T,\eta)$ over η. This does, however, considerably increase the computational effort, and experiments suggest (see the discussion following the proof in Appendix A) that the extra computational effort is not worth the improvement as compared to $\eta = 0.4$. Thus, in the following, the bound

$$\hat{\delta}(\lambda,T) := \begin{cases} 1 - e^{\lambda T} & \text{if } \lambda \text{ is real,} \\ \hat{\delta}(\lambda,T,0.4) & \text{otherwise} \end{cases}$$

will be employed.

In the general case, when A is in BDOUT form, then $\delta(A,T)$ is also block diagonal and the jth block is given by $\delta(D_j + N_j, T)$, D_j and N_j being the jth blocks on the

diagonal of D and N, respectively. Therefore, without loss of generality, it can be assumed that the BDOUT form has one block only, i.e., $k = 1$.

The following theorem is the main result of this section. The proof is given in Appendix B.

THEOREM 1. *Let $A \in \mathbb{C}^{n\times n}$ and $T > 0$. Assume that $A = D + N$, where*

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}, \qquad N = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix},$$

$\lambda_i = \alpha_i + j\omega_i$, $\omega_i \in \mathbb{R}$, and $\alpha_n \le \alpha_{n-1} \le \cdots \le \alpha_1 < 0$. Define

$$\hat{\delta}(D,T) := \begin{bmatrix} \hat{\delta}(\lambda_1,T) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \hat{\delta}(\lambda_n,T) \end{bmatrix},$$

$$\varepsilon_k(\lambda_i,T) := \begin{cases} e^{\alpha_i T}T^k & \text{if} \quad T \le \frac{-k}{\alpha_i}, \\ \left(\frac{-k}{\alpha_i\, e}\right)^k & \text{else} \quad (e = 2.781\ldots), \end{cases}$$

$$\varepsilon_k(D,T) := \begin{bmatrix} \varepsilon_k(\lambda_1,T) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \varepsilon_k(\lambda_n,T) \end{bmatrix},$$

$$\hat{\delta}(A,T) := \hat{\delta}(D,T) + \sum_{k=1}^{n-1} \frac{\varepsilon_k(D,T)\cdot |N|^k}{k!}.$$

Then

$$\delta(A,T) \le \hat{\delta}(A,T).$$

3.2. Using the second derivative. Theorem 1 can also be applied to the second derivative of g, giving the bounds

$$\underline{K}_i := \ddot{g}(t_i) - |CA^2| \cdot \hat{\delta}(A,T) \cdot \left|e^{At_i}B\right|, \tag{6}$$

$$\overline{K}_i := \ddot{g}(t_i) + |CA^2| \cdot \hat{\delta}(A,T) \cdot \left|e^{At_i}B\right|, \tag{7}$$

satisfying

$$\underline{K}_i \le \ddot{g}(t) \le \overline{K}_i, \quad t \in [t_i, t_{i+1}].$$

These bounds can be employed to estimate the approximation error in numerical integration (see the following section). Additionally, they can be used to improve the bounds $\underline{G}_i$ and $\overline{G}_i$ for g, as will be described next.

PROPOSITION 1. *Let g be twice differentiable in a neighborhood of $[a,b]$ and let $\underline{K}$, $\overline{K}$ be such that*

$$\underline{K} \le \ddot{g}(t) \le \overline{K}, \quad t \in [a,b].$$

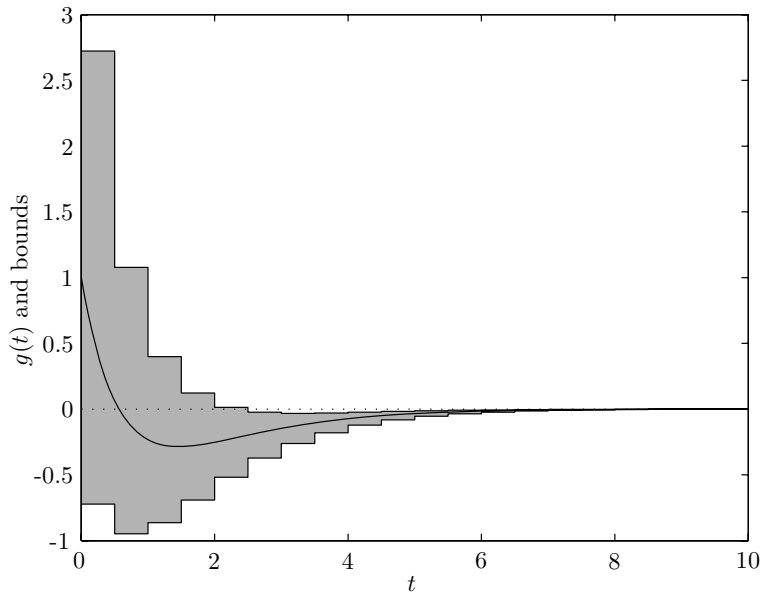


FIG. 1. *Impulse response $g(t)$ of (8) along with upper and lower bounds determined using Theorem 1.*

Then

$$\underline{H} \le g(t) \le \overline{H}, \quad t \in [a, b],$$

where

$$\underline{H} := \begin{cases} \min\{g(a), g(b)\} & \text{if} \quad \overline{K} \le 0 \quad \text{or} \quad \tau(\overline{K}) \notin [a, b], \\ \gamma(\overline{K}) & \text{else}, \end{cases}$$

$$\overline{H} := \begin{cases} \max\{g(a), g(b)\} & \text{if} \quad \underline{K} \ge 0 \quad \text{or} \quad \tau(\underline{K}) \notin [a, b], \\ \gamma(\underline{K}) & \text{else}, \end{cases}$$

$$\tau(K) := \frac{a+b}{2} - \frac{g(b) - g(a)}{K(b-a)},$$

$$\gamma(K) := \frac{g(b) - g(a)}{b-a}\,\tau(K) + \frac{g(a)b - g(b)a}{b-a} + \frac{K}{2}\,(\tau(K) - b)\,(\tau(K) - a).$$

The proof of Proposition 1 is given in Appendix C.

3.3. Example. Consider the system given by

$$A = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad C = [1 \;\; -1 \;\; 1]. \tag{8}$$

Figures 1 and 2 show the bounds for $g(t)$ and $\ddot{g}(t)$, respectively, obtained using the results of section 3.1 for $T = 0.5$. The bounds in Figure 1 are quite poor, which motivates the extension presented in section 3.2. The bounds on the curvature in Figure 2 are still quite rough, but lead to excellent results when combined with Proposition 1; see Figure 3.

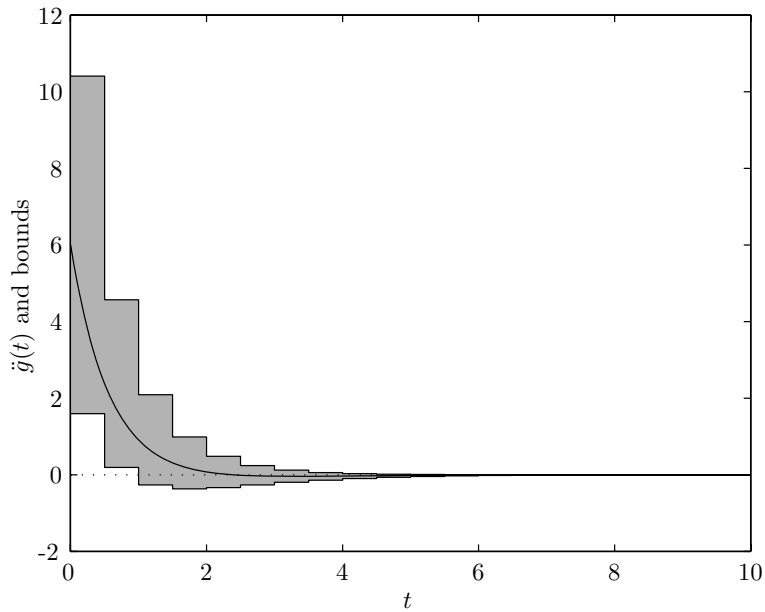


FIG. 2. *Second derivative* $\ddot{g}(t)$ *of the impulse response of* (8) *along with upper and lower bounds determined using Theorem* 1.

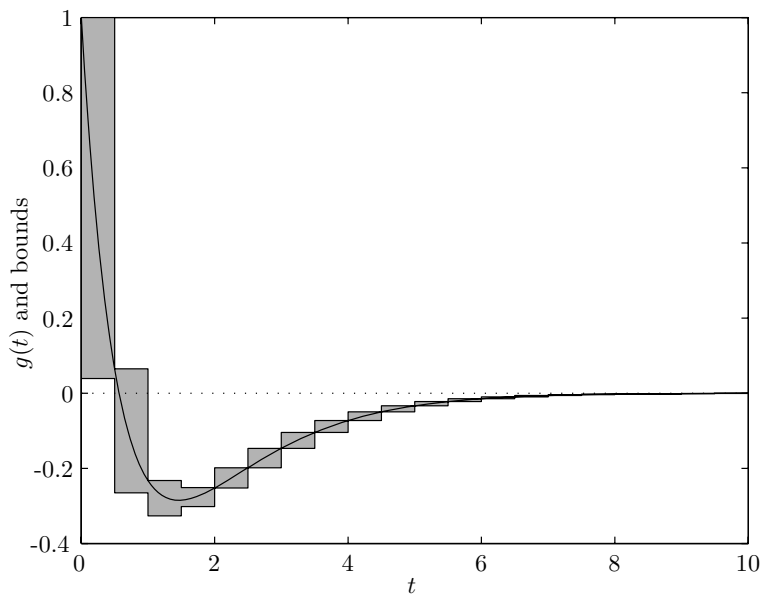


FIG. 3. *Impulse response* $g(t)$ *of* (8) *along with upper and lower bounds determined using Proposition* 1 *and Theorem* 1.

The bounds in Figure 3 fail to be tight only near the local extremum of $g(t)$. This property can be observed in many examples. In our algorithm, the bounds are merely used to detect sign changes in $g(t)$, so that nontight bounds near extrema are generally no problem at all. In the present example, the bounds in Figure 3 restrict

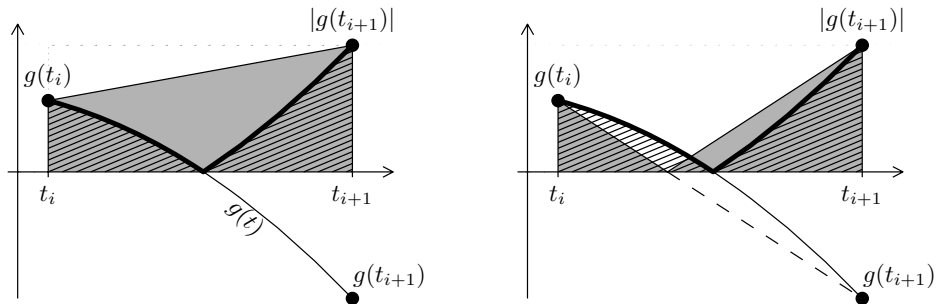


FIG. 4. *Standard trapezoidal rule for the absolute value function (left) and its modification (right). The shaded area corresponds the approximation η_i for the exact integral (striped area).*

sign changes to the interval $0.5 \le t \le 1$, which is the best result that can be expected. It should be noted that explicit bounds like those in Figure 1 are not employed in the algorithm. Figure 1 is included in this paper merely for comparison with Figure 3.

4. A modified trapezoidal rule for $|g(t)|$. This section is devoted to the computation of $\int_{t_i}^{t_{i+1}} |g(t)|\,dt$ with guaranteed error bounds. The traditional estimate

$$\int_{t_i}^{t_{i+1}} |g(t)|\,dt \approx \frac{1}{2}\left(|g(t_i)| + |g(t_{i+1})|\right) \cdot (t_{i+1} - t_i)$$

of the trapezoidal rule has two deficiencies, in case $g(t_i)$ and $g(t_{i+1})$ have different signs: First, the standard error bounds [12] are not applicable, since $|g(t)|$ is not differentiable, in general. Second, it is intuitively clear that a much better estimate η_i is obtained if a straight line interpolating $g(t_i)$ and $g(t_i + 1)$ instead of $|g(t_i)|$ and $|g(t_{i+1})|$ is employed; see Figure 4.

Both deficiencies are removed by the following proposition. It shows that the modified estimate η_i obeys the same error bounds as the trapezoidal rule for differentiable functions. The proof is given in Appendix D.

PROPOSITION 2. *Let g be twice differentiable in a neighborhood of $[a, b]$ and let K be an upper bound for the second derivative, i.e., let K satisfy $|\ddot{g}(t)| \le K$ for all $t \in [a, b]$. Then $|\int_a^b |g(t)|\,dt - \eta| \le \varepsilon$, where*

$$\eta := \begin{cases} |g(a) + g(b)| \cdot \dfrac{b-a}{2} & \text{if} \quad g(a)g(b) \ge 0, \\ \dfrac{g(a)^2 + g(b)^2}{|g(a) - g(b)|} \cdot \dfrac{b-a}{2} & \text{if} \quad g(a)g(b) < 0, \end{cases}$$

$$\varepsilon := \frac{K}{12}(b-a)^3.$$

As an example, consider the impulse response of the system given by (8) in the interval $[a, b]$ with $a = 0.5$, $b = 1$; see Figure 3. The second derivative obeys the bound $K = 4.5694$; see Figure 2. Proposition 2 leads to $\eta = 0.0490$ and $\varepsilon = 0.0476$ (all numbers rounded to 4 digits). The results in section 3 exclude sign changes outside $[a, b]$ (see Figure 3). Thus, outside $[a, b]$, the integral can be exactly computed using (4). Hence we have $|\int_0^{10} |g(t)|\,dt - 0.9425| \le 0.0476$.

Proposition 2 shows that the modified trapezoidal rule gives rise to quadratic convergence. This can be seen as follows. Suppose the interval $[a, b]$ is subdivided into m intervals of length $(b-a)/m$. The corresponding enhanced estimate for the integral obeys the error bound

$$\tilde{\varepsilon} = \sum_{i=1}^{m} \frac{K}{12} \left(\frac{b-a}{m} \right)^3 = \frac{1}{m^2} \varepsilon,$$

which quadratically drops with $1/m$.

Even faster convergence is possible if the bounds $\underline{H}$, $\overline{H}$, and K for $g(t)$ and $\ddot{g}(t)$ are recomputed for each subinterval. First, the new bounds K_i will generally be smaller than K. Second, more efficient bounds $\underline{H}_i$ and $\overline{H}_i$ will in general lead to lots of intervals where the integral can be exactly computed (see the basic algorithm in section 2).

5. The truncation error. As described in section 2, numerical integration will be performed over a finite interval $[0, T_e]$, $T_e = t_N$, only. The resulting truncation error $E := \int_{T_e}^{\infty} |g(t)|\, dt$ tends to zero, as $T_e \to \infty$. The following proposition presents a computable bound for E, based on an exponential weighting idea [5]. Its proof is given in Appendix E.

PROPOSITION 3. *Let $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times 1}$, $C \in \mathbb{R}^{1 \times n}$, and $T_e > 0$ be given. Assume that A is stable and denote by $\alpha(A)$ the spectral abscissa of A, i.e.,*

$$\alpha(A) := \max \{ \mathrm{Re}(\lambda_i) : \lambda_i \text{ eigenvalue of } A \}.$$

Choose $\lambda \in \mathbb{R}$ such that $\alpha(A) < -\lambda < 0$ and let $W \in \mathbb{R}^{n \times n}$ be the solution of the Lyapunov equation

$$(A + \lambda I)^T W + W(A + \lambda I) + C^T C = 0.$$

Then

$$E \le \overline{E} := (2\lambda)^{-\frac{1}{2}} \sum_{i=1}^{n} W_{ii}^{\frac{1}{2}} \cdot |\xi_i|,$$

where W_{ii} is the ith entry on the diagonal of W and ξ_i is the ith entry of $e^{AT_e} B$.

In most cases, the "midpoint" $\lambda = -\alpha(A)/2$ gives the best result. Note that W depends on the problem data only and that ξ_i can be obtained from a simulation of the system.

As an example, consider the system given by (8) along with $T_e = 10$ (see Figure 3). Proposition 3 leads to the bound $\overline{E} = 5.45 \cdot 10^{-4}$ for the truncation error.

An alternative bound for the truncation error is obtained using Corollary 5.2 in [8]. In all examples generated so far, Proposition 3 has led to superior results.

6. The adaptive algorithm. The basic algorithm in section 2 assumes the discretization points $t_1, \ldots, t_N$ to be given and leads to a corresponding overall error $\varepsilon = \varepsilon_1 + \cdots + \varepsilon_N$. The finer the discretization, the smaller the discretization errors ε_i. The greater the endpoint $T_e = t_N$, the smaller the truncation error e_N.

In this section, a tolerance $\delta > 0$ is assumed to be given, and it is shown how to adaptively add discretization points such that $\varepsilon < \delta$. The approach is to follow the following steps.

Adaptive algorithm.

Step 1: Based on rough estimates of the settling time of $g(t)$, initial discretization points $t_1, \ldots, t_N$ are determined (see section 6.1 for details).

Step 2: An estimate η for the norm and associated errors $\varepsilon_1, \ldots, \varepsilon_N$ are computed using the basic algorithm sketched in section 2.

Step 3: If $\varepsilon_N > \delta/2$, i.e., the endpoint T_e is too small, then a new endpoint T_e^{new} is determined based on ε_N, δ, and the bounds in section 5 (see section 6.2 for details). The number N is increased and new discretization points are added to the list obtained so far. The estimate η for the norm and the errors ε_i are updated as in Step 2. Step 3 is repeated until $\varepsilon_N \leq \delta/2$.

Step 4: Let $t_1, \ldots, t_N$ be the discretization points and let $\varepsilon_1, \ldots, \varepsilon_N$ be the associated errors which are obtained so far. If $\varepsilon \geq \delta$, then intervals $[t_i, t_{i+1}]$ with large error ε_i are chosen. In these intervals, additional discretization points are inserted such that the new error ε is reduced below δ; see section 6.3 for details.

The adaptive algorithm is guaranteed to terminate after a finite number of steps. It computes an estimate η of the L_1-norm and an associated error in bound ε such that $\varepsilon < \delta$ and (3) is satisfied.

Since $|\xi_i|$ drops exponentially with increasing T_e, Proposition 3 shows that Step 3 of the algorithm converges exponentially. Step 4 converges quadratically by Proposition 2. Thus the overall convergence of the algorithm is quadratic.

If a tolerance δ_{rel} for the relative error is given instead of the tolerance δ on the absolute error, then the H_∞-norm can be used to be translate δ_{rel} to δ, since $||\Sigma||_{H_\infty} \leq ||\Sigma||_{L_1}$.

6.1. The initial grid. The initial discretization points $t_1, \ldots, t_N$ are chosen linearly spaced between $t = 0$ and $t = T_e$. The endpoint T_e is determined from the upper bound $\overline{E}$ of the truncation error in Proposition 3. Using the rough estimate $|\xi_i| \approx e^{\alpha(A)T_e}||B||$ and solving the aim $\overline{E} \approx \delta/2$ for T_e leads to

$$T_e = \max\left\{\frac{1}{\alpha(A)}, \frac{1}{\alpha(A)} \ln\left(\frac{\delta\lambda^{\frac{1}{2}}}{2^{\frac{1}{2}}||B||\Sigma W_{ii}^{\frac{1}{2}}}\right)\right\}.$$

The number N of initial discretization points is fixed to $N = 300$. Of course, more efficient heuristics are possible to choose N depending on δ, T_e, and the curvature of $g(t)$.

6.2. The new discretization endpoint. In Step 3 of the adaptive algorithm, the truncation error ε_N is reduced by increasing the discretization endpoint T_e. An efficient strategy for increasing T_e is obtained by using Proposition 3 as follows. From the simulation in Step 2, the true values $\xi_i = \xi_i^{old}$ and the true upper bound $\varepsilon_N = \overline{E}^{old}$ are obtained for the old endpoint T_e^{old}. If $\overline{E}^{old} > \delta/2$, then the estimate $|\xi_i| \approx e^{\alpha(A)T_e}||B||$ (see section 6.1) was too conservative. A better estimate is given by $|\xi_i| \approx e^{\alpha(A)T_e}\beta_i$, where β_i is obtained from $\left|\xi_i^{old}\right| = e^{\alpha(A)T_e^{old}} \cdot \beta_i$. Using this new approximation in the formula for $\overline{E}$ (see Proposition 3) and solving the aim $\overline{E} \approx \delta/2$ for T_e leads to the following new estimate T_e^{new} for the discretization endpoint:

$$T_e^{new} = T_e^{old} + \frac{1}{\alpha(A)} \ln\left(\frac{\delta}{2\overline{E}^{old}}\right).$$

If $T_e^{new} - T_e^{old}$ is smaller than a threshold κ, then $T_e^{new} = T_e^{old} + \kappa$ is chosen, in order to guarantee convergence of the algorithm.

6.3. Refinement of the discretization. In this section, the details of the refinement in Step 4 of the adaptive algorithm are elaborated. Without loss of generality it is assumed that the intervals $[t_i, t_{i+1}]$ are sorted such that $\varepsilon_1 \geq \varepsilon_2 \geq \cdots \geq \varepsilon_{N-1}$. The idea is to choose k and m and to divide each of the intervals $[t_i, t_{i+1}]$, $1 \leq i \leq k$, into m subintervals of length $(t_{i+1} - t_i)/m$. This leads to new estimates $\tilde{\eta}_i$ and to improved error bounds $\tilde{\varepsilon}_i$, $1 \leq i \leq k$.

The discussion following Proposition 2 shows that $\tilde{\varepsilon}_i = \varepsilon_i / m^2$. The overall integration error is then given by

$$\tilde{\varepsilon} := \frac{1}{m^2} \left(\sum_{i=1}^{k} \varepsilon_i \right) + \left(\sum_{i=k+1}^{N} \varepsilon_i \right).$$

Solving the requirement $\tilde{\varepsilon} \leq \delta$ for m leads to a lower bound for m depending on k. An optimal k minimizing $k \cdot m$ is easily obtained by direct search.

7. Numerical issues. To estimate the computational time, let M be the total number of intervals to be analyzed by the algorithm and note that the computational effort is dominated by those operations which are to be performed in each of the M intervals. For the sake of a simplified presentation it is assumed that the discretization points t_i are equally spaced, i.e., $t_{i+1} - t_i = T$.

The computations in the interval $[t_i, t_{i+1}]$ involve the determination of

$$x_{i+1} = e^{At_{i+1}} B = e^{AT} x_i, \quad x_1 = B, \tag{9}$$

$g(t_i) = Cx_i$, $\ddot{g}(t_i) = CA^2 x_i$, and $\underline{K}_i$, $\overline{K}_i$ (according to (6), (7)). This requires $n^2 + 3n$ multiplications. The computation of $\underline{H}_i$ and $\overline{H}_i$ according to Proposition 1 takes 3 to 13 multiplications, depending on which case applies. The estimate η_i (including ε_i) according to Proposition 2 requires 1 to 4 multiplications, whereas the computation of the exact value according to (4) takes n multiplications.

In summary, the total number of multiplications per discretization point t_i is given by $n^2 + 5n$, where it is assumed that the mean number of multiplications according to Propositions 1 and 2 are both smaller than n. Considering the fact that the computation of the impulse response requires $n^2 + n$ multiplications per discretization point, it can be concluded that the cost for the computation of the L_1-norm and for an accurate simulation of the impulse response are in the same order of magnitude.

In addition to the above estimates, the computational time is roughly proportional to the number M of intervals to be analyzed. In the adaptive algorithm, this number crucially depends on the shape of the impulse response: Smooth impulse responses with few zeros generally lead to small M. The following section contains examples with considerable oscillations in the impulse response, which is the "worst case" for the adaptive algorithm.

When implemented in finite arithmetic, rounding errors are inevitable. A complete rounding error analysis is beyond the scope of this paper. However, mostly standard components of numerical linear algebra and integration are employed, whose numerical properties are well understood: The rounding error analysis in [11] can be applied to the computation of x_i, $g(t_i)$, $\ddot{g}(t_i)$, $\underline{K}_i$, $\overline{K}_i$, and η_i according to (4). Moreover, the estimate η in Proposition 2 and the bound $\hat{\delta}$ in Theorem 1 can be reliably computed, basically because no cancellation of leading digits is involved [12]. Finally, the numerical properties of the Lyapunov equation in Proposition 3 are well known [11]. A complete error analysis of the algorithm, including the bounds $\underline{H}_i$ and $\overline{H}_i$ according to Proposition 1, is still to be done.

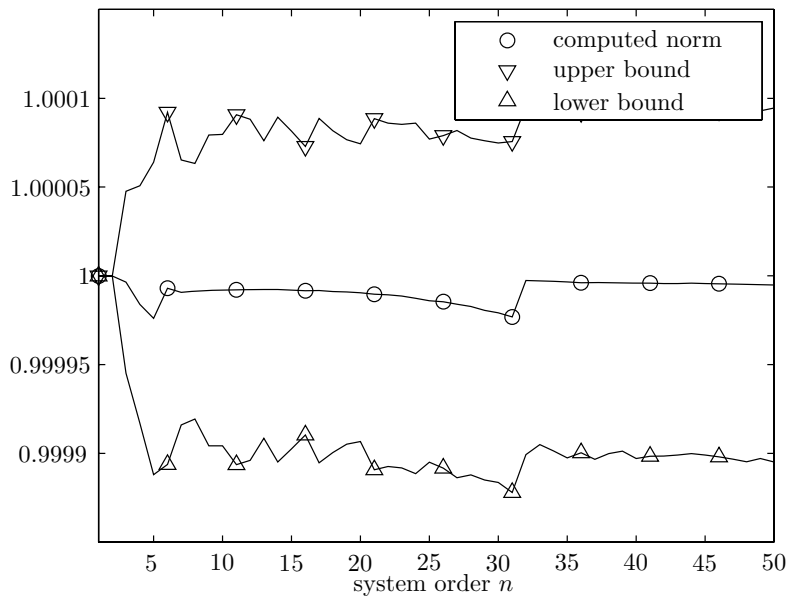


FIG. 5. *Computed L_1-norm for Example* 1.

8. Examples. In this section, the adaptive algorithm of section 6 is applied to examples. The timing results refer to a MATLAB implementation on a 1.8 GHz PC.

Example 1. Consider a system given by the transfer function

$$G(s) = \prod_{i=1}^{n} \frac{1}{T_i s + 1}, \quad T_i = i^2.$$

Figure 5 shows the computed L_1-norm along with the computed (guaranteed) upper and lower bounds for $n = 1, 2, \ldots, 50$ and tolerance $\delta = 10^{-4}$. Figure 6 shows the corresponding computing time. It increases quadratically with n, as expected from section 7. In summary, the algorithm works well for this medium-sized problem.

Example 2. Consider the system given by the transfer function

$$G(s) = \frac{(s-1)^n}{(s+1)^{n+2}}$$

for $n = 1, 2, \ldots$. This is a challenging example, because its BDOUT form (see section 3) consists of one block only and its impulse response oscillates heavily (see Figure 7). For $n = 25$, the L_1-norm $||\Sigma|| = 3.2469$ is computed within a relative tolerance of $\delta_{rel} = 10^{-4}$ in 0.3 seconds. For $n = 50$, the computation of the L_1-norm $||\Sigma|| = 4.3960$ takes 14 seconds.

Example 3. Consider the second order system given by the transfer function

$$F(s) = \frac{1}{s^2 + 2ds + 1}, \quad d = 0.1.$$

Its L_1-norm can be efficiently computed using the closed formulas in Appendix F. Ignoring rounding errors, the computed norm is exact. This allows us to check the result of the adaptive algorithm in section 6 for consistency. Figure 8 shows the relative error bound computed by the adaptive algorithm for various values of the relative

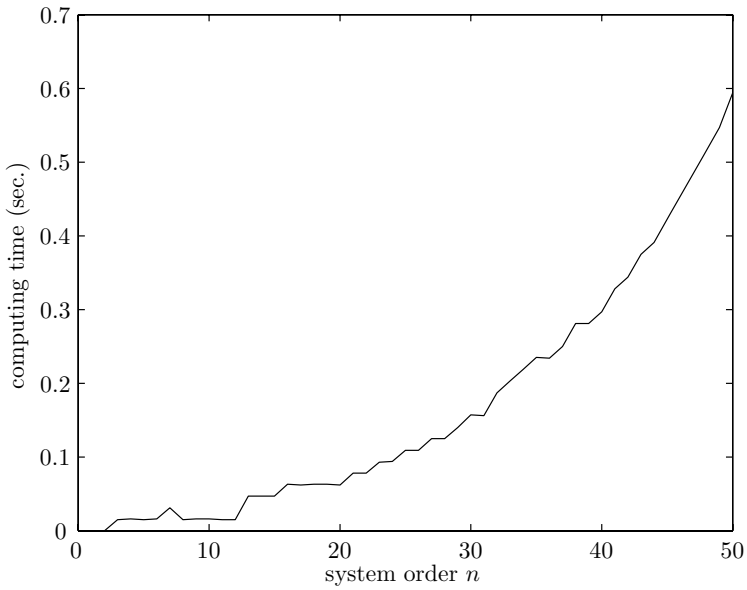


FIG. 6. *Computing time for Example* 1.

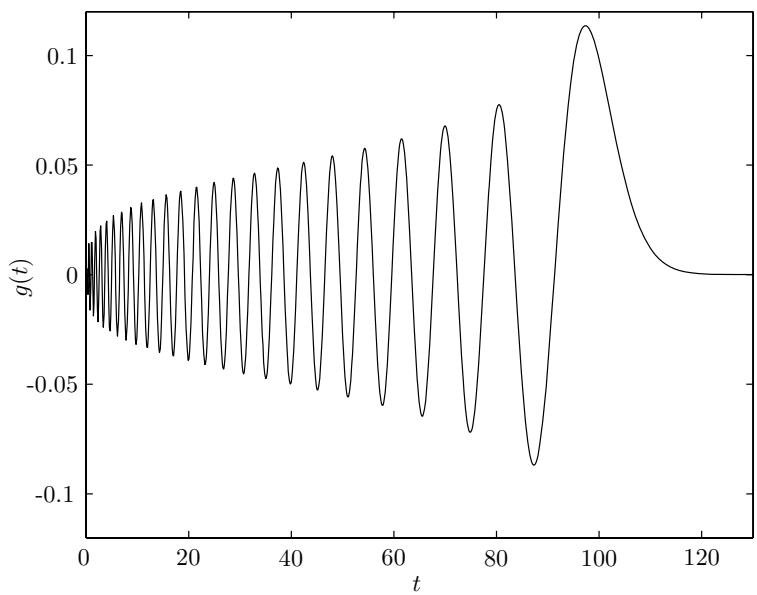


FIG. 7. *Impulse response of the system in Example* 2 *for* $n = 50$.

tolerance δ_{rel}. The true relative error (computed using Appendix F) shows that the algorithm works well even for small δ_{rel}. Figure 9 shows how the computing time behaves towards the tolerance. Thus, for moderate values of δ_{rel} ($\delta_{rel} \geq 10^{-6}$, say), the computing time stays essentially constant. For smaller values of δ_{rel}, the computing time increases. As expected from the quadratic convergence of the algorithm, a decrease of two decades in δ_{rel} results in a increase of one decade in computing time.

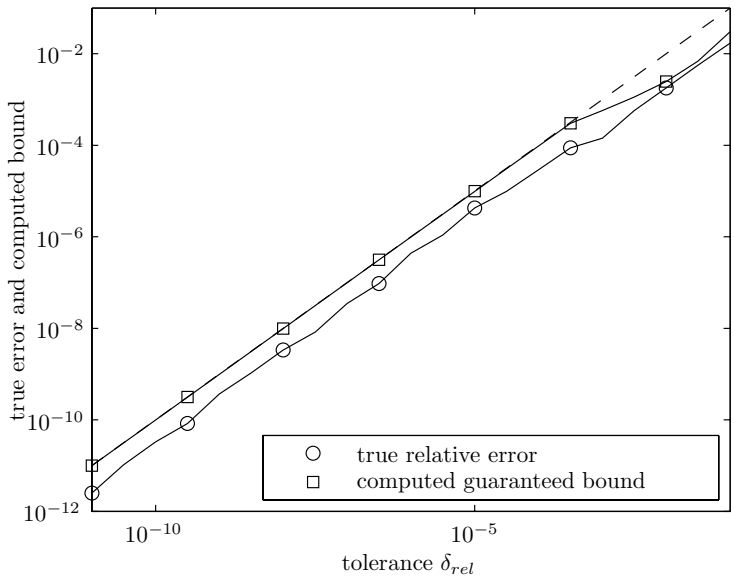


FIG. 8. *Relative error in Example* 3.

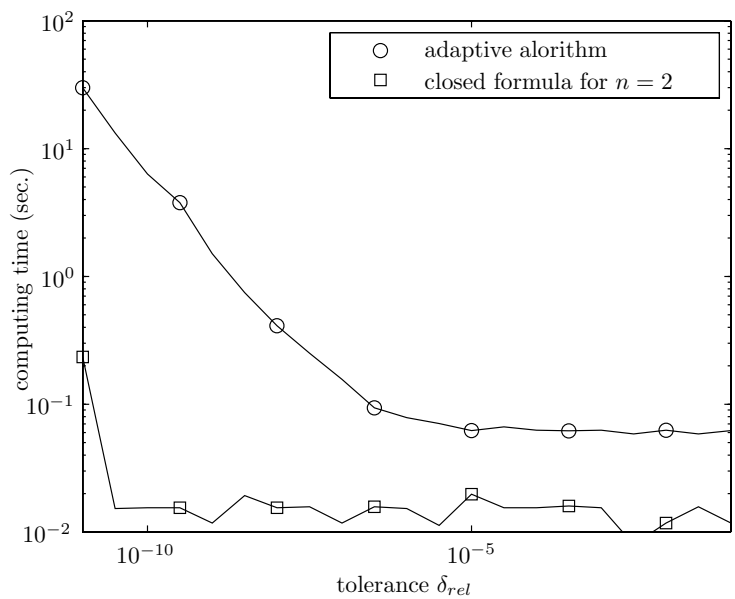


FIG. 9. *Computing time in Example* 3.

9. Multivariable systems. In this section, extensions to multivariable systems with feedthrough are discussed. Accordingly, let the system be given by matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$. Denoting the columns of B by B_k, the rows of C by C_ℓ, and the entries of D by $D_{\ell k}$, the L_1-norm is given by

$$||\Sigma|| := \max_{\ell=1,\dots,p} \sum_{k=1}^{m} \left(\int_0^\infty |C_\ell e^{At} B_k| \, dt + |D_{\ell k}| \right); \tag{10}$$

see [3]. Thus the computation of the multivariable L_1-norm amounts to $m \cdot p$ computations for single-input single-output systems. By exploiting the structure of (10), the computational cost can, however, be reduced. For the sake of illustration, two possible savings are sketched in the following.

The first one is due to the fact that the computational effort is dominated by the cost of determining x_{i+1} via (9), which does not depend on outputs. Assuming that identical discretization points t_i are employed for all outputs in Steps 1 and 2 of the adaptive algorithm and following the reasoning in section 7, the computational cost per discretization point and input can be reduced from $p(n^2 + 5n)$ to $n^2 + 5pn$.

Another saving concerns Step 4 of the adaptive algorithm in the multivariable case. This is due to the fact that the overall L_1-norm is given by the maximum over the L_1-norms for the outputs; see (10). Suppose that, in some iteration of an adaptive algorithm, estimates for the p single-output multiple-input norms and associated rough error bounds are available. Suppose further that the upper bound for the ℓth output is smaller than the lower bound for the qth output. Under these assumptions, the ℓth output can be discarded from further consideration. This reasoning generally leads, especially in applications which widely spread single-output norms, to a considerable reduction of outputs to be handled.

10. Conclusions. The L_2- and H_∞-norms can be efficiently computed by exploiting their characterizations in terms of Lyapunov equations and Hamiltonian matrices, respectively. Similar characterizations are not available for the L_1-norm. Therefore, the numerical computation of the L_1-norm is more involved. The algorithm of the present paper is based on the numerical integration of the absolute value of the impulse response. It is shown to be quadratically convergent and appears to be the first serious algorithm for the computation of the L_1-norm.

Although the paper does not include a complete rounding error analysis, it is expected that the algorithm is numerically reliable, because it is based on standard components from numerical linear algebra and numerical integration. The examples with state dimension up to 50 suggest that the algorithm works well for medium-sized problems.

It is important to note that the algorithm also determines an associated guaranteed error bound for the computed norm. This is an extra feature as compared to Hamiltonian algorithms for the computation of the H_∞-norm.

As far as computational speed is concerned, the L_1-algorithm of this paper cannot, however, compete with algorithms for the computation of L_2- and H_∞-norms.

Appendix A. Proof of Lemma 1. The first inequality of the lemma follows from

$$|e^{\lambda t} - 1| \le |\lambda| t + \frac{1}{2} |\lambda|^2 t^2 + \cdots \le e^{|\lambda|\tau} - 1, \quad t \le \tau. \tag{11}$$

As for the second inequality, note that $t_1 < t_2$ and consider the function

$$f(t) := |e^{\lambda t} - 1|^2 = e^{2\alpha t} + 1 - 2e^{\alpha t} \cos(\omega t).$$

Using (11), it follows that $f(t) \le (e^{|\lambda| t_1} - 1)^2 = \eta^2$, $t \le t_1$. Moreover, for $t_1 \le t \le t_2$, we have

$$f(t) \le f_2(t) := e^{2\alpha t_1} + 1 - 2e^{\alpha t} \cos(\omega t).$$

This proves the second inequality, since f_2 is monotonically increasing for $0 \le t \le t_2$. The third inequality follows from $f_2(t) \le f_2(t_2)$, $t \ge t_2$. ∎

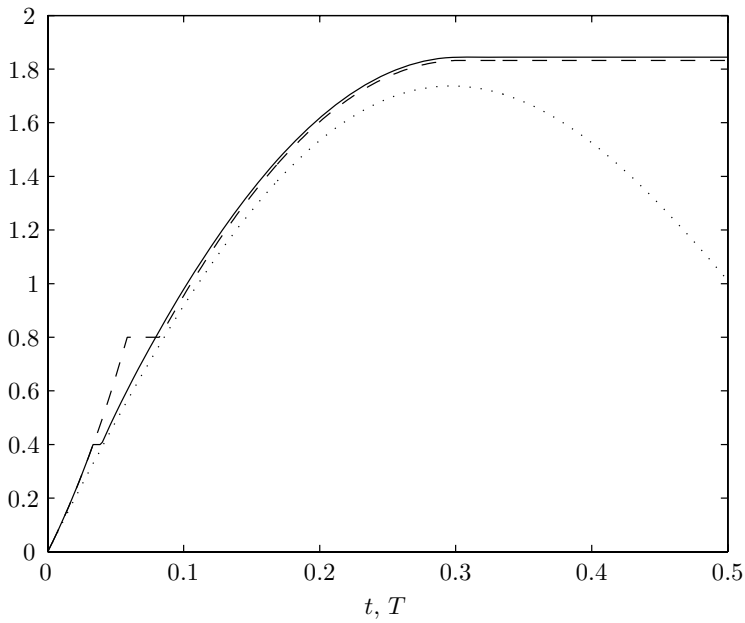


FIG. 10. *Function* $|e^{\lambda t} - 1|$ *(dotted) and bounds* $\hat{\delta}(\lambda, T, \eta)$ *for* $\eta = 0.4$ *(solid) and* $\eta = 0.8$ *(dashed).*

The idea behind Lemma 1 can be illustrated using the function $f(t)$, which is monotonically increasing between 0 and its first local maximum t_2'. Since t_2' is also the global maximum, the best bound is given by

$$\delta_{best}(\lambda, T) := \begin{cases} |e^{\lambda T} - 1|, & T \le t_2', \\ |e^{\lambda t_2'} - 1|, & T \ge t_2'. \end{cases}$$

Unfortunately, there is no analytical expression available for t_2'. In the proof of Lemma 1, $f(t)$ is replaced by $f_2(t)$ using some fixed t_1 and $\eta = e^{|\lambda| t_1} - 1$. Decreasing t_1 improves the bound for values of T which are slightly above t_1 but deteriorates the bound for larger values of T. Figure 10 shows the function $|e^{\lambda t} - 1|$ for $\lambda = -1 + 10j$ along with the bounds for $\eta = 0.4$ and $\eta = 0.8$. It shows that the compromise $\eta = 0.4$ leads to acceptable values for all T.

Appendix B. Proof of Theorem 1. Following along the same lines as those in [13], it can be shown that $|e^{At} - I| \le \delta(D, T) + \sum_{k=1}^{n-1} (|e^{Dt}|t^k)|N|^k/k!$. Hence the proof follows from Lemma 1 and

$$\max_{0 \le t \le T} e^{\alpha t} t^k = \begin{cases} e^{\alpha T} T^k & \text{if } T \le \dfrac{k}{-\alpha}, \\ \left(\dfrac{k}{-\alpha e} \right)^k & \text{else,} \end{cases}$$

which is valid for $\alpha < 0$, $T > 0$, and $k \le 1$. ∎

Appendix C. Proof of Proposition 1. The linear function

$$p(t) = \frac{g(b) - g(a)}{b - a}\, t + \frac{g(a)b - g(b)a}{b - a} \tag{12}$$

interpolates $g(t)$ at $t = a$ and $t = b$. Therefore, using the standard representation of the error in polynomial interpolation [12],

$$g(t) = p(t) + \frac{1}{2}\,(t-a)\,(t-b)\,\ddot{g}(\xi), \tag{13}$$

where $t, \xi \in [a,b]$. Thus $h(t) := p(t) + \frac{1}{2}(t-a)(t-b)\overline{K}$ satisfies $g(t) \geq h(t)$, $t \in [a,b]$. Let $t \in [a,b]$. If $\overline{K} \leq 0$, then $h(t) \geq p(t)$ and hence $g(t) \geq p(t) \geq \min\{g(a), g(b)\}$. If $\overline{K} \geq 0$, then $h(t)$ has an unique minimum at $t = \tau(\overline{K})$. If this minimum is outside the interval $[a,b]$, then again $g(t) \geq \min\{g(a), g(b)\}$, $t \in [a,b]$. If $\tau(\overline{K}) \in [a,b]$, then $g(t) \geq h(t) \geq h(\tau(\overline{K}))$, $t \in [a,b]$, which proves the lower bound $\underline{H}$. The upper bound $\overline{H}$ can be derived along the same lines. □

Appendix D. Proof of Proposition 2. Consider the function $p(t)$ defined in (12). Straightforward computations show that $\int_a^b |p(t)|\,dt = \eta$. By (13) the integration error can be estimated as follows:

$$\left| \int_a^b |g(t)|\,dt - \eta \right| \leq \int_a^b |g(t) - p(t)|\,dt \leq -\frac{K}{2} \int_a^b (t-a)(t-b)\,dt.$$

The proof now follows from the standard Newton–Cotes formula for the trapezoidal rule [12]. □

Appendix E. Proof of Proposition 3. The triangle inequality implies

$$E \leq \sum_{i=1}^{n} \left(\int_0^\infty \left| Ce^{At} e_i \right| dt \right) |\xi_i|.$$

The identity $e^{At} = e^{(A+\lambda I)t} \cdot e^{-\lambda t}$ and the Hölder inequality lead to $\int_0^\infty \left| Ce^{At} e_i \right| dt \leq N_1 \cdot N_2$, where $N_1^2 = \int_0^\infty \left| Ce^{(A+\lambda I)t} e_i \right|^2 dt$ and $N_2^2 = \int_0^\infty (e^{-\lambda t})^2\,dt = 1/(2\lambda)$. Thus the proof follows by expressing the L_2-norm N_1 in terms of the Lyapunov equation; see, e.g., [14]. □

Appendix F. Second order systems. In this section, closed formulas for the L_1-norm of second order systems are derived. It is shown how to efficiently compute the L_1-norm while avoiding any simulations. The formulas depend on parameters $\alpha_1, \alpha_2, \lambda_1, \lambda_2, \alpha, \beta, \lambda, \omega$ of the impulse response (see below). It should be noted that these parameters can be easily computed from a state-space model of the system. The general case is separated into three cases as follows.

F.1. Real disjoint eigenvalues. In this case, the impulse response reads $g(t) = \alpha_1 e^{\lambda_1 t} + \alpha_2 e^{\lambda_2 t}$, $t \geq 0$, where $\alpha_1 \in \mathbb{R}$, $\alpha_2 \in \mathbb{R}$, $\lambda_1 < 0$, $\lambda_2 < 0$, $\lambda_1 \neq \lambda_2$. Straightforward computations show that $g(t)$ changes its sign if and only if $\alpha_1 \neq 0$, $\alpha_2 \neq 0$, $\alpha_1 \alpha_2 < 0$, $t_0 > 0$, where $t_0 := \ln(\frac{-\alpha_2}{\alpha_1})/(\lambda_1 - \lambda_2)$. If these conditions are satisfied, t_0 is the unique zero of $g(t)$ and the L_1-norm is given by

$$||g||_{L_1} = \left| \int_0^{t_0} g(t)\,dt \right| + \left| \int_{t_0}^\infty g(t)\,dt \right| = \left| a - \frac{\alpha_1}{\lambda_1} - \frac{\alpha_2}{\lambda_2} \right| + |a|,$$

where $a := \frac{\alpha_1}{\lambda_1} e^{\lambda_1 t_0} + \frac{\alpha_2}{\lambda_2} e^{\lambda_2 t_0}$. If $g(t)$ does not change sign, then, obviously,

$$||g||_{L_1} = \left| \int_0^\infty g(t)\,dt \right| = \left| \frac{\alpha_1}{\lambda_1} + \frac{\alpha_2}{\lambda_2} \right|.$$

F.2. Repeated eigenvalues. In this case, the impulse response reads $g(t) = \alpha e^{\lambda t} + \beta t e^{\lambda t}$. Identifying the possible zero of this function as in section F.1, leads to the result

$$||g||_{L_1} = \begin{cases} |a| + \left| a - \dfrac{\alpha}{\lambda} + \dfrac{\beta}{\lambda^2} \right| & \text{if } \beta \neq 0,\ \dfrac{-\alpha}{\beta} > 0, \\ \left| \dfrac{\alpha}{\lambda} - \dfrac{\beta}{\lambda^2} \right| & \text{else,} \end{cases}$$

where $a := \frac{1}{\lambda} e^{\lambda t_0} \left(\alpha + \beta t_0 - \frac{\beta}{\lambda} \right)$.

F.3. Complex eigenvalues. In this case, the impulse response reads $g(t) = e^{\lambda t}(\alpha \sin(\omega t) + \beta \cos(\omega t))$, where $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$, $\lambda < 0$, $\omega > 0$. Without loss of generality it is assumed that $(\alpha, \beta) \neq 0$. The function $g(t)$ has infinitely many zeros $t_0, t_1, t_2, \ldots$ in $[0, \infty]$ of the form $t_k = s_0 + k\pi/\omega$, where

$$s_0 = \begin{cases} \dfrac{\pi}{2\omega} & \text{if } \alpha = 0, \\ \dfrac{1}{\omega} \arctan\left(\dfrac{-\beta}{\alpha} \right) & \text{if } \alpha \neq 0,\ \dfrac{-\beta}{\alpha} \geq 0, \\ \dfrac{1}{\omega} \left(\arctan\left(\dfrac{-\beta}{\alpha} \right) + \pi \right) & \text{if } \alpha \neq 0,\ \dfrac{-\beta}{\alpha} < 0. \end{cases}$$

Thus

$$||g||_{L_1} = |N_0| + \sum_{k=1}^{\infty} |N_k|,$$

where

$$N_0 = \int_0^{s_0} g(t)\, dt; \quad N_k = \int_{t_{k-1}}^{t_k} g(t)\, dt, \quad k = 1, 2, 3, \ldots.$$

Elementary computations show that $N_0 = d - \frac{\beta\lambda - \alpha\omega}{\lambda^2 + \omega^2}$ and $|N_k| = |d|(1+q)q^{k-1}$, $k \geq 1$, where

$$d = \frac{(\alpha\lambda + \beta\omega)\sin(\omega s_0) + (\beta\lambda - \alpha\omega)\cos(\omega s_0)}{\lambda^2 + \omega^2} e^{\lambda s_0}, \quad q = e^{\lambda\pi/\omega},$$

which implies

$$||g||_{L_1} = \left| d - \frac{\beta\lambda - \alpha\omega}{\lambda^2 + \omega^2} \right| + |d| \frac{1+q}{1-q}. \qquad \square$$

REFERENCES

[1] C. A. Bavely and G. W. Stewart, *An algorithm for computing reducing subspaces by block diagonalization*, SIAM J. Numer. Anal., 16 (1979), pp. 359–367.

[2] S. Boyd and V. Balakrishan, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its L_∞-norm*, Systems Control Lett., 15 (1990), pp. 1–7.

[3] S. P. Boyd and C. H. Barratt, *Linear Controller Design: Limits of Performance*, Prentice–Hall, Englewood Cliffs, NJ, 1991.

[4] N. Bruinsma and M. Steinbuch, *A fast algorithm to compute the H_∞-norm of a transfer function matrix*, Systems Control Lett., 14 (1990), pp. 287–293.
[5] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.
[6] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum, *Feedback Control Theory*, Macmillan, New York, 1992.
[7] D. Hinrichsen, B. Kelb, and A. Linnemann, *An algorithm for the computation of the structured complex stability radius*, Automatica J. IFAC, 25 (1989), pp. 771–775.
[8] H. Khalil, *Nonlinear Systems*, Prentice–Hall, Upper Saddle River, NJ, 2002.
[9] M. Konstantinov, P. Petkov, J. Kawelke, A. Linnemann, D. W. Gu, and I. Postlethwaite, *Sensitivity of system norms*, Internat. J. Control, 72 (1999), pp. 84–95.
[10] A. Linnemann and J. Kawelke, *H_∞-norm and frequency-response computations—sensitivity and a reliable algorithm*, IMA J. Math. Control Inform., 16 (1999), pp. 249–259.
[11] P. H. Petkov, N. D. Christov, and M. M. Konstantinov, *Computational Methods for Linear Control Systems*, Prentice–Hall, New York, 1991.
[12] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.
[13] C. Van Loan, *The sensitivity of the matrix exponential*, SIAM J. Numer. Anal., 14 (1977), pp. 971–981.
[14] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice–Hall, Upper Saddle River, NJ, 1996.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2071–2195

EXACT CONTROLLABILITY FOR THE TIME DEPENDENT TRANSPORT EQUATION*

MICHAEL V. KLIBANOV† AND MASAHIRO YAMAMOTO‡

Abstract. The exact controllability theorem for the time dependent transport equation is proved.

Key words. exact controllability, transport equation, Carleman estimate

AMS subject classifications. 93B05, 82C70, 82B40

DOI. 10.1137/060652804

1. Introduction.

Let $\Omega \subset R^N$ be a strictly convex bounded domain with the boundary $\partial\Omega \in C^1$. The assumption of strict convexity of Ω makes the arguments simpler in many places, and in order to concentrate on the main issue, the exact controllability, we make this assumption here. Let $S^{N-1} \subset \mathbb{R}^N$ be the unit sphere and ν be the unit vector. Denote

$$W = \Omega \times S^{N-1} \times (0,T), \quad \Gamma = \Gamma(T) = \partial\Omega \times S^{N-1} \times (0,T),$$

$$\Gamma_+ = \Gamma_+(T) = \{(x,t,\nu) \in \Gamma : (n(x),\nu) > 0\},$$

$$\Gamma_- = \Gamma_-(T) = \{(x,t,\nu) \in \Gamma : (n(x),\nu) \le 0\},$$

where $(,)$ is the scalar product and $n(x)$ is the unit outward normal vector to $\partial\Omega$ at x.

In this paper, we consider the homogeneous transport equation:

$$Mu := u_t + (\nu, \nabla u) + a(x,t,\nu)u - \int_{S^N-1} g(x,t,\nu,\mu)u(x,t,\mu)d\sigma_\mu = 0 \text{ in } W \tag{1.1}$$

(e.g., [3], [5]). Here $\nu \in S^{N-1}$ is the unit vector of particle velocity, $u(x,t,\nu)$ is the density of particle flow, a is an absorption coefficient, and g is a scattering indicatrix.

In this publication, the exact controllability theorem for the time dependent transport equation (1.1) is proved for the first time. The transport equation governs diffusion processes, as long as they are linear ones, e.g., propagation of neutrons (see the classic book of Case and Zweifel [3]). A particularly interesting example is propagation of the near infrared light (originated by lasers) in a diffuse background, such as human tissues. The latter has applications in medical optical imaging; see, e.g., the review papers by Das, Liu, and Alfano [4] and Wang, Li, and Jiang [37]. Therefore

*Received by the editors February 23, 2006; accepted for publication (in revised form) May 24, 2007; published electronically December 12, 2007.
http://www.siam.org/journals/sicon/46-6/65280.html

†Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223 (mklibanv@email.uncc.edu). This author's work was supported by, or in part by, the U.S. Army Research Laboratory and U.S. Army Research Office under contract/grant W911NF-05-1-0378. This author's work was also partially supported by NATO under grant PDD(CP)-(PST.NR.CLG 980631).

‡Department of Mathematical Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro, Tokyo 153, Japan (myama@ms.u-tokyo.ac.jp). This author's work was partially supported by grant 15340027 from The Japan Society for the Promotion of Science, as well as by grant 17654019 from The Japan Ministry of Education, Cultures, Sports and Technology.

the transport equation plays an important role in the diffusion theory. We refer the reader to such classical books of physics as the books of Case and Zweifel [3], Dautray and Lions [5], Ishimaru [16], and Landau and Lifshitz [23]. See Larsen and Keller [24] as well as [5] concerning the diffusion approximation. We also refer the reader to the review paper of Ukai [36] and to his book [35]. It is stated in section 1.3 of [3] that the transport equation is actually the equation of the balance and that it is a linearized Boltzmann equation; see, e.g., [36] for the Boltzman equation. Moreover, one can relate the transport equation to the equations of fluid dynamics such as the Euler and the Navier–Stokes equations through an asymptotic expansion of a solution of the Boltzmann equation; e.g., see pp. 42–44 in [36]. As for related physical backgrounds, see [23], especially, p. 89, and [24]. Thus the exact controllability for the transport equation is as important as it is for the equation of parabolic type.

There are many publications on the exact controllability, and the authors are unable to review all of them here. The following is an incomplete list of publications, and the reader can consult the references therein. The papers of Russell [32] and Seidman [33] are early works. Lions has introduced the duality method in [27], [28], [29] (also, see Komornik [22]). We can further list early works: for hyperbolic equations, see, e.g., Bardos, Lebeau, and Rauch [2], Lasiecka and Triggiani [25], Triggiani [34], and, e.g., Zuazua [38] for a plate equation. Exact controllability results are obtained for a variety of partial differential equations; see e.g., Eller and Masters [8] for Maxwell's equations, and Fursikov [9], Fursikov and Imanuvilov [10], Imanuvilov [14], and Imanuvilov and Yamamoto [15] for parabolic equations.

Our proof of the exact controllability consists of two conventional stages. On the first stage the so-called continuous observability estimate is established, i.e., the Lipschitz stability estimate for the time dependent transport equation with the lateral boundary data on the lateral side of the time cylinder (Theorem 1.2). This estimate is an important ingredient of the duality method, which is applied on the second stage. The continuous observability estimate is proved using the method of Carleman estimates. Note that, unlike other techniques, this method enables us to prove the Lipschitz stability for a "transport inequality" (Theorem 1.3). This is because Carleman estimates enable one to "suppress" low order terms which depend on t (i.e., the third and fourth terms in Mu given by (1.1)) by the principal part of the operator. Although our proof is based on the two conventional stages, the execution of each stage needs independent considerations.

First step. By integration by parts and the method of characteristics, one can prove the continuous observability estimate in the case where $a(x,t,\nu)$ and $g(x,t,\nu,\mu)$ in (1.1) are independent of t and the proof is not difficult. Our Carleman estimate can treat also the t-dependent case equally. In many cases, such as nuclear fission, since the physical properties are changed in t, it is natural to assume that the coefficients in (1.1) should depend on t. See [35], [36] for the related physical backgrounds. Moreover, our method by a Carleman estimate can yield the continuous observability estimate for the transport inequality, to which other methods in [2], [28], [29], etc., are not applicable. The continuous observability estimate for the transport inequality should be discussed, for example, if the nonhomogeneous term cannot be determined precisely, but we have to estimate the energy.

Second step. Naturally we have to consider the weak solution to the nonhomogeneous boundary value problem of (1.1). However, there are very few papers in English treating this subject. Note that in Douglis [6] and Ukai [36], the homogeneous boundary value problem of (1.1) is mainly discussed. Bardos [1, pp. 205–208] treats a nonhomogeneous boundary value problem, and the result in [1] may be able

to shorten our argument at this step after some modification in the cylindrical domain $\Omega \times (0,T)$ in (x,t). However, here we define the weak solution in a way that is different and more compatible with the duality argument, which is essential for the exact controllability.

The Lipschitz stability estimate for the transport equation was recently established by Klibanov and Pamyatnykh [20]. However, it is necessary to modify the proof of [20] here for the following three reasons. The first and the most important is linked with the weighted scalar product (1.9) in Theorem 1.2 with the weight function $|\cos(n,\nu)|$. Weight functions were not considered in [20]. The delicacy here is due to the fact that this weight function is vanishing at a set $S \subset \Gamma_-$. It is well known, however, that the presence of zeros of weight functions in Hilbert spaces usually causes complications in the analysis. Because of this, we need to carefully evaluate the boundary terms in the pointwise Carleman estimate for the principal part of the differential operator of the transport equation, which was not done in [20]. The second reason is that the result of [20] was established for solutions $u \in C^1$, whereas we need to work with weak solutions $u \in L^2$ of the transport equation. The latter causes significant additional complications; see Remark 2.1. Third, the Lipschitz stability estimate was proved in [20] in the entire time cylinder, and this estimate is similar to estimate (1.10) in our case. However, in addition to (1.10), we need to obtain an estimate at the top $\{t = T\}$ of the time cylinder; see (1.11).

The method of Carleman estimates was applied for the proof of the continuous observability estimate by Klibanov and Malinsky [18]. They have done this for the case of hyperbolic equations with the constant principal part and low order terms, $w_{tt} = \Delta w + lot$, where "*lot*" stands for lower order terms. In Kazemi and Klibanov [17], the idea of [18] was applied to a more general case of hyperbolic inequalities, $|w_{tt} - \Delta w| \leq A(|\nabla w| + |w_t| + |w| + |f(x,t)|), A = const. > 0$, and the case when one boundary condition is given only at a part of the boundary was considered. One of the auxiliary results of the book of Klibanov and Timonov [19] is an extension of the method of [17] and [18] to the case of a more general hyperbolic inequality, $|a(x)w_{tt} - \Delta w| \leq A(|\nabla w| + |w_t| + |w| + |f(x,t)|)$, with some restrictions imposed on the positive function $a(x)$. The method of [17] and [18] enabled one to establish the exact controllability for the hyperbolic equations with lower order terms; see, e.g., the review paper of Gulliver et al. [11]. The previously applied method of multipliers was working (at least at the time of [17] and [18]) only under the assumption $lot = 0$; see Ho [12] for the first publication of the method of multipliers. Currently Carleman estimates are widely used in control theory for proofs of continuous observability results; see, e.g., [9], [10], [11], [14], and [15]. In this paper we modify the idea of [17], [18], [19] for the case of the transport equation.

In order to take into account the nonzero boundary condition, we derive a pointwise Carleman estimate, as was originated in the book of Lavrent'ev, Romanov, and Shishatskiĭ [26]. Another popular method of deriving of Carleman estimates is that of Hörmander [13]. This method is well suited for the so-called unique continuation theorems, which establishes that certain zero boundary conditions correspond only to the zero solution. However, it cannot be applied in our case. The reason is that one of the requirements of the method of [13] is the zero boundary condition, while our goal is to estimate the solution via a nonzero boundary condition.

All functions considered in this paper are real valued. Thus, Hilbert spaces here contain only real valued functions. Let $z_1, z_2 \in \overline{\Omega}$ be two points such that

$$|z_1 - z_2| = \max_{x,y \in \overline{\Omega}} |x - y| .$$

Without loss of generality we assume that $0 = (z_1 + z_2)/2$. Clearly, $0 \in \Omega$. Denote

$$R = \max_{x \in \overline{\Omega}} |x| .$$

For a function $g(x)$ with $x \in \mathbb{R}^N$, denote $g_i = \partial g / \partial x_i$ whenever the differentiation is appropriate.

In (1.1), we assume that

$$a \in C^1(\overline{W}), \quad g \in C^1(\overline{W} \times S^{N-1}). \tag{1.2}$$

It seems that the weaker assumptions $a \in C(\overline{W}), g \in C(\overline{W} \times S^{N-1})$ may be sufficient. Still, we prefer to use (for brevity) a little bit stronger assumption (1.2) to introduce the definition of the weak solution, which in turn relies on Theorem 2.1. In this paper we consider the following problem.

Exact controllability problem. Consider the zero initial condition

$$u \mid_{t=0} = 0 \tag{1.3}$$

and the boundary condition

$$u \mid_{\Gamma_-} = p(x, t, \nu), \tag{1.4}$$

where $p \in L^2_{\cos}(\Gamma_-)$ (see below for the definition of the Hilbert space $L^2_{\cos}(\Gamma_-)$). We assume that the weak solution $u \in C([0, T]; L^2(\Omega \times S^{N-1}))$ of the problem (1.1)–(1.4) can be defined (Theorem 2.2). Let $u_T(x, \nu) \in L^2(\Omega \times S^{N-1})$ be an arbitrary function. Find a boundary condition (i.e., boundary control) $p = p(u_T) \in L^2_{\cos}(\Gamma_-)$ such that the resulting function $u(x, t, \nu)$ is such that

$$u(x, T, \nu) = u_T(x, \nu). \tag{1.5}$$

Here we can interpret T as the "steering time." Our main result is the following theorem.

THEOREM 1.1. *Let Ω be a strictly convex bounded domain with $\partial\Omega \in C^1$ and $T > 2R$. Then for any function $u_T(x, \nu) \in L^2(\Omega \times S^{N-1})$ there exists a control function $p = p(u_T) \in L^2_{\cos}(\Gamma_-)$ such that if the function u is the weak solution of the initial boundary value problem* (1.1)–(1.4), *then* (1.5) *holds.*

In this paper, without loss of generality, we can assume that the initial value of the controlled system (1.1) is zero. In fact, let Theorem 1.1 be proved and let $u = u(x, t, \nu)$ be the weak solution to (1.1), (1.4) and $u|_{t=0} = u_0$ for a $u_0 \in L^2(\Omega \times S^{n-1})$. For given u_0 and u_T, we have to find a control function $p \in L^2_{\cos}(\Gamma_-)$ such that the function u satisfies (1.5). Let v be the weak solution to $Mv = 0$ in W, $v|_{t=0} = u_0$, and $v|_{\Gamma_-} = 0$. Setting $w = u - v$, we have $Mw = 0$ in W, $w|_{t=0} = 0$ and $w|_{\Gamma_-} = p$. Therefore by Theorem 1.1 in the case of the zero initial condition, which is assumed to be solved, for w we can find $p \in L^2_{\cos}(\Gamma_-)$ such that $w(x, T, \nu) = u_T(x, \nu) - v(x, T, \nu)$. This control p steers u from u_0 at $t = 0$ to u_T at $t = T$.

Consider the weak solution of the adjoint transport equation

$$M^* v := v_t + (\nu, \nabla v) - a(x, t, \nu) v + \int_{S^{N-1}} g(x, t, \mu, \nu) v(x, t, \mu) d\sigma_\mu = 0 \quad \text{in } W, \tag{1.6}$$

$$v(x, T, \nu) = v_0(x, \nu) \in L^2(\Omega \times S^{N-1}), \tag{1.7}$$

$$v \mid_{\Gamma_+} = 0. \tag{1.8}$$

We introduce the weighted scalar product as

$$\langle p, q\rangle = \int_{\Gamma_-} p(x,t,\nu)q(x,t,\nu)\cdot|\cos(n,\nu)|dS_x dt d\sigma_\nu. \tag{1.9}$$

By Lemma 2.1 (below), (1.9) is a scalar product which generates a Hilbert space, which we denote by $L^2_{\cos}(\Gamma_-)$. Note that

$$L^2(\Gamma_-)\subset L^2_{\cos}(\Gamma_-) \text{ and } \|p\|_{L^2_{\cos}(\Gamma_-)}\le \|p\|_{L^2(\Gamma_-)} \quad \forall p\in L^2(\Gamma_-);$$

that is, the $L^2_{\cos}(\Gamma_-)-$ norm is weaker than the $L^2(\Gamma_-)$-norm. The necessity of the introduction of the weighted space $L^2_{\cos}(\Gamma_-)$ can be seen from (3.13) (section 3). In order to prove Theorem 1.1, we combine the duality argument with the following continuous observability result.

THEOREM 1.2. *Assume that Ω is a strictly convex bounded domain with $\partial\Omega\in C^1$ and $T>2R$. Let the function v be the weak solution of the adjoint problem* (1.6)–(1.8) *in the sense of Definition* 2.1. *Let $v\mid_{\Gamma_-}:=(Kv_0)(x,t,\nu)\in L^2_{\cos}(\Gamma_-)$ be the generalized trace of the function v on Γ_- (Definition* 3.1*). Then the following Lipschitz stability estimates are valid:*

$$\|v\|_{L^2(W)}\le C\|Kv_0\|_{L^2_{\cos}(\Gamma_-)}, \tag{1.10}$$

$$\|v_0\|_{L^2(\Omega\times S^{n-1})}\le C\|Kv_0\|_{L^2_{\cos}(\Gamma_-)}, \tag{1.11}$$

where the positive constant $C=C(\Omega,T,\|a\|_{C(\overline{W})},\|g\|_{C(\overline{W}\times S^{N-1})})$ depends only on numbers R, T and norms $\|a\|_{C(\overline{W})}$ and $\|g\|_{C(\overline{W}\times S^{N-1})}$.

Theorem 1.2 is derived from the following theorem, which asserts a continuous observability estimate for the corresponding transport inequality.

THEOREM 1.3. *Assume that Ω is a strictly convex bounded domain with $\partial\Omega\in C^1$ and $T>2R$. Suppose that the function $v\in C^1_{tv grad}(\overline{W})$ satisfies the "transport inequality"*

$$|v_t+(\nu,\nabla v)|\le M\left[|v|+\int_{S^{N-1}}|v(x,t,\mu)|d\sigma_\mu+|f(x,t,\nu)|\right] \text{ in } W, \tag{1.12}$$

where M is a positive constant and the function $f\in L^2(W)$. Then

$$\|v\|_{L^2(W)}\le M_1[\|v\mid_\Gamma\|_{L^2_{\cos}(\Gamma)},+\|f\|_{L_2(W)}], \tag{1.13}$$

$$\|v(x,t_0,\nu)\|_{L^2(\Omega\times S^{N-1})}\le M_1[\|v\mid_\Gamma\|_{L^2_{\cos}(\Gamma)},+\|f\|_{L_2(W)}] \quad \forall t_0\in[0,T], \tag{1.14}$$

where the positive constant $M_1=M_1(\Omega,T,M)$ depends only on numbers Ω,T, and M.

The space $C^1_{tv grad}(\overline{W})$ is defined in subsection 2.2, and we note that $C^1_{tv grad}(\overline{W})\not\subset C^1(\overline{W})$. For $v\in C^1(\overline{W})$, the proof of the Carleman estimate can also be done in a traditional way with the commutator, but in our case we need some device (see (4.6)). Although the smoothness condition $v\in C^1_{tv grad}(\overline{W})$ of this theorem can be relaxed, to save space we are not doing this here, because Theorem 1.2 is more important for our goal. In this paper $C=C(\Omega,T,\|a\|_{C(\overline{W})},\|g\|_{C(\overline{W}\times S^{N-1})})$ and $M_1=M_1(\Omega,T,M)$ denote different positive constants depending on parameters listed. The conditions of Theorem 1.1 are assumed to be satisfied below. In section 2 we introduce the weak solutions of the problems (1.1)–(1.4) and (1.6)–(1.8). In section 3 we prove Theorem 1.1, assuming that Theorem 1.2 is valid. In section 4 we prove Theorems 1.2 and 1.3.

2. Strong and weak solutions.

2.1. The space $L^2_{\cos}(\Gamma_-)$ of controls. In this subsection we prove the following lemma.

LEMMA 2.1. *$L^2_{\cos}(\Gamma_-)$ is a Hilbert space.*

Proof. First, we have to prove that $\langle p, p\rangle = 0 \Leftrightarrow p = 0$. We set $\Phi := \{(x,\nu) \in \partial\Omega \times S^{N-1} : \cos(n,\nu) = 0\}$. It is sufficient to prove that

$$meas\,(\Phi) = 0.$$

Since the boundary $\partial\Omega \in C^1$ class, we can locally represent $\partial\Omega$ by $\{x_1 = 0\}$ via choosing suitable coordinates. Therefore, without loss of generality, we can assume that $\partial\Omega = \cup_{j=1}^J \partial_j\Omega$ and in each Γ_j, we set $n(x) = (1,0,\dots,0)^T$ by changing variables. Here the superscript "T" means the transpose. Then $\cos(n(x),\nu) = 0$ is equivalent to $\nu = (0,\nu_2,\dots,\nu_N)^T$ with $\sum_{j=2}^N \nu_j^2 = 1$. Hence $\Phi \cap (\partial_j\Omega \times S^{N-1}) \subset \partial_j\Omega \times S^{N-2}$; that is, $\Phi\cap(\partial_j\Omega\times S^{N-1})$ is a $(2N-3)$-dimensional hypersurface in the $(2N-2)$-dimensional space. Hence $meas(\Phi\cap(\partial_j\Omega\times S^{N-1})) = 0$. Therefore $meas(\Phi\cap(\cup_{j=1}^J\partial_j\Omega\times S^{N-1})) = meas(\Phi) = 0$. Thus we see that (1.9) defines a scalar product.

Second, we prove the completeness of $L^2_{\cos}(\Gamma_-)$. Let $\{p_k\}_{k=1}^\infty \subset L^2_{\cos}(\Gamma_-)$ be a Cauchy sequence in the norm of $L^2_{\cos}(\Gamma_-)$; that is, $\lim_{k,\ell\to\infty} \|p_k - p_\ell\|_{L^2_{\cos}(\Gamma_-)} = 0$. Then $\{p_k|\cos(n,\nu)|^{\frac12}\}_{k=1}^\infty$ is a Cauchy sequence in $L^2(\Gamma_-)$. Hence, there exists a function $\widetilde{p} \in L^2(\Gamma_-)$ such that

$$\lim_{k\to\infty}\int_{\Gamma_-} |\widetilde{p} - p_k|\cos(n,\nu)|^{\frac12}|^2 dS_x dt d\sigma_\nu = 0.$$

Denote

$$q = \frac{\widetilde{p}}{|\cos(n,\nu)|^{\frac12}}.$$

Hence, $q \in L^2_{\cos}(\Gamma_-)$. Therefore,

$$\lim_{k\to\infty}\int_{\Gamma_-}\left|\frac{\widetilde{p}}{|\cos(n,\nu)|^{\frac12}} - p_k\right|^2 |\cos(n,\nu)| dS_x dt d\nu = \lim_{k\to\infty}\|q - p_k\|^2_{L^2_{\cos}(\Gamma_-)} = 0,$$

which proves the completeness of $L^2_{\cos}(\Gamma_-)$. ∎

2.2. Strong solution. Consider the case when (1.3) is replaced with

$$u\,|_{t=0} = f(x,\nu). \tag{2.1}$$

For the exact controllability, we need a weak solution. However, there are few publications where weak solutions with nonzero boundary values in $L^2_{\cos}(\Gamma_-)$ are introduced. For complete accounts, see Chapter 2 in Ukai [35], which is a monograph in Japanese, and Bardos [1], whose argument can be modified in our case. As for the weak solution with the homogenous boundary data, there are rich references, for example, Bardos [1], Douglis [6], and Ukai [36]. In principle, the L^2-solution of the transport equation with nonhomogeneous boundary value in $L^2_{\cos}(\Gamma_-)$ can be defined by the transposition method (see Chapter 3 in Lions and Magenes [30] or pp. 46–50 in Lions [29]). However, it is convenient for our goal—the exact controllability to define the weak solution via density arguments—and we give self-contained descriptions for

completeness. For this, we use a result of Prilepko and Ivankov [31] about strong solutions.

We first assume that

$$f \in C^\infty(\overline{\Omega} \times S^{N-1}) \text{ and } f(x,\nu) \in C_0^\infty(\Omega) \quad \forall \nu \in S^{N-1}, \tag{2.2}$$

(2.3)

$$p \in C^\infty(\overline{\Gamma}_-), \quad p(x,t,\nu) := p_{x,\nu}(t) \in C_0^\infty(0,T) \quad \text{for every appropriate pair } (x,\nu).$$

Following [31], we introduce the functional space

$$C^1_{t\nu grad}(\overline{W}) = \left\{ u(x,t,\nu) : u, u_t, \frac{d}{ds}u(x+s\nu,t,\nu) \mid_{s=0} \in C(\overline{W})\ \forall \nu \in S^{N-1} \right\},$$

$$\|u\|_{C^1_{t\nu grad}(\overline{W})} = \|u\|_{C(\overline{W})} + \|u_t\|_{C(\overline{W})} + \left\| \frac{d}{ds}u(x+s\nu,t,\nu) \mid_{s=0} \right\|_{C(\overline{W})}.$$

Rewrite (1.1) in a different form:

$$u_t + \frac{d}{ds}u(x+s\nu,t,\nu) \mid_{s=0} + a(x,t,\nu)u - \int_{S^{N-1}} g(x,t,\nu,\mu)u(x,t,\mu)d\sigma_\mu = 0 \quad \text{in } W. \tag{2.4}$$

Equations (1.1) and (2.4) are not equivalent. If, for example, the derivatives $u_t, u_i \in C(\overline{W})$ and the function u satisfies (2.4), then this function also satisfies (1.1). However, if a function $u \in C^1_{t\nu grad}(\overline{W})$ satisfies (2.4), but not all of its derivatives u_i exist, then this function might not be a solution of (1.1). Therefore (2.4) is more general than (1.1). Theorem 2.1 is a simplified version of Theorem 1.1 of [31].

THEOREM 2.1. *Let $\Omega \subset \mathbb{R}^N$ be a strictly convex bounded domain with $\partial\Omega \in C^1$. Assume that conditions* (1.2), (2.2), *and* (2.3) *hold. Then there exists unique solution $u \in C^1_{t\nu grad}(\overline{W})$ of the problem* (1.4), (2.1), (2.4).

Remark 2.1. In the view of our goal, a significant complication linked with Theorem 2.1 is an insufficient smoothness guaranteed by this theorem. In other words, we cannot work directly with the "individual" derivatives u_i, because their existence is not guaranteed. Rather, we need to work with the directional derivatives $du(x+s\nu,t,\nu)/ds \mid_{s=0}$. Furthermore, we cannot even claim that such a directional derivative equals $(\nu, \nabla u)$. The key idea, which helps to overcome these complications, is the introduction of an orthogonal matrix A_{ν_0} in (2.6).

In this section we will relax smoothness conditions (2.2), (2.3). We need these minimal conditions in order to introduce the weak solution. On the other hand, we need the weak solution for the duality argument.

2.3. Weak solution.

LEMMA 2.2 (energy conservation). *Suppose that conditions* (1.2), (2.2), *and* (2.3) *are fulfilled. Let the function $u \in C^1_{t\nu grad}(\overline{W})$ be a solution of the problem* (1.4), (2.1), (2.4). *Denote*

$$E(u,t) = \int_\Omega \int_{S^{N-1}} |u(x,t,\nu)|^2 d\sigma_\nu dx.$$

Then there exists a positive constant $C = C(\Omega, T, \|a\|_{C(\overline{W})}, \|g\|_{C(\overline{W}\times S^{N-1})})$ such that for any two numbers $t_1, t_2 \in [0,T]$

$$E(u,t_2) \le C[E(u,t_1) + \|p\|^2_{L^2_{\cos}(\Gamma_-)}]. \tag{2.5}$$

Proof. Fix an arbitrary vector $\nu_0 \in S^{N-1}$. Let $A_{\nu_0} = (a^{ij}_{\nu_0})_{i,j=1}^N$ be an orthogonal matrix such that

$$A_{\nu_0}\nu_0 = \widetilde{\nu}_0 := (1, 0, 0, \dots, 0)^T. \tag{2.6}$$

Let

$$y = A_{\nu_0}x. \tag{2.7}$$

Denote $A_{\nu_0}\Omega = \{y = Ax : x \in \Omega\}$. Also, for any point $y = A_{\nu_0}x \in \partial(A_{\nu_0}\Omega)$ (hence, $x \in \partial\Omega$) let $\widetilde{n}(y) = A_{\nu_0}n(x)$ be the unit outward normal vector at the point y. Denote

$$\widetilde{u}(y,t,\eta) = u\left(A^{-1}_{\nu_0}y, t, A^{-1}_{\nu_0}\eta\right) \quad \forall \eta \in S^{N-1}, \tag{2.8}$$

$$\widetilde{a}(y,t,\widetilde{\nu}_0) = a\left(A^{-1}_{\nu_0}y, t, A^{-1}_{\nu_0}\widetilde{\nu}_0\right), \tag{2.9}$$

$$\widetilde{g}(y,t,\widetilde{\nu}_0,\eta) = g\left(A^{-1}_{\nu_0}y, t, A^{-1}_{\nu_0}\widetilde{\nu}_0, A^{-1}_{\nu_0}\eta\right) \quad \forall \eta \in S^{N-1}. \tag{2.10}$$

In the new coordinates, noting that $\widetilde{u}(y,t,\widetilde{\nu}_0) = u(x,t,\nu_0)$ and

$$u(x + s\nu_0, t, \nu_0) = u(A^{-1}_{\nu_0}(y + s\widetilde{\nu}_0), t, \nu_0) = \widetilde{u}(y + s\widetilde{\nu}_0, t, \widetilde{\nu}_0)$$
$$= \widetilde{u}(y_1 + s, y_2, \dots, y_n, t, \widetilde{\nu}_0),$$

we have

$$\frac{d}{ds}u(x + s\nu_0, t, \nu_0) \mid_{s=0} = \widetilde{u}_{y_1}(y,t,\widetilde{\nu}_0). \tag{2.11}$$

Hence, setting $\nu = \nu_0$ in (1.1), we obtain

$$(\widetilde{u}_t + \widetilde{u}_{y_1} + \widetilde{a}\widetilde{u})(y,t,\widetilde{\nu}_0) - \int_{S^{N-1}} \widetilde{g}(y,t,\widetilde{\nu}_0,\eta)\widetilde{u}(y,t,\eta)d\sigma_\eta = 0. \tag{2.12}$$

Since $u \in C^1_{t\nu grad}(\overline{W})$, we have by (2.11)

$$\widetilde{u}(y,t,\widetilde{\nu}_0), \widetilde{u}_t(y,t,\widetilde{\nu}_0), \widetilde{u}_{y_1}(y,t,\widetilde{\nu}_0) \in C(\overline{A_{\nu_0}\Omega} \times [0,T]). \tag{2.13}$$

Actually, the goal of the transformation (2.7)–(2.10) was to obtain (2.12) with $\widetilde{u}_{y_1}(y,t,\widetilde{\nu}_0) \in C(A_{\nu_0}\overline{\Omega} \times [0,T])$.

Multiply both sides of (2.12) by the function $\widetilde{u}(y,t,\widetilde{\nu}_0)$. We obtain for this vector $\widetilde{\nu}_0$

$$[(\widetilde{u}^2)_t + (\widetilde{u}^2)_{y_1}](y,\tau,\nu_0) = -2\widetilde{a}\widetilde{u}^2(y,\tau,\nu_0) + 2\widetilde{u}(y,\tau,\nu_0) \cdot \int_{S^{N-1}} \widetilde{g}(y,\tau,\widetilde{\nu}_0,\eta)\widetilde{u}(y,\tau,\eta)d\sigma_\eta,$$

where $\tau \in (0,T)$. Integrating this equality with respect to $(y,\tau) \in A_{\nu_0}\Omega \times (t_1,t)$, $t \in (t_1,T)$, we obtain

$$\int_{A_{\nu_0}\Omega} \widetilde{u}^2(y,t,\widetilde{\nu}_0)dy + \int_{t_1}^t \int_{\partial(A_{\nu_0}\Omega)} \cos(\widetilde{n},y_1)\widetilde{u}^2(y,\tau,\widetilde{\nu}_0)dS_y d\tau$$
$$= \int_{A_{\nu_0}\Omega} \widetilde{u}^2(y,t_1,\widetilde{\nu}_0)dy - 2\int_{t_1}^t \int_{A_{\nu_0}\Omega} \widetilde{a}\widetilde{u}^2(y,\tau,\widetilde{\nu}_0)dyd\tau$$
$$+ 2\int_{t_1}^t \int_{A_{\nu_0}\Omega} \widetilde{u}(y,\tau,\widetilde{\nu}_0)\left[\int_{S^{N-1}} \widetilde{g}(y,\tau,\widetilde{\nu}_0,\eta)\widetilde{u}(y,\tau,\eta)d\sigma_\eta\right]dyd\tau.$$

Changing variables "backwards," $x = A_{\nu_0}^{-1} y$, and noting that by (2.6) $\cos(\widetilde{n}(y), y_1) = \cos(n(x), \nu_0)$, we obtain

$$\begin{aligned}(2.14)\qquad &\int_\Omega u^2(x,t,\nu_0)dx + \int_{t_1}^t \int_{\partial\Omega} \cos(n,\nu_0) u^2(x,\tau,\nu_0) dS_x d\tau \\ &= \int_\Omega u^2(x,t_1,\nu_0)dx - 2\int_{t_1}^t \int_\Omega (au^2)(x,\tau,\nu_0)dxd\tau \\ &\quad + 2\int_{t_1}^t \int_\Omega u(x,\tau,\nu_0)\left[\int_{S^{N-1}} g(x,\tau,\nu_0,\eta)u(x,\tau,\eta)d\sigma_\eta\right]dxd\tau.\end{aligned}$$

Let

$$\Gamma(t_1,t) = \Gamma \cap \{(x,t,\nu) : t \in (t_1,t)\},$$

$$\Gamma_-(t_1,t) = \Gamma_- \cap \Gamma(t_1,t) \quad \text{and} \quad \Gamma_+(t_1,t) = \Gamma_+ \cap \Gamma(t_1,t).$$

Recalling that $\nu_0 \in S^{N-1}$ is an arbitrary vector, we can now integrate (2.14) with respect to $\nu_0 \in S^{N-1}$. We obtain

$$\begin{aligned}(2.15)\qquad &\int_{S^{N-1}}\int_\Omega u^2(x,t,\nu)dxd\sigma_\nu + \int_{\Gamma(t_1,t)} \cos(n(x),\nu)u^2(x,\tau,\nu)dS_x d\sigma_\nu d\tau \\ &= \int_{S^{N-1}}\int_\Omega u^2(x,t_1,\nu)dxd\sigma_\nu - 2\int_{t_1}^t\int_{S^{N-1}}\int_\Omega (au^2)(x,\tau,\nu)\,dxd\sigma_\nu d\tau \\ &\quad + 2\int_{t_1}^t\int_{S^{N-1}}\int_\Omega u(x,\tau,\nu)\left[\int_{S^{N-1}} g(x,\tau,\nu,\eta)u(x,\tau,\eta)d\sigma_\eta\right]dxd\sigma_\nu d\tau.\end{aligned}$$

Since $\Gamma(t_1,t) = \Gamma_-(t_1,t) \cup \Gamma_+(t_1,t)$, $\Gamma_-(t_1,t) \subset \Gamma_-(T)$, and $\cos(n(x),\nu) > 0$ on $\Gamma_+(t_1,t)$, we have

$$\begin{aligned}&\int_{\Gamma(t_1,t)} \cos(n,\nu)u^2(x,\tau,\nu)dS_x d\sigma_\nu d\tau \geq \int_{\Gamma_-(t_1,t)} \cos(n,\nu)u^2(x,\tau,\nu)dS_x d\sigma_\nu d\tau \\ &\geq \int_{\Gamma_-(T)} \cos(n,\nu)u^2(x,\tau,\nu)dS_x d\sigma_\nu d\tau = -\int_{\Gamma_-(T)} |\cos(n,\nu)|u^2(x,\tau,\nu)dS_x d\sigma_\nu d\tau \\ &= -\|u\|^2_{L^2_{\cos}(\Gamma_-)} = -\|p\|^2_{L^2_{\cos}(\Gamma_-)}.\end{aligned}$$

Also,

$$\begin{aligned}&\left|\int_{t_1}^t\int_{S^{N-1}}\int_\Omega u(x,\tau,\nu)\left[\int_{S^{N-1}} g(x,\tau,\nu,\eta)u(x,t,\eta)d\sigma_\eta\right]dxd\sigma_\nu d\tau\right| \\ &\leq C\left|\int_{t_1}^t\int_{S^{N-1}}\int_\Omega |u(x,\tau,\nu)|\left(\int_{S^{N-1}}|u(x,t,\eta)|d\sigma_\eta\right)dxd\sigma_\nu d\tau\right| \\ &= C\int_{t_1}^t\int_\Omega\left(\int_{S^{N-1}}|u(x,\tau,\nu)|d\sigma_\nu \cdot \int_{S^{N-1}}|u(x,\tau,\mu)|d\sigma_\mu\right)dxd\tau \\ &\leq C\int_{t_1}^t\int_\Omega\int_{S^{N-1}} u^2(x,\tau,\nu)d\sigma_\nu dxd\tau.\end{aligned}$$

At the last inequality, we used the Cauchy–Schwarz inequality. Hence we obtain from (2.15)

$$E(u,t) \le E(u,t_1) + \|p\|^2_{L^2_{\cos}(\Gamma_-)} + C\int_{t_1}^{t} E(u,\tau)d\tau.$$

Hence the Gronwall inequality leads to (2.5). □

THEOREM 2.2. *Let $\Omega \subset \mathbb{R}^N$ be a strictly convex bounded domain with $\partial\Omega \in C^1$ and let conditions* (1.2) *hold. Let $f \in L^2(\Omega \times S^{N-1})$ and $p \in L^2_{\cos}(\Gamma_-)$ be two arbitrary functions. Consider two functional sequences $\{f_k\}_{k=1}^{\infty}$, $\{p_k\}_{k=1}^{\infty}$ satisfying conditions* (2.2) *and* (2.3) *and such that*

$$\lim_{k\to\infty} \|f_k - f\|_{L^2(\Omega\times S^{N-1})} = \lim_{k\to\infty} \left\| p_k - p\sqrt{|\cos(n,\nu)|} \right\|_{L^2(\Gamma_-)} = 0.$$

Let $u_k \in C^1_{tv grad}\left(\overline{W}\right)$ be the solution of the boundary value problem (1.4), (2.1), (2.4) *with the initial condition f_k and the boundary condition p_k. Then there exists a function $u \in C([0,T]; L^2(\Omega \times S^{N-1}))$ such that $\lim_{k\to\infty} \|u_k - u\|_{C([0,T];L^2(\Omega\times S^{N-1}))} = 0$. Inequality* (2.5) *holds for this function u, and*

$$\|u\|_{C([0,T];L^2(\Omega\times S^{N-1}))} \le C\left(\|f\|_{L^2(\Omega\times S^{N-1})} + \|p\|_{L^2_{\cos}(\Gamma_-)}\right). \tag{2.16}$$

For any pair of functions $f \in L^2\left(\Omega \times S^{n-1}\right)$ and $p \in L^2_{\cos}\left(\Gamma_-\right)$ the resulting function u is independent of functional sequences $\{f_k\}_{k=1}^{\infty}$ and $\{p_k\}_{k=1}^{\infty}$.

Proof. The existence of a sequence $\{p_k\}_{k=1}^{\infty}$ satisfying conditions (2.3) and such that $\lim_{k\to\infty}\|p_k - p\sqrt{\cos(n,\nu)}\|_{L^2(\Gamma_-)} = 0$ follows from the fact that the function $p\sqrt{|\cos(n,\nu)|} \in L^2(\Gamma_-)$ and the set of functions satisfying (2.3) is dense in $L^2(\Gamma_-)$. Also, since the $L^2_{\cos}(\Gamma_-)$-norm is weaker than the $L^2(\Gamma_-)$-norm and $\{p_k\}_{k=1}^{\infty}$ is a Cauchy sequence in $L^2(\Gamma_-)$, it follows that the sequence $\{p_k\}_{k=1}^{\infty} \subset L^2_{\cos}(\Gamma_-)$ and it is a Cauchy sequence in $L^2_{\cos}(\Gamma_-)$. Since functions $u_k \in C^1_{tv grad}\left(\overline{W}\right)$, it follows that $u_k \in C([0,T]; L^2(\Omega \times S^{N-1}))$. Thus, setting $t_1 = 0$, $u = u_k - u_\ell$, $f = f_k - f_\ell$, and $p = p_k - p_\ell$ in (2.5) and taking the maximum in t, we see that $\{u_k(x,t,\nu)\}_{k=1}^{\infty}$ is a Cauchy sequence in the space $C([0,T]; L^2(\Omega \times S^{N-1}))$. Hence we define the function $u(x,t,\nu)$ as $u := \lim_k \to \infty u_k$ in $C([0,T]; L^2(\Omega \times S^{N-1}))$. Since (2.5) holds for functions u_k, it also holds for the function u, which implies (2.16). The independence of the function u on specific sequences $\{f_k\}_{k=1}^{\infty}$ *and* $\{p_k\}_{k=1}^{\infty}$ follows from (2.5). □

DEFINITION 2.1. *Let $\Omega \subset \mathbb{R}^N$ be a strictly convex bounded domain with $\partial\Omega \in C^1$ and let functions a and g satisfy conditions* (1.2). *Let the function $u \in C([0,T]; L^2(\Omega \times S^{N-1}))$ be the one obtained as the limit described in Theorem* 2.2. *Then we call this function u the* weak solution *of the initial boundary value problem* (1.1), (1.4), (2.1) *with the initial condition $f \in L^2\left(\Omega \times S^{N-1}\right)$ and the boundary condition $p \in L^2_{\cos}\left(\Gamma_-\right)$.*

By (2.16) the limit u is independent of choices of sequences f_k, p_k. Thus the following theorem follows from Theorem 2.2 and Definition 2.1.

THEOREM 2.3. *Let $\Omega \subset \mathbb{R}^N$ be a strictly convex bounded domain with $\partial\Omega \in C^1$ and let conditions* (1.2) *hold. Then for each pair of functions $f \in L^2\left(\Omega \times S^{n-1}\right)$ and $p \in L^2_{\cos}\left(\Gamma_-\right)$ the weak solution $u \in C([0,T]; L^2(\Omega \times S^{N-1}))$ of the problem* (1.1), (1.4), (2.1) *exists and is unique, and* (2.16) *holds.*

Consider now the adjoint problem (1.6)–(1.8). Similarly to (2.4) and following the same considerations, we rewrite (1.6) as

(2.17)
$$v_t + \frac{d}{ds} v\left(x + s\nu, t, \nu\right)|_{s=0} - a(x,t,\nu)v + \int_{S^{N-1}} g(x,t,\mu,\nu)v(x,t,\mu)d\sigma_\mu = 0 \quad \text{in } W.$$

The following result follows immediately from Lemma 2.2 and Theorems 2.1–2.3 via the change of variables $t \Leftrightarrow \tau = T - t$.

THEOREM 2.4. *Let $\Omega \subset \mathbb{R}^N$ be a strictly convex bounded domain with $\partial\Omega \in C^1$ and let conditions* (1.2) *hold. Suppose that in* (1.7) *the function $v_0 \in C^\infty\left(\overline{\Omega} \times S^{N-1}\right)$ and $v_0(x,\nu) \in C_0^\infty(\Omega)$ for all $\nu \in S^{N-1}$. Let the function $v \in C^1_{tv grad}\left(\overline{W}\right)$ be a solution of the problem* (1.7), (1.8), (2.17) *(Theorem* 2.1*). Denote*

$$E(v,t) = \int_\Omega \int_{S^{N-1}} |v(x,t,\nu|^2 \, d\nu dx.$$

Then there exists a positive constant $C = C\left(\Omega, T, \|a\|_{C\left(\overline{W}\right)}, \|g\|_{C\left(\overline{W} \times S^{N-1}\right)}\right)$ such that for any two numbers $t_1, t_2 \in [0,T]$

$$E(v,t_2) \leq CE(v,t_1). \tag{2.18}$$

Thus, the solution $v \in C^1_{tv grad}\left(\overline{W}\right)$ of the problem (1.7), (1.8), (2.17) *both exists and is unique. Next, assume that $v_0 \in L^2\left(\Omega \times S^{N-1}\right)$ is an arbitrary function. Let $\{v_{0k}\}_{k=1}^\infty \subset C^\infty\left(\overline{\Omega} \times S^{N-1}\right)$ be a sequence such that $v_{0k}(x,\nu) \in C_0^\infty(\Omega)$ for all $\nu \in S^{N-1}$ and $\lim_{k\to\infty} \|v_{0k} - v_0\|_{L_2(\Omega \times S^{N-1})} = 0$. Let $\{v_k\}_{k=1}^\infty \in C^1_{tv grad}\left(\overline{W}\right)$ be the sequence of solutions of the initial boundary value problem* (1.7), (1.8), (2.17) *with initial conditions $v_k|_{t=0} = v_{0k}$. Then there exists a function $v = v(x,t,\nu) \in C([0,T]; L^2(\Omega \times S^{N-1}))$ such that $\lim_k \to \infty \|v_k - v\|_{C([0,T];L^2(\Omega \times S^{N-1}))} = 0$. For any given function $v_0 \in L^2\left(\Omega \times S^{N-1}\right)$ the function v is independent of the functional sequence $\{v_{0k}\}_{k=1}^\infty$. Furthermore,*

$$\|v\|_{C([0,T];L^2(\Omega \times S^{N-1}))} \leq C\|v_0\|_{L^2(\Omega \times S^{N-1})}. \tag{2.19}$$

DEFINITION 2.2. *We call the function $v \in C([0,T]; L^2(\Omega \times S^{N-1}))$ constructed in Theorem* 2.4 *the "weak solution" of the adjoint problem* (1.6)–(1.8).

Therefore, the following corollary follows from Theorem 2.4.

COROLLARY 2.1. *For any function $v_0 \in L_2\left(\Omega \times S^{N-1}\right)$ there exists a unique weak solution $v \in C([0,T]; L^2\left(\Omega \times S^{N-1}\right))$ to* (1.6)–(1.8). *Estimate* (2.19) *holds for this function v.*

3. Proof of Theorem 1.1. In this section we prove Theorem 1.1, assuming that Theorem 1.2 holds. By Theorem 2.3, for any function $p \in L^2_{\cos}(\Gamma_-)$ there exists a unique weak solution u of the problem (1.1), (1.3), and (1.4). Also, by Corollary 2.1 for any function $v_0 \in L^2(\Omega \times S^{N-1})$ there exists a unique weak solution v of the problem (1.6)–(1.8).

3.1. Generalized trace of the weak solution of the adjoint problem (1.6)–(1.8). For the weak solution v of the problem (1.6)–(1.8), we define in this subsection a generalized trace of the function $v|_{\Gamma_-} \in L^2_{\cos}(\Gamma_-)$. Consider the case when the function p in (1.4) satisfies conditions (2.3). In addition, assume for a while that the function v_0 in (1.7) satisfies the following two conditions:

$$v_0 \in C^\infty\left(\overline{\Omega} \times S^{N-1}\right) \tag{3.1}$$

and

$$v_0(x,\nu) \in C_0^\infty(\Omega) \quad \forall \nu \in S^{N-1}. \tag{3.2}$$

Therefore, Theorems 2.1 and 2.3 guarantee that unique solutions $u, v \in C^1_{tvgrad}\left(\overline{W}\right)$ exist for the following two initial boundary value problems:

$$u_t + \frac{d}{ds} u\left(x + s\nu, t, \nu\right) |_{s=0} + a(x,t,\nu)u - \int_{S^{N-1}} g(x,t,\nu,\mu)u(x,t,\mu)d\sigma_\mu = 0 \quad \text{in } W, \tag{3.3}$$

$$u\,|_{t=0} = 0, \tag{3.4}$$

$$u\,|_{\Gamma_-} = p(x,t,\nu) \tag{3.5}$$

and the adjoint problem

$$v_t + \frac{d}{ds} v\left(x + s\nu, t, \nu\right) |_{s=0} - a(x,t,\nu)v + \int_{S^{N-1}} g(x,t,\mu,\nu)v(x,t,\mu)d\sigma_\mu = 0 \quad \text{in } W, \tag{3.6}$$

$$v(x,T,\nu) = v_0\left(x,\nu\right), \quad (x,\nu) \in \Omega \times S^{N-1}, \tag{3.7}$$

$$v\,|_{\Gamma_+} = 0. \tag{3.8}$$

By (3.4), we have

$$\int_0^T u_t v dt = (uv)\left(x,T,\nu\right) - \int_0^T u v_t dt. \tag{3.9}$$

Fix an arbitrary vector $\nu_0 \in S^{N-1}$. Let A_{ν_0} be an orthogonal matrix satisfying (2.6). Introduce again notations (2.7)–(2.10). In addition, let $\widetilde{v}(y,t,\widetilde{\nu}_0) = v(A^{-1}_{\nu_0}y, t, A^{-1}_{\nu_0}\nu_0)$. Since in the new coordinates

$$\frac{d}{ds} u\left(x + s\nu_0, t, \nu_0\right) |_{s=0} = \widetilde{u}_{y_1}\left(y,t,\widetilde{\nu}_0\right)$$

and

$$\frac{d}{ds} v\left(x + s\nu_0, t, \nu_0\right) |_{s=0} = \widetilde{v}_{y_1}\left(y,t,\widetilde{\nu}_0\right),$$

(3.3), (3.6), and (3.9) imply that

$$\int_{A_{\nu_0}\Omega} \int_0^T \left\{ \left(-\widetilde{u}_{y_1} - \widetilde{a}\widetilde{u}\right)(y,t,\widetilde{\nu}_0) + \int_{S^{N-1}} \widetilde{g}(y,t,\widetilde{\nu}_0,\eta)\widetilde{u}(y,t,\eta)d\sigma_\eta \right\} \widetilde{v}(y,t,\widetilde{\nu}_0)dtdy$$

$$= \int_{A_{\nu_0}\Omega} \int_0^T \left\{ \left(= \widetilde{v}_{y_1} - \widetilde{a}\widetilde{v}\right)(y,t,\widetilde{\nu}_0) + \int_{S^{N-1}} \widetilde{g}(y,t,\eta,\widetilde{\nu}_0)\widetilde{v}(y,t,\eta)d\sigma_\eta \right\} \widetilde{u}(y,t,\widetilde{\nu}_0)dtdy \tag{3.10}$$

$$+ \int_{A_{\nu_0}\Omega} \left(\widetilde{u}\widetilde{v}\right)(y,T,\widetilde{\nu}_0)dy.$$

For an arbitrary vector $\nu \in S^{N-1}$ denote

$$\partial\Omega_-(\nu) = \{x \in \partial\Omega : (n(x), \nu) \le 0\}, \quad \partial\Omega_+(\nu) = \{x \in \partial\Omega : (n(x), \nu) > 0\}.$$

Hence, by (3.5) and (3.8)

$$u(x,t,\nu) = p(x,t,\nu) \text{ for } x \in \partial\Omega_-(\nu), \qquad v(x,t,\nu) = 0 \text{ for } x \in \partial\Omega_+(\nu). \tag{3.11}$$

Let $\widetilde{p}(y,t,\widetilde{\nu}_0) = p\left(A_{\nu_0}^{-1}y, t, A_{\nu_0}^{-1}\widetilde{\nu}_0\right)$. Integrating by parts, we obtain for the two terms in (3.10)

$$\int_{A_{\nu_0}\Omega} (-\widetilde{u}_{y_1} \cdot \widetilde{v})(y,t,\widetilde{\nu}_0)dy - \int_{A_{\nu_0}\Omega} (\widetilde{u} \cdot \widetilde{v}_{y_1})(y,t,\widetilde{\nu}_0)dy$$

$$= -\int_{\partial\left(A_{\nu_0}\Omega\right)} (\widetilde{u} \cdot \widetilde{v})(y,t,\widetilde{\nu}_0)\cos(\widetilde{n}, y_1)\, dS_y.$$

Change variables "backwards" ($y \Leftrightarrow x = A_{\nu_0}^{-1}y$, $\nu_0 = A_{\nu_0}^{-1}\widetilde{\nu}_0$) in the last integral and note again that by (2.6) $\cos(\widetilde{n}(y), y_1) = \cos(n(x), \nu_0)$. Hence, using (3.11), we obtain

$$-\int_{\partial\left(A_{\nu_0}\Omega\right)} (\widetilde{p} \cdot \widetilde{v})(y,t,\widetilde{\nu}_0)\cos(\widetilde{n}, y_1)\, dS_y = -\int_{\partial\Omega} (u \cdot v)(x,t,\nu_0)\cos(n,\nu_0)\, dS_x$$

$$= -\int_{\partial\Omega_-(\nu_0)} (p \cdot v)(x,t,\nu_0)\cos(n,\nu_0)\, dS_x.$$

Hence, integrating with respect to $t \in (0,T)$, we obtain

$$\int_0^T \int_\Omega (-\widetilde{u}_{y_1} \cdot \widetilde{v})(y,t,\widetilde{\nu}_0)dydt - \int_0^T \int_\Omega (\widetilde{u} \cdot \widetilde{v}_{y_1})(y,t,\widetilde{\nu}_0)dydt$$

$$= -\int_0^T \int_{\partial\Omega_-(\nu_0)} p(x,t,\nu_0)v(x,t,\nu_0)\cos(n,\nu_0)\, dS_x dt. \tag{3.12}$$

Changing variables "backwards" in the rest of the integrals of (3.10), substituting (3.12), integrating with respect to $\nu_0 \in S^{N-1}$, and noting that

$$\int_{S^{N-1}} \int_0^T \int_{\partial\Omega_-(\nu)} (\dots) dS_x dt d\sigma_\nu = \int_{\Gamma_-} (\dots) dS_x dt d\sigma_\nu,$$

we obtain

$$-\int_{\Gamma_-} p(x,t,\nu)v(x,t,\nu)\cos(n,\nu)\, dS_x dt d\sigma_\nu = \int_{S^{N-1}} \int_\Omega u(x,T,\nu)v(x,T,\nu)dxd\sigma_\nu. \tag{3.13}$$

Since

$$-\cos(n,\nu) = |\cos(n,\nu)| \quad \text{on } \Gamma_-,$$

(1.9) implies that (3.13) can be rewritten as

$$\langle p, v\rangle = \int_{S^{N-1}} \int_\Omega u(x,T,\nu)v(x,T,\nu)dxd\sigma_\nu. \tag{3.14}$$

For all functions $p \in L^2_{\cos}(\Gamma_-)$, we define the linear operator L by

$$Lp = u(x,T,\nu) \quad \text{for } (x,\nu) \in \Omega \times S^{N-1}, \tag{3.15}$$

where $u \in C([0,T]; L^2(\Omega \times S^{N-1}))$ is the weak solution of the problem (3.3)–(3.5) (Theorem 2.3). Also, for all functions v_0 satisfying conditions (3.2), we define the linear operator K by

$$Kv_0 = v(x,t,\nu) \quad \text{for } (x,t,\nu) \in \Gamma_-,$$

where the function $v \in C^1_{tv grad}\left(\overline{W}\right)$ is the strong solution of the boundary value problem (3.6)–(3.8) (Theorem 2.1).

It follows from (3.15) and (2.16) that

$$\|u(x,T,\nu)\|_{L^2(\Omega\times S^{N-1})} = \|Lp\|_{L^2(\Omega\times S^{N-1})} \le C\,\|p\|_{L^2_{\cos}(\Gamma_-)} \quad \forall p \in L^2_{\cos}(\Gamma_-).$$

Hence, the linear operator $L : L^2_{\cos}(\Gamma_-) \to L^2\left(\Omega \times S^{N-1}\right)$ is bounded:

$$\|Lp\|_{L^2(\Omega\times S^{N-1})} \le \|L\|\,\|p\|_{L^2\cos(\Gamma_-)} \quad \forall p \in L^2_{\cos}(\Gamma_-). \tag{3.16}$$

Let $[,]$ be the scalar product in the Hilbert space $L^2\left(\Omega \times S^{N-1}\right)$. Hence, it follows from (3.7), (3.13), (3.14), and (1.9) that for all functions $p \in L^2_{\cos}(\Gamma_-)$ and all functions v_0 satisfying conditions (3.2) the following equality holds:

$$\langle p, Kv_0\rangle = [Lp, v_0]. \tag{3.17}$$

Let $v_0(x,\nu)$ be an arbitrarily given function satisfying conditions (3.1) and (3.2). Denote $\widetilde{p} = Kv_0$. Then by (3.17)

$$\|Kv_0\|^2_{L^2_{\cos}(\Gamma_-)} = \langle Kv_0, Kv_0\rangle = \langle \widetilde{p}, Kv_0\rangle = [L\widetilde{p}, v_0]. \tag{3.18}$$

Using (3.16), (3.18), and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\|Kv_0\|^2_{L^2_{\cos}(\Gamma_-)} &= \langle Kv_0, Kv_0\rangle = [L\widetilde{p}, v_0]\\ &\le \|L\| \cdot \|\widetilde{p}\|_{L^2_{\cos}(\Gamma_-)} \cdot \|v_0\|_{L^2(\Omega\times S^{N-1})}\\ &= \|L\| \cdot \|Kv_0\|_{L^2_{\cos}(\Gamma_-)} \cdot \|v_0\|_{L^2(\Omega\times S^{N-1})}.\end{aligned}$$

Hence

$$\|Kv_0\|_{L^2_{\cos}(\Gamma_-)} \le \|L\|\,\|v_0\|_{L^2(\Omega\times S^{N-1})}. \tag{3.19}$$

Since the set of functions v_0 satisfying conditions (3.1) and (3.2) is dense in $L^2(\Omega \times S^{N-1})$, we can uniquely extend the bounded operator K, which was originally defined on the set of functions satisfying (3.1) and (3.2), to the bounded operator defined on the whole space $L^2(\Omega \times S^{N-1})$. We denote this extension by the same notation: $Kv_0 = v|_{\Gamma_-}$, where v is the weak solution of the initial boundary value problem (3.6)–(3.8). Hence, it follows from (3.19) that $K : L^2(\Omega \times S^{N-1}) \to L^2_{\cos}(\Gamma_-)$ is a bounded linear operator.

DEFINITION 3.1. *Let $v_0 \in L^2(\Omega \times S^{N-1})$ be an arbitrary function and let v be the weak solution of the adjoint problem* (3.6)–(3.8) *(Definition* 2.2*). We call the function $Kv_0 \in L^2_{\cos}(\Gamma_-)$ the generalized trace of the function v on Γ_-.*

3.2. Application of the theory of closed range operators. It follows from the above that inequality (3.19) holds for any function $v_0 \in L^2(\Omega \times S^{N-1})$. This means that (3.17) holds for all functions $p \in L^2_{\cos}(\Gamma_-)$ and all functions $v_0 \in L^2\left(\Omega \times S^{N-1}\right)$. Therefore, (3.17) implies that

$$L^* = K. \tag{3.20}$$

We now apply estimate (1.11) of Theorem 1.2. By (1.11), we have $\|v_0\|_{L^2(\Omega \times S^{N-1})} \leq C\|Kv_0\|_{L^2_{\cos}(\Gamma_-)}$. Hence, combining this estimate with (3.20), we obtain

$$\|v_0\|_{L^2(\Omega \times S^{N-1})} \leq C\,\|L^* v_0\|_{L^2_{\cos}(\Gamma_-)} \quad \forall v_0 \in L^2(\Omega \times S^{N-1}). \tag{3.21}$$

The estimate (3.21) implies that the operator $L^* = K$ is one-to-one and its range $R(L^*) = L^*(L^2(\Omega \times S^{N-1})) \subset L^2_{\cos}(\Gamma_-)$ is closed. It now follows immediately from Lemma 3, p. 488, of the classic book of Dunford and Schwartz [7] that the operator $L : L^2_{\cos}(\Gamma_-) \to L^2\left(\Omega \times S^{N-1}\right)$ is surjective; i.e., its range is $R(L) = L^2(\Omega \times S^{N-1})$.

In other words, we have proved that for any function $u_T(x,\nu) \in L^2\left(\Omega \times S^{N-1}\right)$ one can find a control function $p \in L^2_{\cos}(\Gamma_-)$ such that $Lp = u(x,T,\nu) = u_T(x,\nu)$, where $u(x,t,\nu) \in C([0,T]; L^2(\Omega \times S^{N-1}))$, is the weak solution of the initial boundary value problem (3.3)–(3.5). Thus, the proof of Theorem 1.1 is complete. □

4. Proofs of Theorems 1.2 and 1.3. Recall that by the definition of the number R (see introduction)

$$|x| \leq R \quad \forall x \in \overline{\Omega}. \tag{4.1}$$

4.1. Carleman estimate. Consider the function

$$\psi(x,t) = |x|^2 - \alpha\left(t - \frac{T}{2}\right)^2, \qquad \alpha = const. \in (0,1). \tag{4.2}$$

The Carleman weight function is defined as

$$\varphi(x,t) = \exp[\lambda\psi(x,t)],$$

where $\lambda > 1$ is a parameter. Let $c = const. \in (0,R)$. Denote

$$G_c = \{(x,t) \in \Omega \times \mathbb{R} : \psi(x,t) > c^2\}. \tag{4.3}$$

Clearly,

$$G_{c_1} \subset G_{c_2} \text{ if } c_1 > c_2. \tag{4.4}$$

The boundary ∂G_c of the domain G_c consists of two parts, $\partial G_c = \partial_1 G_c \cup \partial_2 G_c$, where

$$\partial_1 G_c = \{(x,t) \in \overline{\Omega} \times \mathbb{R} : x \in \partial\Omega\ \} \quad \text{and} \quad \partial_2 G_c = \overline{G}_c \cap \left\{\psi(x,t) = c^2\right\}. \tag{4.5}$$

Hence, $\partial_1 G_c$ is a part of the boundary $\partial\Omega \times \mathbb{R}$ of the time cylinder $\Omega \times \mathbb{R}$, and $\partial_2 G_c$ is a part of the level surface (hyperboloid) of the function $\psi(x,t)$.

LEMMA 4.1. *Let $T > 2R$. Denote $\alpha(R,T) := (2R/T)^2$. Then for all $\alpha \in [\alpha(R,T),1)$ and for all $c \in (0,R)$ the domain $G_c \neq \varnothing$ and $G_c \subset \Omega \times (0,T)$.*

Proof. The following implication follows from (4.5):

$$\partial_1 G_c \subset \{\partial\Omega \times (0,T)\} \Rightarrow G_c \subset \Omega \times (0,T).$$

On the other hand, by (4.2), (4.3), and (4.5)

$$\partial_1 G_c \subset \{\partial\Omega \times (0,T)\} \Leftrightarrow \max_{\partial\Omega}[\psi(x,T)] < c^{2\cdot}.$$

By (4.1) and (4.2)

$$R^2 - \alpha\frac{T^2}{4} < c^2 \Rightarrow \max_{\partial\Omega}[\psi(x,T)] = \max_{\partial\Omega}[\psi(x,0)] < c^{2\cdot}.$$

Since $T > 2R$, the number $\alpha(R,T) = (2R/T)^2 \in (0,1)$. On the other hand, for all $\alpha \in [\alpha(R,T),1)$

$$R^2 - \alpha\frac{T^2}{4} \le R^2 - \alpha(R,T)\frac{T^2}{4} = R^2 - \left(\frac{2R}{T}\right)^2\frac{T^2}{4} = 0 < c^2.$$

Also, since all points of the segment of the straight line connecting points z_1 and z_2 belong to the domain Ω, it follows that for all $c \in (0,R)$

$$[G_c \cap \{t = T/2\}] \cap \Omega = \{x \in \Omega : |x| > c\} \neq \varnothing. \qquad \square$$

In any Carleman estimate for a differential operator, only the principal part of this operator is considered. In other words, a Carleman estimate for a differential operator is independent of its lower order terms including the integral term in (1.1), which is the advantage of our method. As for the lower order terms, they are incorporated at a later stage when either a uniqueness or stability result is proved for a corresponding Cauchy problem. Hence, we denote

$$L_0 u = u_t + \frac{d}{ds}u(x + s\nu, t, \nu)\,|_{s=0} \quad \forall \nu \in S^{N-1}. \tag{4.6}$$

Because of the insufficient smoothness guaranteed by Theorem 2.1 (Remark 2.1), it is convenient to formulate the Carleman estimate for the operator L_0 in terms of the above vector $\widetilde{\nu}_0 = (1,0,0,\dots,0)^T = A_{\nu_0}\nu_0$ in (2.6), where $\nu_0 \in S^{N-1}$ is an arbitrarily chosen unit vector.

LEMMA 4.2 (pointwise Carleman estimate). *Let $T > 2R$ and in* (4.2) *let the constant $\alpha \in [\alpha(R,T),1)$ (Lemma* 4.1*). Then for all values of the parameter $\lambda > 1$ and for all functions $u \in C^1_{tv grad}(\overline{W})$, the following pointwise Carleman estimate holds:*

$$(L_0 u)^2\varphi^2 \ge 2\lambda(1-\alpha)u^2\varphi^2 + \nabla\cdot U + V_t \quad \forall (x,t) \in G_c, \quad \forall \nu \in S^{N-1}, \tag{4.7}$$

where the vector function (U,V) can be estimated as

$$|(U,V)| \le C\lambda u^2\varphi^2 \tag{4.8}$$

and the vector function U is such that

$$\left|\int_{\partial_1 G_c}(U,n)\,dS_x dt\right| \le C\lambda\int_{\partial_1 G_c}|\cos(n,\nu)|\,u^2\varphi^2 dS_x dt \quad \forall \nu \in S^{N-1}. \tag{4.9}$$

Proof. By Lemma 4.1, $G_c \subset \Omega \times (0,T)$. Fix an arbitrary vector $\nu_0 \in S^{N-1}$. Let $A_{\nu_0} = \left(a^{ij}_{\nu_0}\right)_{ij=1}^n$ be an orthogonal matrix such that (2.6) is fulfilled. Introduce again notations (2.7) and (2.8). Then (4.2) and (4.6) imply that

$$L_0\widetilde{u} = \widetilde{u}_t + \widetilde{u}_{y_1}, \quad \psi(y,t) = |y|^2 - \alpha\left(t - \frac{T}{2}\right)^2, \quad \varphi(y,t) = \exp[\lambda\psi(y,t)].$$

Hence, the resulting domain $\widetilde{G}_c$ has the same form as the original domain G_c. Denote $v = \widetilde{u} \cdot \exp\left[\lambda\psi(y,t)\right] = \widetilde{u} \cdot \varphi$. Then

$$\widetilde{u} = v \exp\left\{\lambda\left[\alpha\left(t - \frac{T}{2}\right)^2 - |y|^2\right]\right\},$$

$$\widetilde{u}_{y_1} = \left(v_{y_1} - 2\lambda y_1 v\right) \exp\left[-\lambda\psi(y,t)\right],$$

$$\widetilde{u}_t = \left[v_t + 2\lambda\alpha\left(t - \frac{T}{2}\right) v\right] \exp\left[-\lambda\psi(y,t)\right].$$

Hence, for this vector ν_0

$$\begin{aligned}
(L_0 u)^2 \varphi^2 &= \left\{\left(v_t + v_{y_1}\right) - 2\lambda\left[y_1 - \alpha\left(t - \frac{T}{2}\right)\right] v\right\}^2 \\
&\geq -4\lambda\left[y_1 - \alpha\left(t - \frac{T}{2}\right)\right] v \left(v_t + v_{y_1}\right) \\
&= \left\{-2\lambda\left[y_1 - \alpha\left(t - \frac{T}{2}\right)\right] v^2\right\}_t - 2\lambda\alpha v^2 \\
&\quad + \left\{-2\lambda\left[y_1 - \alpha\left(t - \frac{T}{2}\right)\right] v^2\right\}_{y_1} + 2\lambda v^2 \\
&= 2\lambda\left(1 - \alpha\right)\widetilde{u}^2\varphi^2 + \nabla_y \cdot \widetilde{U} + V_t.
\end{aligned}$$

Thus,

$$(L_0\widetilde{u})^2\varphi^2 \geq 2\lambda\left(1 - \alpha\right)\widetilde{u}^2\varphi^2 + \nabla_y \cdot \widetilde{U} + \widetilde{V}_t, \tag{4.10}$$

where

$$\nabla_y \cdot \widetilde{U} = \left\{-2\lambda\left[y_1 - \alpha\left(t - \frac{T}{2}\right)\right]\widetilde{u}^2\varphi^2\right\}_{y_1} \tag{4.11}$$

and

$$\widetilde{V}_t = \left\{-2\lambda\left[y_1 - \alpha\left(t - \frac{T}{2}\right)\right]\widetilde{u}^2\varphi^2\right\}_t. \tag{4.12}$$

The backwards change of variables $y \to x$ will replace $\nabla_y \cdot \widetilde{U}$ with $\nabla_x \cdot U$ and $\widetilde{V}_t$ with V_t. Hence, (4.10) is equivalent to (4.6). It is clear from (4.11) and (4.12) that estimate (4.8) holds for the vector function (U, V).

Thus, in order to finish the proof, we now need to prove (4.9). Consider the integral

$$\int_{\widetilde{G}_c} \nabla_y \cdot \widetilde{U} dy dt.$$

Obviously,

$$\int_{\widetilde{G}_c} \nabla_y \cdot \widetilde{U} dy dt = \int_{G_c} \nabla_x \cdot U dx dt.$$

By the Gauss theorem, we have

$$\int_{\widetilde{G}_c} \nabla_y \cdot \widetilde{U} dy dt = \int_{\partial_1 \widetilde{G}_c} \left(\widetilde{U}, \widetilde{n}\right) dS_{y,t} + \int_{\partial_2 \widetilde{G}_c} \left(\widetilde{U}, \widetilde{n}\right) dS_{y,t}. \tag{4.13}$$

In order to prove (4.9), we estimate from the above the first integral on the right-hand side of (4.13). Using (4.11) and recalling again that by (2.6) $\cos(\widetilde{n}(y), y_1) = \cos(n(x), \nu_0)$, where $x = A_{\nu_0}^{-1} y$, we obtain

$$\left| \int_{\partial_1 \widetilde{G}_c} \left(\widetilde{U}, \widetilde{n}\right) dS_y dt \right| = \left| 2\lambda \int_{\partial_1 \widetilde{G}_c} \cos(\widetilde{n}, y_1) \left[y_1 - \alpha \left(t - \frac{T}{2} \right) \right] \widetilde{u}^2 \varphi^2 dS_y dt \right|$$

$$\leq C\lambda \int_{\partial_1 \widetilde{G}_c} |\cos(\widetilde{n}, y_1)| \, \widetilde{u}^2 \varphi^2 dS_y dt = C\lambda \int_{\partial_1 G_c} |\cos(n, \nu_0)| \, u^2 \varphi^2 dS_y dt.$$

On the other hand,

$$\int_{\partial_1 \widetilde{G}_c} \left(\widetilde{U}, \widetilde{n}\right) dS_y dt = \int_{\partial_1 G_c} (U, n) \, dS_x dt.$$

Hence,

$$\left| \int_{\partial_1 G_c} (U, n) \, dS_x dt \right| \leq C\lambda \int_{\partial_1 G_c} |\cos(n, \nu_0)| \, u^2 \varphi^2 dS_y dt,$$

which proves (4.9) for $\nu = \nu_0$. Since $\nu_0 \in S^{N-1}$ is an arbitrary vector and (4.10) holds, the proof is complete. □

4.2. Proof of Theorem 1.3. Since $T > 2R$, $\sqrt{5R^2 + T^2} > 3R$. Hence, we choose a number $\varepsilon = \varepsilon(\Omega)$ so small that

$$0 < \varepsilon \leq \min\left(\frac{R}{3}, \frac{\sqrt{5R^2 + T^2} - 3R}{4} \right) \tag{4.14}$$

and

$$\{|x| < 3\varepsilon\} \subset \Omega. \tag{4.15}$$

From now on we set $\alpha = [1 + \alpha(R, T)]/2$ in the function ψ in (4.2), for the sake of the definiteness, where the number $\alpha(R, T) = (2R/T)^2 \in (0, 1)$ was chosen in Lemma 4.1. Choose the number $\delta = \delta(\varepsilon) = \varepsilon/20$. Since $\varepsilon/2 + 3\delta \in (0, R)$, by Lemma 4.1 and (4.4)

$$G_{\varepsilon/2+3\delta} \neq \varnothing \text{ and } G_{\varepsilon/2+3\delta} \subset G_{\varepsilon/2+2\delta} \subset G_{\varepsilon/2+\delta} \subset G_{\varepsilon/2} \subset \Omega \times (0, T). \tag{4.16}$$

Introduce the "cut-off" function $\chi(x, t) \in C^1(\overline{\Omega} \times [0, T])$ such that $0 \leq \chi \leq 1$ and

$$\chi(x, t) = \begin{cases} 1 & \text{in } G_{\varepsilon/2+2\delta}, \\ 0 & \text{in } \{\Omega \times (0, T)\} \backslash G_{\varepsilon/2+\delta}. \end{cases} \tag{4.17}$$

Let the function $v \in C^1_{t\nu grad}\left(\overline{W}\right)$ be a solution of the adjoint transport equation (1.6). For $(x, t, \nu) \in \Gamma(T)$ let the function $q(x, t, \nu)$ be its boundary value, $v|_{x \in \partial\Omega} := q(x, t, \nu)$. Denote

$$w(x, t, \nu) = v(x, t, \nu) \chi(x, t). \tag{4.18}$$

Then

$$L_0 w = w_t + \frac{d}{ds} w\left(x+\nu s, t, \nu\right)\mid_{s=0}$$

$$= \chi\left(v_t + \frac{d}{ds} v\left(x+\nu s, t, \nu\right)\mid_{s=0}\right) + v\left(\chi_t + \sum_{i=1}^{n} \nu_i \chi_i\right).$$

Therefore, using (1.12), we obtain

$$\left| w_t + \frac{d}{ds} w\left(x+\nu s, t, \nu\right)\mid_{s=0}\right| \le M_1\left[|w| + \int_{S^{N-1}} |w(x,t,\mu)|\, d\sigma_\mu + |f|\right]$$

$$+ M_1\left(1-\chi\right)\left[|v| + \int_{S^{N-1}} |v(x,t,\mu)|\, d\sigma_\mu\right].$$

Here and below $M_1 = M_1\left(M, \Omega, T\right)$ denotes different positive constants depending only on M, Ω, and T. Square both sides of the latter equality, multiply by the function $\varphi(x,t)$, integrate over $G_{\varepsilon/2}$, and apply the Carleman estimate of Lemma 4.2 to the resulting left-hand side. Note that derivatives $\chi_t, \chi_i, i = 1, \dots, n$, are bounded and differ from zero only in the domain $G_{\varepsilon/2+\delta} \backslash G_{\varepsilon/2+2\delta}$. Hence, (4.8) implies that the corresponding vector function $(U, V) = 0$ on $\partial_2 G_{\varepsilon/2}$. Also, $V\cos(n,t) = 0$ on $\partial_1 G_{\varepsilon/2}$. Hence, the Gauss theorem and (4.9) imply that

$$\left|\int_{G_{\varepsilon/2}} \left(\nabla\cdot U + V_t\right) dxdt\right| = \left|\int_{\partial_1 G_{\varepsilon/2}} \left(U, n\right) dS_x dt\right| \le C\lambda \int_{\partial_1 G_{\varepsilon/2}} \left|\cos\left(n,\nu\right)\right| v^2 \varphi^2 dS_x dt$$

$$= C\lambda \int_{\partial_1 G_{\varepsilon/2}} \left|\cos\left(n,\nu\right)\right| q^2 \varphi^2 dS_x dt.$$

Thus, we obtain for all $\nu \in S^{N-1}$

$$2\lambda(1-\alpha)\int_{G_{\varepsilon/2}} w^2\varphi^2 dxdt \tag{4.19}$$

$$\le M_1\left[\int_{G_{\varepsilon/2}} \left(|w|^2 + \int_{S^{N-1}} w^2 d\sigma_\mu + f^2\right)\varphi^2 dxdt\right]$$

$$+ M_1 \int_{G_{\varepsilon/2}} \left(1-\chi\right) v^2\varphi^2 dxdt + M_1 \int_{G_{\varepsilon/2}} \left(1-\chi\right) \int_{S^{N-1}} v^2\left(x,t,\mu\right) d\sigma_\mu$$

$$+ C\lambda \int_{\partial_1 G_{\varepsilon/2}} \left|\cos\left(n,\nu\right)\right| q^2\varphi^2 dS_x dt.$$

For each $c \in (0, R)$ denote $H_c = G_c \times S^{N-1}$, $M_c = \partial_1 G_c \times S^{N-1}$, and $dh = dxdtd\sigma_\nu$. Integrate (4.19) with respect to $\nu \in S^{N-1}$. Noticing that

$$\int_{H_{\varepsilon/2}} \left(\int_{S^{N-1}} w^2\left(x,t,\mu\right) d\sigma_\mu\right)\varphi^2 dh = A_N \cdot \int_{H_{\varepsilon/2}} w^2\varphi^2 dh,$$

where A_N is the area of the unit sphere S^{N-1}, we obtain

(4.20)
$$2\lambda(1-\alpha)\int_{H_{\varepsilon/2}} w^2\varphi^2 dh \le M_1\left(\int_{H_{\varepsilon/2}}\left(w^2+f^2\right)\varphi^2 dh + \int_{H_{\varepsilon/2}}(1-\chi)v^2\varphi^2 dh\right) + M_1\lambda\int_{M_{\varepsilon/2}}|\cos(n,\nu)|\, q^2\varphi^2 dS_x dt d\sigma_\nu.$$

Choose $\lambda_0 = \lambda_0(C) > 1$ such that $C/(\lambda_0(1-\alpha)) < 1$. Then

$$M_1\int_{H_{\varepsilon/2}} w^2\varphi^2 dh \le \lambda(1-\alpha)\int_{H_{\varepsilon/2}} w^2\varphi^2 dh \quad \forall\lambda > \lambda_0.$$

Hence, (4.20) leads to

(4.21)
$$\lambda\int_{H_{\varepsilon/2}} w^2\varphi^2 dh \le M_1\int_{H_{\varepsilon/2}}(1-\chi)v^2\varphi^2 dh + M_1\lambda\int_{M_{\varepsilon/2}}|\cos(n,\nu)|\, q^2\varphi^2 dS_x dt d\sigma_\nu + M_1\int_{H_{\varepsilon/2}} f^2\varphi^2 dh \quad \forall\lambda > \lambda_0.$$

Estimate from the below the left-hand side of inequality (4.21). By (4.17) and (4.18) $w = v$ in $H_{\varepsilon/2+2\delta}$. Also, by (4.16) $H_{\varepsilon/2+3\delta} \subset H_{\varepsilon/2+2\delta} \subset H_{\varepsilon/2+\delta} \subset H_{\varepsilon/2}$ and by (4.3) $\varphi^2(x,t) \ge \exp\left[2\lambda\left(\varepsilon/2+3\delta\right)^2\right]$ in $H_{\varepsilon/2+3\delta}$. Hence,

(4.22)
$$\lambda\int_{H_{\varepsilon/2}} w^2\varphi^2 dh \ge \lambda\int_{H_{\varepsilon/2+3\delta}} w^2\varphi^2 dh = \lambda\int_{H_{\varepsilon/2+3\delta}} v^2\varphi^2 dh \ge \lambda\exp\left[2\lambda\left(\varepsilon/2+3\delta\right)^2\right]\int_{H_{\varepsilon/2+3\delta}} v^2 dh.$$

Estimate now the right-hand side of inequality (4.21) from the above. Since by (4.17) $1-\chi(x,t) = 0$ in $G_{\varepsilon/2+2\delta}$, we have

$$\sup_{H_{\varepsilon/2}}\left[(1-\chi)\varphi^2\right] \le \exp\left[2\lambda\left(\varepsilon/2+2\delta\right)^2\right].$$

Hence,

(4.23)
$$\int_{H_{\varepsilon/2}}\int(1-\chi)v^2\varphi^2 dh \le \exp\left[2\lambda\left(\varepsilon/2+2\delta\right)^2\right]\int_{H_{\varepsilon/2}} v^2 dh.$$

Therefore (4.21)–(4.23) imply that

(4.24)
$$\lambda\exp\left[2\lambda\left(\varepsilon/2+3\delta\right)^2\right]\int_{H_{\varepsilon/2+3\delta}} v^2 dh \le M_1\exp\left[2\lambda\left(\varepsilon/2+2\delta\right)^2\right]\cdot\int_W v^2 dh + M_1\lambda\int_\Gamma|\cos(n,\nu)|q^2\varphi^2 dS_x dt s\sigma_\nu + M_1\int_{H_{\varepsilon/2}} f^2\varphi^2 dh.$$

Let $m = \sup_{G_{\varepsilon/2}} [\psi(x,t)]$. Then (4.24) leads to

$$\lambda \exp\left[2\lambda\left(\varepsilon/2+3\delta\right)^2\right] ||v||^2_{L^2(H_{\varepsilon/2+3\delta})} \leq M_1 \exp\left[2\lambda\left(\varepsilon/2+2\delta\right)^2\right] ||v||^2_{L^2(W)} + M_1\lambda e^{2\lambda m}||q||^2_{L^2_{\cos}(\Gamma)} + M_1 e^{2\lambda m}\left\|f\right\|^2_{L^2(W)},$$

where the Hilbert space $L^2_{\cos}(\Gamma)$ is defined similarly with $L^2_{\cos}(\Gamma_-)$. Dividing this inequality by $\lambda \exp\left[2\lambda(\varepsilon/2+3\delta)^2\right]$, we obtain

(4.25)

$$||v||^2_{L^2(H_{\varepsilon/2+3\delta})} \leq M_1 \exp\left[-2\lambda\delta\left(\varepsilon+5\delta\right)\right] ||v||^2_{L^2(W)} + M_1 e^{2\lambda m}\left[||q||^2_{L^2_{\cos}(\Gamma)} + \left\|f\right\|^2_{L^2(W)}\right].$$

An inconvenience of the domain $H_{\varepsilon/2+3\delta}$ for our goal is that

$$H_{\varepsilon/2+3\delta} \cap \{t = T/2\} \subset \left(\Omega \times S^{N-1}\right), \quad \text{but} \quad \left(\Omega \times S^{N-1}\right) \diagdown H_{\varepsilon/2+3\delta} \cap \{t = T/2\} \neq \varnothing.$$

Thus, we now "shift" this domain. Choose an x_0 such that $|x_0| = 3\varepsilon/2$. By (4.15) $x_0 \in \Omega$. Consider the domain

$$G_{\varepsilon/2}(x_0) = \left\{(x,t) \in \Omega \times \mathbb{R} : |x-x_0|^2 - \alpha\left(t - \frac{T}{2}\right)^2 > \left(\frac{\varepsilon}{2}\right)^2\right\} \tag{4.26}$$

$$= \left\{(x,t) \in \Omega \times \mathbb{R} : \psi\left(x - x_0, t\right) > \left(\frac{\varepsilon}{2}\right)^2\right\},$$

which is obtained by a shift of the domain $G_{\varepsilon/2}$. We now prove that

$$G_{\varepsilon/2}(x_0) \subset \Omega \times (0,T). \tag{4.27}$$

Indeed, since by (4.1)

$$\max_{x \in \overline{\Omega}} |x - x_0| \leq |x| + |x_0| \leq R + \frac{3}{2}\varepsilon,$$

using (4.2), we obtain

$$\max_{x \in \partial\Omega} \left[\psi(x - x_0, T)\right] \leq \left(R + \frac{3}{2}\varepsilon\right)^2 - \frac{R^2}{2} - \frac{T^2}{8}.$$

It follows from (4.14) that

$$\left(R + \frac{3}{2}\varepsilon\right)^2 - \frac{R^2}{2} - \frac{T^2}{8} < \left(\frac{\varepsilon}{2}\right)^2.$$

Hence,

$$\max_{x \in \partial\Omega} \left[\psi(x - x_0, T)\right] < \left(\frac{\varepsilon}{2}\right)^2,$$

which proves (4.27). Also since $\delta = \delta(\varepsilon) = \varepsilon/20$, it follows from (4.26) that $(0, T/2) \in G_{\varepsilon/2+3\delta}(x_0) \cap \{t = T/2\}$, which proves that

$$G_{\varepsilon/2+3\delta}(x_0) \cap [\Omega \times (0,T)] \neq \varnothing.$$

Hence, the Carleman estimate of Lemma 4.2 is valid for the domain $G_{\varepsilon/2}(x_0)$. Thus, similarly to (4.25), we obtain

$$||v||^2_{L^2\left(H_{\varepsilon/2+3\delta}(x_0)\right)} \leq M_1 \exp\left[-2\lambda\delta\left(\varepsilon+5\delta\right)\right]||v||^2_{L^2(W)} + M_1 e^{2\lambda m}\left[||q||^2_{L^2_{\cos}(\Gamma)} + \|f\|^2_{L^2(W)}\right], \tag{4.28}$$

where $H_{\varepsilon/2+3\delta}\left(x_0\right) = G_{\varepsilon/2+3\delta}(x_0) \times S^{N-1}$.

It follows from (4.2)–(4.4) and (4.26) that

$$G_{\varepsilon/2+3\delta} \cap \{t = T/2\} = \left\{|x| > \frac{\varepsilon}{2} + 3\delta\right\} \cap \Omega \tag{4.29}$$

and

$$G_{\varepsilon/2+3\delta}\left(x_0\right) \cap \{t = T/2\} = \left\{|x - x_0| > \frac{\varepsilon}{2} + 3\delta\right\} \cap \Omega. \tag{4.30}$$

Consider the ball $B\left(0, \varepsilon/2 + 4\delta\right)$,

$$B\left(0, \frac{\varepsilon}{2} + 4\delta\right) = \left\{x : |x| < \frac{\varepsilon}{2} + 4\delta\right\} = \left\{x : |x| < \frac{7}{10}\varepsilon\right\} = B\left(0, \frac{7}{10}\varepsilon\right).$$

By (4.15) $B\left(0, \varepsilon/2 + 4\delta\right) \subset \Omega$. We now prove that $B\left(0, \varepsilon/2 + 4\delta\right) \subset G_{\varepsilon/2+3\delta}\left(x_0\right) \cap \{t = T/2\}$. Let $x \in B\left(0, \varepsilon/2 + 4\delta\right)$ be an arbitrary point of the ball B. Then

$$|x - x_0| \geq |x_0| - |x| = \frac{3}{2}\varepsilon - |x| > \frac{3}{2}\varepsilon - \frac{\varepsilon}{2} - 4\delta = \varepsilon - 4\delta.$$

Since $\delta = \varepsilon/20$, we have $\varepsilon - 4\delta > \varepsilon/2 + 3\delta$. Hence,

$$|x - x_0| > \varepsilon - 4\delta > \frac{\varepsilon}{2} + 3\delta \quad \forall x \in B\left(0, \frac{\varepsilon}{2} + 4\delta\right).$$

Hence, by (4.30) $B\left(0, \varepsilon/2 + 4\delta\right) \subset \left\{G_{\varepsilon/2+3\delta}(x_0) \cap \{t = T/2\}\right\}$. Hence,

$$\left\{|x| \leq \frac{\varepsilon}{2} + 3\delta\right\} \subset B\left(0, \frac{\varepsilon}{2} + 4\delta\right) \subset \left\{G_{\varepsilon/2+3\delta}(x_0) \cap \{t = T/2\}\right\}. \tag{4.31}$$

Recall that by (4.16) and (4.26)

$$\left\{G_{\varepsilon/2+3\delta} \cap \{t = T/2\}\right\} \subset \Omega, \left\{G_{\varepsilon/2+3\delta}(x_0) \cap \{t = T/2\}\right\} \subset \Omega. \tag{4.32}$$

Therefore, (4.29)–(4.32) lead to

$$\Omega = \left(G_{\varepsilon/2+3\delta} \cup G_{\varepsilon/2+3\delta}\left(x_0\right)\right) \cap \{t = T/2\}.$$

Hence, there exists a number $\eta \in (0, T/2)$ such that the layer

$$E_\eta = \left\{(x,t) : x \in \Omega, \left|t - \frac{T}{2}\right| < \eta\right\} \subset \left(G_{\varepsilon/2+3\delta} \cup G_{\varepsilon/2+3\delta}\left(x_0\right)\right).$$

Hence, estimates (4.25) and (4.28) imply that

$$||v||^2_{L^2(E_\eta \times S^{N-1})} \leq M_1 \exp\left[-2\lambda\delta\left(\varepsilon+5\delta\right)\right]||v||^2_{L^2(W)} + M_1 e^{2\lambda m}\left[||q||^2_{L^2_{\cos}(\Gamma)} + \|f\|^2_{L^2(W)}\right].$$

Hence, by the mean value theorem there exists a number $t_1 \in (T/2-\eta, T/2+\eta)$ such that

$$||v(x,t_1,\nu)||^2_{L^2(\Omega\times S^{N-1})} \leq \frac{M_1}{2\eta}\exp\left[-2\lambda\delta(\varepsilon+5\delta)\right]||v||^2_{L^2(W)} + \frac{M_1}{2\eta}e^{2\lambda m}\left[||q||^2_{L^2_{\cos}(\Gamma)} + \|f\|^2_{L^2(W)}\right].$$

That is, with a new constant M_1

$$||v(x,t_1,\nu)||^2_{L^2(\Omega\times S^{N-1})} \leq M_1\exp\left[-2\lambda\delta(\varepsilon+5\delta)\right]||v||^2_{L^2(W)} + M_1 e^{2\lambda m}\left[||q||^2_{L^2_{\cos}(\Gamma)} + \|f\|^2_{L^2(W)}\right]. \tag{4.33}$$

This inequality and the energy estimate (2.16) lead to (with a new constant M_1)

$$||v||^2_{L^2(W)} \leq M_1\exp\left[-2\lambda(\varepsilon+5\delta)\right]||v||^2_{L^2(W)} + M_1 e^{2\lambda m}\left[||q||^2_{L^2_{\cos}(\Gamma)} + \|f\|^2_{L^2(W)}\right]. \tag{4.34}$$

Choose $\lambda \geq \lambda_0$ such that $C\exp\left[-2\lambda\delta(\varepsilon+5\delta)\right] < 1/2$. Then (4.34) implies that for this λ

$$\|v\|_{L^2(W)} \leq M_1\left[||q||_{L^2_{\cos}(\Gamma)} + \|f\|_{L^2(W)}\right]. \tag{4.35}$$

Using (4.33) and (4.35), we obtain that

$$\|v(x,t_1,\nu)\|_{L^2(\Omega\times S^{N-1})} \leq M_1\left[||q||_{L^2_{\cos}(\Gamma)} + \|f\|_{L^2(W)}\right].$$

Hence, using (2.16), we obtain that

$$\|v(x,t_0,\nu)\|_{L^2(\Omega\times S^{N-1})} \leq M_1\left[||q||_{L^2_{\cos}(\Gamma)} + \|f\|_{L^2(W)}\right] \quad \forall t_0 \in [0,T]. \tag{4.36}$$

□

4.3. Proof of Theorem 1.2. Estimates (4.35) and (4.36) are valid for any strong solution $v \in C^1_{tv grad}\left(\overline{W}\right)$ of the adjoint transport equation (1.6) with the boundary condition $v\,|_{x\in\partial\Omega} = q(x,t,\nu)$. Clearly, these estimates are also valid for any strong solution $u \in C^1_{tv grad}\left(\overline{W}\right)$ of the original equation (1.1) with the boundary condition $u\,|_{x\in\partial\Omega} = q(x,t,\nu)$. Recall now that Theorem 1.2 is concerned with the weak solution of the problem (1.6)–(1.8). Hence, by (1.8) we need to assume that $v\,|_{\Gamma_+} = 0$ and $f = 0$. Recalling that $v\,|_{\Gamma_-} := (Kv_0)(x,t,\nu)$, we see that (4.35) and (4.36) lead to (1.10) and (1.11), respectively. Thus, Theorem 1.2 is valid for strong solutions $v \in C^1_{tv grad}(\overline{W})$ of (1.6) with the boundary condition (1.8).

Consider now an arbitrary function $v_0 \in L^2\left(\Omega\times S^{N-1}\right)$, and let the function $v \in L^2(W)$ be the weak solution of the problem (1.6)–(1.8). Let $\{v_{0k}\}_{k=1}^{\infty}$ be a sequence of functions satisfying conditions (3.1), (3.2) and such that

$$\lim_{k\to\infty}\|v_0 - v_{0k}\|_{L^2(\Omega\times S^{N-1})} = 0. \tag{4.37}$$

Let $\{v_k\}_{k=1}^{\infty} \subset C^1_{tv grad}\left(\overline{W}\right)$ be the corresponding sequence of solutions of the problem (1.6)–(1.8) with the initial condition $v_k\,|_{t=T} = v_{0k}$. Then by Theorem 2.4

$$\lim_{k\to\infty}\|v - v_k\|_{L^2(W)} = 0.$$

In addition, functions $p_k := v_k|_{\Gamma_-} := Kv_{0k} \in L^2_{\cos}(\Gamma_-)$, and by Definition 3.1 of the generalized trace of the weak solution, we see that

$$\lim_{k\to\infty} \|p - p_k\|_{L^2\cos(\Gamma_-)} = 0. \tag{4.38}$$

Replacing v_0 with v_{0k} in estimates (1.10) and (1.11), we see that

$$\|v_k\|_{L^2(W)} \le C \|Kv_{0k}\|_{L^2_{\cos}(\Gamma_-)} \tag{4.39}$$

and

$$\|v_{0k}\|_{L^2(\Omega\times S^{N-1})} \le C \|Kv_{0k}\|_{L^2_{\cos}(\Gamma_-)}. \tag{4.40}$$

Since $K : L^2(\Omega \times S^{N-1}) \to L^2_{\cos}(\Gamma_-)$ is a bounded linear operator (see section 3 after (3.19)), the passage to limits in (4.37)–(4.40) yields (1.10) and (1.11) for the weak solution v. □

Acknowledgment. The authors thank the anonymous referees for their invaluable comments.

REFERENCES

[1] C. Bardos, *Problèmes aux limites pour les équations aux dérivées partielles du premier ordre à coefficients réels; théorèmes d'approximation; applications à l'équations de tranport*, Ann. Sci. École Norm. Sup. (4), 3 (1970), pp. 185–233.

[2] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[3] K. M. Case and P. F. Zweifel, *Linear Transport Theory*, Addison-Wesley, Reading, MA, 1967.

[4] B. B. Das, F. Liu, and R. R. Alfano, *Time-resolved fluorescence and photon migration studies in biomedical and model random media*, Rep. Progr. Phys., 60 (1997), pp. 227–292.

[5] R. Dautray and J. L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 6, Springer-Verlag, Berlin, 1993.

[6] A. Douglis, *The solutions of multidimensional generalized transport equations and their calculation by difference methods*, in Numerical Solution of Partial Differential Equations, Academic Press, New York, 1966, pp. 197–256.

[7] N. Dunford and J. T. Schwartz, *Linear Operators. Part* I. *General Theory*, John Wiley & Sons, New York, 1964.

[8] M. Eller and J. E. Masters, *Exact controllability of electromagnetic fields in a general region*, Appl. Math. Optim., 45 (2002), pp. 99–123.

[9] A. V. Fursikov, *Optimal Control of Distributed Systems*, AMS, Providence, RI, 2000.

[10] A. V. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Seoul National University, Seoul, 1996.

[11] R. Gulliver, I. Lasiecka, W. Littman, and R. Triggiani, *The case for differential geometry in the control of single and coupled PDEs: The structural acoustic chamber*, in IMA Vol. Math. Appl. 137, Springer-Verlag, Berlin, 2004, pp. 73–181.

[12] L. F. Ho, *Observabilité frontière de l'équation des ondes*, C.R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 443–446.

[13] L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1963.

[14] O. Yu. Imanuvilov, *Boundary controllability of parabolic equations*, Sbornik Math., 186 (1995), pp. 879–900.

[15] O. Yu. Imanuvilov and M. Yamamoto, *Carleman inequalities for parabolic equations in Sobolev spaces of negative order and exact controllability for semilinear parabolic equations*, Publ. Res. Inst. Math. Sci., 39 (2003), pp. 227–274.

[16] A. Ishimaru, *Wave Propagation and Scattering in Random Media*, Academic Press, New York, 1978.

[17] M. Kazemi and M. V. Klibanov, *Stability estimates for ill-posed Cauchy problem involving hyperbolic equations and inequalities*, Appl. Anal., 50 (1993), pp. 93–102.

[18] M. V. Klibanov and J. Malinsky, *Newton-Kantorovich method for 3-dimensional potential inverse scattering problem and stability of the hyperbolic Cauchy problem with time dependent data*, Inverse Problems, 7 (1991), pp. 577–595.

[19] M. V. Klibanov and A. A. Timonov, *Carleman Estimates for Coefficient Inverse Problems and Numerical Applications*, VSP, Utrecht, The Netherlands, 2004.

[20] M. V. Klibanov and S. E. Pamyatnykh, *Lipschitz stability of a non-standard problem for the non-stationary transport equation via Carleman estimate*, Inverse Problems, 22 (2006), pp. 881–890.

[21] M. V. Klibanov and M. Yamamoto, *Exact Controllability for the Non-Stationary Transport Equation*, Preprint 06-37, http://www.ma.utexas.edu/mp_arc/index-06.html#end (23 February 2006).

[22] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, John Wiley & Sons, Chichester, UK, 1994.

[23] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics. Volume* 10. *Physical Kinetics*, Pergamon Press, New York, 1981.

[24] E. W. Larsen and J. B. Keller, *Asymptotic solution of neutron transport problems for small mean free paths*, J. Math. Phys., 15 (1974), pp. 75–81.

[25] I. Lasiecka and R. Triggiani, *Exact controllability of the wave equation with Neumann boundary control*, Appl. Math. Optim., 19 (1989), pp. 243–290.

[26] M. M. Lavrent'ev, V. G. Romanov, and S. P. Shishatskiĭ, *Ill-Posed Problems of Mathematical Physics and Analysis*, AMS, Providence, RI, 1986.

[27] J. L. Lions, *Contrôlabilité exacte des systèmes distribués*, C. R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 471–475.

[28] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[29] J. L. Lions, *Contrôlabilité Exacte Perturbations et Stabilisations de Systèmes Distribués*, Masson, Paris, 1988.

[30] J. L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, 1972.

[31] A. I. Prilepko and A. L. Ivankov, *Inverse problems for the determination of a coefficient and the right side of a nonstationary multivelocity transport equation with overdetermination at a point*, Differ. Equ., 21 (1985), pp. 88–96.

[32] D. I. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–211.

[33] T. J. Seidman, *Observation and prediction for the heat equation*, III, J. Differential Equations, 20 (1976), pp. 18–27.

[34] R. Triggiani, *Exact boundary controllability on* $L_2(\Omega) \times H^{-1}(\Omega)$ *of the wave equation with Dirichlet boundary control acting on a portion of the boundary*, Appl. Math. Optim., 18 (1988), pp. 241–277.

[35] S. Ukai, *Transport Equations*, Sangyo-tosyo, Tokyo, 1976 (in Japanese).

[36] S. Ukai, *Solutions of Boltzmann equations*, in Patterns and Waves, North-Holland, Amsterdam, 1986, pp. 37–96.

[37] G. Wang, Y. Li, and M. Jiang, *Uniqueness theorems in bioluminescence tomography*, Med. Phys., 31 (2004), pp. 2289–2299; Erratum: Med. Phys., 32 (2005), p. 3059.

[38] E. Zuazua, *Contrôlabilité exacte et un temps arbitrairement petit de quelques modèles de plaques*, in Contrôlabilité Exacte Perturbations et Stabilisations de Systèmes Distribués, Masson, Paris, 1988, pp. 465–491.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2096–2119

THE MULTI-AGENT RENDEZVOUS PROBLEM. PART 1: THE SYNCHRONOUS CASE*

J. LIN†, A. S. MORSE‡, AND B. D. O. ANDERSON§

Abstract. This paper is concerned with the collective behavior of a group of $n > 1$ mobile autonomous agents, labelled 1 through n, which can all move in the plane. Each agent is able to continuously track the positions of all other agents currently within its "sensing region," where by an agent's *sensing region* we mean a closed disk of positive radius r centered at the agent's current position. The *multi-agent rendezvous problem* is to devise "local" control strategies, one for each agent, which without any active communication between agents cause all members of the group to eventually rendezvous at a single unspecified location. This paper describes a solution to this problem consisting of individual agent strategies which are mutually synchronized in the sense that all depend on a common clock.

Key words. cooperative control, distributed control, multi-agent systems

AMS subject classifications. 93C65, 93C85, 93C55

DOI. 10.1137/040620552

1. Introduction.

Current interest in cooperative control has led to the development of a number of distributed control algorithms capable of causing large groups of mobile autonomous agents to perform useful tasks [11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. Of particular interest here are provably correct algorithms which solve what we shall refer to as the "multi-agent rendezvous problem." This problem, which was considered previously in [19, 1], is concerned with the collective behavior of a group of $n > 1$ mobile autonomous agents, labelled 1 through n, which can all move in the plane. Each agent is able to continuously track the positions of all other agents currently within its "sensing region," where by an agent's *sensing region* we mean a closed disk of positive radius r centered at the agent's current position. The *multi-agent rendezvous problem* is to devise "local" control strategies, one for each agent, which without any active communication between agents cause all members of the group to eventually rendezvous at a single unspecified location.

In this paper, as in [1], we consider distributed strategies which guide each agent toward rendezvous by performing a sequence of "stop-and-go" maneuvers. A *stop-and-go maneuver* takes place within a time interval consisting of two consecutive subintervals. The first, called a *sensing period*, is an interval of fixed length during which the agent is stationary. The second, called a *maneuvering period*, is an interval of variable length during which the agent moves from its current position to its next

*Received by the editors December 9, 2004; accepted for publication (in revised form) May 29, 2007; published electronically December 12, 2007. A preliminary version of this paper appeared in [11]. The research of the first two authors was supported by the US Army Research Office, the US National Science Foundation, and by a gift from Xerox Corporation.

http://www.siam.org/journals/sicon/46-6/62055.html

†Xerox Corporation, 800 Phillips Road, MS:0128-30E, Webster, NY 14580. Current address: 21/F Central Plaza, 227 Huangpi Bei Lu, Shanghai 200003, People's Republic of China (jie.lin@aya.yale.edu).

‡Yale University, PO Box 208267, New Haven, CT 06520 (morse@sysc.eng.yale.edu).

§Australian National University & National ICT Australia Ltd, Locked bag 8001, Canberra ACT 2601, Australia (Brian.Anderson@nicta.com.au). This author's research was supported by National ICT Australia, which is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through the Backing Australia's Ability initiative and the ICT Centre of Excellence Program.

"way-point" and again comes to rest. Successive way-points for each agent are chosen to be within r_M units of each other, where r_M is a prespecified positive distance no larger than r. It is assumed that there has been chosen for each agent i a positive number τ_{M_i}, called a *maneuver time*, which is large enough so that the required maneuver for agent i from any one way-point to the next can be accomplished in at most τ_{M_i} seconds. Since our interest here is exclusively with devising *high level* strategies which dictate when and where agents are to move, we will use point models for agents and shall not deal with how maneuvers are actually carried out or with how vehicle collisions are to be avoided.

In this paper we describe a family of stop-and-go strategies which solves the problem. The family includes the specific strategies proposed in [1] and consists of agent strategies which are mutually synchronized in the sense that all depend on a common clock. In a sequel to this paper [12] we propose and analyze families of strategies which also solve the problem, but without the need for synchronization.

In the synchronous case treated here, the kth maneuvering periods of all n agents begin at the same time $\bar{t}_k$. The kth way-point of each agent is a function of the positions of its "registered neighbors" at time $\bar{t}_k$. Agent i's registered neighbors at time $\bar{t}_k$ are all those other agents positioned within its sensing region at time $\bar{t}_k$. This notion of a neighbor induces a *symmetric* relation on the agent group since agent j is a registered neighbor of agent i at time $\bar{t}_k$ just in case agent i is a registered neighbor of agent j at the same time. Because of this it is possible to characterize neighbor relationships at time $\bar{t}_k$ with a simple graph whose vertices represent agents and whose edges represent existing neighbor relationships (see section 2.2). Although the neighbor relation is symmetric, it is clearly not transitive. On the other hand, if agent i is at the same position as neighbor j at time $\bar{t}_k$, then any registered neighbor of agent j at time $\bar{t}_k$ must certainly be a registered neighbor of agent i at the same time. It is precisely because of this *weak transitivity* property that one can infer a *global* condition of the entire agent group from a *local* condition of one agent and its neighbors. In particular, if the graph characterizing neighbor relationships at time $\bar{t}_k$ is connected, and any one agent is at the same position as all of its neighbors, then the weak transitivity property guarantees at once that all n agents have rendezvoused at time $\bar{t}_k$.

One way to ensure that a neighbor graph is connected at time $\bar{t}_k$, assuming it is connected when the rendezvousing process begins, is to constrain each agent's way-points to be positioned in such a way so that no agent can lose any of its registered neighbors when it moves from one way-point to the next. This can be accomplished using a clever idea taken from [1]. An immediate consequence is that each agent's set of registered neighbors is nondecreasing and, because of this, ultimately converges to a fixed neighbor set for $\bar{t}_k$ sufficiently large.

A second local constraint is to require the way-point of each agent i at the beginning of its kth maneuvering period to lie in the "local" convex hull $\mathcal{H}_i(k)$ of agent i's own position at time $\bar{t}_k$ and the sensed positions of its registered neighbors at the same time. It is quite easy to prove that doing this causes the global convex hull $\mathcal{H}(k+1)$ of all n agent positions at time $\bar{t}_{k+1}$ to be contained in the corresponding global convex hull $\mathcal{H}(k)$ at time $\bar{t}_k$.

A third constraint is to stipulate that for each i, the only condition under which agent i's kth way-point can be positioned at a corner of $\mathcal{H}_i(k)$ is when $\mathcal{H}_i(k)$ is a single point. The global implication of doing this is that the diameter of $\mathcal{H}(k+1)$ must either be strictly smaller than the diameter of $\mathcal{H}(k)$ or every agent must be at

the same position as all of its registered neighbors at time $\bar{t}_k$—and this is true whether or not the graph characterizing neighbor relationships at time $\bar{t}_k$ is connected.

In section 4, a more or less standard Lyapunov-based argument is used to prove that if the preceding constraints are adopted by all agents and if the graph characterizing initial neighbor positions is connected, then all n agents will eventually rendezvous at a single point. Not surprisingly, the Lyapunov function used for this purpose is the diameter of the global convex hull. However, although connectivity of the graph characterizing initial neighbor positions is sufficient for rendezvousing, it is not necessary. An example illustrating this is given in section 3.2. The example deals with the situation when the initial neighbor graph consists of two connected components, with one "encircling" the other in a suitably defined sense.

2. The synchronous agent system. In the synchronous case treated in this paper, the maneuvering times for all agents are all the same positive value τ_M. Along any trajectory of the system to be considered, the real time axis can be partitioned into a sequence of consecutive time intervals $[0, t_1), [t_1, t_2), \ldots [t_{k-1}, t_k), \ldots$, each of length at least τ_M. Each interval consists of a sensing period followed by a maneuvering period of fixed length τ_M. All agents function in synchronization in the sense that all are at rest during sensing periods and all can maneuver only during maneuvering periods. In particular, all agents actions are synchronized to the time sequence $\bar{t}_1, \bar{t}_2, \ldots \bar{t}_k \ldots$, where $\bar{t}_k$ denotes the real time $t_k - \tau_M$ at which the kth maneuvering period begins. Agent i's *registered neighbors* at the beginning of its kth maneuvering period, $[\bar{t}_k, t_k)$, are those agents, except for agent i, which are within agent i's sensing region at time $\bar{t}_k$. Note that this definition is a symmetric relation on the set of all agents; i.e., if agent i is a registered neighbor of agent j at the beginning of maneuvering period k, then agent j is a registered neighbor of agent i at the beginning of the same maneuvering period.

2.1. Pairwise motion constraint. A pair of agents which are registered neighbors at the beginning of maneuvering period k are said to satisfy the *pairwise motion constraint* during the period if the positions to which they move at time t_k are both within a closed disk of diameter r centered at the mean of their registered positions at time $\bar{t}_k$. The definition implies that any two agents which are registered neighbors at the beginning of maneuvering period k will be registered neighbors at the beginning of maneuvering period $k+1$ if they satisfy the pairwise motion constraint during the kth maneuvering period. We are interested in strategies possessing this property and accordingly make the following assumption.

Cooperation assumption. During each maneuvering period k, each pair of agents which are registered neighbors at the beginning of the period restrict their motions to satisfy the pairwise motion constraint.

Agent i's *kth way-point* is the point to which agent i is to move at time t_k. Thus if $x_i(t)$ denotes the position of agent i at time t represented in a world coordinate system, then $x_i(t_k)$ and agent i's kth way-point are one and the same. The rule which determines each such way-point is a function depending only on the number and relative positions of agent i's registered neighbors. In particular, if agent i has m_i registered neighbors at time $\bar{t}_k$, positioned relative to agent i at points

$$z_j \stackrel{\Delta}{=} x_{i_j}(\bar{t}_k) - x_i(\bar{t}_k), \quad j \in \{1, 2, \ldots, m_i\}, \tag{1}$$

then agent i's kth way-point is

$$x_i(t_{k-1}) + u_{m_i}(z_1, z_2, \ldots, z_{m_i}), \tag{2}$$

where $u_0 = 0$, $u_m : \mathbb{D}^m \to \mathbb{D}_M$, $m \in \{1, \dots, n-1\}$, and $\mathbb{D}$ and $\mathbb{D}_M$ are the closed disks of radii r and r_M, respectively, centered at the origin in $\mathbb{R}^2$. In other words, if agent i has no registered neighbors at time $\bar{t}_k$ (i.e., $m_i = 0$), it does not move during the kth maneuvering period. On the other hand, if agent i has $m_i > 0$ neighbors at time $\bar{t}_k$ with relative positions $z_1, z_2, \dots, z_{m_i}$, then agent i moves to the position $x_i(t_{k-1}) + u_{m_i}(z_1, z_2, \dots, z_{m_i})$ at time t_k. Thus

$$x_i(t_k) = x_i(t_{k-1}) + u_{m_i(\bar{t}_k)}(x_{i_1}(\bar{t}_k) - x_i(\bar{t}_k),\ x_{i_2}(\bar{t}_k) - x_i(\bar{t}_k), \dots,\ x_{i_{m_i(\bar{t}_k)}}(\bar{t}_k) - x_i(\bar{t}_k)). \tag{3}$$

In what follows we will explain how the u_m are defined. At the very least we will require each to be a continuous function.

2.2. Definition of u_m. We have already defined $u_0 = 0$. To define u_m for $m > 0$ it is necessary to take into account the pairwise motion constraint. Toward this end, for each $z \in \mathbb{D}$, let $\mathcal{C}(z)$ denote the closed disk of diameter r centered at the point $\frac{1}{2}z$. More generally, for each $\{z_1, z_2, \dots, z_m\} \in \mathbb{D}^m$, let

$$\mathcal{C}(z_1, z_2, \dots, z_m) = \bigcap_{j=1}^{m} \mathcal{C}(z_j). \tag{4}$$

Note that 0 is in each $\mathcal{C}(z_i)$ and, moreover, that each such $\mathcal{C}(z_i)$ is closed and strictly convex. Consequently $\mathcal{C}(z_1, z_2, \dots, z_m)$ is either the singleton $\{0\}$ or a strictly convex, closed set containing 0. We can now define u_m to be any continuous function on $\mathbb{D}^m$ satisfying

$$u_m(z_1, z_2, \dots, z_m) \in \mathbb{D}_M \cap \mathcal{C}(z_1, z_2, \dots, z_m) \cap \langle 0, z_1, z_2, \dots, z_m \rangle \ \forall \{z_1, z_2, \dots, z_m\} \in \mathbb{D}^m, \tag{5}$$

where $\langle 0, z_1, z_2, \dots, z_m \rangle$ is the convex hull of the points $0, z_1, z_2, \dots, z_m$. The u_m are further required to have the property that

$$u_m(z_1, z_2, \dots, z_m) \neq \text{a corner}^1 \text{ of } \langle 0, z_1, z_2, \dots, z_m \rangle \tag{6}$$

unless $z_1 = z_2 = \cdots = z_m = 0$. In other words, u_m is required to be (i) a continuous function on $\mathbb{D}^m$ which maps each $\{z_1, z_2, \dots, z_m\} \in \mathbb{D}^m$ into $\mathbb{D}_M \cap \mathcal{C}(z_1, z_2, \dots, z_m) \cap \langle 0, z_1, z_2, \dots, z_m \rangle$ and (ii) a function with the property that $u_m(z_1, z_2, \dots, z_m)$ is not a corner of $\langle 0, z_1, z_2, \dots, z_m \rangle$ unless $z_1 = z_2 = \cdots = z_m = 0$. Examples of functions satisfying these conditions will be given in what follows.

2.3. Target points. One way to go about defining specific u_m which are continuous and which satisfy (5) and (6) is by first defining what we shall refer to as a "target point." By a *target point* we mean a continuous function $\tau : \mathbb{D}^m \to \langle 0, z_1, z_2, \dots, z_m \rangle$ defined in such a way that for each $\{z_1, z_2, \dots, z_m\} \in \mathbb{D}^m$ for which 0 is a corner of $\langle 0, z_1, z_2, \dots, z_m \rangle$, the segment of the line from 0 to $\tau(z_1, z_2, \dots, z_m)$ which lies within $\mathcal{C}(z_1, z_2, \dots, z_m)$ has positive length. For should it be possible to define such a τ, one could satisfy (5) and (6) as well as the continuity requirement with a control of the form

$$u_m = g(z_1, z_2, \dots, z_m)\tau(z_1, z_2, \dots, z_m),$$

[1]Recall that a point x in a polytope $\mathbb{P}$ in $\mathbb{R}^m$ is a *corner* if the only points y and z in $\mathbb{P}$ for which x is a convex combination are $y = z = x$.

where $g : \mathbb{D}^m \to \mathbb{R}$ is any continuous, positive definite function satisfying

$$g < \max_{(0,1]} \left\{ \mu : \mu\tau \in \mathbb{D}_M \bigcap \mathcal{C}(z_1, z_2, \ldots, z_m) \right\}.$$

Note that $g\tau \in \langle 0, z_1, z_2, \ldots, z_m \rangle$ for all $g \in [0,1]$ because $0 \in \langle 0, z_1, z_2, \ldots, z_m \rangle$. The role of g is therefore to scale down the magnitude of τ enough to ensure that $g\tau$ is in the constraint set $\mathbb{D}_M \bigcap \mathcal{C}(z_1, z_2, \ldots, z_m)$.

It might be thought that one could choose for τ the centroid of $\langle 0, z_1, z_2, \ldots, z_m \rangle$ or perhaps the average of the z_i and 0, namely

$$\tau \triangleq \frac{1}{m+1} \sum_{i=1}^{m} z_i.$$

Both candidate definitions satisfy the requirement that $\tau(z_1, z_2, \ldots, z_m)$ must be a point in $\langle 0, z_1, z_2, \ldots, z_m \rangle$. Unfortunately, simple examples show that the centroid definition does not necessarily yield a function which satisfies the continuity requirement, while the averaging definition may lead to a function which fails to satisfy the requirement that when 0 is a corner of $\langle 0, z_1, z_2, \ldots, z_m \rangle$, the segment of the line from 0 to $\tau(z_1, z_2, \ldots, z_m)$ which lies within $\mathcal{C}(z_1, z_2, \ldots, z_m)$ has positive length. For example, the centroid of the convex hull of the points $(0,0)$, $z_1 = (0,1)$, and $z_2 = (p,1)$ is at $(\frac{p}{3}, \frac{2}{3})$ for $p > 0$ and at $(0, \frac{1}{2})$ for $p = 0$ so the centroid is discontinuous at $p = 0$. As a counterexample to the use of coordinate averaging to define a target point, note that the average of the four points located at $(0,0)$, $z_1 = (-r, 0)$, $z_2 = (\frac{2r}{3}, \frac{r}{2})$, and $z_3 = (\frac{r}{3}, \frac{r}{2})$ is at $(0, \frac{r}{4})$, while the constraint set $\mathcal{C}(z_1, z_2, z_3)$ determined by these points must be contained in the constraint disk $\mathcal{C}(z_1)$. Since the line $\mathcal{L}$ from $(0,0)$ to $(0, \frac{r}{4})$ is tangent to this disk at the origin, the intersection of $\mathcal{L}$ with $\mathcal{C}(z_1, z_2, z_3)$ is just the point $(0,0)$ and consequently not a line segment of positive length.

In what follows we shall approach the problem of defining τ in a slightly different way. We begin by stating the following proposition which provides a simple condition on $\tau(\cdot)$, which, if satisfied, automatically implies satisfaction of the requirement that when 0 is a corner of $\langle 0, z_1, z_2, \ldots, z_m \rangle$, the segment of the line from 0 to $\tau(z_1, z_2, \ldots, z_m)$ which lies within $\mathcal{C}(z_1, z_2, \ldots, z_m)$ has positive length.

PROPOSITION 1. *Let $z_1, z_2, \ldots, z_m$ be a set of $m > 0$ points in $\mathbb{D}$ which are not all 0. If 0 is a corner of $\langle 0, z_1, z_2, \ldots, z_m \rangle$ and z is any nonzero point in $\mathbb{D}$ within r units of each point in $\{z_1, z_2, \ldots, z_m\}$, then the segment of the line from 0 to z which lies in $\mathcal{C}(z_1, z_2, \ldots, z_m)$ has positive length.*

The proofs of this and subsequent propositions and lemmas are in section 6.

Proposition 1 suggests the following approach for defining a target point. First, for each $z \in \mathbb{D}$, let $\mathcal{D}(z)$ denote a closed disk of radius r centered at z. More generally, for any set of $m > 0$ points $z_1, z_2, \ldots, z_m$ in $\mathbb{D}$, write

$$\mathcal{D}(z_1, z_2, \ldots, z_m) = \bigcap_{i=1}^{m} \mathcal{D}(z_i).$$

By construction, each point in $\mathcal{D}(z_1, z_2, \ldots, z_m)$ is within r units of each point in $\{z_1, z_2, \ldots, z_m\}$ and conversely. Thus $0 \in \mathcal{D}(z_1, z_2, \ldots, z_m)$ because $z_i \in \mathbb{D}$, $i \in \{1, 2, \ldots, m\}$.

Second, note that if $z_1, z_2, \ldots, z_m$ is any set of $m > 0$ points in $\mathbb{D}$ which are not all zero and for which 0 is a corner of $\langle 0, z_1, z_2, \ldots, z_m \rangle$, then by Proposition 1 the segment of the line from 0 to any nonzero point $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$ which lies in

$\mathcal{C}(z_1, z_2, \ldots, z_m)$ must have positive length. It follows that any continuous function $\tau : \mathbb{D}^m \to \langle 0, z_1, z_2, \ldots, z_m \rangle$ which satisfies

$$\tau(z_1, z_2, \ldots, z_m) \in \mathbb{D} \bigcap \mathcal{D}(z_1, z_2, \ldots, z_m) \bigcap \langle 0, z_1, z_2, \ldots, z_m \rangle$$

and which is nonzero whenever 0 is a corner of $\langle 0, z_1, z_2, \ldots, z_m \rangle$ and $z_1, z_2, \ldots, z_m$ are not all zero fulfills all the conditions required to be a target point. In what follows we will show that there are at least two different ways to so define τ.

2.3.1. The centroid of $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$. In order for the centroid of $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$ to be a target point, it must depend continuously on the z_i and, in addition, must have the property that it is nonzero for any set of m points in $\mathbb{D}$ which are not all zero and for which 0 is a corner of $\langle 0, z_1, z_2, \ldots, z_m \rangle$. These properties are guaranteed by the following two propositions.

PROPOSITION 2. *Let $z_1, z_2, \ldots, z_m$ be a set of $m > 0$ points in $\mathbb{D}$ which are not all 0. Then the centroid of $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$ is in $\langle 0, z_1, z_2, \ldots, z_m \rangle$. If, in addition, 0 is a corner of $\langle 0, z_1, z_2, \ldots, z_m \rangle$, then $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$ has a nonempty interior, and the centroid of $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$ cannot be at 0.*

PROPOSITION 3. *The function which assigns to each set of $m > 0$ points $z_1, z_2, \ldots, z_m$ in $\mathbb{D}$ the centroid of $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$ is continuous.*

Examination of the proof of Proposition 3 given in section 6 reveals that the continuity of the centroid of $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$ depends crucially on the fact that the centroid is at 0 whenever the area of $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$ is zero. This property is not shared by the centroid of $\langle 0, z_1, z_2, \ldots, z_m \rangle$, and it is for this reason that the centroid of $\langle 0, z_1, z_2, \ldots, z_m \rangle$ is not a continuous function of the z_i.

It turns out that Propositions 2 and 3 both hold if the set $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$ is replaced throughout by the constraint set $\mathbb{D} \cap \mathcal{C}(z_1, z_2, \ldots, z_m)$. This can be shown using essentially the same proofs of the propositions as those given in the appendix. What this means then is that the centroid of $\mathbb{D} \cap \mathcal{C}(z_1, z_2, \ldots, z_m)$ is also a valid target point.

2.3.2. The center of the smallest circle containing $\langle 0, z_1, z_2, \ldots, z_m \rangle$. It is also possible to define τ to be the center of the smallest circle containing $\langle 0, z_1, z_2, \ldots, z_m \rangle$. To understand why this is so, let us note first that for any set of points $z_i \in \mathbb{D}$, $i \in \{1, 2, \ldots, m\}$, the set of points $\mathcal{Q} \triangleq \{0, z_1, \ldots, z_m\}$ is contained in a circle of radius r centered at 0. It follows that the center of this circle is at most r units from every point in $\mathcal{Q}$. This suggests that one might choose for $\tau(z_1, z_2, \ldots, z_m)$ the center $\tau_C(z_1, z_2, \ldots z_m)$ of the smallest circle containing $\mathcal{Q}$ or, equivalently, $\langle 0, z_1, z_2, \ldots, z_m \rangle$, since $\tau_C(z_1, z_2, \ldots, z_m)$ would have to be within r units of every point in $\mathcal{Q}$. It is known that there is such a smallest circle [17] and that if the z_i are not all zero, $\tau_C(z_1, z_2, \ldots z_m)$ is either the midpoint between two of the points in $\mathcal{Q}$ or a point within the interior of a triangle formed from at least one set of three points in $\mathcal{Q}$ [1]. In either case it is clear that $\tau_C(z_1, z_2, \ldots z_m) \in \langle 0, z_1, z_2, \ldots, z_m \rangle$ and, if the z_i are not all zero and 0 is a corner of $\langle 0, z_1, z_2, \ldots, z_m \rangle$, that $\tau_C(z_1, z_2, \ldots z_m)$ is nonzero as well. Furthermore it can be shown that $\tau_C(z_1, z_2, \ldots z_m)$ depends continuously on the z_i [18]. In other words, $\tau_C(z_1, z_2, \ldots z_m)$ satisfies all the conditions required to be a target point. This elegant choice for τ is the one proposed in [1].

3. Main results. Define $t_0 = 0$. Note that because agents do not move during sensing periods, for $k \geq 1$ the position of each agent at time t_{k-1} is the same as its

position at time $\bar{t}_k$. Thus (3) can be rewritten as

$$
\begin{aligned}
x_i(t_k) = x_i(t_{k-1}) & \\
& + u_{m_i(t_{k-1})}(x_{i_1}(t_{k-1}) - x_i(t_{k-1}),\ x_{i_2}(t_{k-1}) \\
& - x_i(t_{k-1}), \ldots,\ x_{i_{m_i(t_{k-1})}}(t_{k-1}) - x_i(t_{k-1})),
\end{aligned} \tag{7}
$$

where $m_i(t_{k-1}) \triangleq m_i(\bar{t}_k)$. Because of this, the system just defined admits the model of a nonlinear discrete-time system with state $x(t_k) = \text{column } \{x_1(t_k), x_2(t_k), \ldots x_n(t_k)\}$ evolving on the time set $t_0, t_1, \ldots t_k, \ldots$. Analysis of this system depends on the relationships between neighbors and how they evolve with time. These relationships can be conveniently described by a simple, undirected graph with vertex set $\{1, 2, \ldots, n\}$ which is defined so that (i, j) is one of the graph's edges just in case agents i and j are registered neighbors at the beginning of maneuvering period k. Since these relationships can change from one maneuvering period to the next, so can the graph which describes them. In what follows we use the symbol $\mathcal{P}$ to denote a suitably defined set, indexing the class of all simple graphs $\mathbb{G}_p$ on n vertices. Let us partially order the set $\{\mathbb{G}_p : p \in \mathcal{P}\}$ by agreeing to say that $\mathbb{G}_p$ is contained in $\mathbb{G}_q$ if the edge set of $\mathbb{G}_p$ is a subset on the edge set of $\mathbb{G}_q$. It is natural then to define the *union* of a collection of such graphs, $\{\mathbb{G}_{p_1}, \mathbb{G}_{p_2}, \ldots, \mathbb{G}_{p_m}\}$, to be the simple graph $\mathbb{G}$ with vertex set $\{1, 2, \ldots, n\}$ and edge set equaling the union of the edge sets of all of the graphs in the collection.

Let $\sigma(k)$ denote the index of the graph in $\{\mathbb{G}_p : p \in \mathcal{P}\}$ which describes the relationship between registered neighbors at the beginning of maneuvering period k. Because of the cooperation assumption, we know that each agent keeps all of its registered neighbors as the system evolves. What this means is the sequence of graphs $\mathbb{G}_{\sigma(1)}, \mathbb{G}_{\sigma(2)}, \ldots, \mathbb{G}_{\sigma(k)}, \ldots$ forms the ascending chain

$$
\mathbb{G}_{\sigma(1)} \subset \mathbb{G}_{\sigma(2)} \subset \cdots \subset \mathbb{G}_{\sigma(k)} \subset \cdots. \tag{8}
$$

Because $\{G_p : p \in \mathcal{P}\}$ is a finite set, the chain must converge to the graph

$$
\mathbb{G} \triangleq \bigcup_{k=1}^{\infty} \mathbb{G}_{\sigma(k)} \tag{9}
$$

in a finite number of steps. Since the sequence of graphs stops changing in a finite number of steps, rendezvousing at a single point can only occur if $\mathbb{G}$ is a complete graph. There is, however, no a priori guarantee that, along a particular trajectory, $\mathbb{G}$ will turn out to be complete. On the other hand, it is clear that $\mathbb{G}$ will always be at least connected if the initial graph $\mathbb{G}_{\sigma(1)}$ in the ascending chain is. It turns out that connectivity of $\mathbb{G}_{\sigma(1)}$ implies not only that $\mathbb{G}$ is connected but also that the types of distributed control strategies just described actually cause all agents to rendezvous at a single point.

3.1. Rendezvousing.

THEOREM 1. *Let $u_0 = 0 \in \mathbb{D}_M$ and for each $m \in \{1, 2, \ldots, n-1\}$, let $u_m : \mathbb{D}^m \to \mathbb{D}_M$ be any continuous function satisfying* (5) *and* (6). *For each set of initial agent positions $x_1(0), x_2(0), \ldots, x_n(0)$, each agent's position $x_i(t)$ converges to a unique point $p_i \in \mathbb{R}^2$ such that for each $i, j \in \{1, 2, \ldots, n\}$, either $p_i = p_j$ or $||p_i - p_j|| > r$. Moreover, if agents i and j are registered neighbors at any time t, then $p_i = p_j$.*

The proof of this theorem is given in section 4.

Theorem 1 states that the strategies under consideration cause all agents' positions to converge to points in the plane with the property that each two such points are either equal to each other or separated by a distance greater than r units. The theorem further states that if two agents are ever registered neighbors of each other, then their positions converge to the same point. We are led to the following corollary.

COROLLARY 1. *If the graph characterizing registered neighbors at the beginning of period* 1 *is connected, then the positions of all* n *agents converge to a common point in the plane.*

It is quite straightforward to extend these results to the leader-follower case when the rendezvous point is specified at the outset. This can be accomplished by simply fixing one additional agent (i.e., a virtual agent) at the desired rendezvous point and letting the remaining n agents maneuver just as before. With initial graph connectivity of all $n+1$ agent positions, convergence to the position of the virtual agent is then assured.

A more interesting case occurs when two virtual agents are fixed at distinct points in the plane. In this case it can be shown that with initial connectivity of the $(n+2)$-agent graph, all n agents will eventually move to positions on the line connecting the two virtual agents and will distribute themselves in a predictable manner depending on only the number of agents, r, and the distance between the two fixed, virtual agents. This behavior will be explored in greater depth in another paper dealing with forming formations using distributed control.

3.2. Trapping. While the graph connectivity hypothesis of Corollary 1 is sufficient for rendezvousing, it is not necessary. For example, suppose that the $\mathbb{G}_{\sigma(1)}$ has a connected component $\mathbb{G}_C$ which contains a simple closed cycle whose vertices are $i_1, i_2, \ldots, i_m$. Then in the plane, the geometric form obtained by connecting by a straight line the initial position of each agent $i_j \in \{i_1, i_2, \ldots, i_m\}$ with its registered neighbors with labels in $\{i_1, i_2, \ldots, i_m\}$ will be a simple, closed, polygon $\mathbb{P}$. It turns out that if the initial positions of all agents whose labels are not in the vertex set of $\mathbb{G}_C$ are within $\mathbb{P}$, then rendezvous will necessarily occur. While this conclusion might appear to be an obvious consequence of the established property that agents $i_j \in \{i_1, i_2, \ldots, i_m\}$ eventually rendezvous at a point, actually proving that this is true is not so straightforward. There are two reasons for this. First, there is no guarantee that the polygon $\mathbb{P}(k)$ formed by the positions at time t_k of agents $i_j \in \{i_1, i_2, \ldots, i_m\}$ will remain simple as the system evolves, even if it is initially; thus just what it means for an agent to be "inside" of $\mathbb{P}(k)$ requires a more sophisticated notion of interior than the obvious one for a simple closed curve in the plane, and this in turn complicates the analysis. Second, it is quite possible that an agent initially positioned inside of $\mathbb{P}(0)$ will be outside of $\mathbb{P}(k)$ for some $k > 0$. In what follows we explain how to overcome both of these difficulties and in so doing we establish a rendezvousing result along the lines just described.

We begin by reviewing the concept of a "winding number" and what it means for a point to be inside of a closed curve in $\mathbb{R}^2$. Let $\kappa : [0,1] \to \mathbb{R}^2$ be any continuous closed curve and let y be any point in $\mathbb{R}^2$ which does not lie on κ. The *winding number* of y with respect to κ, written $\mathrm{wn}(\kappa, y)$, is the number of times a point p traversing κ encircles y in a counterclockwise direction as p makes a full circuit of κ. Points not on κ with nonzero winding numbers are inside of κ, while those with a winding number of zero are outside of κ. There is a well-known formula for $\mathrm{wn}(\kappa, y)$, involving the integral around a closed contour $\tilde{\kappa} : [0,1] \to \mathbb{C}$ in the complex plane [15]. $\tilde{\kappa}$ is a

representation of κ resulting from the assignment to each vector $x = [\,a \quad b\,]'$ in $\mathbb{R}^2$ the associated complex number $\tilde{x} \triangleq a + jb$. In this setting, $\mathrm{wn}(\kappa, y)$ is given by the contour integral

$$\mathrm{wn}(\kappa, y) = \frac{1}{2\pi j} \oint_{\tilde{\kappa}} \frac{dz}{z - \tilde{y}}.$$

We will use this formula in what follows to prove Lemma 8.

The closed curves of interest here are of a specific type determined by finite point sets in $\mathbb{R}^2$. In particular, let us note that any ordered set of $m > 0$ points $\{y_1, y_2, \ldots, y_m\}$ in $\mathbb{R}^2$ uniquely determines a continuous, piecewise linear, closed curve $c : [0, m] \to \mathbb{R}^2$ defined so that

$$c(t) = (t + 1 - i)y_{i+1} + (i - t)y_i, \qquad i - 1 \le t \le i, \qquad i \in \{1, 2, \ldots, m\},$$

where $y_{m+1} = y_1$. An ordered set $\{y_1, y_2, \ldots, y_m\}$ of three or more such points is called a *cycle* if $||y_{i+1} - y_i|| \le r,\ i \in \{1, 2, \ldots, m\}$; in what follows we denote such a cycle by $\lceil y_1, y_2, \ldots, y_m \rceil$. A point $z \in \mathbb{R}^2$ is called an *interior point* of $\lceil y_1, y_2, \ldots, y_m \rceil$ if it is an interior point of the closed, piecewise linear curve c determined by $\{y_1, y_2, \ldots, y_m\}$.

A point $z \in \mathbb{R}^2$ is said to be *linked* to a nonempty set of vectors $\{y_1, y_2, \ldots, y_m\}$ in $\mathbb{R}^2$ if for some $i \in \{1, 2, \ldots, m\}$, $||z - y_i|| \le r$. More generally, z is *connected* to $\{y_1, y_2, \ldots, y_m\}$ through a set of vectors $\{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^2$ if there exists a subset $\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$ with $x_{i_k} \in \{y_1, y_2, \ldots, y_m\}$ such that $||z - x_{i_1}|| \le r$ and $||x_{i_{s-1}} - x_{i_s}|| \le r,\ i \in \{2, 3, \ldots k\}$. The following corollary to Theorem 1 will be proved later in this section.

COROLLARY 2. *Suppose that the set of initial positions* $\{x_1(0), x_2(0), \ldots, x_n(0)\}$ *of the n agents contains a cycle* $\lceil x_{i_1}(0), x_{i_2}(0), \ldots, x_{i_m}(0) \rceil$. *Then all agents initially positioned inside the cycle eventually rendezvous at one point with all agents with positions initially connected to the cycle through* $\{x_1(0),\ x_2(0), \ldots,\ x_n(0)\}$.

In what follows we use the abbreviated notation $\mathbf{C}(k) \triangleq \lceil x_{i_1}(t_k), x_{i_2}(t_k), \ldots, x_{i_m}(t_k) \rceil, k \ge 0$, and say that a vector x is connected to $\mathbf{C}(k)$ whenever x is connected to $\mathbf{C}(k)$ through $\{x_1(t_k),\ x_2(t_k), \ldots,\ x_n(t_k)\}$. Note that Corollary 2 does not require agents initially positioned inside of $\mathbf{C}(0)$ to be connected to $\mathbf{C}(0)$. It is natural to say that such "disconnected" agents are ultimately *trapped* by those agents whose initial positions comprise $\mathbf{C}(0)$. This particular group behavior is accordingly referred to as "trapping."

Consider the situation hypothesized in Corollary 2. We already know from Theorem 1 that all agents with positions initially connected to $\mathbf{C}(0)$ eventually rendezvous at a single point. So what remains to be shown is that all agents at initial positions interior to $\mathbf{C}(0)$ but not connected to it also rendezvous at the same point. To do this it is enough to show that each such initially disconnected internal agent eventually moves at some finite time t_K to a position which is connected to $\mathbf{C}(K)$—for once this happens, Theorem 1 can be applied with a start time of t_K, thereby enabling one to conclude that the agent under consideration will eventually rendezvous at the same point as the agents with positions initially connected to $\mathbf{C}(0)$. Carrying out this program relies on three key propositions which follow and which are proved in section 6.

PROPOSITION 4. *The interior of any cycle* $\lceil y_1, y_2, \ldots, y_m \rceil$ *in* $\mathbb{R}^2$ *is contained in its convex hull* $\langle y_1, y_2, \ldots, y_m \rangle$.

This proposition is used as follows. Note that because all agents initially positioned at points comprising $\mathbf{C}(0)$ eventually rendezvous at a single point, the diameter

of the convex hull $\langle x_{i_1}(t_k), x_{i_2}(t_k), \ldots, x_{i_m}(t_k)\rangle$ must eventually become smaller than r and remain so for all future time. What this and Proposition 4 therefore imply is that any agent whose position remains inside of $\mathbf{C}(k)$ for all time must at some finite time $t_{\bar{k}}$ reach a position connected to $\mathbf{C}(\bar{k})$. Unfortunately not every agent initially positioned at a point inside of and disconnected from $\mathbf{C}(0)$ can be counted on to be so accommodating. We will deal with this situation by proving that when such an agent first leaves $\mathbf{C}(k)$—say at time $t_{\bar{k}}$—it automatically moves to a position connected to $\mathbf{C}(\bar{k})$. Let A be the label of such an agent and let $x_A(t_{\bar{k}})$ denote its position at time $t_{\bar{k}}$. Below we shall argue using the following proposition that all of agent A's registered neighbors at the beginning of maneuvering period $\bar{k}-1$ are inside of $\mathbf{C}(\bar{k}-1)$ at time $t_{\bar{k}-1}$.

PROPOSITION 5. *Let $\lceil y_1, y_2, \ldots, y_m\rceil$ be a cycle in $\mathbb{R}^2$ which contains a point z which is not linked to $\lceil y_1, y_2, \ldots, y_m\rceil$. Then any point within r units of z is either inside of $\lceil y_1, y_2, \ldots, y_m\rceil$ or is linked to $\lceil y_1, y_2, \ldots, y_m\rceil$.*

We've assumed that $x_A(t_{\bar{k}})$ is not inside of $\mathbf{C}(\bar{k})$, and that $x_A(t_{\bar{k}-1})$ is inside of $\mathbf{C}(\bar{k}-1)$ and not connected to $\mathbf{C}(\bar{k}-1)$. Clearly $x_A(t_{\bar{k}-1})$ is not linked to $\mathbf{C}(\bar{k}-1)$. From Proposition 5 it follows that all of agent A'*s* registered neighbors at the beginning of maneuvering period $k-1$ are at positions at time $\bar{t}_{\bar{k}-1}$ {and consequently time $t_{\bar{k}-1}$} which are either inside of $\mathbf{C}(\bar{k}-1)$ or linked to $\mathbf{C}(\bar{k}-1)$. If any registered neighbor's position were connected to $\mathbf{C}(\bar{k}-1)$, then $x_A(t_{\bar{k}-1})$ would be connected to $\mathbf{C}(\bar{k}-1)$, which we have explicitly assumed is not the case. Therefore none of A'*s* registered neighbors is connected (or therefore linked) to $\mathbf{C}(\bar{k}-1)$ at time $t_{\bar{k}-1}$; moreover, all must be inside of $\mathbf{C}(\bar{k}-1)$ because of Proposition 5.

To show that under these conditions, $x_A(t_{\bar{k}})$ is necessarily connected to $\mathbf{C}(\bar{k})$, we will make use of the following concept. Let us agree to call a cycle $\lceil \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m\rceil$ a *successor* of a given cycle $\lceil y_1, y_2, \ldots, y_m\rceil$ if, in addition to the cycle requirement that $||\bar{y}_{i+1} - \bar{y}_i|| \le r,\ \ i \in \{1, 2, \ldots, n\}$, the inequalities $||\bar{y}_i - y_i|| \le r,\ ||\bar{y}_{i+1} - y_i|| \le r$, and $||\bar{y}_i - y_{i+1}|| \le r$ all hold for $i \in \{1, 2, \ldots, m\}$. Observe that each cycle in the sequence $\lceil y_1, y_2, \ldots, y_m\rceil, \lceil \bar{y}_1, y_2, \ldots, y_m\rceil, \lceil \bar{y}_1, \bar{y}_2, \ldots, y_m\rceil, \ldots, \lceil \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m\rceil$ is a successor of the cycle which precedes it. It is easy to verify that for each $k \ge 0$, $\mathbf{C}(k+1)$ is a successor of $\mathbf{C}(k)$.

PROPOSITION 6. *Let $\lceil \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m\rceil$ be a successor of a given cycle $\lceil y_1, y_2, \ldots, y_m\rceil$ in $\mathbb{R}^2$. Suppose that $z_1, z_2, \ldots, z_k$ are $k > 0$ interior points of $\lceil y_1, y_2, \ldots, y_m\rceil$ which are not linked to $\lceil y_1, y_2, \ldots, y_m\rceil$ and which satisfy $||z_1 - z_i|| \le r$, $i \in \{2, 3, \ldots, k\}$. Then each point in the convex hull $\langle z_1, z_2, \ldots, z_k\rangle$ is either an interior point of $\lceil \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m\rceil$ or is linked to $\lceil \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m\rceil$.*

Recall that the strategy under consideration puts $x_A(t_{\bar{k}})$ at a point in the convex hull of the set consisting of $x_A(t_{\bar{k}-1})$ and the positions at time $t_{\bar{k}-1}$ of agent A's registered neighbors. Proposition 6 therefore implies that $x_A(t_{\bar{k}})$ must be either inside of $\mathbf{C}(\bar{k})$ or linked to it. Since we have ruled out the former by assumption, $x_A(t_{\bar{k}})$ is linked and therefore connected to $\mathbf{C}(\bar{k})$ as claimed. This completes the proof of Corollary 2. □

4. Analysis. The aim of this section is to establish the correctness of Theorem 1. Towards this end, let $\{\{x_1(t_k), x_2(t_k), \ldots, x_n(t_k)\} : k \ge 1\}$ be a system trajectory determined by (7) and any initial set of agent positions. Let k^* denote the value of k for which the ascending chain shown in (8) converges to the limit graph $\mathbb{G}$ in (9). Thus for $t_k \ge t_{k^*}$, the neighbors of each agent do not change. For each $i \in \{1, 2, \ldots, n\}$, let $\{i_1, i_2, \ldots, i_{m_i}\}$ denote the set of indices labelling the neighbors of agent i. For simplicity, we will deal only with the case when each agent has at least one neighbor.

This means that all m_i are positive integers. These assumptions imply that for $k \geq k^*$, the system under consideration will have a state $\{x_1(t_k), x_2(t_k), \ldots, x_n(t_k)\}$ taking values in the space

$$(10)\, \mathcal{X} = \{\{x_1, x_2, \ldots x_n\} : ||x_j - x_i|| \leq r,\ \ j \in \{i_1, i_2, \ldots, i_{m_i}\}, i \in \{1, 2, \ldots, n\}\}.$$

4.1. Error system. To analyze system behavior it is convenient to introduce a suitably defined "error system." For $\{x_1, x_2, \ldots, x_n\} \in \mathcal{X}$, define

$$e_i = x_i - x_n, \quad i \in \{1, 2, \ldots, n\}, \tag{11}$$

and note that $e_n = 0$. Let $e \stackrel{\Delta}{=} \{e_1, e_2, \ldots, e_{n-1}\}$. In view of (10) and the fact that $e_j - e_i = x_j - x_i$ for all $i, j \in \{1, 2, \ldots, n\}$, we see that e takes values in the closed space

$$\mathcal{E} = \{\{e_1, e_2, \ldots e_{n-1}\} : e_n = 0, ||e_j - e_i|| \leq r, \quad j \in \{i_1, i_2, \ldots, i_{m_i}\},\ i \in \{1, 2, \ldots, n\}\}. \tag{12}$$

Note that

$$x_{i_j}(t_{k-1}) - x_i(t_{k-1}) = e_{i_j}(t_{k-1}) - e_i(t_{k-1}), \quad j \in \{1, 2, \ldots, m_i\}, \quad i \in \{1, 2, \ldots, n\}.$$

It follows that the update equation (7) for x_i can be written as

$$x_i(t_k) = x_i(t_{k-1}) + f_i(e(k-1)), \quad k \geq k^*, \tag{13}$$

where $f_i : \mathcal{E} \to \mathbb{D}$ is the continuous function

$$\{e_1, e_2, \ldots, e_{n-1}\} \longmapsto u_{m_i}(e_{i_1} - e_i,\ e_{i_2} - e_i, \ldots,\ e_{i_{m_i}} - e_i)|_{e_n = 0}.$$

In view of (13) and the definition of the e_i,

$$(14)\, e_i(t_k) = e_i(t_{k-1}) + f_i(e(t_{k-1})) - f_n(e(t_{k-1})), \quad k > k^*,\ i \in \{1, 2, \ldots, n-1\}.$$

This enables us to define the *error system*

$$e(t_k) = e(t_{k-1}) + f(e(t_{k-1})), \quad k > k^*, \tag{15}$$

where $f(e) = \{f_1(e) - f_n(e),\ f_2(e) - f_n(e), \ldots,\ f_{n-1}(e) - f_n(e)\}$.

4.2. Proving convergence in the style of Lyapunov. In what follows, we will prove that under certain conditions $e(t_k) \to 0$ as $k \to \infty$. We will do this using the positive definite function $V : \mathcal{E} \to \mathbb{R}$ defined by

$$V(e) = \text{dia}\{e_1, e_2, \ldots, e_{n-1}, 0\}, \tag{16}$$

where for any set of vectors $y_1, y_2, \ldots, y_m$ in $\mathbb{R}^2$, dia$\{y_1, y_2, \ldots, y_m\}$ denotes the diameter[2] of $\langle y_1, y_2, \ldots, y_m \rangle$. The following proposition is central to the proof of Theorem 1.

PROPOSITION 7. *The difference function $\Delta : \mathcal{E} \to \mathbb{R}$ defined by*

$$\Delta(e) = V(e + f(e)) - V(e) \tag{17}$$

is negative semidefinite. Moreover, if $\mathbb{G}$ is connected, then Δ is negative definite.

[2]Recall that the *diameter* of a closed set $\mathcal{S} \subset \mathbb{R}^2$ is the maximum of $||s_1 - s_2||$ over all $s_1, s_2 \in \mathcal{S}$.

Proof of Theorem 1. In general the graph $\mathbb{G}$ to which the ascending chain (8) converges for some finite $k = k^*$ consists of a finite set of connected components. Suppose that $\mathbb{G}_c$ is any one of these. To prove Theorem 1 it is enough to show that the positions of those agents whose indices are the vertices of $\mathbb{G}_c$ converge to a common point. For simplicity we will do this only for the case when $\mathbb{G}_c = \mathbb{G}$, since, except for notation, the proof is essentially the same even if $\mathbb{G}_c \neq \mathbb{G}$.

By hypothesis $n > 1$. Note that if $e(t_k) = 0$ for some $k = \bar{k}$, then all agents are in the same position at time $t_{\bar{k}}$; moreover, any such position will remain fixed for all $t \geq t_{\bar{k}}$ because $f(0) = 0$. Therefore to complete the proof it is enough to show that $e(t_k)$ tends to 0 as $k \to \infty$.

Let $V : \mathcal{E} \to \mathbb{R}$ be defined as in (16). Note that

$$V(e(t_k)) = \operatorname{dia}\{x_1(t_k), x_2(t_k), \ldots, x_n(t_k)\} \tag{18}$$

because the diameter of a convex set in $\mathbb{R}^2$ is invariant under translation of the set. From this and Proposition 7, it follows that the difference function

$$\Delta(e(t_k)) = V(e(t_k) + f(e(t_k))) - V(e(t_k))$$

is nonpositive for $k \geq k^*$. Thus $V(e(t_k))$ is a monotone nonincreasing function of k for $k \geq k^*$. Since for $k \geq k^*$, $V(e(t_k))$ is bounded above by $V(e(t_{k^*}))$ and below by 0, there must exist a finite limit

$$V^* \triangleq \lim_{k \to \infty} V(e(t_k)).$$

We claim that $V^* = 0$. To prove this claim, suppose that it is false. Then $V^* > 0$. Let $\mathcal{B}$ denote the set of all points $e \in \mathcal{E}$ such that $V^* \leq V(e) \leq V(e(t_{k^*}))$. Note that $\mathcal{B}$ is closed and bounded because $V(\cdot)$ is continuous and $\mathcal{E}$ is closed. Moreover, $0 \notin \mathcal{B}$ because $V(\cdot)$ is positive definite and bounded away from zero on $\mathcal{B}$. By Proposition 7, $\Delta(\cdot)$ is negative definite. Therefore for all $e \in \mathcal{B}$, $\Delta(e) < 0$. From this, the compactness of $\mathcal{B}$, and the continuity of $\Delta(\cdot)$, it follows that

$$\mu \triangleq \max_{e \in \mathcal{B}} \Delta(e)$$

is a finite negative number. Since $e(t_k) \in \mathcal{B}$ for $k \geq k^*$, it must therefore be true that

$$V(e(t_{k+1})) - V(e(t_k)) = \Delta(e(t_k)) \leq \mu, \quad k \geq k^*.$$

Thus by summing,

$$V(e(t_k)) \leq V(e(t_{k^*})) + (k - k^*)\mu, \quad k \geq k^*.$$

Therefore, for k sufficiently large, $V(e(t_k))$ must be negative because $\mu < 0$. But this is impossible because $V(\cdot)$ is positive definite. Hence V^* cannot be positive. □

The proof just given is basically a standard Lyapunov argument[3] applied to the system (17). It is worth pointing out here that the continuity of $\Delta(\cdot)$ is crucial to the proof as is the fact that $\mathcal{E}$ is closed. If $\mathcal{E}$ were not a closed set, the preceding proof would break down because one could not conclude that $\mathcal{B}$ is closed. The closure of $\mathcal{E}$ is

[3]It is worth noting that a similar proof could also be crafted using recent results by Moreau which appeared in [16] after this paper was submitted in December, 2004.

a direct consequence of the fact that sensing regions are defined to be closed sets. The continuity of $\Delta(\cdot)$ is a consequence of the requirement that the $u_m(\cdot)$ be continuous functions. In summary, for the present analysis to go through, it is essential that sensing regions be closed sets and that the $u_m(\cdot)$ be continuous functions. Whether or not these requirements can be relaxed by approaching convergence differently remains to be seen.

5. Concluding remarks. In this paper we have reconsidered the multi-agent rendezvous problem originally posed in [1] and have described several alternate synchronous solutions. We have provided an example which shows that rendezvousing can in some cases be guaranteed to occur even if the graph characterizing initial relations is not initially connected. In a sequel to this paper [12] we will explain how rendezvousing can be achieved asynchronously, without assuming that the agents share a common clock.

Since this paper and its sequel [12] were written, a number of papers on rendezvous have appeared. We refer the reader to [3] for additional references and for interesting new results on the rendezvous problem posed in higher dimensional spaces and with more general assumptions about sensing and communications.

6. Appendix. The proof of Proposition 1 depends on the following lemma.

LEMMA 1. *Let $z_1, z_2, \ldots, z_m$ be a set of $m > 0$ points in $\mathbb{D}$ which are not all 0. If 0 is a corner of $\langle 0, z_1, z_2, \ldots, z_m\rangle$, then the constraint set $\mathcal{C}(z_1, z_2, \ldots, z_m)$ has a nonempty interior.*

Proof of Lemma 1. Suppose that $\mathcal{C}(z_1, z_2, \ldots, z_m)$ has an empty interior in which case $\mathcal{C}(z_1, z_2, \ldots, z_m)$ is the singleton $\{0\}$. It will be enough to prove that 0 is not a corner of $\langle 0, z_1, z_2, \ldots, z_m\rangle$. In what follows we shall assume, without loss of generality, that 0 is on the boundary of each disk in the intersection; for if there were any disks in the intersection which contained the origin in their interiors, all such disks could be removed from the intersection without changing what the intersection equals.

To proceed, let us note first that $m > 1$ because each $\mathcal{C}(z_i)$ has a nonempty interior and, by hypothesis, the intersection $\mathcal{C}(z_1, z_2, \ldots, z_m)$ does not. Next observe that since $m > 1$ and $\mathcal{C}(z_1)$ has a nonempty interior, there must be a least integer $j \in \{2, 3, \ldots, m\}$ such that $\mathcal{I} \triangleq \bigcap_{i=1}^{j-1} \mathcal{C}(z_j)$ has a nonempty interior and $\mathcal{I} \cap \mathcal{C}(z_j)$ contains just the origin. The intersection of any positive number of disks from $\{\mathcal{C}(z_1), \mathcal{C}(z_2), \ldots, \mathcal{C}(z_m)\}$ is either the origin or a convex set with a nonempty interior; moreover, the latter will always be a strictly convex set whose edges are arcs from circles bounding disks in the intersection and whose corners are intersections of such arcs. It follows that $\mathcal{C}(z_j)$ must either be tangent at the origin to an arc which is from a circle bounding some disk $\mathcal{C}(z_k) \in \{\mathcal{C}(z_1), \mathcal{C}(z_2), \ldots, \mathcal{C}(z_{j-1})\}$ or $\mathcal{C}(z_j)$'s boundary must pass through a corner of $\mathcal{I}$ at the origin. If the former is true, then z_k must equal $-z_j$. Since $z_j \neq 0$, this means that the origin is halfway between z_k and z_j on the line connecting these two points. Hence the origin cannot be a corner of the polytope $\langle 0, z_1, z_2, \ldots, z_m\rangle$.

Now suppose that the boundary of $\mathcal{C}(z_j)$ passes through a corner of $\mathcal{I}$ at the origin. Let $\mathcal{C}(z_k)$ and $\mathcal{C}(z_l)$ denote two disks in $\{\mathcal{C}(z_1), \mathcal{C}(z_2), \ldots, \mathcal{C}(z_{j-1})\}$ whose intersection at the origin determines this corner. Under these conditions, $z_k + z_l \neq 0$—for if $z_k + z_l = 0$, then $\mathcal{C}(z_k)$ and $\mathcal{C}(z_l)$ would be tangent, and $\mathcal{I}$ would consequently contain just the origin. Moreover, the intersection $\mathcal{C}(z_j) \cap \mathcal{C}(z_k) \cap \mathcal{C}(z_l)$ must consist of just the origin—for if this were not so, then $\mathcal{I} \cap \mathcal{C}(z_j)$ would have a nonempty interior since $\mathcal{I}$ coincides locally, in an open neighborhood of 0, with $\mathcal{C}(z_k) \cap \mathcal{C}(z_l)$.

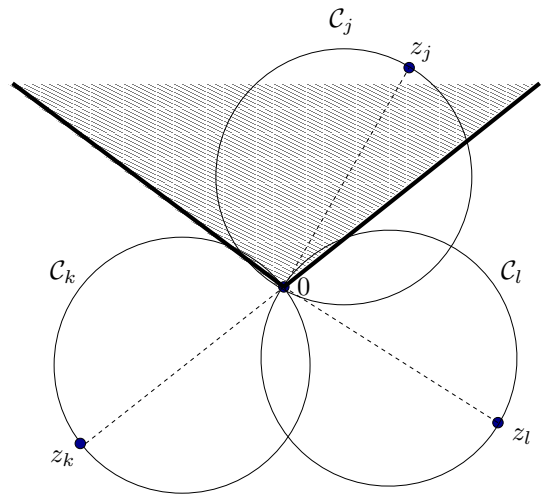


FIG. 1. *Three constraint disks whose intersection is the origin.*

As illustrated in Figure 1, the requirement that $\mathcal{C}(z_j) \cap \mathcal{C}(z_k) \cap \mathcal{C}(z_l)$ consist of just the origin implies that $\mathcal{C}(z_j)$ must be positioned in such a way so that it intersects only at the origin with a cone of points determined by tangents to $\mathcal{C}(z_k)$ and $\mathcal{C}(z_l)$ at the origin. This means that z_j must lie within the opposing cone shown in grey in Figure 1. Hence the origin is within the interior of the convex hull of z_j, z_k, and z_l. Therefore the origin cannot be a corner of $\langle 0, z_1, z_2, \ldots, z_m \rangle$. □

Proof of Proposition 1. Lemma 1 and the hypothesis that 0 is a corner of $\langle 0, z_1, z_2, \ldots, z_m \rangle$ imply that $\mathcal{C}(z_1, z_2, \ldots, z_m)$ has a nonempty interior. From this and the hypothesis that $z \neq 0$, it follows that if 0 is an interior point of $\mathcal{C}(z_1, z_2, \ldots, z_m)$, then a line segment with the required property must exist.

Suppose next that 0 is on the boundary of $\mathcal{C}(z_1, z_2, \ldots, z_m)$. To complete the proof it is clearly enough to show that the line from 0 to z passes through the interior of each disk $\mathcal{C}(z_j)$ for which 0 is a boundary point. Let $\mathcal{C}(z_j)$ be such a disk in which case $||z_j|| = r$. Suppose that the line from 0 to z does not pass through the interior $\mathcal{C}(z_j)$. This means that $z_j' z \leq 0$ and thus that $||z||^2 - 2z_j' z + ||z_j||^2 \geq ||z||^2 + ||z_j||^2$. Since $||z - z_j||^2 = ||z||^2 - 2z_j' z + ||z_j||^2$ and $||z_j|| = r$, it follows that

$$||z - z_j||^2 \geq ||z||^2 + r^2.$$

But $||z||^2 > 0$ because $z \neq 0$, so

$$||z - z_j|| > r.$$

This contradicts the hypothesis that z is within r units of each point in $\{z_1, z_2, \ldots, z_m\}$. Therefore the line from 0 to z must pass through the interior $\mathcal{C}(z_j)$. □

The proof of Proposition 2 depends on the following two lemmas.

LEMMA 2. *Let $z_1, z_2, \ldots, z_m$ be a set of $m > 0$ points in $\mathbb{D}$ which are not all zero. Let $\mathcal{E}(x, y)$ be an edge of $\langle 0, z_1, z_2, \ldots, z_m \rangle$ with distinct corners x and y. Write $\mathcal{L}(x, y)$ for the line passing through x and y, and let $\mathcal{S}(x, y)$ denote the closed half-plane bounded by this line whose intersection with $\langle 0, z_1, z_2, \ldots, z_m \rangle$ is $\mathcal{E}(x, y)$. If z is any point in $\mathcal{S}(x, y)$ which is also in $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$, then the reflection of z about $\mathcal{L}(x, y)$ is also in $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$.*

Proof of Lemma 2. Note that $\langle 0, z_1, z_2, \ldots, z_m\rangle$ is contained in the half-plane obtained by reflecting $\mathcal{S}(x, y)$ about $\mathcal{L}(x, y)$. Because of this, for each $w \in \mathcal{S}$

$$||\bar{w} - q|| \leq ||w - q|| \quad \forall q \in \langle 0, z_1, z_2, \ldots, z_m\rangle,$$

where $\bar{w}$ is the reflection of w about $\mathcal{L}(x, y)$. In particular, this implies that

$$||\bar{z} - z_i| \leq ||z - z_i||, \quad i \in \{0, 1, 2, \ldots, m\}, \tag{19}$$

where $z_0 = 0$ and $\bar{z}$ is the reflection of z about $\mathcal{L}(x, y)$. But

$$||z - z_i|| \leq r, \quad i \in \{0, 1, 2, \ldots, m\}, \tag{20}$$

because $z \in \mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$. From (19) and (20) it follows that $\bar{z} \in \mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$. □

LEMMA 3. *Let $\mathcal{L}$ be a line in $\mathbb{R}^2$ which divides a given closed set $\mathcal{D}$ into closed subsets $\mathcal{P}$ and $\mathcal{Q}$ with $\mathcal{Q}$ convex. If the reflection of $\mathcal{P}$ about $\mathcal{L}$ is a subset of $\mathcal{Q}$, then the centroid of $\mathcal{D}$ is in $\mathcal{Q}$.*

Proof. Let $\bar{\mathcal{P}}$ denote the reflection of $\mathcal{P}$ about $\mathcal{L}$. By hypothesis, $\bar{\mathcal{P}} \subset \mathcal{Q}$. Then write $\mathcal{Q} - \bar{\mathcal{P}}$ for the complement of $\bar{\mathcal{P}}$ in $\mathcal{Q}$. By symmetry, the centroid of $\mathcal{P} \cup \bar{\mathcal{P}}$ is in $\mathcal{L} \subset \mathcal{Q}$. Meanwhile, the centroid of $\mathcal{Q} - \bar{\mathcal{P}}$ must also be in $\mathcal{Q}$ because $\mathcal{Q}$ is convex and $\mathcal{Q} - \bar{\mathcal{P}} \subset \mathcal{Q}$. Thus the centroid of $\mathcal{D}$ must be in $\mathcal{Q}$ because it is the average of the centroids of $\mathcal{P} \cup \bar{\mathcal{P}}$ and $\mathcal{Q} - \bar{\mathcal{P}}$ weighted by the areas of $\mathcal{P} \cup \bar{\mathcal{P}}$ and $\mathcal{Q} - \bar{\mathcal{P}}$, respectively. □

Proof of Proposition 2. Write $\mathcal{D}$ for $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$ and let $\{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_k\}$ denote the set of edges of $\langle 0, z_1, z_2, \ldots, z_m\rangle$. For each such edge $\mathcal{E}_i$, let $\mathcal{L}_i$ denote the line in $\mathbb{R}^2$ containing $\mathcal{E}_i$ and write $\mathcal{S}_i$ for the closed half-plane bounded by this line whose intersection with $\langle 0, z_1, z_2, \ldots, z_m\rangle$ is $\mathcal{E}_i$. Let $\bar{\mathcal{S}}_i$ denote the reflection of $\mathcal{S}_i$ about $\mathcal{L}_i$. In view of Lemma 2,

$$\overline{\mathcal{S}_i \cap \mathcal{D}} \subset \mathcal{D}, \quad i \in \{1, 2, \ldots, k\},$$

where $\overline{\mathcal{S}_i \cap \mathcal{D}}$ is the reflection of $\mathcal{S}_i \cap \mathcal{D}$ about $\mathcal{L}_i$. Since $\overline{\mathcal{S}_i \cap \mathcal{D}}$ is also a subset of $\bar{\mathcal{S}}_i$,

$$\overline{\mathcal{S}_i \cap \mathcal{D}} \subset \bar{S}_i \cap \mathcal{D}, \quad i \in \{1, 2, \ldots, k\}. \tag{21}$$

Moreover, by de Morgan's rule

$$\{\mathcal{S}_i \cap \mathcal{D}\} \cup \{\bar{S}_i \cap \mathcal{D}\} = \mathcal{D}, \quad i \in \{1, 2, \ldots, k\},$$

because $\mathcal{S}_i \cup \bar{\mathcal{S}}_i = \mathbb{R}^2$, $i \in \{1, 2, \ldots, k\}$. Thus for each $i \in \{1, 2, \ldots\}$, $\mathcal{L}_i$ divides $\mathcal{D}$ into two closed convex regions, namely $\mathcal{S}_i \cap \mathcal{D}$ and $\bar{S}_i \cap \mathcal{D}$. From this, (21), and Lemma 3 it follows that

$$\text{centroid}\{\mathcal{D}\} \in \bar{\mathcal{S}}_i \cap \mathcal{D}, \quad i \in \{1, 2, \ldots, k\}.$$

Therefore

$$\text{centroid}\{\mathcal{D}\} \in \bigcap_{i=1}^{k} \{\bar{\mathcal{S}}_i \cap \mathcal{D}\}. \tag{22}$$

But

$$\bigcap_{i=1}^{k} \{\bar{\mathcal{S}}_i \cap \mathcal{D}\} = \langle 0, z_1, z_2, \ldots, z_m\rangle \bigcap \mathcal{D} \tag{23}$$

because

$$\langle 0, z_1, z_2, \ldots, z_m \rangle = \bigcap_{i=1}^{k} \bar{\mathcal{S}}_i.$$

From (22) and (23) it follows that centroid$\{\mathcal{D}\} \in \langle 0, z_1, z_2, \ldots, z_m \rangle$.

Now suppose that 0 is a corner of $\langle 0, z_1, z_2, \ldots, z_m \rangle$. Then in view of Lemma 1, $\mathbb{D} \cap \mathcal{C}(z_1, z_2, \ldots, z_m)$ has a nonempty interior. To prove that $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$ also has a nonempty interior, it is therefore enough to show that

$$\mathcal{C}(z_1, z_2, \ldots, z_m) \subset \mathcal{D}(z_1, z_2, \ldots, z_m). \tag{24}$$

Recall that for $z \in \mathbb{D}$, $\mathcal{D}(z) = \{x : ||z - x|| \leq r\}$ and $\mathcal{C}(z) = \{x : ||\frac{1}{2}z - x|| \leq \frac{r}{2}\}$. Thus for $x \in \mathcal{C}(z)$

$$||z - x|| = \left\| \frac{1}{2}z - x + \frac{1}{2}z \right\| \leq \left\| \frac{1}{2}z - x \right\| + \left\| \frac{1}{2}z \right\| \leq \frac{r}{2} + \frac{1}{2}||z|| \leq r;$$

thus $x \in \mathcal{D}(z)$. Hence $\mathcal{C}(z) \subset \mathcal{D}(z)$, $z \in \mathbb{D}$, from which (24) follows.

To prove that the centroid of $\mathcal{D}$ is not at 0, it is enough to show that it is not at 0 whenever it lies on an edge of $\langle 0, z_1, z_2, \ldots, z_m \rangle$ which contains 0. Accordingly, let $\mathcal{E}_j$ be an edge of $\langle 0, z_1, z_2, \ldots, z_m \rangle$ which contains both 0 and the centroid of $\mathcal{D}$. Since the centroid of $\mathcal{D}$ lies in $\mathcal{L}_j$, and both $\mathcal{S}_j \cap \mathcal{D}$ and $\bar{\mathcal{S}}_j \cap \mathcal{D}$ have nonempty interiors, it must be true that

$$\text{area}\{\mathcal{S}_j \cap \mathcal{D}\}d = \text{area}\{\bar{\mathcal{S}}_j \cap \mathcal{D}\}\bar{d},$$

where d is the distance from the centroid of $\mathcal{S}_j \cap \mathcal{D}$ to the point closest on $\mathcal{L}_j$ and $\bar{d}$ is correspondingly the distance from the centroid of $\bar{\mathcal{S}}_j \cap \mathcal{D}$ to the point closest on $\mathcal{L}_j$. But area$\{\overline{\mathcal{S}_j \cap \mathcal{D}}\}$ = area$\{\mathcal{S}_j \cap \mathcal{D}\}$, so

$$\text{area}\{\overline{\mathcal{S}_j \cap \mathcal{D}}\}d = \text{area}\{\bar{\mathcal{S}}_j \cap \mathcal{D}\}\bar{d}. \tag{25}$$

We claim that

$$\overline{\mathcal{S}_j \cap \mathcal{D}} = \bar{\mathcal{S}}_j \cap \mathcal{D}. \tag{26}$$

To establish this claim, we first note that $\overline{\mathcal{S}_j \cap \mathcal{D}} \subset \bar{\mathcal{S}}_j \cap \mathcal{D}$ because (21) holds for all $i \in \{1, 2, \ldots, k\}$. Thus to prove (26) it is enough to show that the complement of $\overline{\mathcal{S}_j \cap \mathcal{D}}$ in $\bar{\mathcal{S}}_j \cap \mathcal{D}$, denoted by $\mathcal{W}$, is empty. Towards this end, suppose that $\mathcal{W}$ is nonempty and has a nonempty interior. Since $\bar{\mathcal{S}}_j \cap \mathcal{D} = \{\overline{\mathcal{S}_j \cap \mathcal{D}}\} \cup \mathcal{W}$ and $\{\overline{\mathcal{S}_j \cap \mathcal{D}}\} \cap \mathcal{W}$ is empty, it must be true that

$$\text{area } \{\bar{\mathcal{S}}_j \cap \mathcal{D}\}\bar{d} = \text{area } \{\overline{\mathcal{S}_j \cap \mathcal{D}}\}d_1 + \text{area } \{\mathcal{W}\}d_2,$$

where d_1 is the distance from the centroid of $\{\overline{\mathcal{S}_j \cap \mathcal{D}}\}$ to the point closest on $\mathcal{L}_j$ and d_2 is correspondingly the distance from the centroid of $\mathcal{W}$ to the point closest on $\mathcal{L}_j$. But $\overline{\mathcal{S}_j \cap \mathcal{D}}$ is the reflection of $\mathcal{S}_j \cap \mathcal{D}$ about $\mathcal{L}_j$, and thus $d_1 = d$. Therefore

$$\text{area } \{\bar{\mathcal{S}}_j \cap \mathcal{D}\}\bar{d} = \text{area } \{\overline{\mathcal{S}_j \cap \mathcal{D}}\}d + \text{area } \{\mathcal{W}\}d_2.$$

This and (25) imply that area $\{\mathcal{W}\}d_2 = 0$. But $d_2 \neq 0$ because we have assumed that $\mathcal{W}$ has a nonempty interior. This implies that area $\{\mathcal{W}\} = 0$, which contradicts the hypothesis that $\mathcal{W}$ has a nonempty interior. Therefore $\mathcal{W}$ has an empty interior.

To show that $\mathcal{W}$ is actually empty or, equivalently, that (26) holds, it is enough to prove that the interior of $\bar{\mathcal{S}}_j \cap \mathcal{D}$ is contained in $\overline{\mathcal{S}_j \cap \mathcal{D}}$. For if this is true, then (26) holds, because both sets are closed and convex with nonempty interiors and $\overline{\mathcal{S}_j \cap \mathcal{D}} \subset \bar{\mathcal{S}}_j \cap \mathcal{D}$.

Suppose that the interior of $\bar{\mathcal{S}}_j \cap \mathcal{D}$ is not contained in $\overline{\mathcal{S}_j \cap \mathcal{D}}$. Then there must be a point p in the interior of $\bar{\mathcal{S}}_j \cap \mathcal{D}$ which is not in $\overline{\mathcal{S}_j \cap \mathcal{D}}$. Since $\{p\}$ and $\overline{\mathcal{S}_j \cap \mathcal{D}}$ are disjoint and each is a closed, convex set, there must be a line $\bar{\mathcal{L}}$ which separates the two and intersects neither. From this it is clear that there is an open set $\mathcal{N}_p \subset \bar{\mathcal{S}}_j \cap \mathcal{D}$ which contains p and which does not intersect $\bar{\mathcal{L}}$. It follows that $\mathcal{N}_p$ and $\overline{\mathcal{S}_j \cap \mathcal{D}}$ are disjoint and thus that $\mathcal{N}_p \subset \mathcal{W}$. But this is impossible because $\mathcal{W}$ has no interior. Therefore $\mathcal{W}$ is empty. We have therefore proved that $\mathcal{D}$ is *symmetric* about $\mathcal{L}_j$ in the sense that (26) holds.

Since $\mathcal{D}$ has a nonempty interior, its boundary consists of circular arcs resulting from the intersection of $m+1$ disks of radius r. Let $\mathcal{A}$ denote a circular arc of positive length which lies in $\mathcal{S}_j$ and which comprises part of the boundary of $\mathcal{D}$. In view of $\mathcal{D}$'s symmetry about $\mathcal{L}_j$ as defined by (26), the reflection of $\mathcal{A}$ about $\mathcal{L}_j$, namely $\bar{\mathcal{A}}$, must be a circular arc of positive length which lies in $\bar{\mathcal{S}}_j$ and which comprises part of the boundary of $\mathcal{D}$. Let x and y be points in $\{0, z_1, z_2, \ldots, z_m\}$ which define disks $\mathcal{D}(x)$ and $\mathcal{D}(y)$ whose boundaries contain $\mathcal{A}$ and $\bar{\mathcal{A}}$, respectively. Clearly the reflection of $\mathcal{D}(x)$ about $\mathcal{L}_j$ must equal $\mathcal{D}(y)$, which implies that $\bar{x} = y$. Thus $\bar{x} \in \langle 0, z_1, z_2, \ldots, z_m \rangle$. Since either $\bar{x}$ or x must be in $\mathcal{S}_j$, at least one of these two points must be in $\mathcal{S}_j \cap \langle 0, z_1, z_2, \ldots, z_m \rangle$, which is equal to $\mathcal{E}_j$. This can only occur if $\bar{x} = x$. In summary we have shown that if $\mathcal{A}$ is any circular arc of positive length comprising part of the boundary of $\mathcal{D}$, and if x is any point in $\{0, z_1, z_2, \ldots, z_m\}$ which defines a disk $\mathcal{D}(x)$ whose boundary contains $\mathcal{A}$, then x must be in $\mathcal{E}_j$.

Now let y be the nonzero endpoint of the edge $\mathcal{E}_j$, let $\mathcal{A}$ be any circular arc of positive length comprising part of the boundary of $\mathcal{D}$, and let $x_{\mathcal{A}}$ be any point in $\{0, z_1, z_2, \ldots, z_m\}$ which defines a disk $\mathcal{D}(x_{\mathcal{A}})$ whose boundary contains $\mathcal{A}$. As we have just shown, $x_{\mathcal{A}} \in \mathcal{E}_j$. This means there must be a number $\lambda \in [0,1]$ such that $x_{\mathcal{A}} = \lambda y$. Let z be any point in $\mathbb{D} \cap \mathcal{D}(y)$. Then by definition $||z|| \le r$ and $||y - z|| \le r$. Therefore

$$\begin{aligned} ||x_{\mathcal{A}} - z|| &= ||\lambda y - z|| = ||\lambda(y-z) - (1-\lambda)z|| \\ &\le ||\lambda(y-z)|| + ||(1-\lambda)z|| \le \lambda||y-z|| + (1-\lambda)||z|| \le r, \end{aligned}$$

so $z \in \mathcal{D}(x_{\mathcal{A}})$. Since z was chosen arbitrarily,

$$\mathbb{D} \cap \mathcal{D}(y) \subset \mathcal{D}(x_{\mathcal{A}}).$$

This containment holds for each disk $\mathcal{D}(x_{\mathcal{A}})$ whose boundary contains a circular arc $\mathcal{A}$ of positive length comprising part of the boundary of $\mathcal{D}$. Since the intersection of the $\mathcal{D}(x_{\mathcal{A}})$ over all such $\mathcal{A}$ is $\mathcal{D}$, it must therefore be true that

$$\mathbb{D} \cap \mathcal{D}(y) \subset \mathcal{D}. \tag{27}$$

On the other hand, $\mathcal{D} \subset \mathcal{D}(y)$ since $y \in \langle 0, z_1, z_2, \ldots, z_m \rangle$. Thus $\mathcal{D} \subset \mathbb{D} \cap \mathcal{D}(y)$. This and (27) thus imply that

$$\mathbb{D} \cap \mathcal{D}(y) = \mathcal{D}.$$

It follows that the centroid of $\mathcal{D}$ must be the centroid of $\mathbb{D} \cap \mathcal{D}(y)$. But the centroid of two intersection disks with the same radius must be at the midpoint between their centers. Therefore the centroid of $\mathcal{D}$ is at $\frac{1}{2}y$ which is not 0. □

Proof of Proposition 3. In what follows we write z for the n-tuple $\{z_1, z_2, \ldots, z_m\} \in \mathbb{D}^m$, and $\mathcal{S}(z)$ for the intersection $\mathbb{D} \cap \mathcal{D}(z_1, z_2, \ldots, z_m)$. Thus for $x, y \in \mathbb{D}^m$, $\mathcal{S}(x) \cap \mathcal{S}(y) = \mathbb{D} \cap \mathcal{D}(x_1, x_2, \ldots, x_m) \cap \mathcal{D}(y_1, y_2, \ldots, y_m)$. For $x \in \mathbb{D}^m$, let $\alpha(\mathcal{S}(x))$ and $\sigma(\mathcal{S}(x))$ denote, respectively, the area and centroid of $\mathcal{S}(x)$. Note that $\sigma(\mathcal{S}(x)) = 0$ whenever $\alpha(\mathcal{S}(x)) = 0$. This crucial property (which is not true for polygons) is a consequence of the fact that $\mathcal{S}(x)$ is either strictly convex with nonempty interior or the singleton 0.

It will first be shown that $z \longmapsto \sigma(\mathcal{S}(z))$ is continuous at each point $x \in \mathbb{D}^m$ at which $\alpha(\mathcal{S}(x)) = 0$. Let x be any such point. Clearly $\sigma(\mathcal{S}(x)) = 0$. Let $\epsilon > 0$ be fixed. Since $z \longmapsto \text{diameter}(\mathcal{S}(z))$ is continuous on $\mathbb{D}^m$, there must be a number $\delta > 0$ such that $\text{diameter}(\mathcal{S}(z)) \leq \epsilon$ whenever $||z - x|| \leq \delta$. But both 0 and $\sigma(\mathcal{S}(z))$ are points in $\mathcal{S}(z)$ for all $z \in \mathbb{D}^m$. Hence $||\sigma(\mathcal{S}(z))|| \leq \text{diameter}(\mathcal{S}(z)),\ z \in \mathbb{D}^m$; thus $|\sigma(\mathcal{S}(z))| \leq \epsilon$ whenever $||z - x|| \leq \delta$. Therefore $z \longmapsto \sigma(\mathcal{S}(z))$ is continuous at each point $x \in \mathbb{D}^m$ at which $\alpha(\mathcal{S}(x)) = 0$.

It will now be shown that $z \longmapsto \sigma(\mathcal{S}(z))$ is continuous at each point $x \in \mathbb{D}^m$ at which $\alpha(\mathcal{S}(x)) > 0$. Let x be such a point. Pick $\epsilon > 0$ and define

$$\bar{\epsilon} = \frac{\epsilon}{\epsilon + 4r} \alpha(\mathcal{S}(x)). \tag{28}$$

Since $z \longmapsto \alpha(\mathcal{S}(z))$ and $z \longmapsto \alpha(\mathcal{S}(x) \cap \mathcal{S}(z))$ are continuous on $\mathbb{D}^m$ and $\alpha(\mathcal{S}(x) \cap \mathcal{S}(x)) = \alpha(\mathcal{S}(x))$, there must be a number $\delta > 0$ such that

$$|\alpha(\mathcal{S}(x)) - \alpha(\mathcal{S}(z))| \leq \bar{\epsilon} \quad \text{and} \quad |\alpha(\mathcal{S}(x)) - \alpha(\mathcal{S}(x) \cap \mathcal{S}(z))| \leq \bar{\epsilon} \tag{29}$$

whenever $||z - x|| \leq \delta$. Fix z at any such value. To complete the proof it is enough to show that

$$||\sigma(\mathcal{S}(x)) - \sigma(\mathcal{S}(z))|| \leq \epsilon. \tag{30}$$

From the first inequality in (29), $\alpha(\mathcal{S}(z)) \geq \alpha(\mathcal{S}(x)) - \bar{\epsilon}$. But from (28), $\alpha(\mathcal{S}(x)) - \bar{\epsilon} = \frac{4r\bar{\epsilon}}{\epsilon}$ and thus

$$\alpha(\mathcal{S}(z)) \geq \frac{4r\bar{\epsilon}}{\epsilon}. \tag{31}$$

In general

$$\mathcal{S}(x) = (\mathcal{S}(x) \cap \mathcal{S}(z)) \cup \mathcal{X} \quad \text{and} \quad \mathcal{S}(z) = (\mathcal{S}(x) \cap \mathcal{S}(z)) \cup \mathcal{Z}, \tag{32}$$

where $\mathcal{X}$ and $\mathcal{Z}$ are the complements of $\mathcal{S}(x) \cap \mathcal{S}(z)$ in $\mathcal{S}(x)$ and $\mathcal{S}(z)$, respectively. If $\mathcal{S}(x) \cap \mathcal{S}(z)$ is a strictly proper subset of $\mathcal{S}(x)$ (respectively, $\mathcal{S}(z)$), then $\mathcal{X}$ (respectively, $\mathcal{Z}$) is a subset with nonempty interior; in this case $\alpha(\mathcal{X})$ and $\sigma(\mathcal{X})$ (respectively, $\alpha(\mathcal{Z})$ and $\sigma(\mathcal{Z})$) are well defined. If, on the other hand, $\mathcal{S}(x) \cap \mathcal{S}(z)$ equals $\mathcal{S}(x)$ (respectively, $\mathcal{S}(z)$), then $\mathcal{X}$ (respectively, $\mathcal{Z}$) is the empty set; in this case $\alpha(\mathcal{X})$ (respectively, $\alpha(\mathcal{Z})$) is zero, and $\sigma(\mathcal{X})$ (respectively, $\sigma(\mathcal{Z})$) is taken to be the 0 vector in $\mathbb{R}^2$.

In view of (32)

$$\alpha(\mathcal{S}(x)) = \alpha(\mathcal{S}(x) \cap \mathcal{S}(z)) + \alpha(\mathcal{X}), \tag{33}$$

$$\alpha(\mathcal{S}(z)) = \alpha(\mathcal{S}(x) \cap \mathcal{S}(z)) + \alpha(\mathcal{Z}), \tag{34}$$

$$\alpha(\mathcal{S}(x))\sigma(\mathcal{S}(x)) = \alpha(\mathcal{S}(x) \cap \mathcal{S}(z))\sigma(\mathcal{S}(x) \cap \mathcal{S}(z)) + \alpha(\mathcal{X})\sigma(\mathcal{X}), \tag{35}$$

$$\alpha(\mathcal{S}(z))\sigma(\mathcal{S}(z)) = \alpha(\mathcal{S}(x) \cap \mathcal{S}(z))\sigma(\mathcal{S}(x) \cap \mathcal{S}(z)) + \alpha(\mathcal{Z})\sigma(\mathcal{Z}). \tag{36}$$

Subtracting (33) from (34) and (35) from (36), one obtains

$$\alpha(\mathcal{S}(z)) - \alpha(\mathcal{S}(x)) = \alpha(\mathcal{Z}) - \alpha(\mathcal{X}) \tag{37}$$

and

$$\alpha(\mathcal{S}(z))\sigma(\mathcal{S}(z)) - \alpha(\mathcal{S}(x))\sigma(\mathcal{S}(x)) = \alpha(\mathcal{Z})\sigma(\mathcal{Z}) - \alpha(\mathcal{X})\sigma(\mathcal{X}), \tag{38}$$

respectively. Using (37) to eliminate $\alpha(\mathcal{Z})$ from (38), there results

$$\alpha(\mathcal{S}(z))\sigma(\mathcal{S}(z)) - \alpha(\mathcal{S}(x))\sigma(\mathcal{S}(x)) = \alpha(\mathcal{X})\{\sigma(\mathcal{Z}) - \sigma(\mathcal{X})\} + \{\alpha(\mathcal{S}(z)) - \alpha(\mathcal{S}(x))\}\sigma(\mathcal{Z}),$$

which can be rewritten as

$$\sigma(\mathcal{S}(z)) - \sigma(\mathcal{S}(x)) = \frac{1}{\alpha(\mathcal{S}(z))} \times \{\alpha(\mathcal{X})\{\sigma(\mathcal{Z}) - \sigma(\mathcal{X})\} + \{\alpha(\mathcal{S}(z)) - \alpha(\mathcal{S}(x))\}\{\sigma(\mathcal{Z}) - \sigma(\mathcal{S}(x))\}\}. \tag{39}$$

Since the centroids of $\mathcal{Z}, \mathcal{X}, \mathcal{S}(z)$, and $\mathcal{S}(x)$ are all in $\mathbb{D}$, it must be true that the norm of each is bounded above by r. This and (40) imply that

$$||\sigma(\mathcal{S}(z)) - \sigma(\mathcal{S}(x))|| \leq \left\| \frac{1}{\alpha(\mathcal{S}(z))} \right\| \{2r||\alpha(\mathcal{X})|| + 2r||\alpha(\mathcal{S}(z)) - \alpha(\mathcal{S}(x))||\}. \tag{40}$$

But $\|\frac{1}{\alpha(\mathcal{S}(z))}\| \leq \frac{\epsilon}{4r\bar{\epsilon}}$ because of (31); moreover, $||\alpha(\mathcal{X})|| \leq \bar{\epsilon}$ because of (33) and the second inequality in (29). From these inequalities, the first inequality in (29), and (40), it follows that (30) is true. □

Proof of Proposition 4. Note first that each point on the piecewise linear curve c determined by the points $y_1, y_2, \ldots, y_m$ is on a line connecting two of these points. It follows that each point on c is contained in $\langle y_1, y_2, \ldots, y_m \rangle$. Let y be an interior point of $\lceil y_1, y_2, \ldots, y_m \rceil$; in other words, $\mathrm{wn}(y, c) \neq 0$. Because of this, c must encircle y at least once. Since y is an interior point, any line of sufficient length which passes through y must intersect c at a minimum of two distinct points. Since points on c are in $\langle y_1, y_2, \ldots, y_m \rangle$, y must therefore be in $\langle y_1, y_2, \ldots, y_m \rangle$ as well. □

The proof of Proposition 5 depends on the following fact.

LEMMA 4. *Let a, b, c, d be four points in the plane positioned so that the line from a to b intersects the line from c to d, and so that $||a - b|| \leq r$ and $||c - d|| \leq r$. Then*

$$\min\{||a - d||,\ ||b - c||\} \leq r \quad \text{and} \quad \min\{||a - c||,\ ||b - d||\} \leq r. \tag{41}$$

Proof of Lemma 4. Let e denote any point at which the line from a to b intersects the line from c to d. Since $a-d = (a-e)+(e-d)$ and $c-b = (c-e)+(e-b)$, we can use the triangle inequality to get $||a-d|| \leq ||a-e||+||e-d||$ and $||c-b|| = ||c-e||+||e-d||$, respectively. Adding these inequalities yields

$$||a - d|| + ||c - b|| \leq ||a - e|| + ||e - d|| + ||c - e|| + ||e - b||.$$

But because a, b, and e are colinear and c, d, and e are colinear, $||a - e|| + ||e - b|| = ||a - b||$ and $||c - e|| + ||e - d|| = ||c - d||$, respectively. Therefore

$$||a - d|| + ||c - b|| \leq ||a - b|| + ||c - d|| \leq 2r.$$

It follows that either $||a-d|| \leq r$ or $||c-b|| \leq r$. By the same reasoning, either $||a-c|| \leq r$ or $||d-b|| \leq r$. Therefore (41) is true. □

Proof of Proposition 5. Suppose w is within r units of z and is not interior to $\lceil y_1, y_2, \ldots, y_m \rceil$. Then the line connecting z and w must intersect the line from y_j to y_{j+1} for some $j \in \{1, 2, \ldots, m\}$. Since $||y_j - y_{j+1}|| \leq r$, Lemma 4 can be applied with $a = z$, $b = w$, $c = y_j$, and $d = y_{j+1}$. It follows from (41) that either z or w is linked to $\lceil y_1, y_2, \ldots, y_m \rceil$. Therefore w is so linked. □

The proof of Proposition 6 depends on the following lemmas.

LEMMA 5. *Let a, b, and c be three points in the plane such that $||a-b|| \leq r$ and $||a-c|| \leq r$. Any point in the convex hull of a, b, and c is within at most r units of both a and either b or c.*

Proof of Lemma 5. Let $d = \frac{1}{2}(b+c)$. Since $d - a = \frac{1}{2}\{(a-b)+(a-c)\}$, it must be true that $||d-a|| \leq \frac{1}{2}\{||a-b|| + ||a-c||\}$. From this and the hypotheses $||a-b|| \leq r$ and $||a-c|| \leq r$, it follows that $||d-a|| \leq r$. Moreover, from the triangle inequality, $||b-c|| \leq ||b-a|| + ||a-c||$. Therefore $||b-c|| \leq 2r$. Since d is the midpoint between b and c, $||b-d|| \leq r$ and $||c-d|| \leq r$. Thus the sets $\langle a, b, d\rangle$ and $\langle a, c, d\rangle$ each have diameter no greater than r. Since $\langle a, b, c\rangle = \langle a, b, d\rangle \cup \langle a, c, d\rangle$, it follows that any point in $\langle a, b, c\rangle$ must be in $\langle a, b, d\rangle$ or $\langle a, c, d\rangle$ and consequently within r units of a and either b or c. □

LEMMA 6. *Suppose that $z_1, z_2, \ldots, z_k$ are $k > 0$ interior points of a given cycle $\lceil y_1, y_2, \ldots, y_m \rceil$ which are not linked to $\lceil y_1, y_2, \ldots, y_m \rceil$ and which satisfy $||z_1 - z_i|| \leq r$, $i \in \{2, 3, \ldots, k\}$. Then each point in the convex hull $\langle z_1, z_2, \ldots, z_k\rangle$ is an interior point of $\lceil y_1, y_2, \ldots, y_m \rceil$.*

Proof of Lemma 6. Note first that if there is any point $z \in \langle z_1, z_2, \ldots, z_k\rangle$ which is not an interior point of $[y_1, y_2, \ldots, y_m]$, then z would have to be either on or outside of the piecewise linear curve c determined by $y_1, y_2, \ldots, y_m$; in either case this would mean that the line connecting z to any point in $\{z_1, z_2, \ldots, z_k\}$ would have to intersect c since, by assumption, $z_1, z_2, \ldots, z_k$ are interior points of c. Since any such line is contained in $\langle z_1, z_2, \ldots, z_k\rangle$, the convex hull itself would have to intersect c. Thus to prove the lemma it is enough to show that $\langle z_1, z_2, \ldots, z_k\rangle$ does not intersect c. To do this it is sufficient to show that for each pair of points $z_i, z_j \in \{z_1, z_2, \ldots, z_k\}$, the line ℓ_{ij} from z_i to z_j does not intersect c. To do this we suppose the contrary, namely that there is a pair of points $z_i, z_j \in \{z_1, z_2, \ldots, z_k\}$ such that ℓ_{ij} intersects c. Suppose this intersection occurs on the line ℓ between y_q and y_{q+1}.

First consider the case when either z_i or z_j equals z_1 in which case $||z_i - z_j|| \leq r$. To prove that ℓ_{ij} does not intersect ℓ, it is sufficient to prove that for any $s \in \{2, 3, \ldots, k\}$, ℓ_{1s} and ℓ do not intersect. Suppose that for some such s such an intersection exists. Since $||y_q - y_{q+1}|| \leq r$ and $||z_1 - z_s|| \leq r$, Lemma 4 applies with $a = z_1$, $b = z_s$, $c = y_q$, and $d = y_{q+1}$. It follows from (41) that either z_1 or z_s is within r units of either y_q or y_{q+1}. This means that either z_1 or z_s is linked to $\lceil y_1, y_2, \ldots, y_m \rceil$, which is a contradiction. Therefore for any $s \in \{2, 3, \ldots, k\}$, ℓ_{1s} and ℓ do not intersect. In particular, ℓ_{ij} does not intersect ℓ if either z_i or z_j equals z_1.

Now suppose that neither z_i nor z_j equals z_1. From what has just been shown we can conclude that ℓ does not intersect either ℓ_{1i} or ℓ_{1j}. Since ℓ is assumed to intersect ℓ_{ij}, either y_q or y_{q+1} must be in the convex hull $\langle z_1, z_i, z_j\rangle$. But $||z_1 - z_i|| \leq r$ and $||z_1 - z_j|| \leq r$; thus from Lemma 5 we can conclude that either y_q or y_{q+1} must be within r units of z_1. But this is a contradiction of the hypothesis that z_1 is not linked to $\lceil y_1, y_2, \ldots, y_m \rceil$. Hence ℓ_{ij} and ℓ do not intersect. □

LEMMA 7. *For any four points a, b, c, d in $\mathbb{R}^2$, the set $\langle a, b, d\rangle \cup \langle a, c, d\rangle \cup \langle b, c, d\rangle$ is convex.*

Proof of Lemma 7. For the case when $d \in \langle a,b,c\rangle$, the union $\langle a,b,d\rangle \cup \langle a,c,d\rangle \cup \langle b,c,d\rangle$ equals $\langle a,b,c\rangle$ which is convex. Suppose therefore that $d \notin \langle a,b,c\rangle$. Then the line through at least one of the bounding edges of $\langle a,b,c\rangle$—say the edge from b to c—must separate d and $\langle a,b,c\rangle$. We claim that the four- (or less) corner polygon $\mathbb{P} = \langle a,b,d\rangle \cup \langle a,c,d\rangle$ is convex. This certainly must be true if either $c \in \langle a,b,d\rangle$ or $b \in \langle a,c,d\rangle$, since in either case $\mathbb{P}$ would be a polygon with at most three corners. On the other hand, if neither of these cases holds, then the line segment from b to c must lie totally within $\mathbb{P}$. Thus in this case the line segment between any pair of corners of $\mathbb{P}$ must lie completely within $\mathbb{P}$. Since any four- (or less) corner polygon in the plane with this property is necessarily convex, $\mathbb{P}$ is convex. Finally we note that $\langle b,c,d\rangle \subset \mathbb{P}$ because b, c, and d are in $\mathbb{P}$. Thus $\langle a,b,d\rangle \cup \langle a,c,d\rangle \cup \langle b,c,d\rangle = \mathbb{P}$ so $\langle a,b,d\rangle \cup \langle a,c,d\rangle \cup \langle b,c,d\rangle$ is convex as claimed. ∎

LEMMA 8. *Let $\kappa : [0,1] \to \mathbb{R}^2$ be any continuous closed curve, and let a and b be any two distinct points on κ. Let κ_1 be the closed curve consisting of the segment of κ from a to b together with the straight line segment from b to a. Let κ_2 be the closed curve consisting of the segment of κ from b to a together with the straight line segment from a to b. Then for any point $y \in \mathbb{R}^2$ which is not on κ or on the line from a to b,*

$$\mathrm{wn}(y,\kappa) = \mathrm{wn}(y,\kappa_1) + \mathrm{wn}(y,\kappa_2). \tag{42}$$

Proof. Write ϕ_1 and ϕ_2 for the segments of κ from a to b and b to a, respectively, and let ℓ_1 and ℓ_2 denote the line segments from a to b and b to a, respectively. Then

$$\begin{aligned}\mathrm{wn}(y,\kappa_1) + \mathrm{wn}(y,\kappa_2) &= \frac{1}{2\pi j}\left\{\oint_{\tilde{\kappa_1}} \frac{dz}{z-\tilde{y}} + \oint_{\tilde{\kappa_2}} \frac{dz}{z-\tilde{y}}\right\} \\ &= \frac{1}{2\pi j}\left\{\int_{\tilde{\phi_1}} \frac{dz}{z-\tilde{y}} + \int_{\tilde{\ell_1}} \frac{dz}{z-\tilde{y}} + \int_{\tilde{\phi_2}} \frac{dz}{z-\tilde{y}} + \int_{\tilde{\ell_2}} \frac{dz}{z-\tilde{y}} + \right\}.\end{aligned}$$

But

$$\int_{\tilde{\ell_1}} \frac{dz}{z-\tilde{y}} + \int_{\tilde{\ell_2}} \frac{dz}{z-\tilde{y}} = 0,$$

so

$$\mathrm{wn}(y,\kappa_1) + \mathrm{wn}(y,\kappa_2) = \frac{1}{2\pi j}\left\{\int_{\tilde{\phi_1}} \frac{dz}{z-\tilde{y}} + \int_{\tilde{\phi_2}} \frac{dz}{z-\tilde{y}}\right\} = \frac{1}{2\pi j}\oint_{\tilde{\kappa}} \frac{dz}{z-\tilde{y}}$$

from which (42) follows. ∎

LEMMA 9. *Let $\lceil y_1, y_2, \ldots, y_m\rceil$ and $\lceil \bar{y}_1, y_2, \ldots, y_m\rceil$ be cycles such that $||y_1 - \bar{y}_1|| \leq r$. If z is an interior point of $\lceil y_1, y_2, \ldots, y_m\rceil$, then either $||z - \bar{y}_1|| \leq r$ or z is an interior point of $\lceil \bar{y}_1, y_2, \ldots, y_m\rceil$ or both.*

Proof of Lemma 9. Suppose z is not an interior point of $\lceil \bar{y}_1, y_2, \ldots, y_m\rceil$. It is enough to prove that $||z - \bar{y}_1|| \leq r$. Towards this end let c, $\bar{c}$, c_1, c_2, and $\bar{c}_2$ denote the piecewise linear closed curves determined by the ordered point sets $\{y_1, y_2, \ldots, y_m\}$, $\{\bar{y}_1, y_2, \ldots, y_m\}$, $\{y_2, y_3, \ldots, y_m\}$, $\{y_1, y_2, y_m\}$, and $\{\bar{y}_1, y_2, y_m\}$, respectively.

Suppose first that z is inside or on c_2; that is, $z \in \langle y_1, y_2, y_m\rangle$. By Lemma 7, $\langle y_1, y_2, \bar{y}_1\rangle \cup \langle y_1, y_m, \bar{y}_1\rangle \cup \langle y_2, y_m, \bar{y}_1\rangle$ is a convex set. Thus $\langle y_1, y_2, y_m\rangle \subset \langle y_1, y_2, \bar{y}_1\rangle \cup \langle y_1, y_m, \bar{y}_1\rangle \cup \langle y_2, y_m, \bar{y}_1\rangle$ because y_1, y_2, and y_m are all in the union. Therefore $z \in \langle y_1, y_2, \bar{y}_1\rangle \cup \langle y_1, y_m, \bar{y}_1\rangle \cup \langle y_2, y_m, \bar{y}_1\rangle$. We have assumed $||y_1 - \bar{y}_1|| \leq r$. Moreover, $||\bar{y}_1 - y_2|| \leq r$ and $||\bar{y}_1 - y_m|| \leq r$ because $\lceil \bar{y}_1, y_2, \ldots, y_m\rceil$ is assumed to be a cycle.

Thus no matter whether z is in $\langle y_1, y_2, \bar{y}_1\rangle$, $\langle y_1, y_m, \bar{y}_1\rangle$, or $\langle y_2, y_m, \bar{y}_1\rangle$, Lemma 5 applies, and it can be concluded that $||z - \bar{y}_1|| \leq r$ as claimed.

Consider next the case when z is outside of c_2; in other words, $\text{wn}(z, c_2) = 0$. Since z is not on c_2, it is clearly not on the line segment from y_2 to y_m. Therefore Lemma 8 can be applied to c, c_1, and c_2 providing that $\text{wn}(z, c) = \text{wn}(z, c_1) + \text{wn}(z, c_2)$. But by assumption z is an interior point of $\lceil y_1, y_2, \ldots, y_m\rceil$, so $\text{wn}(z, c) \neq 0$. Therefore

$$\text{wn}(z, c_1) \neq 0. \tag{43}$$

By assumption, z is not an interior point of $\lceil \bar{y}_1, y_2, \ldots, y_m\rceil$. Thus z must be either on $\bar{c}$ or outside of $\bar{c}$. If z is on $\bar{c}$, then it is linked to $\{\bar{y}_1, y_2, \ldots, y_m\}$. On the other hand, if z is outside of $\bar{c}$, then $\text{wn}(z, \bar{c}) = 0$. Moreover, in this case $\text{wn}(z, \bar{c}) = \text{wn}(z, c_1) + \text{wn}(z, \bar{c}_2)$ because of Lemma 8. From this and (43), it follows that $\text{wn}(z, \bar{c}_2) \neq 0$. Thus z is inside of $\bar{c}_2$. But $\lceil \bar{y}_1, y_2, \ldots, y_m\rceil$ is a cycle, so $||\bar{y}_1 - y_2|| \leq r$ and $||y_m - \bar{y}_1|| \leq r$. From this and Lemma 5, it follows that $||z - \bar{y}_1|| \leq r$ as claimed. □

Proof of Proposition 6. Consider the sequence of cycles $\lceil y_1, y_2, \ldots, y_m\rceil$, $\lceil \bar{y}_1, y_2, \ldots, y_m\rceil$, $\lceil \bar{y}_1, \bar{y}_2, \ldots, y_m\rceil, \ldots, \lceil \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m\rceil$, each being a successor of the one before it. Let z be any point in $\langle z_1, z_2, \ldots, z_k\rangle$. By Lemma 6, z is an interior point of $\lceil y_1, y_2, \ldots, y_m\rceil$. Therefore by Lemma 9, either $||z - \bar{y}_1|| \leq r$ or z is an interior point of $\lceil \bar{y}_1, y_2, \ldots, y_m\rceil$. If the former is true, then z is clearly linked to $\lceil \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m\rceil$. On the other hand, if the latter is true, Lemma 9 can again be used, this time to reach the conclusion that either $||z - \bar{y}_2|| \leq r$ or z is an interior point of $\lceil \bar{y}_1, \bar{y}_2, \ldots, y_m\rceil$ which is not linked to $\lceil \bar{y}_1, \bar{y}_2, \ldots, y_m\rceil$. Continuing this process a finite number of times completes the proof. □

The proof of Proposition 7 is a simple consequence of the following lemmas.

LEMMA 10. *Let $\mathcal{S}$ be a closed, bounded convex set in $\mathbb{R}^m$. If x and y are vectors in $\mathcal{S}$ for which*

$$||x - y|| = \text{diameter}(\mathcal{S}), \tag{44}$$

then x and y are corners of $\mathcal{S}$.

Proof of Lemma 10. Suppose (44) holds. It is enough to show that y is a corner of $\mathcal{S}$. Suppose that it is not. Then there must be *distinct* vectors x_1 and x_2 in $\mathcal{S}$ and a number $\alpha \in (0, 1)$ for which $y = \alpha x_1 + (1 - \alpha)x_2$. In view of (44) and the definition of $\text{dia}(\mathcal{S})$, the function $f(\lambda) \triangleq ||x - \lambda x_1 - (1 - \lambda)x_2||^2$ must attain its maximum on $[0, 1]$ at the interior point $\lambda = \alpha$. But this is impossible because $f(\lambda)$ is a nonconstant, convex function of λ. Therefore, by contradiction y must be a corner of $\mathcal{S}$. □

LEMMA 11. *Let $\{x_1, x_2, \ldots, x_n\} \in \mathcal{X}$ be fixed. Then*

$$\text{dia}\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\} \leq \text{dia}\{x_1, x_2, \ldots, x_n\}, \tag{45}$$

where for $i \in \{1, 2, \ldots, n\}$,

$$\bar{x}_i = x_i + u_{m_i}(x_{i_1} - x_i,\ x_{i_2} - x_i, \ldots,\ x_{i_{m_i}} - x_i). \tag{46}$$

Moreover, if $\mathbb{G}$ is connected, then either the inequality in (45) *is strict or $x_1 = x_2 = \cdots = x_n$.*

Proof of Lemma 11. By definition, for $m > 0$, $u_m(\cdot)$ maps the vectors $z_i \in \mathbb{D}$, $i \in \{1, 2, \ldots, m\}$, into a point $\bar{z}$ in the convex hull $\langle 0, z_1, z_2, \ldots z_m\rangle$; moreover, $\bar{z}$ is not a corner of $\langle 0, z_1, z_2, \ldots, z_m\rangle$ unless $z_1 = z_2 = \cdots = z_m = 0$. In the present context this means that for $i \in \{1, 2, \ldots, n\}$, $x_i + u_{m_i}(\cdot)$ maps the vectors $x_{i_j} - x_i \in \mathbb{D}, j \in$

$\{1, 2, \ldots, m_i\}$, into the point $\bar{x}_i$ in the convex hull $\langle x_i, x_{i_1}, x_{i_2}, \ldots, x_{i_{m_i}}\rangle$; moreover, $\bar{x}_i$ is not a corner of $\langle x_i, x_{i_1}, x_{i_2}, \ldots, x_{i_{m_i}}\rangle$ unless $x_i = x_{i_1} = x_{i_2} = \cdots = x_{i_{m_i}}$. Since each $\langle x_i, x_{i_1}, x_{i_2}, \ldots, x_{i_{m_i}}\rangle$ is a subset of $\langle x_1, x_2, \ldots, x_n\rangle$, it must be true that

$$\langle \bar{x}_1, \bar{x}_2, \ldots \bar{x}_n\rangle \subset \langle x_1, x_2, \ldots x_n\rangle. \tag{47}$$

Moreover, $\bar{x}_i$ is not a corner of $\langle x_1, x_2, \ldots, x_n\rangle$ unless $x_i = x_{i_1} = x_{i_2} = \cdots = x_{i_{m_i}}$. Inequality (45) is a direct consequence of (47).

Now suppose that $\mathbb{G}$ is connected and that the x_i are not all equal. Then for each $i \in \{1, 2, \ldots, n\}$, there is at least one $i_j \in \{i_1, i_2, \ldots, i_{m_i}\}$ for which $x_{i_j} \neq x_i$. This means that it cannot be true that $x_i = x_{i_1} = x_{i_2} = \cdots = x_{i_{m_i}}$ for any value of $i \in \{1, 2, \ldots, n\}$. Therefore $\bar{x}_i,\ i \in \{1, 2, \ldots, n\}$, is not a corner of $\langle x_1, x_2, \ldots, x_n\rangle$. From this and Lemma 10 it follows that the inequality in (45) is strict. □

It is worth noting that (47) establishes that the sequence of convex hulls of agent positions generated on successive steps must form a descending chain of convex sets. As a consequence, one can conclude at once that the sequence has a limit set $\mathcal{H}$ into which all agents must eventually move and remain. While this fact does not depend upon the $u_m(\cdot)$ being continuous, the fact that $\mathcal{H}$ is actually a single point does.

Proof of Proposition 7. Note that (11) implies that

$$\operatorname{dia}\{x_1, x_2, \ldots, x_n\} = \operatorname{dia}\{e_1, e_2, \ldots, e_{n-1}, 0\}$$

because the diameter of a convex set is invariant under translation. Therefore

$$V(e) = \operatorname{dia}\{x_1, x_2, \ldots, x_n\}. \tag{48}$$

Next observe that Lemma 11 says that

$$\operatorname{dia}\{x_1 + f_1(e), x_2 + f_2(e), \ldots, x_n + f_n(e)\} \leq \operatorname{dia}\{x_1, x_2, \ldots, x_n\} \tag{49}$$

with the inequality being strict if $\mathbb{G}$ is connected. But

$$\begin{aligned}
&\operatorname{dia}\{x_1 + f_1(e), x_2 + f_2(e), \ldots, x_n + f_n(e)\} \\
&\quad = \operatorname{dia}\{e_1 + f_1(e) - f_n(e), e_2 + f_2(e) - f_n(e), \ldots, x_{n-1} + f_{n-1}(e) - f_n(e), 0\} \\
&\quad = V(e + f(e)).
\end{aligned}$$

From this, (49), and (48) it is clear that

$$V(e + f(e)) - V(e)$$

is a negative semidefinite function and actually a negative definite function if $\mathbb{G}$ is connected. □

REFERENCES

[1] H. Ando, Y. Oasa, I. Suzuki, and M. Yamashita, *Distributed memoryless point convergence algorithm for mobile robots with limited visibility*, IEEE Trans. Robotics Automation, 15 (1999), pp. 818–828.

[2] D. E. Chang and J. E. Marsden, *Gyroscopic forces and collision avoidance with convex obstacles*, in New Trends in Nonlinear Dynamics and Control, and Their Applications, Springer-Verlag, Berlin, 2003, pp. 145–159.

[3] J. Cortes, S. Martinez, and F. Bullo, *Robust rendezvous for mobile autonomous agents via proximity graphs in arbitrary dimensions*, IEEE Trans. Automat. Control, 51 (2006), pp. 1289–1296.
[4] J. P. Desai, J. P. Ostrowski, and V. Kumar, *Modeling and control of formations of nonholonomic mobile robots*, IEEE Trans. Robotics Automation, 17 (2001), pp. 905–908.
[5] T. Eren, P. N. Belhumeur, B. D. O. Anderson, and A. S. Morse, *A framework for maintaining formations based on rigidity*, in Proceedings of the 15th IFAC World Congress, Barcelona, Spain, 2002, pp. 2752–2757.
[6] T. Eren, P. N. Belhumeur, and A. S. Morse, *Closing ranks in vehicle formations based on rigidity*, in Proceedings of the 41st IEEE Conference on Decision and Control, 2002, pp. 2959–2964.
[7] J. A. Fax and R. M. Murray, *Graph Laplacians and Vehicle Formation Stabilization*, CDS Technical Report 01-007, California Institute of Technology, Pasadena, CA, 2001.
[8] J. A. Fax and R. M. Murray, *Graph Laplacians and stabilization of vehicle formations*, in Proceedings of the 15th IFAC World Congress, Barcelona, Spain, 2002.
[9] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.
[10] N. Leonard and E. Friorelli, *Virtual leaders, artificial potentials and coordinated control of groups*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001.
[11] J. Lin, A. S. Morse, and B. D. O. Anderson, *The multi-agent rendezvous problem*, in Proceedings of the 42nd IEEE Conference on Decision and Control, 2003, pp. 1508–1513.
[12] J. Lin, A. S. Morse, and B. D. O. Anderson, *The multi-agent rendezvous problem. Part* 2: *The asynchronous case*, SIAM J. Control Optim., 46 (2007), pp. 2120–2147.
[13] Z. Lin, M. Brouche, and B. Francis, *Local Control Strategies for Groups of Mobile Autonomous Agents*, in Proceedings of the 42nd IEEE Conference on Decision and Control, 2003, pp. 1006–1011.
[14] Y. Liu, K. M. Passino, and M. Polycarpou, *Stability analysis of one-dimensional asynchronous swarms*, IEEE Trans. Automat. Control, 48 (2003), pp. 1848–1854.
[15] J. E. Marsden, *Basic Complex Analysis*, W. H. Freeman, San Francisco, 1973.
[16] L. Moreau, *Stability of multi-agent systems with time-dependent communication links*, IEEE Trans. Automat. Control, 50 (2005), pp. 169–182.
[17] H. Rademacher and O. Toeplitz, *The Enjoyment of Mathematics*, Princeton University Press, Princeton, NJ, 1957.
[18] R. K. Sundaram, *A First Course in Optimization Theory*, Cambridge University Press, Cambridge, UK, 1996.
[19] I. Suzuki and M. Yamashita, *Distributed autonomous mobile robots: Formation of geometric patterns*, SIAM J. Comput., 28 (1999), pp. 1347–1363.
[20] H. Tanner, A. Jadbabaie, and G. Pappas, *Distributed Coordination Strategies for Groups of Mobile Autonomous Agents*, Technical report, ESE Department, University of Pennsylvania, Philadelphia, 2002.
[21] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, *Novel type of phase transition in a system of self-driven particles*, Phys. Rev. Lett., 75 (1995), pp. 1226–1229.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2120–2147

THE MULTI-AGENT RENDEZVOUS PROBLEM. PART 2: THE ASYNCHRONOUS CASE*

J. LIN†, A. S. MORSE‡, AND B. D. O. ANDERSON§

Abstract. This paper is concerned with the collective behavior of a group of $n > 1$ mobile autonomous agents, labelled 1 through n, which can all move in the plane. Each agent is able to continuously track the positions of all other agents currently within its "sensing region," where by an agent's *sensing region* we mean a closed disk of positive radius r centered at the agent's current position. The *multi-agent rendezvous problem* is to devise "local" control strategies, one for each agent, which without any active communication between agents cause all members of the group to eventually rendezvous at a single unspecified location. This paper describes a family of unsynchronized strategies for solving the problem. Correctness is established appealing to the concept of "analytic synchronization."

Key words. cooperative control, distributed control, multi-agent systems, asynchronous systems

AMS subject classifications. 93C65, 93C85, 93C55

DOI. 10.1137/040620564

1. Introduction.

This paper is concerned with the collective behavior of a group of $n > 1$ mobile autonomous agents, labelled 1 through n, which can all move in the plane. Each agent is able to continuously track the positions of all other agents currently within its "sensing region," where by an agent's *sensing region* we mean a closed disk of positive radius r centered at the agent's current position. The *multi-agent rendezvous problem* is to devise "local" control strategies, one for each agent, which without any active communication between agents cause all members of the group to eventually rendezvous at a single unspecified location.

The rendezvous problem, which is also sometimes called a "gathering problem," has been studied before assuming that all agents possess either unlimited visibility (e.g., $r = \infty$) [4] or a common sense of direction [9, 5] or both; see [5] for additional references. The problem has also been addressed before without making either of these assumptions [1, 8]. This paper also treats the case in which individual agents have limited visibility and distinct frames of reference. What distinguishes this work from that in [1, 8] is that individual agents clocks are taken to be unsynchronized. These three features, namely limited sensing, no common frame of reference or sense of direction, and no common clock, are of obvious practical importance but have apparently not been dealt with before at the same time as components of one multi-

*Received by the editors December 9, 2004; accepted for publication (in revised form) May 29, 2007; published electronically December 21, 2007. A preliminary version of this paper appears in [7]. The research of the first two authors was supported by the US Army Research Office, the US National Science Foundation, and by a gift from Xerox Corporation.

http://www.siam.org/journals/sicon/46-6/62056.html

†Xerox Corporation, 800 Phillips Road, MS:0128-30E, Webster, NY 14580. Current address: 21/F Central Plaza, 227 Huangpi Bei Lu, Shanghai 200003, People's Republic of China (jie.lin@aya.yale.edu).

‡Yale University, PO Box 208267, New Haven, CT 06520 (morse@sysc.eng.yale.edu).

§Australian National University and National ICT Australia Ltd., Locked Bag 8001, Canberra ACT 2601, Australia (Brian.Anderson@nicta.com.au). This author's research was supported by National ICT Australia, which is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through the Backing Australia's Ability initiative and the ICT Centre of Excellence Program.

agent rendezvous problem.

As in [1, 8], we consider distributed strategies which guide each agent toward rendezvous by performing a sequence of "stop-and-go" maneuvers. A *stop-and-go maneuver* takes place within a time interval consisting of two consecutive subintervals. The first, called a *sensing period*, is an interval of fixed length during which the agent is stationary. The second, called a *maneuvering period*, is an interval of variable length during which the agent moves from its current position to its next "way-point" and again comes to rest. Successive way-points for each agent are chosen to be within r_M units of each other, where r_M is a prespecified positive distance no larger than r. It is assumed that there has been chosen for each agent i a positive number τ_{M_i}, called a *maneuver time*, which is large enough so that the required maneuver for agent i from any one way-point to the next can be accomplished in at most τ_{M_i} seconds. Since our interest here is exclusively with devising *high level* strategies which dictate when and where agents are to move, we will use point models for agents and shall not deal with how maneuvers are actually carried out or with how vehicle collisions are to be avoided.

In the synchronous case treated in [1, 8], the kth maneuvering period of each agent is synchronized to begin at the same time $\bar{t}_k$ as the kth maneuvering period of every other agent. Agent i's registered neighbors at the beginning of its kth maneuvering period are taken to be all those other agents positioned within agent i's sensing region at the beginning of the period. Because of synchronization, this notion of a registered neighbor induces a *symmetric* relation on the agent group in that agent j is a registered neighbor of agent i at the beginning of maneuvering period k just in case agent i is a registered neighbor of agent j at the same time. As a result, it is possible to characterize neighbor relationships at time $\bar{t}_k$ with a simple graph whose vertices represent agents and whose edges represent existing neighbor relationships [8]. Although the neighbor relation is symmetric, it is clearly not transitive. On the other hand, if agent i is at the same position as neighbor j at time $\bar{t}_k$, then any registered neighbor of agent j at time $\bar{t}_k$ certainly must be a registered neighbor of agent i at the same time. It is precisely because of this *weak transitivity* property that one can infer a *global* condition of the entire synchronized agent group from a *local* condition of one agent and its neighbors. In particular, if the graph characterizing neighbor relationships at time $\bar{t}_k$ is connected, and any one agent is at the same position as all of its neighbors, then the weak transitivity property guarantees at once that all n agents have rendezvoused at time $\bar{t}_k$.

Our aim in this paper is to relax the synchronization requirement. In particular we will not require synchronization of the start times of the maneuvering periods of different agents. To accomplish this it is necessary to modify somewhat what is meant by a registered neighbor of agent i at time $\bar{t}_{ik}$ where for the asynchronous case under consideration, $\bar{t}_{ik}$ denotes the time at which agent i's kth maneuvering period begins. Our definition is guided by considerations discussed above for the synchronous case. For example, the new definition is crafted to retain versions of the symmetry and weak transitivity properties of the registered neighbor relation inherent in the synchronous case. Doing this is challenging, because unlike the synchronous case, the times each agent registers its neighbors and its neighbors' positions are not synchronized with the times its neighbors do the same thing.

Exactly the same way-point update rules considered in the synchronous case [8] are adopted for the asynchronous case. Thus the only functional differences between the two cases are the definitions of registered neighbors and registered neighbor po-

sitions. Of course in the asynchronous case, way-point updates are computed asynchronously, whereas in the synchronous case they are not.

Not surprisingly, the analysis of the asynchronous version of the problem is considerably more challenging than that of the synchronous version. For example, while it is more or less obvious in the synchronous case that the proposed way-point update rules causes all agents to retain their neighbors as the system evolves [8], proving that this is also true in the asynchronous case involves a number of steps.

Just as in the synchronous case, it is possible to characterize neighbor relationships with a graph. This is done in section 3 by first merging together into a single ordered time set the distinct "event times" $\bar{t}_{ik}$, $i \in \{1, 2, \ldots, n\}$, $k \geq 1$, generated by all n agents. The elements of this set are then relabelled as $t_1, t_2, \ldots$ in such a way so that $t_j < t_{j+1}$, $j \in \{1, 2, \ldots\}$. With this notation, agent i's registered neighbors at its kth event time $\bar{t}_{ik}$ are its registered neighbors at time $t_{P_i(k)}$, where $P_i(k)$ denotes that value of p for which $t_p = \bar{t}_{ik}$. For each $i \in \{1, 2, \ldots, n\}$, the domain of definition of agent i's registered neighbors is then extended from the set $\{t_{P_i(k)} : k \geq 1\}$ to the set $\{t_p : p \geq P_i(1)\}$ by stipulating that for values of t_p which are between two successive event times of agent i, say between $\bar{t}_{ik}$ and $\bar{t}_{i(k+1)}$, agent i's registered neighbors are the same as its registered neighbors at time $\bar{t}_{ik}$. This means that registered neighbors of each agent are defined at each time $t_p \geq t_{\bar{p}}$, where $\bar{p} \triangleq \max\{P_1(1), P_2(1), \ldots P_n(1)\}$. Because of this, it is possible to describe neighbor relationships with a directed graph with vertex set $\{1, 2, \ldots, n\}$ and directed edge set defined so that (i, j) is a directed edge from vertex i to vertex j just in case agent j is a registered neighbor of agent i at event time t_s. The main result of this paper (Corollary 1) is that if this graph is ever strongly connected, then rendezvous of all n agents will eventually occur.

Establishing the correctness of Corollary 1 requires the analysis of the asymptotic behavior of the *asynchronous* process which describes the n-agent system. Despite the apparent complexity of this process, it is possible to capture its salient features using a suitably defined *synchronous* discrete-time, hybrid dynamical system $\mathbb{S}$. We call the sequence of steps involved in defining $\mathbb{S}$ *analytic synchronization*. Analytic synchronization is applicable to any finite family of continuous or discrete-time dynamical processes $\{\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_n\}$ under the following conditions. First, each process $\mathbb{P}_i$ must be a dynamical system whose inputs consist of functions of the states of the other processes as well as signals which are exogenous to the entire family. Second, each process $\mathbb{P}_i$ must have associated with it an ordered sequence of event times $\{t_{i1}, t_{i2}, \ldots\}$ defined in such a way so that the state of $\mathbb{P}_i$ at event time $t_{i(k_i+1)}$ is uniquely determined by values of the exogenous signals and states of the $\mathbb{P}_j$, $j \in \{1, 2, \ldots, n\}$, at event times t_{jk_j} which occur prior to $t_{i(k_i+1)}$ but in the finite past. Event time sequences for different processes need not be synchronized. Analytic synchronization is a procedure for creating a single synchronous process for purposes of analysis which captures the salient features of the original n asynchronously functioning processes. As a first step, all n event time sequences are merged into a single ordered sequence of even times $\mathcal{T}$. This clever idea has been used before in [2] to study the convergence of totally asynchronous iterative algorithms. Second, the "synchronized" state of $\mathbb{P}_i$ is then defined to be the original of $\mathbb{P}_i$ at $\mathbb{P}_i$'s event times $\{t_{i1}, t_{i2}, \ldots\}$ plus possibly some additional state variables; at values of $t \in \mathcal{T}$ between event times t_{ik_i} and $t_{i(k_i+1)}$, the synchronized state of $\mathbb{P}_i$ is taken to be the same as the value of its original state at time t_{ik}. Although it is not always possible to carry out all of these steps, in this case it is. What ultimately results is a synchronous dynamical system evolving on $\mathcal{T}$ with a state composed of the synchronized states of the n individual processes under con-

sideration. The definition of $\mathbb{S}$ in section 4.1 illustrates the analytic synchronization procedure.

2. The asynchronous agent system. The strategy analyzed in [1, 8] cannot be regarded as truly distributed because each agent's decisions must be synchronized to a common clock shared by all other agents in the group. In what follows we redefine the strategies so that a common clock is not required. To do this it will be necessary to modify somewhat what is meant by a registered neighbor and by a registered neighbor's position.

For each agent i, the real time axis can be partitioned into a sequence of time intervals $[0, t_{i1}), [t_{i1}, t_{i2}), \ldots, [t_{i(k-1)}, t_{ik}), \ldots$, each of length at most $\tau_D + \tau_{M_i}$, where τ_D is a number greater than τ_{M_i} called a *dwell time*. Each interval $[t_{i(k-1)}, t_{ik})$ consists of a *sensing period* $[t_{i(k-1)}, \bar{t}_{ik})$ of fixed length τ_D during which agent i is stationary, followed by a maneuvering period $[\bar{t}_{ik}, t_{ik})$ of length at most τ_{M_i} during which agent i moves from its current position to its next way-point. Although all agents use the same dwell time, they operate asynchronously in the sense that the time sequences $t_{i1}, t_{i2}, \ldots$, $i \in \{1, 2, \ldots, n\}$, are uncorrelated. Thus each agent's strategy can be implemented independent of the rest, without the need for a common clock.

2.1. Registered neighbors. Because of the asynchronous nature of the control strategies under consideration, care must be exercised in defining what is meant by a registered neighbor if one is to end up with something similar to the symmetry property of the neighbor relationship defined in the synchronous case. For the asynchronous case, agent i's *registered neighbors* at time $\bar{t}_{ik}$ (i.e., at the beginning of its kth maneuvering period $[\bar{t}_{ik}, t_{ik})$) are taken to be those agents which are fixed at one position within agent i's sensing region for at least $\tau_S > 0$ seconds during agent i's kth sensing period $\mathcal{S}_i(k) \triangleq [t_{i(k-1)}, \bar{t}_{ik})$. Here τ_S is a positive number called a *sensing time*. For reasons to be made clear below, we shall require τ_S to satisfy

$$\tau_S \leq \frac{1}{2}(\tau_D - \tau_{M_i}) \qquad \forall i \in \{1, 2, \ldots, n\}. \tag{1}$$

Note that this implies that $\tau_D > \tau_{M_i}$, $i \in \{1, 2, \ldots, n\}$, which means that the n agents spend more time between successive way-points sensing their neighbors' positions than they do maneuvering between successive way-points. For any agent j, there may be more than one distinct interval of length at least τ_S within $\mathcal{S}_i(k)$ during which agent j is stationary. Let t^* denote the end time of the last of these. For purposes of calculation, agent i takes the *registered position* of agent j at the beginning of its kth maneuvering period to be the actual position of agent j at *registration time* t^*. To attain a symmetry-like property for the asynchronous case, it is necessary to make sure that the *registration interval* $[t^* - \tau_S,\ t^*)$ lies within one of agent j's sensing periods. One way to guarantee this is to require each agent to keep moving during each of its maneuvering periods except possibly for brief periods which are each shorter than τ_S. Another way is to equip each agent with a signaling device (such as a light in the case of visual sensing) which is on just in case the agent is in one of its sensing periods. In what follows we will assume that registration of each agent j during one of agent i's sensing periods always occurs at the end of a registration interval $[t^* - \tau_S,\ t^*)$ which also lies within one of agent j's sensing periods. Note that this and the requirement that agent j be stationary during its sensing periods together imply that agent j's registered position $x_j(t^*)$ is equal to $x_j(\bar{t}_{jk^*})$, where k^* is the sensing/maneuvering interval of agent j during which registration takes place.

2.1.1. Neighbor characterization. Prompted by the preceding, let us agree to say that for each $i, j \in \{1, 2, \dots, n\}$, agent j's *pth* sensing period $\mathcal{S}_j(p)$ *strongly overlaps* agent i's kth sensing period $\mathcal{S}_i(k)$ if the overlap $\mathcal{S}_j(p) \cap \mathcal{S}_i(k)$ is a nonempty interval of length at least τ_S seconds. In what follows we write $\mathcal{S}_j(p) \cap \mathcal{S}_i(k) \succ \tau_S$ whenever $\mathcal{S}_i(k)$ and $\mathcal{S}_j(p)$ strongly overlap. Let us note that because all sensing periods of all agents are τ_D seconds long, the largest number of sensing periods of any one agent which a given sensing period of agent i can overlap is two. On the other hand, each sensing period of agent i must strongly overlap at least one sensing period of every other agent. To understand why this is so, note first that the maximal possible amount of time between two successive sensing periods of agent j is τ_{M_j}, but τ_{M_j} is bounded above by $\tau_D - 2\tau_S$ because of (1). Thus the maximal possible amount of time between two successive sensing periods of agent j is no greater than $\tau_D - 2\tau_S$. Given this and the fact that all sensing periods are τ_D seconds long, it follows that each sensing period of agent i must strongly overlap at least one sensing period of each agent j.

It is possible to be more explicit about which sensing periods of agent j overlap $\mathcal{S}_i(k)$. For each $i, j \in \{1, 2, \dots, n\}$ and each $k \geq 1$, let $\lceil \bar{t}_{ik} \rceil_j$ denote the smallest integer q such that $\bar{t}_{jq} \geq \bar{t}_{ik}$. In other words, $\lceil \bar{t}_{ik} \rceil_j$ is the unique integer for which $\bar{t}_{ik} \in (\bar{t}_{j(q-1)}, \bar{t}_{jq}]$. Set $q = \lceil \bar{t}_{ik} \rceil_j$. In view of the definition of $\lceil \cdot \rceil_j$ and the preceding discussion it is clear that the only sensing periods of agent j which $\mathcal{S}_i(k)$ can overlap are $\mathcal{S}_j(q-1)$ and $\mathcal{S}_j(q)$; moreover, $\mathcal{S}_i(k)$ must strongly overlap one of these. There are three possible situations which might occur. In the first, shown in Figure 1(a), the only sensing period of agent j which overlaps $\mathcal{S}_i(k)$ is $\mathcal{S}_j(q-1)$; in this case the length of the overlap is $\tau_D - (\bar{t}_{ik} - \bar{t}_{j(q-1)})$, and this length will always be greater than or equal to τ_S. Therefore in this situation, $\mathcal{S}_i(k)$ and $\mathcal{S}_j(q-1)$ strongly overlap. For the second situation, shown in Figure 1(b), the only sensing period of agent j which overlaps $\mathcal{S}_i(k)$ is $\mathcal{S}_j(q)$; in this case the length of the overlap is $\tau_D - (\bar{t}_{jq} - \bar{t}_{ik})$, and this length will also always be greater than or equal to τ_S. Therefore in this situation $\mathcal{S}_i(k)$ and $\mathcal{S}_j(q)$ strongly overlap. The only other possible situation that can occur, which is shown in Figure 1(c), is when $\mathcal{S}_i(k)$ is overlapped by both $\mathcal{S}_j(q-1)$ and $\mathcal{S}_j(q)$. In this case the lengths of the first and second overlapping intervals are $\tau_D - (\bar{t}_{j(q-1)} - \bar{t}_{ik})$ and $\tau_D - (\bar{t}_{ik} - \bar{t}_{jq})$, respectively, and at least one of these lengths will always be greater than or equal to τ_S. Thus in this situation, $\mathcal{S}_i(k)$ strongly overlaps $\mathcal{S}_j(q-1)$ or $\mathcal{S}_j(q)$ or both. We summarize.

LEMMA 1 (overlaps). *Let i and j be distinct integers in $\{1, 2, \dots, n\}$. Let $\bar{t}_{ik}$ be fixed and define $q = \lceil \bar{t}_{ik} \rceil_j$. The only possible sensing periods of agent j which $\mathcal{S}_i(k)$ can overlap are $\mathcal{S}_j(q-1)$ and $\mathcal{S}_j(q)$; moreover, $\mathcal{S}_i(k)$ must strongly overlap at least one of these. In addition,*

1. *$\mathcal{S}_i(k) \cap \mathcal{S}_j(q) \succ \tau_S$ if and only if $\bar{t}_{jq} - \bar{t}_{ik} \leq \tau_D - \tau_S$;*
2. *$\mathcal{S}_i(k) \cap \mathcal{S}_j(q-1) \succ \tau_S$ if and only if $\bar{t}_{ik} - \bar{t}_{j(q-1)} \leq \tau_D - \tau_S$.*

Note that for agent j to be a registered neighbor of agent i at the beginning of agent i's kth maneuvering period, it is necessary and sufficient that agent j be "within range of agent i" (i.e., within agent i's sensing region) during a sensing period of agent j which strongly overlaps $\mathcal{S}_i(k)$. Consider again Figure 1 where $q = \lceil \bar{t}_{ik} \rceil_j$. In the situation depicted in Figure 1(a), agent j will be a registered neighbor of agent i just in case $||x_i(\bar{t}_{ik}) - x_j(\bar{t}_{j(q-1)})|| \leq r$; moreover, if this condition holds, $x_j(\bar{t}_{j(q-1)})$ will be the registered position of agent j. Similarly in the situation shown in Figure 1(b), agent j will be a registered neighbor of agent i just in case $||x_j(\bar{t}_{jq}) - x_i(\bar{t}_{ik})|| \leq r$; moreover, if this condition holds, $x_j(\bar{t}_{jq})$ will be the registered position of agent j.

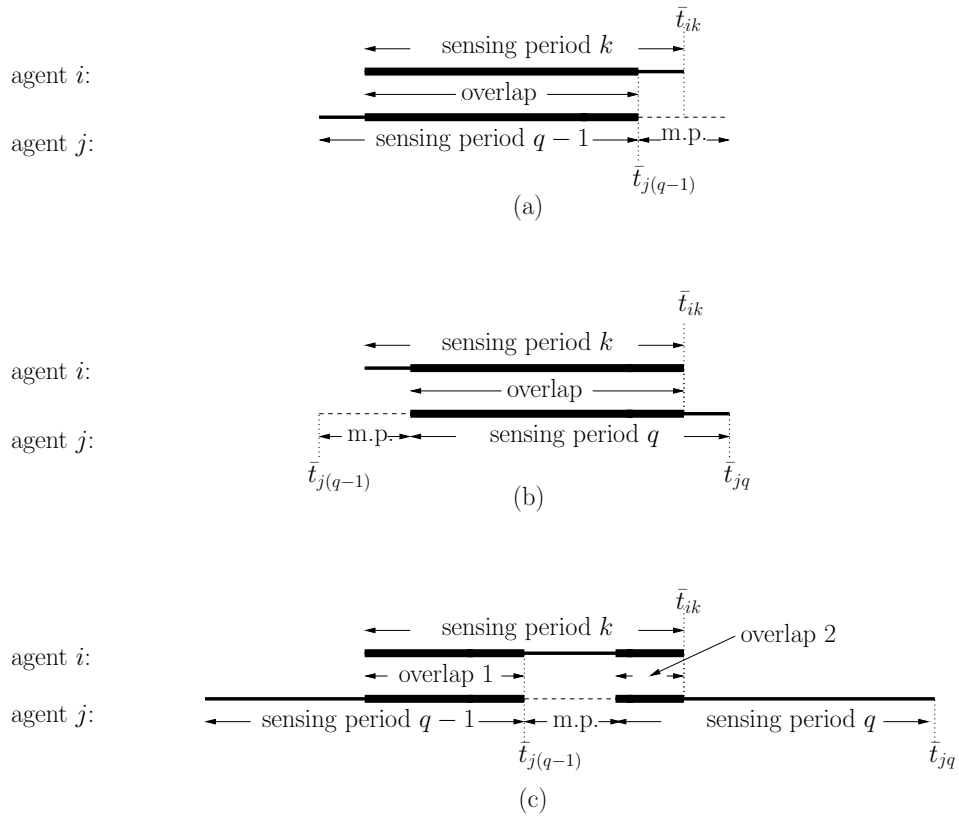


FIG. 1. *Sensing period overlaps.*

The remaining situation shown in Figure 1(c) is slightly more complicated. If, on the one hand, the length of the second overlap is greater than or equal to τ_S and $||x_j(\bar{t}_{jq}) - x_i(\bar{t}_{ik})|| \leq r$, then agent j will be a registered neighbor of agent i with registered position $x_j(\bar{t}_{jq})$. If either of these two conditions fails to hold, if the length of the first overlap is greater than or equal to τ_S, and if $||x_i(\bar{t}_{ik}) - x_j(\bar{t}_{j(q-1)})|| \leq r$, then agent j will be a registered neighbor of agent i, and $x_j(\bar{t}_{j(q-1)})$ will be its registered position. The following proposition summarizes these observations.

PROPOSITION 1 (neighbor characterization). *Let $i, j \in \{1, 2, \ldots, n\}$ and $\bar{t}_{ik}$ be fixed and let $q = \lceil \bar{t}_{ik} \rceil_j$. Then agent j is a registered neighbor of agent i at the beginning of agent i's kth maneuvering period if and only if at least one of the following is true.*

(A) $\mathcal{S}_i(k) \cap \mathcal{S}_j(q) \succ \tau_S$ *and* $||x_j(\bar{t}_{jq}) - x_i(\bar{t}_{ik})|| \leq r$.

(B) $\mathcal{S}_i(k) \cap \mathcal{S}_j(q-1) \succ \tau_S$ *and* $||x_i(\bar{t}_{ik}) - x_j(\bar{t}_{j(q-1)})|| \leq r$.

Moreover, if (A) *is true, then $x_j(\bar{t}_{jq})$ is the registered position of agent j at the beginning of agent i's kth maneuvering period, and if* (A) *is not true while* (B) *is, then $x_j(\bar{t}_{j(q-1)})$ is the registered position of agent j at the beginning of agent i's kth maneuvering period.*

2.1.2. Neighbor relationship symmetry. The definition of a registered neighbor determines a relationship between agents similar to the symmetric relationship determined by the definition of a registered neighbor in the synchronous case [8]. Suppose that agent j is a registered neighbor of agent i at the beginning of agent i's kth maneuvering period. In view of Proposition 1, either condition (A) or condition

(B) must hold. Suppose first that condition (A) is true. Then $\mathcal{S}_i(k)$ strongly overlaps $\mathcal{S}_j(q)$, and agent i is in range of agent j during the overlap. There are two cases to consider. First, it is possible that $\mathcal{S}_i(k+1)$ also strongly overlaps $\mathcal{S}_j(q)$ for at least τ_S time units and agent i is in range of agent j during this overlap; in this case agent i would be a registered neighbor of agent j at time $\bar{t}_{jq}$, and $x_i(\bar{t}_{i(k+1)})$ would be its registered position. Second, it is possible either that $\mathcal{S}_i(k+1)$ does not strongly overlap $\mathcal{S}_j(q)$ or that agent i is not in range of agent j during this overlap; in this case agent i would be a registered neighbor of agent j at time $\bar{t}_{jq}$, and $x_i(\bar{t}_{ik})$ would be its registered position. Thus in summary, if condition A is true, then agent i will be a registered neighbor of agent j at time $\bar{t}_{jq}$ with registered position which could be either $x_i(\bar{t}_{ik})$ or $x_i(\bar{t}_{i(k+1)})$.

Suppose next that condition (A) does not hold. In view of Proposition 1, condition (B) must hold. In other words, $\mathcal{S}_i(k)$ must strongly overlap $\mathcal{S}_j(q-1)$, and agent i must be in range of agent j during this overlap. In view of Lemma 1, this must be the last sensing period of agent i with these properties because we have assumed that condition (A) does not hold. Therefore agent i must be a registered neighbor of agent j at time $\bar{t}_{j(q-1)}$, and $x_i(\bar{t}_{ik})$ must be its registered position. We summarize.

PROPOSITION 2 (neighbor relationship symmetry). *Suppose that agent j is a registered neighbor of agent i at the beginning of agent i's kth maneuvering period. Let $q = \lceil \bar{t}_{ik} \rceil_j$. If condition* (A) *of Proposition* 1 *holds, then agent i is a registered neighbor of agent j at the beginning of agent j's qth maneuvering period with either $x_i(\bar{t}_{ik})$ or $x_i(\bar{t}_{i(k+1)})$ as its registered position. If condition* (A) *of Proposition* 1 *does not hold, then condition* (B) *must hold and agent i is a registered neighbor of agent j at the beginning of agent j's $(q-1)$st maneuvering period with registered position $x_i(\bar{t}_{ik})$.*

2.1.3. Motion constraint. In the synchronous case treated in [1], each agent's way-points are constrained to positions defined in such a way so that no agent can lose any of its neighbors as it moves from one way-point to the next. This is accomplished by adopting a clever idea proposed in [1] which we call the *pairwise motion constraint.* Neighbor retention can also be achieved in the asynchronous case by enforcing the following constraint. Agent i is said to satisfy the *motion constraints induced by its neighbors* if for each $j \in \{1, 2, \ldots, n\}$ for which $j \neq i$ and each $k \in \{1, 2, \ldots\}$ for which agent j is a registered neighbor of agent i at the beginning of maneuvering period k, the position to which agent i moves at the end of the period is within a closed disk of diameter r centered at the mean of agent i's position at the beginning of the period (i.e., at time $\bar{t}_{ik}$) and the registered position of agent j at the beginning of the period. As mentioned, in the synchronous case, satisfaction of the pairwise motion constraint by agent i and neighbor j causes each to retain the other as a neighbor. The following proposition implies that essentially the same thing is true in the asynchronous case when the induced motion constraints are satisfied by agents i and j.

PROPOSITION 3 (neighbor retention). *Suppose that agents i and j satisfy the motion constraints induced by their registered neighbors. If agent j is a registered neighbor of agent i at the beginning of agent i's kth maneuvering period, then agent j is also a registered neighbor of agent i at the beginning of agent i's $(k+1)$st maneuvering period.*

In proving Proposition 3 and several subsequent claims we will make use of the inequalities

$$\bar{t}_{j(q+p)} - \bar{t}_{jq} \geq p\tau_D, \quad p \in \{0, 1, 2, \ldots\}, \quad q \in \{1, 2, \ldots\}, \quad j \in \{1, 2, \ldots, n\}, \tag{2}$$

and

$$\bar{t}_{i(k+1)} - \bar{t}_{ik} \leq 2(\tau_D - \tau_S), \quad k \in \{1, 2, \ldots\}, \quad i \in \{1, 2, \ldots, n\}, \tag{3}$$

which are both direct consequences of the definitions of the sensing and maneuver periods and (1). To justify (2), let us first recall that for each integer $s \geq 1$, $\bar{t}_{js}$ is at the end of agent j's sth sensing period. In addition, agent j's sensing periods do not intersect and are each of length τ_D. It follows that $\bar{t}_{j(s+1)} - \bar{t}_{js} \geq \tau_D$, $s \geq 1$, and thus that (2) is true. To justify (3), note that $\bar{t}_{i(k+1)}$ can be written as $\bar{t}_{i(k+1)} = \bar{t}_{ik} + \tau_D + \tau$, where τ is the length of agent i's kth maneuvering period. Since τ is constrained to satisfy $\tau \leq \tau_{M_i}$, we can write $\bar{t}_{i(k+1)} \leq \bar{t}_{ik} + \tau_D + \tau_{M_i}$. From this and (1) it follows that $\bar{t}_{i(k+1)} \leq \bar{t}_{ik} + \tau_D + (\tau_D - 2\tau_S)$ and thus that (3) is true.

To prove Proposition 3, we will make use of the two conditions characterizing a registered neighbor in Proposition 1. Each of these conditions in turn involves both an overlap requirement and a range requirement. The next lemma provides the needed facts about the way in which two agents' sensing periods overlap. This is followed by Lemma 3 which provides the range information needed to prove Proposition 3 and subsequent claims.

LEMMA 2. *Let i and j be distinct integers in $\{1, 2, \ldots, n\}$. Let $\bar{t}_{ik}$ be fixed and define $q = \lceil \bar{t}_{ik} \rceil_j$. Then*

$$\lceil \bar{t}_{i(k+1)} \rceil_j \in \{q, q+1, q+2\}. \tag{4}$$

1. *If $\lceil \bar{t}_{i(k+1)} \rceil_j = q$, then $\mathcal{S}_i(k+1) \cap \mathcal{S}_j(q) \succ \tau_S$.*
2. *If $\lceil \bar{t}_{i(k+1)} \rceil_j = q+1$, then $\mathcal{S}_i(k+1) \cap \mathcal{S}_j(q) \succ \tau_S$ or $\mathcal{S}_i(k+1) \cap \mathcal{S}_j(q+1) \succ \tau_S$.*
3. *If $\lceil \bar{t}_{i(k+1)} \rceil_j = q+2$, then $\mathcal{S}_i(k) \cap \mathcal{S}_j(q) \succ \tau_S$ and $\mathcal{S}_i(k+1) \cap \mathcal{S}_j(q+1) \succ \tau_S$.*

Moreover, if $\lceil \bar{t}_{i(k+1)} \rceil_j \in \{q+1, q+2\}$, then $\mathcal{S}_i(k)$ and $\mathcal{S}_i(k+1)$ are the only sensing periods of agent i which can strongly overlap $\mathcal{S}_j(q)$.

Proof of Lemma 2. It will be shown first that (4) is true. Since $\bar{t}_{ik} \in (\bar{t}_{j(q-1)} - \bar{t}_{jq}]$ and $\bar{t}_{i(k+1)} > \bar{t}_{ik}$, it must be true that $\bar{t}_{i(k+1)} > \bar{t}_{j(q-1)}$. Thus $\lceil \bar{t}_{i(k+1)} \rceil \geq q$. To prove that $\lceil \bar{t}_{i(k+1)} \rceil_j \leq q+2$, we use (3) and the fact that $\bar{t}_{ik} \leq \bar{t}_{jq}$ to write $\bar{t}_{i(k+1)} \leq 2(\tau_D - \tau_S) + \bar{t}_{jq}$. In view of (2) (with $p = 1$), $2(\tau_D - \tau_S) + \bar{t}_{jq} \leq \tau_D + \bar{t}_{j(q+1)} \leq \bar{t}_{j(q+2)}$. Therefore $\bar{t}_{i(k+1)} \leq \bar{t}_{j(q+2)}$. This means that $\lceil \bar{t}_{i(k+1)} \rceil_j \leq q+2$. Thus (4) is true.

To prove assertion 1, we use (2) with i substituted for j and $p = 1$ to write $\bar{t}_{i(k+1)} \geq \bar{t}_{ik} + \tau_D$. In view of the definition of q, $\bar{t}_{ik} > \bar{t}_{j(q-1)}$. Therefore $\bar{t}_{i(k+1)} - \bar{t}_{j(q-1)} > \tau_D > \tau_D - \tau_S$. The hypothesis $\lceil \bar{t}_{i(k+1)} \rceil = q$ implies that Lemma 1 holds with $k+1$ substituted for k. Thus $\mathcal{S}_i(k+1)$ and $\mathcal{S}_j(q-1)$ cannot overlap because of the lemma's last claim. Since the lemma also states that $\mathcal{S}_i(k+1)$ must strongly overlap either $\mathcal{S}_j(q-1)$ or $\mathcal{S}_j(q)$, it must be true that $\mathcal{S}_i(k+1)$ strongly overlaps $\mathcal{S}_j(q)$. Therefore assertion 1 is true.

Assertion 2 assumes that $\lceil \bar{t}_{i(k+1)} \rceil_j = q+1$. Lemma 1 thus applies with $k+1$ and $q+1$ replacing k and q, respectively. From this it follows that the only sensing periods of agent j which can overlap $\mathcal{S}_1(k+1)$ are $\mathcal{S}_j(q)$ and $\mathcal{S}_j(q+1)$; moreover, $\mathcal{S}_1(k+1)$ must strongly overlap at least one of these. Thus assertion 2 is true.

Assertion 3 assumes that $\lceil \bar{t}_{i(k+1)} \rceil_j = q+2$. Thus $\bar{t}_{j(q+1)} < \bar{t}_{i(k+1)}$. But $\bar{t}_{jq} + \tau_D \leq \bar{t}_{j(q+1)}$ because of (2) (with $p = 1$), and $\bar{t}_{i(k+1)} \leq \bar{t}_{ik} + 2(\tau_D - \tau_S)$ because of (3). Therefore $\bar{t}_{jq} \leq \bar{t}_{ik} + \tau_D - 2\tau_S$. It follows that $\bar{t}_{jq} - \bar{t}_{ik} + \tau_D - \tau_S$. Therefore by the first assertion of Lemma 1, $\mathcal{S}_i(k)$ and $\mathcal{S}_j(q)$ must strongly overlap. It remains to be shown that $\mathcal{S}_i(k+1) \cap \mathcal{S}_j(q+1) \succ \tau_S$ if $\lceil \bar{t}_{i(k+1)} \rceil_j = q+2$. Since $\lceil \bar{t}_{i(k+1)} \rceil_j = q+2$, Lemma 1 applies with $k+1$ and $q+2$ replacing k and q, respectively. Thus, to prove that $\mathcal{S}_i(k+1)$ and $\mathcal{S}_j(q+1)$ also strongly overlap, it is enough to show that $\bar{t}_{i(k+1)} - \bar{t}_{j(q+1)} \leq \tau_D - \tau_S$.

To do this, we first use (3) and the fact that $\bar{t}_{ik} \le \bar{t}_{jq}$ to write $\bar{t}_{i(k+1)} \le \bar{t}_{jq}+2(\tau_D-\tau_S)$. From this and (2) with $p=1$ there follows $\bar{t}_{i(k+1)} \le \bar{t}_{j(q+1)} + \tau_D - 2\tau_S$. Therefore $\bar{t}_{i(k+1)} \le \bar{t}_{j(q+1)} + \tau_D - \tau_S$. Thus $\mathcal{S}_i(k+1) \cap \mathcal{S}_j(q+1) \succ \tau_S$, so assertion 3 is true.

Now suppose that $\lceil \bar{t}_{i(k+1)} \rceil_j \in \{q+1, q+2\}$. Then in either case $\bar{t}_{jq} \le \bar{t}_{i(k+1)}$. Therefore $\bar{t}_{ik} \le \bar{t}_{jq} \le \bar{t}_{i(k+1)}$. If $\bar{t}_{ik} \ne \bar{t}_{jq}$, then $\bar{t}_{ik} < \bar{t}_{jq} \le \bar{t}_{i(k+1)}$, which means that $\lceil \bar{t}_{jq} \rceil = \bar{t}_{i(k+1)}$; thus Lemma 1 applies with k and q replaced by q and $k+1$, respectively. Therefore in this case $\mathcal{S}_i(k)$ and $\mathcal{S}_i(k+1)$ are the only sensing periods of agent i which can strongly overlap $\mathcal{S}_j(q)$. Now suppose that $\bar{t}_{ik} = \bar{t}_{jq}$. This means that $\lceil \bar{t}_{jq} \rceil = \bar{t}_{ik}$; thus Lemma 1 applies with k and q interchanged. Therefore in this case $\mathcal{S}_i(k-1)$ and $\mathcal{S}_i(k)$ are the only sensing periods of agent i which can strongly overlap $\mathcal{S}_j(q)$. To complete the proof, it is enough to show that $\mathcal{S}_i(k-1)$ cannot strongly overlap $\mathcal{S}_j(q)$. Towards this end, first note that $\bar{t}_{ik} \ge \bar{t}_{i(k-1)} + \tau_D$ because of (2). Thus $\bar{t}_{jq} \ge \bar{t}_{i(k-1)} + \tau_D$, so $\bar{t}_{jq} - \bar{t}_{i(k-1)} > +\tau_D - \tau_S$. Therefore $\mathcal{S}_i(k-1)$ cannot strongly overlap $\mathcal{S}_j(q)$ because of Lemma 1. □

LEMMA 3. *Let $q = \lceil \bar{t}_{ik} \rceil_j$. Suppose that agents i and j satisfy the motion constraints induced by their registered neighbors. If agent j is a registered neighbor of agent i at the beginning of agent i's kth maneuvering period, then*

$$||x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{jq^*})|| \le r, \tag{5}$$

$$||x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{j(q^*+1)})|| \le r, \tag{6}$$

where $q^* = q$ if condition (A) *of Proposition* 1 *is true and $q^* = q-1$ if it is not. Moreover, in either case*

$$||x_i(\bar{t}_{ik}) - x_j(\bar{t}_{jq})|| \le r. \tag{7}$$

Proof of Lemma 3. First suppose that agent j is a registered neighbor of agent i at the beginning of maneuvering period k. Thus by Proposition 1, x_{jq^*} is agent j's registered position, and

$$||x_i(\bar{t}_{ik}) - x_j(\bar{t}_{jq^*})|| \le r, \tag{8}$$

where $q^* = q$ if condition A holds and $q^* = q-1$ if it does not. The positions of agent i at the beginning and end of its kth maneuvering period are $x_i(\bar{t}_{ik})$ and $x_i(t_{ik})$, respectively. Therefore since agent i satisfies the motion constraint induced by agent j during this period, $||x_i(t_{ik}) - \frac{1}{2}\{x_i(\bar{t}_{ik}) + x_j(\bar{t}_{jq^*})\}|| \le \frac{r}{2}$. But $x_i(\bar{t}_{i(k+1)}) = x_i(t_{ik})$ because agent i does not move during sensing period $[t_{ik},\ \bar{t}_{i(k+1)})$. This enables us to rewrite the preceding inequality as

$$\left\| x_i(\bar{t}_{i(k+1)}) - \frac{1}{2}\{x_i(\bar{t}_{ik}) + x_j(\bar{t}_{jq^*})\} \right\| \le \frac{r}{2}. \tag{9}$$

Observe that

$$x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{jq^*}) = x_i(\bar{t}_{i(k+1)}) - \frac{1}{2}\{x_i(\bar{t}_{ik}) + x_j(\bar{t}_{jq^*})\} - \frac{1}{2}\left(x_j(\bar{t}_{jq^*}) - x_i(\bar{t}_{ik})\right).$$

Hence

$$\begin{aligned} ||x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{jq^*})|| \le & \left\| x_i(\bar{t}_{i(k+1)}) - \frac{1}{2}\{x_i(\bar{t}_{ik}) + x_j(\bar{t}_{jq^*})\} \right\| \\ & + \left\| \frac{1}{2}\left(x_j(\bar{t}_{jk^*}) - x_i(\bar{t}_{ik})\right) \right\|. \end{aligned}$$

From this, (8), and (9) there follows $||x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{jq^*})|| \leq \frac{r}{2} + \frac{r}{2} = r$. Therefore (5) is true.

It will now be shown that (6) is also true. By Proposition 2, agent i is a registered neighbor of agent j at the beginning of agent j's q^*th maneuvering period, where $q^* = q$ if condition (A) of Proposition 1 holds and $q^* = q - 1$ if it does not. Thus by Proposition 1

$$||x_j(\bar{t}_{jq^*}) - \bar{x}_i|| \leq r, \tag{10}$$

where $\bar{x}_i$ denotes the registered position of agent i at $\bar{t}_{jq^*}$. The positions of agent j at the beginning and end of its q^*th maneuvering period are $x_j(\bar{t}_{jq^*})$ and $x_j(t_{jq^*})$, respectively. Therefore since agent j satisfies the motion constraint induced by agent i during this period, $||x_j(t_{jq^*}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + \bar{x}_i\}|| \leq \frac{r}{2}$. But $x_j(\bar{t}_{j(q^*+1)}) = x_j(t_{jq^*})$ because agent j does not move during sensing period $q^* + 1$. Therefore

$$\left\| x_j(\bar{t}_{j(q^*+1)}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + \bar{x}_i\} \right\| \leq \frac{r}{2}. \tag{11}$$

In view of Proposition 2, $\bar{x}_i$ could be either $x_i(\bar{t}_{ik})$ or $x_i(\bar{t}_{i(k+1)})$ if condition A of Proposition 1 holds, while $\bar{x}_i = x_i(\bar{t}_{ik})$ if it does not. Consider first the case when $\bar{x}_i = x_i(\bar{t}_{ik})$. It is then possible to rewrite (11) as

$$\left\| x_j(\bar{t}_{j(q^*+1)}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + x_i(\bar{t}_{ik})\} \right\| \leq \frac{r}{2}. \tag{12}$$

But

$$\begin{aligned} ||x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{j(q^*+1)})|| &= \left\| x_i(\bar{t}_{i(k+1)}) - \frac{1}{2}\{x_i(\bar{t}_{ik}) + x_j(\bar{t}_{jq^*})\} \right. \\ &\quad \left. -(x_j(\bar{t}_{j(q^*+1)}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + x_i(\bar{t}_{ik})\}) \right\| \\ &\leq \left\| x_i(\bar{t}_{i(k+1)}) - \frac{1}{2}\{x_i(\bar{t}_{ik}) + x_j(\bar{t}_{jq^*})\} \right\| \\ &\quad + \left\| x_j(\bar{t}_{j(q^*+1)}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + x_i(\bar{t}_{ik})\} \right\|. \end{aligned}$$

From this, (9), and (12) it follows that $||x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{j(q^*+1)})|| \leq r$ and thus that (6) holds.

It will now be shown that (6) also holds for the case when $\bar{x}_i = x_i(\bar{t}_{i(k+1)})$ which only occurs when $q^* = q$. Assuming this possibility, (11) can be written as

$$\left\| x_j(\bar{t}_{j(q^*+1)}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + x_i(\bar{t}_{i(k+1)})\} \right\| \leq \frac{r}{2}. \tag{13}$$

Observe that it is possible to write

$$\begin{aligned} x_j(\bar{t}_{j(q^*+1)}) - x_i(\bar{t}_{i(k+1)}) &= x_j(\bar{t}_{j(q^*+1)}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + x_i(\bar{t}_{i(k+1)})\} \\ &\quad - \frac{1}{2}\left(x_i(\bar{t}_{i(k+1)}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + x_i(\bar{t}_{ik})\} \right) \\ &\quad + \frac{1}{4}\left(x_j(\bar{t}_{jq^*}) - x_i(\bar{t}_{ik}) \right). \end{aligned}$$

Clearly

$$||x_j(\bar{t}_{j(q^*+1)}) - x_i(\bar{t}_{i(k+1)})|| \leq \left\| x_j(\bar{t}_{j(q^*+1)}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + x_i(\bar{t}_{i(k+1)})\} \right\| + \frac{1}{2}\left\| x_i(\bar{t}_{i(k+1)}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + x_i(\bar{t}_{ik})\} \right\| + \frac{1}{4}\left\| x_j(\bar{t}_{jq^*}) - x_i(\bar{t}_{ik}) \right\|.$$

Using (8), (9), and (13) we thus obtain $||x_j(\bar{t}_{j(q^*+1)}) - x_i(\bar{t}_{i(k+1)})|| \leq \frac{r}{2} + \frac{r}{4} + \frac{r}{4} = r$. Thus (6) holds in this case too.

In view of (8), (7) is true if $q^* = q$. To prove that (7) also holds if $q^* = q - 1$, we first write

$$x_j(\bar{t}_{j(q^*+1)}) - x_i(\bar{t}_{ik}) = x_j(\bar{t}_{j(q^*+1)}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + x_i(\bar{t}_{ik})\} - \frac{1}{2}\left(x_i(\bar{t}_{ik}) - x_j(\bar{t}_{jq^*})\right).$$

Therefore

$$||x_j(\bar{t}_{j(q^*+1)}) - x_i(\bar{t}_{ik})|| \leq \left\| x_j(\bar{t}_{j(q^*+1)}) - \frac{1}{2}\{x_j(\bar{t}_{jq^*}) + x_i(\bar{t}_{ik})\} \right\| + \left\| \frac{1}{2}\left(x_i(\bar{t}_{ik}) - x_j(\bar{t}_{jq^*})\right) \right\|. \tag{14}$$

But if $q^* = q - 1$, both (8) and (12) hold. From these inequalities and (14) it follows that $||x_j(\bar{t}_{j(q^*+1)}) - x_i(\bar{t}_{ik})|| \leq \frac{1}{r} + \frac{1}{r} = r$ and therefore that (7) is true. □

Proof of Proposition 3. Consider first the case when $\lceil \bar{t}_{i(k+1)} \rceil = q$. If condition (A) of Proposition 1 holds, then $q^* = q$ and

$$||x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{jq})|| \leq r \tag{15}$$

because of (5). On the other hand, if condition (A) of Proposition 1 does not hold, then $q = q^* - 1$ and (15) still holds, in this case because of (6). Since $\lceil \bar{t}_{i(k+1)} \rceil = q$, it must be true that $\mathcal{S}_i(k+1) \cap \mathcal{S}_j(q) \succ \tau_S$ because of Lemma 2. This and (15) mean that condition (A) of Proposition 1 is satisfied with $k+1$ substituted for k. Therefore agent j is a registered neighbor of agent i at $\bar{t}_{i(k+1)}$.

Now suppose that $\lceil \bar{t}_{i(k+1)} \rceil \in \{q+1, q+2\}$. Consider first the case when condition (A) of Proposition 1 holds. Then Lemma 3 applies with $q^* = q$, so

$$||x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{j(q+1)})|| \leq r \tag{16}$$

and

$$||x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{jq})|| \leq r. \tag{17}$$

If $\lceil \bar{t}_{i(k+1)} \rceil = q + 1$, then $\mathcal{S}_i(k+1)$ must strongly overlap either $\mathcal{S}_j(q)$ or $\mathcal{S}_j(q+1)$ because of Lemma 2. In view of (16) and (17), condition (A) of Proposition 1 is satisfied in either situation with $k+1$ substituted for k and $q+1$ substituted for q. Therefore agent j is a registered neighbor of agent i at $\bar{t}_{i(k+1)}$. If $\lceil \bar{t}_{i(k+1)} \rceil = q + 2$, then $\mathcal{S}_i(k+1)$ and $\mathcal{S}_j(q+1)$ still must strongly overlap because of Lemma 2. Thus in

this case condition (B) of Proposition 1 is satisfied with $k+1$ substituted for k and $q+2$ substituted for q. Therefore agent j is a registered neighbor of agent i at $\bar{t}_{i(k+1)}$.

Consider finally the case when condition (A) of Proposition 1 does not hold. Since (7) holds, $\mathcal{S}_i(k)$ and $\mathcal{S}_j(q)$ cannot overlap. Therefore $\lceil t_{ik} \rceil \neq q+2$ because of statement 3 in Lemma 2. Thus $\lceil t_{ik} \rceil = q+1$. In addition, Lemma 2 states that the only sensing periods of agent i which can strongly overlap $\mathcal{S}_j(q)$ are $\mathcal{S}_i(k)$ and $\mathcal{S}_i(k+1)$. Since $\mathcal{S}_j(q)$ must strongly overlap at least one sensing period of agent i, it must be true that

$$\mathcal{S}_j(q) \cap \mathcal{S}_i(k+1) \succ \tau_S. \tag{18}$$

Since condition (A) of Proposition 1 does not hold, condition (B) must hold, because agent j is a neighbor of agent i at $\bar{t}_{ik}$. Thus Lemma 3 applies with $q^* = q-1$, so by (6),

$$||x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{jq})|| \leq r. \tag{19}$$

Since $\lceil \bar{t}_{i(k+1)} \rceil = q+1$, (19) and (18) show that condition (B) of Proposition 1 is satisfied with $k+1$ and $q+1$ substituted for k and q, respectively. □

2.2. Unsynchronized agent strategies. We are interested in strategies which cause agents to retain their registered neighbors. We therefore make the following assumption.

Cooperation assumption. Each agent i satisfies the motion constraints induced by each of its registered neighbors.

Suppose that the cooperation assumption is satisfied. Proposition 3 states that if agent j is a registered neighbor of agent i during maneuvering interval k, then it will also be a registered neighbor of agent i during maneuvering interval $k+1$. In other words, if the cooperation assumption is satisfied, each agent retains all of its prior registered neighbors as the system evolves. Thus if $\mathcal{N}_i(k)$ denotes the set of labels of agent i's neighbors at the beginning of its kth maneuvering period, then $\mathcal{N}_i(k) \subset \mathcal{N}_i(k+1),\ k \geq 1$.

Agent i's *kth way-point* $\bar{x}_i(k)$ is the point to which agent i moves at the end of its kth maneuvering period. Thus if $x_i(t)$ denotes the position of agent i at time t represented in a world coordinate system, then $x_i(t_{ik})$ and agent i's kth way-point are one and the same. The rule which determines $\bar{x}_i(k)$ is essentially the same as that considered previously for the synchronous case in [1, 8], except that now $\bar{x}_i(k)$ depends on agent i's own position at the beginning of its kth maneuvering period and the registered (relative) positions of agent i's registered neighbors at the beginning of the period. In particular, if agent i has m_{ik} registered neighbors at time $\bar{t}_{ik}$ with registered positions $z_1, z_2, \ldots, z_{m_{ik}}$ relative to agent i's, then agent i moves to the position $\bar{x}_i(k) = x_i(t_{i(k-1)}) + u_{m_{ik}}(z_1, \ldots, z_{m_{ik}})$ at the end of the period where

$$z_j = x_{ii_j}(\bar{t}_{ik}) - x_i(t_{i(k-1)}), \quad j \in \{1, 2, \ldots, m_{ik}\}, \tag{20}$$

and $x_{ii_j}(\bar{t}_{ik})$ is the registered position of neighbor i_j at time $\bar{t}_{ik}$. As in [8], $u_0 = 0$, and for $m \in \{1, \ldots, n-1\}$, u_m is a continuous control law mapping $\mathbb{D}^m$ into $\mathbb{D}_M$, where $\mathbb{D}$ and $\mathbb{D}_M$ are the closed disks of radii r and r_M, respectively, centered at the origin in $\mathbb{R}^2$. For $m > 0$, u_m is defined so that the aforementioned neighbor motion constraint is satisfied and, in addition, so that for each $\{z_1, z_2, \ldots, z_m\} \in \mathbb{D}^m$, $u_m(z_1, z_2, \ldots, z_n)$ is in the convex hull of $\{0, z_1, z_2, \ldots, z_m\}$, but not at a corner unless

$z_1 = z_2 = \cdots = z_m = 0$. Examples of $u_m(\cdot)$ satisfying these *control law requirements* can be found in [1, 8].

Since each agent is assumed to move to its kth way-point at the end of its kth maneuvering period, agent i's position at time t_{ik} is given by

$$x_i(t_{ik}) = x_i(t_{i(k-1)}) + u_{m_{ik}}(x_{ii_1}(\bar{t}_{ik}) - x_i(t_{i(k-1)}), \ldots, x_{ii_{m_{ik}}}(\bar{t}_{ik}) - x_i(t_{i(k-1)})). \tag{21}$$

In view of Proposition 1 and (7), the formulas for the $x_{ij}(\bar{t}_{ik})$ can be written as

$$x_{ij}(\bar{t}_{ik}) = \left\{ \begin{array}{ll} x_j(\bar{t}_{jq}) & \text{if } \mathcal{S}_i(k) \cap \mathcal{S}_j(q) \succ \tau_S \\ x_j(\bar{t}_{j(q-1)}) & \text{otherwise} \end{array} \right\}, \quad j \in \mathcal{N}_i(k), \tag{22}$$

where $q = \lceil \bar{t}_{ik} \rceil_j$ and

$$\mathcal{N}_i(k) = \{j : ||x_i(\bar{t}_{ik}) - x_j(\bar{t}_{iq})|| \leq r \text{ and } \mathcal{S}_i(k) \cap \mathcal{S}_j(q) \succ \tau_S\} \bigcup \{j : ||x_i(\bar{t}_{ik}) - x_j(\bar{t}_{i(q-1)})|| \leq r \text{ and } \mathcal{S}_i(k) \cap \mathcal{S}_j(q-1) \succ \tau_S\}. \tag{23}$$

The expressions for the $x_{ij}(\bar{t}_{ik})$ in (22) are a direct consequence of the characterization of registered positions in Proposition 1, the fact that (7) holds whenever $j \in \mathcal{N}_i(k)$, and the implication of Lemma 1 that $\mathcal{S}_i(k) \cap \mathcal{S}_j(q-1) \succ \tau_S$ whenever $\mathcal{S}_i(k) \cap \mathcal{S}_j(q) \not\succ \tau_S$. Of course the neighbor set $\mathcal{N}_i(k)$ and the registration positions x_{ij}, $j \in \mathcal{N}_i(k)$, all depend on i and k.

3. Main results. Note that because agents do not move during sensing periods, for each $i \in \{1, 2, \ldots, n\}$ the positions of agent i at times $t_{i(k-1)}$ and t_{ik} are the same as at times $\bar{t}_{ik}$ and $\bar{t}_{i(k+1)}$, respectively. Thus (21) can also be written as

$$x_i(\bar{t}_{i(k+1)}) = x_i(\bar{t}_{ik}) + u_{m_{ik}}(x_{ii_1}(\bar{t}_{ik}) - x_i(\bar{t}_{ik}), \ldots, x_{ii_{m_{ik}}}(\bar{t}_{ik}) - x_i(\bar{t}_{ik})). \tag{24}$$

The n equations given by (24) for $i \in \{1, 2, \ldots, n\}$ together with (22) and (23) completely describe the evolution of the positions of the n agents under consideration as each maneuvers from way-point to way-point. Just as in the synchronous case, the analysis of these equations depends on the relationships between registered neighbors and how these relationships evolve with time. To characterize these relationships, we first extend the domain of definition of each agent's registered neighbors from its set of maneuvering period start times to a suitably defined set of "event times" common to all n agents. By an *event time* is meant any time $\bar{t}_{ik}$ at which any maneuvering period $[\bar{t}_{ik}, t_{ik})$ of any agent begins. Let $\{\bar{t}_{ik} : i \in \{1, 2, \ldots, n\}, k \geq 1\}$ denote the set of all distinct event times. Label this set's elements as $t_1, t_2, \ldots, t_p, \ldots$ in such a way that $t_p < t_{p+1}$, $j \in \{1, 2, \ldots\}$. For $i \in \{1, 2, \ldots, n\}$, let P_i denote that strictly monotone function from the set of positive integers $\mathcal{I}$ to $\mathcal{I}$ which assigns to $k \in \mathcal{I}$ that value of $p \in \mathcal{I}$ for which $t_p = \bar{t}_{ik}$. Thus with this notation, $t_{P_i(k)} = \bar{t}_{ik}$, so agent i's registered neighbors at its kth event time $t_{P_i(k)}$ are its registered neighbors at time $\bar{t}_{ik}$. For each $i \in \{1, 2, \ldots, n\}$ we extend the domain of definition of agent i's registered neighbors from the set $\{t_{P_i(k)} : k \geq 1\}$ to the set $\{t_p : p \geq P_i(1)\}$ by stipulating that for values of t_p which are between two successive event times of agent i, say between t_{ik} and $t_{i(k+1)}$, agent i's registered neighbors are the same as its registered neighbors at time t_{ik}.

Let $\mathcal{T} \triangleq \{t_{\bar{p}}, t_{\bar{p}+1}, t_{\bar{p}+2}, \ldots\}$ denote the set of all event times greater than or equal to $t_{\bar{p}}$, where $\bar{p} \triangleq \max\{P_1(1), P_2(1), \ldots P_n(1)\}$. Note that the registered neighbors of each agent are defined at each time t_p in $\mathcal{T}$. For each $p \geq \bar{p}$, it is therefore possible to describe neighbor relationships using a directed[1] graph $\mathbb{G}_p$ with vertex set $\{1, 2, \ldots, n\}$ and directed edge set defined so that (i, j) is a directed edge from vertex i to vertex j just in case agent j is a registered neighbor of agent i at event time t_p.

Let us partially order the set of all directed graphs with vertex set $\{1, 2, \ldots, n\}$ by agreeing to say that $\mathbb{G}$ is contained in $\bar{\mathbb{G}}$ if the edge set of $\mathbb{G}$ is a subset on the edge set of $\bar{\mathbb{G}}$. It is natural then to define the *union* of a collection of such graphs to be the directed graph with vertex set $\{1, 2, \ldots, n\}$ and edge set equaling the union of the edge sets of all of the graphs in the collection. Because of the cooperation assumption and Proposition 3, we know that each agent keeps all of its registered neighbors as the system evolves. What this means is the sequence of graphs $\mathbb{G}_{\bar{p}}, \mathbb{G}_{\bar{p}+1}, \ldots, \mathbb{G}_p, \ldots$ forms the ascending chain

$$\mathbb{G}_{\bar{p}} \subset \mathbb{G}_{\bar{p}+1} \subset \cdots \mathbb{G}_p \cdots. \tag{25}$$

Because the set of directed graphs on vertices $\{1, 2, \ldots, n\}$ is a finite set, the chain must converge to the graph

$$\mathbb{G} \triangleq \bigcup_{p=\bar{p}}^{\infty} \mathbb{G}_p \tag{26}$$

in a finite number of steps. More is true. Suppose that agent i has agent j as a registered neighbor at the beginning of one of agent i's maneuvering periods. Then because of Proposition 2, agent i must be a registered neighbor of agent j at the beginning of one of agent j's maneuvering periods. These observations together with the cooperation assumption imply that agents i and j must both eventually become and remain registered neighbors of each other. As a consequence, there must be directed arcs in $\mathbb{G}$ from vertex i to vertex j as well as from vertex j to vertex i. Clearly $\mathbb{G}$ must be a directed graph with the property that for each distinct pair of vertices—say i and j—either there is no directed arc connecting one to the other or there are two directed arcs, one from vertex i to vertex j and the other from vertex j to vertex i. Directed graphs with this property are usually regarded as simple graphs whose edges represent such pairs of directed arcs [6]. In what follows we shall adopt this viewpoint and refer to $\mathbb{G}$ as a simple graph. Our main result is as follows.

THEOREM 1. *Let $u_0 = 0 \in \mathbb{D}_M$ and for each $m \in \{1, 2, \ldots, n-1\}$, let $u_m : \mathbb{D}^m \to \mathbb{D}_M$ be any continuous function satisfying the aforementioned control law requirements. For each set of initial agent positions $x_1(0), x_2(0), \ldots, x_n(0)$, each agent's position $x_i(t)$ converges to a unique point $\pi_i \in \mathbb{R}^2$ such that for each $i, j \in \{1, 2, \ldots, n\}$, either $\pi_i = \pi_j$ or $||\pi_i - \pi_j|| > r$. Moreover, if agent j is a registered neighbor of agent i at the beginning of one of agent i's maneuvering periods, then $\pi_i = \pi_j$.*

This theorem will be proved in section 4.

Theorem 1 states that the strategies under consideration cause all agents' positions to converge to points in the plane with the property that each pair of such points are either equal to each other or separated by a distance greater than r units.

[1]It will soon be clear that the aforementioned symmetry of the neighbor relationship will ultimately enable us to characterize neighbor relationships with a simple, undirected graph as in the synchronous case.

The theorem further states that if one agent is ever a registered neighbor of another, then both converge to the same point. Thus all n agents' positions will converge to a single point if any one directed graph in the ascending chain is strongly connected. We are led to the following corollary.

COROLLARY 1. *If at any event time $t_p \geq t_{\bar{p}}$, the directed graph $\mathbb{G}_p$ characterizing registered neighbors is strongly connected, then positions of all n agents converge to a common point in the plane.*

4. Analysis. The aim of this section is to establish the correctness of Theorem 1. This requires the analysis of the asymptotic behavior of the *asynchronous* process described by (22) and (24) for $i \in \{1, 2, \ldots, n\}$. Despite the apparent complexity of this process, it is possible to capture its salient features for t_s sufficiently large using a suitably defined *synchronous* discrete-time, hybrid dynamical system $\mathbb{S}$. The process of constructing a synchronous process to model the behavior of an asynchronous process is called *analytic synchronization* and has been outlined in the introduction to this paper. In what follows we demonstrate the utility of this idea by applying it to the problem at hand.

4.1. A synchronous model of the asynchronous agent system. It is sufficient to analyze the behavior of the n-agent system for times beyond the time at which each agent's neighbor set stops changing. Analytic synchronization would thus have us define $\mathbb{S}$ to be a synchronous system evolving on the event time set $\{t_p : p \in \mathcal{P}\}$ where $\mathcal{P} = \{p; p \geq p^*\}$ and p^* is the smallest value of $p \geq \bar{p}$ for which the ascending chain shown in (25) has converged to the limit graph $\mathbb{G}$ in (26). To reduce clutter we will instead define $\mathbb{S}$ to be a synchronous discrete-time dynamical system evolving on the index set $\mathcal{P}$. Thus for $p \in \mathcal{P}$, the registered neighbors of each agent do not change. For simplicity, we will only deal with the case when each agent has at least one neighbor for $t_p \geq t_{p^*}$. The position update equation (24) for agent i can thus be written as

$$x_i(\bar{t}_{i(k+1)}) = x_i(\bar{t}_{ik}) + u_{m_i}(x_{ii_1}(\bar{t}_{ik}) - x_i(\bar{t}_{ik}), \ldots, x_{ii_{m_i}}(\bar{t}_{ik}) - x_i(\bar{t}_{ik})), \tag{27}$$

where m_i is a positive number and $\mathcal{N}_i \triangleq \{i_1, i_2, \ldots, i_{m_i}\}$ is the set of indices labelling agent i's registered neighbors. Just as before,

$$x_{ij}(\bar{t}_{ik}) = \begin{cases} x_j(\bar{t}_{jq}) & \text{if } \mathcal{S}_i(k) \cap \mathcal{S}_j(q) \succ \tau_S, \\ x_j(\bar{t}_{j(q-1)}) & \text{otherwise,} \end{cases} \tag{28}$$

and

$$\mathcal{N}_i = \{j : ||x_i(\bar{t}_{ik}) - x_j(\bar{t}_{iq})|| \leq r \text{ and } \mathcal{S}_i(k) \cap \mathcal{S}_j(q) \succ \tau_S\}$$

$$\bigcup \{j : ||x_i(\bar{t}_{ik}) - x_j(\bar{t}_{i(q-1)})|| \leq r \text{ and } \mathcal{S}_i(k) \cap \mathcal{S}_j(q-1) \succ \tau_S\}, \tag{29}$$

where $q = \lceil \bar{t}_{ik} \rceil_j$. Note that it must be true that

$$||x_j(\bar{t}_{jq}) - x_i(\bar{t}_{ik})|| \leq r \tag{30}$$

because of (7). In view of (29) it also must be true that

$$||x_j(\bar{t}_{j(q-1)}) - x_i(\bar{t}_{ik})|| \leq r \text{ if } \mathcal{S}_i(k) \cap \mathcal{S}_j(q) \not\succ \tau_S. \tag{31}$$

Inequalities (30) and (31) are consequences of the assumption that $j \in \mathcal{N}_i$. These inequalities will translate into constraints on the state of $\mathbb{S}$.

4.1.1. Definition of $\mathbb{S}$. We will take as the state space of $\mathbb{S}$ the space $\mathcal{X}$ of all lists $\{y_1, y_2, \ldots, y_n, w_1, w_2, \ldots, w_n\}$ satisfying

$$\left.\begin{array}{l} y_i, w_i \in \mathbb{R}^2, \\ ||y_i - y_j|| \le r \end{array}\right\}, \qquad j \in \mathcal{N}_i, \quad i \in \{1, 2, \ldots, n\}. \tag{32}$$

In what follows we often write y for $\{y_1, y_2, \ldots, y_n\}$ and w for $\{w_1, w_2, \ldots, w_n\}$. We sometimes refer to $\{y_i, w_i\}$ as the state of "node" i. For $i \in \{1, 2, \ldots, n\}$ let P_i^{-1} be a left inverse of P_i and let $\mathcal{P}_i = \mathcal{P} \cap \text{image } P_i$. We now define $\mathbb{S}$ to be a time-varying system with state $\{y, w\}$; for each $i \in \{1, 2, \ldots, n\}$, the state of node i evolves on $\mathcal{P}$ according to update equations defined for $p \in \mathcal{P}_i$ by

$$y_i(p+1) = y_i(p) + u_{m_i}(v_{ii_1}(p) - y_i(p), \ldots, \ v_{ii_{m_i}}(p) - y_i(p)), \tag{33}$$

$$w_i(p+1) = y_i(p), \tag{34}$$

where

$$v_{ij}(p) = \left\{\begin{array}{ll} y_j(p) & \text{if } \mathcal{S}_i(P_i^{-1}(p)) \cap \mathcal{S}_j(\lceil t_p \rceil_j) \succ \tau_S, \\ w_j(p) & \text{otherwise} \end{array}\right\}, \qquad j \in \mathcal{N}_i, \tag{35}$$

and by

$$y_i(p+1) = y_i(p), \tag{36}$$

$$w_i(p+1) = w_i(p) \tag{37}$$

for $p \notin \mathcal{P}_i$. We require that y_i satisfies the *neighbor constraints*

$$||y_i(p) - w_j(p)|| \le r \ \text{ if } \quad \mathcal{S}_i(P_i^{-1}(p)) \cap \mathcal{S}_j(\lceil t_p \rceil_j) \not\succ \tau_S, \ \ p \in \mathcal{P}_i, \quad j \in \mathcal{N}_i. \tag{38}$$

Note that these constraint requirements together with the definition of $\mathcal{X}$ and v_{ij} ensure that $||v_{ij} - y_i(p)|| \le r$ whenever $p \in \mathcal{P}_i$. This in turn is necessary for (33) to make sense because the domain of u_{m_i} is $\mathbb{D}^{m_i}$.

The preceding defines $\mathbb{S}$ to be a synchronous discrete-time dynamical system with state constraints given by (38). The definition depends on the $\mathcal{N}_i$ as well as the n event time sequences $\{\bar{t}_{ik}; k \ge 1\}$. We have assumed that the $\mathcal{N}_i$ are nonempty; in addition, $\mathcal{N}_i \subset \{1, 2, \ldots, i-1, i+1, \ldots, n\}$. As a consequence of Proposition 2 and the assumption that neighbors stop changing, the $\mathcal{N}_i$ all have the following *symmetry property:* If $j \in \mathcal{N}_i$, then $i \in \mathcal{N}_j$. Because of the symmetry property we can associate with the $\mathcal{N}_i$ a simple graph $\mathbb{G}$ with vertex set $\{1, 2, \ldots, n\}$ and edge set defined in such a way that (i, j) is in the edge set just in case $i \in \mathcal{N}_j$ and $j \in \mathcal{N}_i$. Note that this is precisely the same as the simple graph mentioned just before Theorem 1. As for event times, recall that each event time sequence is strictly monotone increasing and that together they all satisfy Lemma 1, (2), and (3). In defining $\mathbb{S}$, these are the only properties of the $\mathcal{N}_i$ and the event times which are assumed.

4.1.2. Validation of $\mathbb{S}$. We claim that $\mathbb{S}$ provides a synchronous model of the asynchronous agent system describe by (27)–(31). The first step in justifying this claim is to define

$$\left.\begin{array}{l} y_i(p) = x_i(\bar{t}_{ik}), \\ w_i(p) = x_i(\bar{t}_{i(k-1)}) \end{array}\right\}, \quad P_i(k-1) < p \le P_i(k), \ \ k \in P_i^{-1}(\mathcal{P}), \tag{39}$$

for $i \in \{1, 2, \ldots, n\}$. Note that y_i has been defined so that it is constant between agent i's event times and agrees with x_i whenever p is such that t_p is within one of agent i's sensing periods.

To justify the claim that $\mathbb{S}$ models (27)–(31), we need to prove that with the $y_i(p)$ and $w_i(p)$ defined by (39), $\{y(p), w(p)\} \in \mathcal{X},\ p \in \mathcal{P}$, and (33)–(38) are satisfied. In view of (30) and the definition of the $y_i(p)$ in (2), it is clear that for $i \in \{1, 2, \ldots, n\}$, $||y_i(p) - y_j(p)|| \leq r, j \in \mathcal{N}_i,\ p \in \mathcal{P}$. Therefore $\{y(p), w(p)\} \in \mathcal{X},\ p \in \mathcal{P}$. It remains to be shown that (33)–(38) are satisfied. To accomplish this, fix $p \in \mathcal{P}$ and suppose that k is that value for which $P_i(k) \leq p < P_i(k+1)$. Set $p_1 = P_i(k)$ and $p_2 = P_i(k+1)$. By definition,

$$y_i(p_1) = x_i(\bar{t}_{ik}), \tag{40}$$

$$w_i(p_1) = x_i(\bar{t}_{i(k-1)}), \tag{41}$$

$$y_i(p_2) = x_i(\bar{t}_{i(k+1)}), \tag{42}$$

$$w_i(p_2) = x_i(\bar{t}_{ik}), \tag{43}$$

$$y_i(s) = y_i(p_2),\ \ p_1 < s \leq p_2, \tag{44}$$

$$w_i(s) = w_i(p_2),\ \ p_1 < s \leq p_2. \tag{45}$$

Suppose first that $p \notin \mathcal{P}_i$ or, equivalently, that $p_1 < p < p_2$. Then $p_1 < p + 1 \leq p_2$, so $y_i(p+1) = y_i(p_2)$ and $w_i(p+1) = w_i(p_2)$ because of (44) and (45), respectively. But $y_i(p) = y_i(p_2)$ and $w_i(p) = w_i(p_2)$ also because of (44) and (45), respectively. It follows that (36) and (37) are true.

Now suppose that $p \in \mathcal{P}_i$ or, equivalently, that $p = p_1$. Then $p_1 < p + 1 \leq p_2$, so $y_i(p+1) = y_i(p_2)$ and $w_i(p+1) = w_i(p_2)$ because of (44) and (45), respectively. It follows from (42) and (43) that

$$y_i(p+1) = x_i(\bar{t}_{i(k+1)}) \tag{46}$$

and

$$w_i(p+1) = x_i(\bar{t}_{ik}). \tag{47}$$

But

$$x_i(\bar{t}_{ik}) = y_i(p) \tag{48}$$

because of (40); thus (34) is true.

Fix $j \in \mathcal{N}_i$ and set $q = \lceil t_p \rceil_j$. To justify (38) and (33) we will need to express $x_j(\bar{t}_{iq})$, $x_j(\bar{t}_{i(q-1)})$, and k in terms of y_j, w_j, and p, respectively. Note first that $t_p = \bar{t}_{ik}$ because $p = p_1$. Thus

$$q = \lceil \bar{t}_{ik} \rceil_j, \tag{49}$$

so $\bar{t}_{j(q-1)} < \bar{t}_{ik} \leq \bar{t}_{jq}$. This means that $P_j(q-1) < P_i(k) \leq P_j(q)$ and thus that $P_j(q-1) < p \leq P_j(q)$. But by definition $y_j(s) = x_j(\bar{t}_{jq})$ and $w_j(s) = x_j(\bar{t}_{j(q-1)})$ for $P_j(q-1) < s \leq P_j(q)$. Therefore

$$x_j(\bar{t}_{jq}) = y_j(p), \tag{50}$$

$$x_j(\bar{t}_{j(q-1)}) = w_j(p). \tag{51}$$

Finally note that

$$k = P_i^{-1}(p) \tag{52}$$

because $P_i(k) = p_1 = p$. It is now clear from (40), (51), and (52) that the inequality in (31) translates into neighbor constraint (38).

In addition, examination of (48)–(52) together with the definitions of $x_{ij}(\bar{t}_{ik})$ and $v_{ij}(p)$ in (28) and (35), respectively, reveals that

$$x_{ij}(\bar{t}_{ik}) = v_{ij}(p). \tag{53}$$

From this and (48) it follows that the expression for $x_i(\bar{t}_{i(k+1)})$ in (27) can be written as

$$x_i(\bar{t}_{i(k+1)}) = y_i(p) + u_{m_i}(v_{ii_1}(p) - y_i(p), \ldots, \ v_{ii_{m_i}}(p) - y_i(p)).$$

This and (46) thus finally justify (33).

By a *trajectory* of $\mathbb{S}$ is meant a sequence of states $\{\{y(p), w(p)\} : p \in \mathcal{P}\}$ which satisfy (33)–(37) as well as the neighbor constraints (38). The preceding proves that the family of such trajectories is nonempty and contains the trajectory which represents the actual agent system under consideration. It turns out that the trajectory representing the actual agent system has an additional property which we will exploit later.

LEMMA 4. *For $i \in \{1, 2, \ldots, n\}$, let $y_i(p)$ and $w_i(p)$ be defined by* (39). *Let $i \in \{1, 2, \ldots, n\}$ and $s \in \mathcal{S}_i$ be fixed. Suppose that for some $j \in \{1, 2, \ldots, n\}$ and $p \in \mathcal{P}_i$,*

$$||y_i(p+1) - y_j(p)|| \leq r, \tag{54}$$

$$||w_i(p+1) - y_j(p)|| \leq r. \tag{55}$$

Then $j \in \mathcal{N}_i$.

Proof of Lemma 4. Since $p \in \mathcal{P}_i$ and P_i is strictly monotone, there is a unique integer k for which $p = P_i(k)$. Let $q = \lceil \bar{t}_{ik} \rceil$. As noted previously in the development leading to (46)–(50), $y_i(p+1) = x_i(\bar{t}_{i(k+1)})$, $w_i(p) = x_i(\bar{t}_{ik})$, and $y_j(p) = x_j(\bar{t}_{jq})$. Thus (54) and (55) translate into $||x_i(\bar{t}_{i(k+1)}) - x_j(\bar{t}_{jq})|| \leq r$ and $||x_i(\bar{t}_{ik}) - x_j(\bar{t}_{jq})|| \leq r$, respectively. Moreover, Lemma 1 states that $\mathcal{S}_i(k)$ must strongly overlap either $\mathcal{S}_j(q)$ or $\mathcal{S}_j(q-1)$. If the former is true, then condition (A) of Proposition 1 is satisfied and thus $j \in \mathcal{N}_i$. Suppose next that $\mathcal{S}_i(k)$ does not strongly overlap $\mathcal{S}_j(q)$. Then $\bar{t}_{i(k+1)} \in \{q, q+1\}$ because of (4) and condition 3 in Lemma 2. If $\bar{t}_{i(k+1)} = q$, then $\mathcal{S}_i(k+1) \cap \mathcal{S}_j(q) \geq \tau_S$ because of condition 1 in Lemma 2. Thus condition (A) of Proposition 1 is satisfied when $k+1$ is substituted for k; thus in this case $j \in \mathcal{N}_i$. Suppose $\bar{t}_{i(k+1)} = q+1$. In view of Lemma 2, $\mathcal{S}_i(k)$ and $\mathcal{S}_i(k+1)$ are the only sensing periods of agent i which can strongly overlap $\mathcal{S}_j(q)$. Since $\mathcal{S}_j(q)$ must be strongly overlapped by at least one of agent i's sensing periods, it must be true that $\mathcal{S}_i(k+1) \cap \mathcal{S}_j(q) \geq \tau_S$. Thus condition (B) of Proposition 1 is satisfied with $k+1$ and $q+1$ substituted for k and q, respectively. Therefore $j \in \mathcal{N}_i$. □

Conditions (54) and (55) do not necessarily imply that $j \in \mathcal{N}_i$ for every trajectory of $\mathbb{S}$. The claim of Lemma 4 is that the implication does indeed hold if the trajectory in question is the one which models the actual agent system.

4.2. Properties of $\mathbb{S}$. In section 4.1 we defined $\mathbb{S}$ and proved that it faithfully models the actual agent system. In this section we derive several important properties of $\mathbb{S}$.

4.2.1. Local convex hulls. In what follows we denote the convex hull of a given set of points $x_1, x_2, \ldots, x_q$ in $\mathbb{R}^2$ by $\langle x_1, x_2, \ldots, x_q \rangle$. We write $\mathcal{H}_i(p)$ for the *ith local convex hull*

$$\mathcal{H}_i(p) = \langle y_i(p), y_{i_1}(p), \ldots, y_{i_{m_i}}(p), w_i(p), w_{i_1}(p), \ldots, w_{i_{m_i}}(p) \rangle,$$

where $\{i_1, i_2, \ldots, i_{m_i}\} = \mathcal{N}_i$. We also write $\mathcal{H}(p)$ for the (global) convex hull

$$\mathcal{H}(p) = \langle y_1(p), y_2(p), \ldots, y_n(p), w_1(p), w_2(p), \ldots, w_n(p) \rangle,$$

and $\mathcal{K}(p)$ for the set of corners of $\mathcal{H}(p)$. Clearly

$$\mathcal{H}_i(p) \subset \mathcal{H}(p), \qquad i \in \{1, 2, \ldots, n\}, \quad p \in \mathcal{P}. \tag{56}$$

This fact plays a role in the proof of the following lemma which establishes a fundamental property of $\mathbb{S}$.

LEMMA 5.

$$\mathcal{H}(p+1) \subset \mathcal{H}(p), \quad p \in \mathcal{P}. \tag{57}$$

Proof of Lemma 5. Fix $i \in \{1, 2, \ldots, n\}$ and note that (33) and the control law requirement that $u_m(z_1, z_2, \ldots, z_m) \in \langle 0, z_1, \ldots, z_m \rangle$, $z_i \in \mathbb{D}$, imply that $y_i(p+1) \in \mathcal{H}_i(p)$, $p \in \mathcal{P}_i$; thus $y_i(p+1) \in \mathcal{H}(p)$, $p \in \mathcal{P}_i$, because of (56). Moreover, $y_i(p+1)$ is also in $\mathcal{H}(p)$ for $p \notin \mathcal{P}_i$ because of (36). Therefore $y_i(p+1) \in \mathcal{H}(p)$ for all $p \in \mathcal{P}$. Similarly, $w_i(p+1) \in \mathcal{H}(p)$, $p \in \mathcal{P}$, because of (34) and (37). Thus $\{y_i(p+1), w_i(p+1)\} \subset \mathcal{H}(p)$, $p \in \mathcal{P}$. Since this holds for all $i \in \{1, 2, \ldots, n\}$, (57) is true. □

4.2.2. Stationary nodes. Let us agree to say that node i is *stationary* at time $p \in \mathcal{P}_i$ if

$$y_i(p) = v_{ii_1}(p) = \cdots = v_{ii_{m_i}}(p).$$

The terminology is prompted by the fact that if node i is stationary at p, then $y_i(p+1) = y_i(p)$; this can be seen from (33) and the control law requirements imposed on u_{m_i}. In addition, the requirement that $u_m(z_1, z_2, \ldots, z_m)$ not be a corner of $\langle 0, z_1, \ldots, z_m \rangle$ unless $z_1 = z_2 = \cdots = z_m = 0$ implies that if $y_i(p+1)$ is a corner of $\langle y_i(p), v_{ii_1}(p), \ldots, v_{ii_{m_i}}(p) \rangle$, then node i must be stationary at p. The following lemma implies that this is also true if $y_i(p+1)$ is a corner of $\mathcal{H}(p)$.

LEMMA 6. *Fix $i \in \{1, 2, \ldots, n\}$ and $\bar{p} \in \mathcal{P}_i$. If $y_i(\bar{p}+1) \in \mathcal{K}(\widehat{p})$ for some $\widehat{p} \le \bar{p}$, then node i must be stationary at each $p \in \mathcal{P}_i \cap \{p : \widehat{p} \le p \le \bar{p}\}$ and*

$$y_i(p) = y_i(\bar{p}+1) \tag{58}$$

for all such p.

Proof of Lemma 6. Let $p_1, p_2, \ldots, p_m$ denote the elements of the set $\mathcal{P}_i \cap \{p : \widehat{p} \le p \le \bar{p}\}$, ordered so that $p_1 < p_2 < \cdots < p_m = \bar{p}$. To prove the lemma it is sufficient to show that the following statements hold for $k \in \{1, 2, \ldots, m\}$:

(i) Node i is stationary at $p_k, p_{k+1}, \ldots, p_m$.

(ii) $y_i(p_k) = y_i(p_{k+1}) = \cdots = y_i(p_m) = y_i(\bar{p}+1)$.

Let $\bar{\mathcal{H}}(p_s) = \langle y_i(p_s), v_{ii_1}(p_s), \ldots, v_{ii_{m_i}}(p_s) \rangle$, $s \in \{1, 2, \ldots, m\}$. Note that u_m must satisfy the control law requirement $u_m(z_1, z_2, \ldots, z_m) \in \langle 0, z_1, \ldots, z_m \rangle$. In view of (33), it must therefore be true that

$$y_i(p_s + 1) \in \bar{\mathcal{H}}(p_s), \quad s \in \{1, 2, \ldots, m\}. \tag{59}$$

Note next that the definition of v_{ij} in (35) implies that $v_{ij}(p_s) \in \{y_j(p_s), w_j(p_s)\}$, $s \in \{1, 2, \dots, m\}$. Therefore $\bar{\mathcal{H}}(p_s) \subset \mathcal{H}_i(p_s)$. But $\mathcal{H}_i(p_s) \subset \mathcal{H}(p_s)$; moreover, $\mathcal{H}(p_s) \subset \mathcal{H}(\widehat{p})$ because of Lemma 5. Thus $\bar{\mathcal{H}}(p_s) \subset \mathcal{H}(\widehat{p})$. This implies that

$$\bar{\mathcal{H}}(p_s) \cap \mathcal{K}(\widehat{p}) \subset \bar{\mathcal{K}}(p_s), \quad s \in \{1, 2, \dots, m\}, \tag{60}$$

where $\bar{\mathcal{K}}(p_s)$ is the corner set of $\bar{\mathcal{H}}(p_s)$.

Recall that $p_m = \bar{p}$. By assumption, $y_i(\bar{p}+1) \in \mathcal{K}(\widehat{p})$. These facts and (59) imply that $y_i(p_m+1) \in \bar{\mathcal{H}}(p_m) \cap \mathcal{K}(\widehat{p})$. Thus $y_i(p_m+1) \in \bar{\mathcal{K}}(p_m)$ because of (60). Therefore node i is stationary at p_m, and because of this, $y_i(p_m+1) = y_i(p_m)$. Thus statements (i) and (ii) above are true for $k = m$. If $m = 1$, the proof is complete.

Suppose next that $m > 1$ and that statements (i) and (ii) hold for all $k \in \{q+1, \dots, m\}$ where q is some integer satisfying $1 < q+1 \le m$. In view of (36), $y_i(p) = y_i(p_{q+1})$ for $p_q < p \le p_{q+1}$. Therefore

$$y_i(p_q + 1) = y_i(p_{q+1}). \tag{61}$$

By hypothesis, (ii) holds for $k = q+1$; thus $y_i(p_q+1) = y_i(\bar{p}+1)$. Therefore $y_i(p_q+1) \in \mathcal{K}(\widehat{p})$. But $y_i(p_q+1) \in \bar{\mathcal{H}}(p_q)$ because of (59). Therefore $y_i(p_q+1) \in \bar{\mathcal{H}}(p_q) \cap \mathcal{K}(\widehat{p})$. From this and (60) it follows that $y_i(p_q+1) \in \bar{\mathcal{K}}(p_q)$. Therefore node i is stationary at p_q, and because of this $y_i(p_q+1) = y_i(p_q)$. Hence $y_i(p_q) = y_i(p_{q+1})$ because of (61). Thus statements (i) and (ii) above are true for $k = \{q, q+1, \dots, m\}$. By induction, statements (i) and (ii) must hold for all $k \in \{1, 2, \dots, m\}$. ∎

4.2.3. Equilibrium states. By an *equilibrium state* of $\mathbb{S}$ we mean a state which does not change under the action of (33)–(37) under any conditions for every value of $p \in \mathcal{P}$. It is easy to see that equilibrium states are precisely those states $\{y, w\} \in \mathcal{X}$ for which

$$y_i = y_{ii_1} = \cdots = y_{ii_{m_i}} = w_i = w_{ii_1} \cdots = w_{ii_{m_i}} \quad \forall i \in \{1, 2, \dots, n\}.$$

Note that each equilibrium state is invariant under the action of (33)–(37) under any and all possible conditions. It is clear that if $\mathbb{S}$ is in an equilibrium state at p, then each node of $\mathbb{S}$ is stationary at p. It is also not difficult to see that if each node of $\mathbb{S}$ is stationary at p, then $\mathbb{S}$ is at an equilibrium state at time $p+1$.

4.2.4. Locally rendezvoused nodes. In what follows we will say node $i \in \{1, 2, \dots, n\}$ has *locally rendezvoused* at time p if $\mathcal{H}_i(p)$ is a single point, i.e., if $y_i(p) = y_{i_1}(p) = \cdots = y_{i_{m_i}}(p) = w_i(p) = w_{i_1}(p) = \cdots = w_{i_{m_i}}(p)$. Note that if a node has locally rendezvoused at p, it must be stationary at p. The following proposition provides a criterion for a node of $\mathbb{S}$ to be locally rendezvoused.

PROPOSITION 4. *Let $p_1 < p_2 < p_3 < p_4$ be four consecutive values of p in $\mathcal{P}_i$. If $y_i(p_4+1) \in \mathcal{K}(p_1)$, then node i is locally rendezvoused at $p = p_3$.*

The proof of Proposition 4 depends on the following lemmas.

LEMMA 7. *Let p_1 and p_2 be two consecutive values of p in $\mathcal{P}_i$. Suppose for some $i \in \{1, 2, \dots, n\}$ that $y_i(p_2+1) \in \mathcal{K}(p_1)$. Then*

$$y_i(p_1) = y_j(p_1), \quad j \in \mathcal{N}_i. \tag{62}$$

Proof of Lemma 7. By hypothesis, $y_i(p_2+1) \in \mathcal{K}(p_1)$. Therefore by Lemma 6, $y_i(p_1) = y_i(p_2)$ and node i is stationary at both p_1 and p_2. Because node i is stationary at p_2, $y_i(p_2) = v_{ij}(p_2)$, $j \in \mathcal{N}_i$. Therefore

$$y_i(p_1) = v_{ij}(p_2), \ j \in \mathcal{N}_i. \tag{63}$$

To justify (62) it is therefore enough to show that

$$v_{ij}(p_2) = y_j(p_1),\ j \in \mathcal{N}_i. \tag{64}$$

For this fix $j \in \mathcal{N}_i$ and define $k = P_i^{-1}(p_1)$ and $q = \lceil \bar{t}_{ik} \rceil_j$. Since $p_1 = P_i(k)$ and $\bar{t}_{j(q-1)} < \bar{t}_{ik} \leq \bar{t}_{jq}$,

$$P_j(q-1) < p_1 \leq P_j(q). \tag{65}$$

Let $\bar{q} = \lceil \bar{t}_{i(k+1)} \rceil_j$. Since $p_2 = P_i(k+1)$ and $\bar{t}_{j(\bar{q}-1)} < \bar{t}_{i(k+1)} \leq \bar{t}_{j\bar{q}}$,

$$P_j(\bar{q}-1) < p_2 \leq P_j(\bar{q}). \tag{66}$$

By Lemma 2, $\bar{q} \in \{q, q+1, q+2\}$. We claim that no matter which value $\bar{q}$ takes,

$$v_{ij}(p_2) \in \{y_j(P_j(q)),\ y_j(P_j(q+1)),\ y_j(P_j(q+2))\}. \tag{67}$$

To justify this claim, consider first the case when $\bar{q} = q$. Then $\mathcal{S}_i(k+1) \cap \mathcal{S}_j(q) \succ \tau_S$ because of Lemma 2. In general $\bar{q} = \lceil t_{p_2} \rceil_j$ because $t_{p_2} = \bar{t}_{i(k+1)}$. Thus in this case $q = \lceil t_{p_2} \rceil_j$. In addition $k+1 = P_i^{-1}(p_2)$. Therefore $\mathcal{S}_i(P_i^{-1}(p_2)) \cap \mathcal{S}_j(\lceil t_{p_2} \rceil_j) \succ \tau_S$. From this and (35) it follows that $v_{ij}(p_2) = y_j(p_2)$. In view of (36), $y_j(p) = y_j(P_j(q))$ for all values of p in the range $P_j(q-1) < p \leq P_j(q)$. But $P_j(q-1) < p_2 \leq P_j(q)$ because of (66). Therefore $y_j(p_2) = y_j(P_j(q))$. Thus $v_{ij}(p_2) = y_j(P_j(q))$ which proves that (67) holds in this case.

Now suppose that $\bar{q} = \{q+1, q+2\}$. In this case $v_{ij}(p_2)$ equals either $y_j(p_2)$ or $w_j(p_2)$ because of (35). In view of (36), $y_j(p) = y_j(P_j(\bar{q}))$ for $P_j(\bar{q}-1) < p \leq P_j(\bar{q})$. From this and (66) it follows that $y_j(p_2) = y_j(P_j(\bar{q}))$. Thus if $v_{ij}(p_2) = y_j(p_2)$, then $v_{ij}(p_2) = y_j(P_j(\bar{q}))$. Since $\bar{q} \in \{q+1, q+2\}$, (67) must hold in this situation. To prove that (67) also holds in the alternative situation, when $v_{ij}(p_2) = w_j(p_2)$, we exploit the relation $w_j(P_j(\bar{q}-1)+1) = y_j(P_j(\bar{q}-1))$ which is valid because of (34). In view of (37), $w_j(p)$ is constant for p in the range $P_j(\bar{q}-1) < p \leq P_j(\bar{q})$. But p_2 is in this range because of (66); clearly $P_j(\bar{q}-1)+1$ is as well. Therefore $w_j(p_2) = w_j(P_j(\bar{q}-1)+1)$. It follows that $w_j(p_2) = y_j(P_j(\bar{q}-1))$. Thus if $v_{ij}(p_2) = w_j(p_2)$, then $v_{ij}(p_2) = y_j(P_j(\bar{q}-1))$. Since $\bar{q} \in \{q+1, q+2\}$, (67) must hold in this situation too. Thus (67) holds under all conditions.

It will now be shown that

$$v_{ij}(p_2) = y_j(P_j(q)). \tag{68}$$

Consider first the situation when $v_{ij}(p_2) = y_j(P_j(s))$, where s is fixed at either value in $\{q+1, q+2\}$. Since node i is stationary at p_2, $v_{ij}(p_2) = y_i(p_2+1)$. Thus $y_j(P_j(s)) = y_i(p_2+1)$. By hypothesis, $y_i(p_2+1) \in \mathcal{K}(p_1)$. Thus $y_j(P_j(s)) \in \mathcal{K}(p_1)$. Moreover, $p_1 \leq P_j(q)$ because of (65). Thus by Lemma 6, $y_j(P_j(s)) = y_j(P_j(q))$. Therefore (68) holds when $v_{ij}(p_2) = y_j(P_j(s))$ for $s \in \{q+1, q+2\}$. In view of (67), the only other possibility is $v_{ij}(p_2) = y_j(P_j(q))$. Therefore (68) is true under all conditions.

It remains to be shown that (64) holds. In view of (36), $y_j(p) = y_j(P_j(q))$ for p in the range $P_j(q-1) < p \leq P_j(q)$. But (65) shows that p_1 is in this range so $y_j(p_1) = y_j(P_j(q))$. From this and (68) it follows that (64) holds. □

LEMMA 8. *For any integers $i \in \{1, 2, \ldots, n\}$ and $k \geq 1$,*

$$P_i(k+1) - P_i(k) \leq 2(n-1). \tag{69}$$

Moreover, for any integer $j \in \{1,2,\dots,n\}$ which is not equal to i, there are at most two successive positive integers $s, s+1$ such that

$$P_i(k) \le P_j(s) < P_j(s+1) \le P_i(k+1). \tag{70}$$

Proof of Lemma 8. Fix $i,j \in \{1,2,\dots,n\}$ and $k>0$. Let s and p be positive integers such that $\bar{t}_{ik} \le \bar{t}_{js} < \bar{t}_{j(s+p)} \le \bar{t}_{i(k+1)}$. These inequalities imply that $\bar{t}_{j(s+p)} - \bar{t}_{js} < \bar{t}_{i(k+1)} - \bar{t}_{ik}$. But $p\tau_D \le \bar{t}_{j(s+p)} - \bar{t}_{js}$ because of (2) and $\bar{t}_{i(k+1)} - \bar{t}_{ik} < 2\tau_D$ because of (3). Therefore $p\tau_D < 2\tau_D$, so $p=1$. Thus there are at most two successive event times $\bar{t}_{js}$ and $\bar{t}_{j(s+1)}$ for which $\bar{t}_{ik} \le \bar{t}_{js} < \bar{t}_{j(s+1)} \le \bar{t}_{i(k+1)}$. Moreover, since $\{j : j \in \{1,2,\dots,n\},\ j \ne i\}$ contains $n-1$ integers, it therefore follows that the number of distinct event times in the set $\{\bar{t}_{js} : j \in \{1,2,\dots,n\},\ j \ne i,\ s \ge 1\}$ which satisfy $\bar{t}_{ik} \le \bar{t}_{js} \le \bar{t}_{i(k+1)}$ does not exceed $2(n-1)$. But $P_i(\cdot)$ and $P_j(\cdot)$ are strictly monotone increasing, and $\bar{t}_{iq} = t_{P_i(q)}$, $\bar{t}_{jq} = t_{P_j(q)}$ for all $q \ge 1$. Therefore (69) is true, and there are at most two successive integers $s, s+1$ for which (70) holds. □

Proof of Proposition 4. By hypothesis, $y_i(p_4+1) \in \mathcal{K}(p_1)$, and $p_1 < p_2 < p_3 < p_4$. Therefore by Lemma 6,

$$y_i(p_2) = y_i(p_3) = y_i(p_4) = y_i(p_4+1), \tag{71}$$

and node i is stationary at p_3 and p_4. In view of (34), $w_i(p_2+1) = y_i(p_2)$. But $w_i(p) = w_i(p_3)$ for $p_2 < p \le p_3$ because of (37), so $w_i(p_2+1) = w_i(p_3)$. Therefore $y_i(p_2) = w_i(p_3)$. From this and (71) it follows that

$$y_i(p_3) = w_i(p_3). \tag{72}$$

By hypothesis $y_i(p_4+1) \in \mathcal{K}(p_1)$. In addition, $y_i(p_4+1) \in \mathcal{H}(p_3)$ because of (71). Thus $y_i(p_4+1) \in \mathcal{K}(p_1) \cap \mathcal{H}(p_3)$. In view of Lemma 5, $\mathcal{H}(p_3) \subset \mathcal{H}(p_1)$. Thus $\mathcal{K}(p_1) \cap \mathcal{H}(p_3) \subset \mathcal{K}(p_3)$. Therefore $y_i(p_4+1) \in \mathcal{K}(p_3)$. Hence by Lemma 7,

$$y_i(p_3) = y_j(p_3), \quad j \in \mathcal{N}_i. \tag{73}$$

In view of (72) and (73), node i will be rendezvoused at p_3 provided that

$$y_j(p_3) = w_j(p_3), \ \ j \in \mathcal{N}_i. \tag{74}$$

It will now be shown that this is true.

Fix $j \in \mathcal{N}_i$ and let $q = \lceil \bar{t}_{ik} \rceil_j$, where $k = P_i^{-1}(p_3)$. Equivalently, q is the unique integer for which $P_j(q-1) < p_3 \le P_j(q)$. In view of (36) and (37), $y_j(p)$ and $w_j(p)$ are constant for p in the range $P_j(q-1) < p \le P_j(q)$. Since both p_3 and $P_j(q-1)+1$ are in this range,

$$y_j(p_3) = y_j(P_j(q-1)+1) \quad \text{and} \quad w_j(p_3) = w_j(P_j(q-1)+1). \tag{75}$$

Note next that $y_i(p_4+1) = y_i(p_4)$ because node i is stationary at p_4. From this and (71) and (73) it follows that $y_i(p_4+1) = y_j(p_3)$. Thus $y_i(p_4+1) = y_j(P_j(q-1)+1)$. Since $y_i(p_4+1) \in \mathcal{K}(p_1)$ it must be true that

$$y_j(P_j(q-1)+1) \in \mathcal{K}(p_1). \tag{76}$$

In view of Lemma 8, there can be at most two consecutive integers in $\mathcal{P}_j$ which are in the set $\{p : P_j(q-1) \le p \le P_j(q)\}$. Since p_3 is one such integer, it must be true that p_1 is not in the set. Therefore $p_1 < P_j(q-1)$. From this, (76), and Lemma 6 it follows that $y_j(P_j(q-1)+1) = y_j(P_j(q-1))$. But $w_j(P_j(q-1)+1) = y_j(P_j(q-1))$ because of (34), so $w_j(P_j(q-1)+1) = y_j(P_j(q-1)+1)$. From this and (75) it follows that $w_j(p_3) = y_j(p_3)$. Therefore (74) is true. □

4.3. Error system. To analyze system behavior it is helpful to use a suitably defined error system $\bar{\mathbb{S}}$ derived from $\mathbb{S}$. Towards this end, for each $p \in \mathcal{P}$ let

$$\left.\begin{array}{l} \bar{y}_i(p) \ = y_i(p) - w_n(p), \\ \bar{w}_i(p) = w_i(p) - w_n(p) \end{array}\right\}, \qquad i \in \{1, 2, \ldots, n\}. \tag{77}$$

Note that

$$\bar{w}_n(p) = 0, \quad p \in \mathcal{P}. \tag{78}$$

Using (33)–(37) we obtain the update equations for $\{\bar{y}_i, \bar{w}_i\}$ defined for $p \in \mathcal{P}_i$ by

$$\bar{y}_i(p+1) = \bar{y}_i(p) + u_{m_i}(\bar{v}_{ii_1}(p) - \bar{y}_i(p), \ldots, \ \bar{v}_{ii_{m_i}}(p) - \bar{y}_i(p)) - \omega(p)\bar{y}_n(p), \tag{79}$$

$$\bar{w}_i(p+1) = \bar{y}_i(p) - \omega(p)\bar{y}_n(p), \tag{80}$$

where

$$\bar{v}_{ij}(p) = \left\{\begin{array}{ll} \bar{y}_j(p) & \text{if } \mathcal{S}_i(P_i^{-1}(p)) \cap \mathcal{S}_j(\lceil t_p \rceil_j) \succ \tau_S, \\ \bar{w}_j(p) & \text{otherwise} \end{array}\right\}, \qquad j \in \mathcal{N}_i, \tag{81}$$

and by

$$\bar{y}_i(p+1) = \bar{y}_i(p) - \omega(p)\bar{y}_n(p), \tag{82}$$

$$\bar{w}_i(p+1) = \bar{w}_i(p) - \omega(p)\bar{y}_n(p) \tag{83}$$

for $p \notin \mathcal{P}_i$. Here $\omega(p) = 1$ if $p \in \mathcal{P}_n$ and $\omega(p) = 0$ otherwise. In terms of error variables, the neighbor constraints given by (38) can be written as

$$||\bar{y}_i(p) - \bar{w}_j(p)|| \leq r \ \text{ if } \ \mathcal{S}_i(P_i^{-1}(p)) \cap \mathcal{S}_j(\lceil t_p \rceil_j) \not\succ \tau_S, \ \ p \in \mathcal{P}_i \quad j \in \mathcal{N}_i. \tag{84}$$

In what follows $\bar{\mathbb{S}}$ denotes the error system defined by (79)–(84). Note that the state of $\bar{\mathbb{S}}$, namely $\{\bar{y}_1(p), \ldots, \bar{y}_n(p), \bar{w}_1(p), \ldots, \bar{w}_{n-1}(p)\}$, takes values in the closed space $\bar{\mathcal{X}}$ of all lists $\{\bar{y}_1, \ldots, \bar{y}_n, \bar{w}_1, \ldots, \bar{w}_{n-1}\}$ satisfying

$$\left.\begin{array}{l} \bar{y}_i, \bar{w}_i \in \mathbb{R}^2, \\ ||\bar{y}_i - \bar{y}_j|| \leq r \end{array}\right\}, \qquad j \in \mathcal{N}_i, \quad i \in \{1, 2, \ldots, n\}. \tag{85}$$

It is possible to describe the preceding state update equations concisely as

$$\bar{x}(p+1) = f(p, \bar{x}(p)), \quad p \in \mathcal{P},$$

where $\bar{x}$ is the state $\{\bar{y}_1, \ldots, \bar{y}_n, \bar{w}_1, \ldots, \bar{w}_{n-1}\}$, $f(p, \cdot) : \bar{\mathcal{X}}(p) \to \bar{\mathcal{X}}$ is the next state map defined by (79)–(83), and $\bar{\mathcal{X}}(p)$ is the set of states in $\bar{\mathcal{X}}$ for which the neighbor constraints (84) hold at time p. It is important to recognize that even though there are infinitely many possible values of p, there are only finitely many distinct $\bar{\mathcal{X}}(p)$ and finitely many distinct $f(p, \cdot)$. Moreover, each $\bar{\mathcal{X}}(p)$ is closed because of (84), and each $f(p, \cdot)$ is continuous on its domain because each $u_m(\cdot)$ is. The following lemma summarizes these observations.

LEMMA 9. *There exist a finite index set $\mathcal{Q}$ and a finite set of continuous functions $F_q : \mathcal{X}_q \to \bar{\mathcal{X}}$ with closed domains such that the following statement is true. For any $p \in \mathcal{P}$ there is a $q \in \mathcal{Q}$ such that $\bar{\mathcal{X}}(p) = \mathcal{X}_q$ and $F_q(\cdot) = f(p, \cdot)$.*

The implication of Lemma 9 is that if $\{\bar{x}(p) : p \in \mathcal{P}\}$ is a trajectory of $\bar{\mathbb{S}}$, then there are indices $q(p) \in \mathcal{Q}$, $p \in \mathcal{P}$ such that

$$\bar{x}(p) = F_{q(p)}F_{q(p-1)} \cdots F_{q(\tau+1)}(\bar{x}(\tau)), \quad p > \tau, \quad p, \tau \in \mathcal{P}. \tag{86}$$

Here $F_{q(p)}F_{q(p-1)} \cdots F_{q(\tau+1)}$ is a "composed function," where by the composition of functions F_s and F_q we mean the function $F_qF_s : \mathcal{X}_{qs} \to \bar{\mathcal{X}}$, whose domain $\mathcal{X}_{qs}$ is the inverse image of $\mathcal{X}_q$ under F_s, and whose action on $\bar{x}$ is $\bar{x} \longmapsto F_q(F_s(\bar{x}))$. Composition is an associative operation, and because of this, the operation extends unambiguously to finite families of F_q. Note that any such composed function $F = F_{q_1}F_{q_2} \cdots F_{q_k}$ has a closed domain on which it is continuous.

Suppose that $\bar{p} > 0$ is fixed. If follows from the preceding that there are $q(p) \in \mathcal{Q}$ such that

$$\bar{x}(p + \bar{p}) = F_{q(p+\bar{p})}F_{q(p+\bar{p}-1)} \cdots F_{q(p+1)}(\bar{x}(p)), \quad p \in \mathcal{P}. \tag{87}$$

It is important to recognize that even though the composed function $F_{q(p+\bar{p})}F_{q(p+\bar{p}-1)} \cdots F_{q(p+1)}(\bar{x}(p))$ depends on p, there can be only a finite number of such composed functions. This is because the family of maps $\{F_q : q \in \mathcal{Q}\}$ is a finite set and because the composed functions in question are all compositions of exactly $\bar{p}$ maps in the family. The following proposition summarizes these observations.

PROPOSITION 5. *Let $\bar{p} > 0$ be fixed. There exist a finite index set $\bar{\mathcal{Q}}$, a finite set of closed subsets $\bar{\mathcal{X}}_q \subset \bar{\mathcal{X}}$, and a finite set of continuous maps $D_q : \bar{\mathcal{X}}_q \to \bar{\mathcal{X}}$, $q \in \bar{\mathcal{Q}}$, with the following property. For each trajectory $\{\bar{x}(p) : p \in \mathcal{P}\}$ of $\bar{\mathbb{S}}$, and each $p \in \mathcal{P}$, there is a $q \in \bar{\mathcal{Q}}$ such that*

$$\bar{x}(\bar{p} + p) = D_q(\bar{x}(p)). \tag{88}$$

4.4. Global rendezvous. It is natural to say that the n nodes of $\mathbb{S}$ have (globally) *rendezvoused* at time p if $\mathcal{H}(p)$ is a single point, i.e., if $y_1(p) = y_2(p) = \cdots = y_n(p) = w_1(p) = w_2(p) = \cdots = w_n(p)$. In view of the definitions of t_p and the y_i and w_i in (39), it is clear that the rendezvousing of all n nodes at time p implies the rendezvousing of all n agents at time t_p. It is also clear that the rendezvousing of all n nodes at time p implies that each node has locally rendezvoused at p. Under certain conditions the converse is also true.

LEMMA 10. *Suppose $\mathbb{G}$ is a connected graph. Suppose in addition that $\{\{y(p), w(p)\} : p \in \mathcal{P}\}$ is the trajectory of $\mathbb{S}$ defined by (39). If for some $i \in \{1, 2, \ldots, n\}$ and $p \in \mathcal{P}_i$, node i is locally rendezvoused, then the n nodes of $\mathbb{S}$ have globally rendezvoused.*

Proof of Lemma 10. Suppose node i is locally rendezvoused at $p \in \mathcal{P}_i$. Then $y_i(p) = y_j(p)$ and $w_i(p) = y_j(p), j \in \mathcal{N}_i$. Moreover, since node i is locally rendezvoused at p, it must be stationary at p. Therefore $y_i(p+1) = y_i(p)$; in addition, $w_i(p+1)) = y_i(p)$ because of (34). Thus $y_i(p+1) = y_j(p)$ and $w_i(p+1) = y_j(p)$, $j \in \mathcal{N}_i$. Fix $j \in \mathcal{N}_i$ and $k \in \mathcal{N}_j$. Then $||y_j(p) - y_k(p)|| \leq r$ because of the definition of $\mathcal{X}$. Therefore $||y_i(p+1) - y_k(p)|| \leq r$ and $||w_i(p+1) - y_k(p)|| \leq r$. It follows from Lemma 4 that $k \in \mathcal{N}_i$. Since this holds for every $k \in \mathcal{N}_j$, it must be true that $\mathcal{N}_j \subset \mathcal{N}_i$. Since j is arbitrary, this must be true for all $j \in \mathcal{N}_i$. Since $\mathbb{G}$ is connected, this can happen only if $\mathbb{G}$ is complete. Thus $\mathcal{N}_i = \{1, 2, \ldots, n\}$ which means that $\mathcal{H}_i(p) = \mathcal{H}(p)$. By hypothesis $\mathcal{H}_i(p)$ is a single point. Therefore $\mathcal{H}_i(p)$ is also a single point, so the n nodes of $\mathbb{S}$ have globally rendezvoused. □

Establishing the preceding result requires one to be able to conclude that if for some $i, j \in \{1, 2, \ldots, n\}$ and some $p \in \mathcal{P}_i$, nodes i and j are in the same "position" in

the sense that $y_i(p) = y_j(p) = w_i(p)$, then $\mathcal{N}_j \subset \mathcal{N}_i$. In words, what this is roughly saying is that if node j is in the same position as node i, then node j's "neighbors" must also be neighbors of node i. This *transitivity property* is not true in general, but it is true if $y(p)$ and $w(p)$ are defined by (39). This is a consequence of Lemma 4.

The following proposition shows that if $\mathcal{H}$ does not change for a sufficiently long period of time, then the n nodes have rendezvoused.

PROPOSITION 6. *Suppose that $\mathbb{G}$ is a connected graph. Suppose in addition that $\{y(p), w(p) : p \in \mathcal{P}\}$ is the trajectory of $\mathbb{S}$ defined by* (39). *Suppose that p_a and p_b are values in $\mathcal{P}$ for which $p_b - p_a \geq 8n$ and*

$$\operatorname{dia}\{\mathcal{H}(p_a)\} = \operatorname{dia}\{\mathcal{H}(p_b)\}. \tag{89}$$

Then the n nodes of $\mathbb{S}$ have rendezvoused at $p = p_b$.

Proof of Proposition 6. Choose $i \in \{1, 2, \ldots, n\}$ so that for some $z \in \mathcal{H}(p_b)$, $||y_i(p_b) - z|| = \operatorname{dia}\{\mathcal{H}(p_b)\}$. Then $y_i(p_b) \in \mathcal{K}(p_b)$. In view of Lemma 5, $\mathcal{H}(p_b) \subset \mathcal{H}(p_a)$. Therefore $y_i(p_b), z \in \mathcal{H}(p_a)$. Moreover, $||y_i(p_b) - z|| = \operatorname{dia}\{\mathcal{H}(p_a)\}$ because of (89); thus

$$y_i(p_b) \in \mathcal{K}(p_a). \tag{90}$$

Let p_4 be the largest value of $p \in \mathcal{P}_i$ such that $p_4 < p_b$. Define $k = P_i^{-1}(p_4) - 3$ so that $P_i(k+3) = p_4$. Then $p_4 < p_b \leq P_i(k+4)$. By (69),

$$p_b - p_4 \leq 2(n-1). \tag{91}$$

In view of (36), $y_i(p)$ is constant for p in the range $p_4 < p \leq P_i(k+4)$. Since both $p_4 + 1$ and p_b are in this range, $y_i(p_4 + 1) = y_i(p_b)$. Thus

$$y_i(p_4 + 1) \in \mathcal{K}(p_a). \tag{92}$$

Define $p_1 = P_i(k)$, $p_2 = P_i(k+1)$, and $p_3 = P_i(k+2)$. Clearly $p_1 < p_2 < p_3 < p_4$. Moreover, $p_{j+1} - p_j \leq 2(n-1)$, $j \in \{1, 2, 3\}$, because of (69). From these inequalities and (91) it follows that $p_b - p_1 \leq 8(n-1)$. By hypothesis, $p_b - p_b \geq 8n$. Therefore $p_a < p_1$. In view of Lemma 5, $\mathcal{H}(p_4) \subset \mathcal{H}(p_1)$ and $\mathcal{H}(p_1) \subset \mathcal{H}(p_a)$. Therefore $\mathcal{H}(p_1) \cap \mathcal{K}(p_a) \subset \mathcal{K}(p_1)$. But $\mathcal{H}(p_4 + 1) \subset \mathcal{H}(p_1)$ because of Lemma 5; thus $y_i(p_4 + 1) \in \mathcal{H}(p_1)$. This and (92) imply that $y_i(p_4 + 1) \in \mathcal{H}(p_1) \cap \mathcal{K}(p_a)$. Therefore $y_i(p_4 + 1) \in \mathcal{K}(p_1)$. From this and Proposition 4 it follows that node i has locally rendezvoused at p_3. Therefore by Lemma 10, the n nodes of $\mathbb{S}$ are rendezvoused at p_3. □

The following theorem is our main convergence result concerning $\mathbb{S}$. The main result of this paper, Theorem 1, is an immediate consequence.

THEOREM 2. *Let $\{\{y(s), w(s)\} : p \in \mathcal{P}\}$ be the trajectory of $\mathbb{S}$ defined by* (39). *If $\mathbb{G}$ is a connected graph, then*

$$\lim_{s \to \infty} \operatorname{dia}\langle y_1(s), y_2(s), \ldots, y_n(s), w_1(s), w_2(s), \ldots, w_n(s)\rangle = 0. \tag{93}$$

Proof of Theorem 2. In what follows we write $x(p)$ for $\{y_1(p), \ldots, y_n(p), w_1(p), \ldots, w_n(p)\}$ and $\bar{x}(p)$ for the error vector $\{\bar{y}_1(p), \ldots, \bar{y}_n(p), \bar{w}_1(p), \ldots, \bar{w}_{n-1}(p)\}$ defined by (77). Let $V : \mathcal{X} \to \mathbb{R}$ denote the diameter function $x \longmapsto \operatorname{dia}\langle y_1, y_2, \ldots, y_n, w_1, w_2, \ldots, w_n\rangle$. Similarly, let $\bar{V} : \bar{\mathcal{X}} \to \mathbb{R}$ denote the diameter function $\bar{x} \longmapsto \operatorname{dia}\langle \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n, \bar{w}_1, \bar{w}_2, \ldots, w_{n-1}, 0\rangle$. Note that

$$V(x(p)) = \bar{V}(\bar{x}(p)). \tag{94}$$

Note in addition that because $0 \in \langle \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n, \bar{w}_1, \bar{w}_2, \ldots, w_{n-1}, 0\rangle$, $\bar{V}$ is radially unbounded, whereas V is not.

As a consequence of Lemma 5, $V(x(p))$ is a monotone nonincreasing function of p. Clearly $V(x(p))$ is bounded below by 0. Moreover, $V(x(p))$ is bounded above by $V(x(0))$ because $V(\cdot)$ is continuous. Therefore there must exist a finite limit

$$V^* = \lim_{p\to\infty} V(x(p)).$$

We claim that $V^* = 0$. To prove this claim, suppose it is false. Then $V^* > 0$. This means that the trajectory $\{x(p) : p \in \mathcal{P}\}$ cannot contain any points in the set $\mathcal{E} = \{x : V(x) = 0\}$. To proceed, fix $\bar{s} > 8n$ and let $\Delta(x(p))$ denote the difference

$$\Delta(x(p)) = V(x(\bar{p} + p)) - V(x(p)). \tag{95}$$

Since $V(x(p))$ is monotone nonincreasing, $\Delta(x(p)) \leq 0, p \in \mathcal{P}$. In the light of Proposition 6 and the fact that $\mathcal{E}$ has no points in common with $\{x(p) : p \in \mathcal{P}\}$, one can conclude that $\Delta(x(p)) \neq 0, p \in \mathcal{P}$. Therefore

$$\Delta(x(p)) < 0, \quad p \in \mathcal{P}. \tag{96}$$

Define $\bar{\Delta}(\bar{x}(p))$ as

$$\bar{\Delta}(\bar{x}(p)) = \bar{V}(\bar{x}(\bar{p} + p)) - \bar{V}(\bar{x}(p)). \tag{97}$$

In view of (94)

$$\Delta(x(p)) = \bar{\Delta}(\bar{x}(p)). \tag{98}$$

Therefore

$$\bar{\Delta}(\bar{x}(p)) < 0, \quad p \in \mathcal{P}. \tag{99}$$

According to Proposition 5, for each $p \in \mathcal{P}$ there is a continuous function D_q such that $\bar{x}(p + \bar{p}) = D_q(x(p))$. Let $\mathcal{W}_q$ denote the set of state pairs $(\bar{x}(p + \bar{p}), \bar{x}(p))$ along the given trajectory of $\bar{\mathbb{S}}$ for which this formula holds. It follows that

$$\{(x(s + \bar{s}), x(s)) : s \in \mathcal{S}\} = \bigcup_{q\in\mathcal{Q}} \mathcal{W}_q$$

and that each $\mathcal{W}_q$ is a closed set. We claim that each $\mathcal{W}_q$ is bounded as well. This is in fact so because of (94), because $\bar{V}$ is radially unbounded, and because $0 \leq V(x(p)) \leq V(x(0)) < \infty$.

For $(\widehat{x}, \bar{x}) \in \mathcal{W}_q$ define $\Delta_q : \mathcal{W}_q \to \mathbb{R}$ so that $(\widehat{x}, \bar{x}) \longmapsto \bar{V}(D_q(\widehat{x})) - V(\bar{x})$. Note that Δ_q is a continuous function on $\mathcal{W}_q$ whose value at each point $(\widehat{x}, \bar{x}) \in \mathcal{W}_q$ agrees with $\bar{\Delta}(\bar{x}(p))$ for some p. It follows from (99) that

$$\Delta_q(\widehat{x}, \bar{x}) < 0, \quad (\widehat{x}, \bar{x}) \in \mathcal{W}_q.$$

Define

$$\mu_q = \sup_{(\widehat{x},\bar{x})\in\mathcal{W}_q} \Delta_q(\widehat{x}, \bar{x}).$$

Since $\mathcal{W}_q$ is compact and Δ_q is negative and continuous on $\mathcal{W}_q$, it must be true that $\mu_q < 0$. Let

$$\mu = \max_{q \in \mathcal{Q}} \mu_i.$$

Since $\mathcal{Q}$ is finite, $\mu < 0$. Clearly

$$\Delta_q(\widehat{x}, \bar{x}) \leq \mu, \quad (\widehat{x}, \bar{x}) \in \mathcal{W}_q, \quad q \in \mathcal{Q}. \tag{100}$$

Note that by construction, for each $p \in \mathcal{S}$ there must be a $q \in \mathcal{Q}$ such that $\bar{\Delta}(\bar{x}(p)) = \Delta_q(\bar{x}(p + \bar{p}), \bar{x}(p))$. From this and (100) it follows that

$$\bar{\Delta}(\bar{x}(p)) \leq \mu, \quad p \in \mathcal{P}.$$

Therefore

$$\Delta(x(p)) \leq \mu, \quad p \in \mathcal{P},$$

because of (98). Note that

$$V(x(p + \bar{p})) - V(x(p)) = \Delta(x(p)) \leq \mu, \quad p \in \mathcal{P}.$$

Thus by summing,

$$V(x(p + k\bar{p})) \leq V(x(p)) + k\mu, \quad k \geq 1.$$

Therefore, for k sufficiently large $V(x(p + k\bar{p}))$ must be negative because $\mu < 0$. But this is impossible because $V(\cdot)$ is positive semidefinite. Hence V^* cannot be positive. This concludes the proof. □

5. Concluding remarks. The analysis used in this paper exploits ideas which appear to have much in common with the embedding process discussed in Chapter 7 of [2] for analyzing "partially asynchronous iterative algorithms." This suggests that the tools developed in [2] may be helpful in further understanding the asynchronous system considered in this paper.

The asynchronous multi-agent rendezvous problem we have considered serves as an example of the type of problem to which the idea of analytic synchronization can be applied. The asynchronous version of the flocking problem considered in [3] provides another. Despite these examples, there are several unsettled issues concerning the analytic synchronization idea. First, it is not clear what the general process is for choosing a state vector. Second, it is also not clear what the exact conditions are on an asynchronously interacting set of dynamical systems for analytic synchronization to be possible. The examples provided by this paper and by [3] may help to more precisely formulate these issues and to lead to their resolution.

REFERENCES

[1] H. Ando, Y. Oasa, I. Suzuki, and M. Yamashita, *Distributed memoryless point convergence algorithm for mobile robots with limited visibility*, IEEE Trans. Robotics Automation, 15 (1999), pp. 818–828.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[3] M. Cao, A. S. Morse, and B. D. O. Anderson, *Agreeing asynchronously*, IEEE Trans. Automat. Control, submitted.

[4] M. Cieliebak, P. Flocchini, G. Prencipe, and N. Santoro, *Solving the robot gathering problem*, in Automata, Languages and Programming (ICALP, 2003), Lecture Notes in Comput. Sci. 2719, Springer, Berlin, 2003, pp. 1181–1196.

[5] P. Flocchini, G. Prencipe, N. Santoro, and P. Widmayer, *Gathering of asynchronous mobile robots with limited visibility*, Theoret. Comput. Sci., 337 (2005), pp. 1–3, 147–168.

[6] C. Godsil and G. Royle, *Algebraic Graph Theory*, Springer, New York, 2001.

[7] J. Lin, A. S. Morse, and B. D. O. Anderson, *The multi-agent rendezvous problem—The asynchronous case*, in Proceedings of the 43rd IEEE Conference on Decision and Control, 2004, pp. 1926–1931.

[8] J. Lin, A. S. Morse, and B. D. O. Anderson, *The multi-agent rendezvous problem. Part* 1: *The synchronous case*, SIAM J. Control Optim., 46 (2007), pp. 2096–2119.

[9] G. Prencipe, *CORDA: Distributed coordination of a set of autonomous mobile robots*, in Proceedings of the European Research Seminar on Advances in Distributed Systems, 2001, pp. 185–190.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2148–2181

THE ANALYSIS OF EXACT CONTROLLABILITY OF NEUTRAL-TYPE SYSTEMS BY THE MOMENT PROBLEM APPROACH*

RABAH RABAH† AND GRIGORY M. SKLYAR‡

Abstract. The problem of exact null-controllability is considered for a wide class of linear neutral-type systems with distributed delay. The main tool of the analysis is the application of the moment problem approach and the theory of the basis property of exponential families. A complete characterization of this problem is given. The minimal time of controllability is specified. The results are based on the analysis of the Riesz basis property of eigenspaces of the neutral-type systems in Hilbert space.

Key words. neutral-type systems, exact controllability, moment problem, Riesz basis, distributed delays

AMS subject classifications. 93B05, 93C23, 93C25

DOI. 10.1137/060650246

1. Introduction.

Many applied problems from physics, mechanics, biology, and other fields can be described by partial differential equations or delay differential equations. This leads to the construction and study of the infinite-dimensional system theory concerning also the systems with control. In this context the problem of controllability for distributed parameter systems leads to the study of the abstract controllability problem in infinite-dimensional spaces, which may be formulated in Hilbert spaces as follows. Consider the abstract system

$$\dot{x} = \mathcal{A}x + \mathcal{B}u, \tag{1.1}$$

where $x(t) \in X, u(t) \in U$, X and U being Hilbert spaces, $\mathcal{A}$ is the generator of a C_0-semigroup $\mathrm{e}^{\mathcal{A}t}$, and $\mathcal{B} \in \mathcal{L}(U, X)$ is a bounded operator. The problem of controllability is to find all the states x_T that can be reached from a fixed initial state (say 0) at a finite time T by the choice of the controls $u(\cdot) \in L_2(0, T; U)$. The mild solution of the system (1.1) is given by

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

The reachability set from 0 at time T is defined by

$$\mathcal{R}_T = \left\{ x : x = \int_0^T \mathrm{e}^{\mathcal{A}t}\mathcal{B}u(t)\mathrm{d}t, \quad u(\cdot) \in L_2(0, T; U) \right\}.$$

*Received by the editors January 18, 2006; accepted for publication (in revised form) June 11, 2007; published electronically December 21, 2007. This work was partially supported by the French-Polish grant Polonium 07599VH, by École Centrale de Nantes, and by the French research center CNRS.

http://www.siam.org/journals/sicon/46-6/65024.html

†IRCCyN, UMR CNRS 6597, École des Mines de Nantes, 4 rue Alfred Kastler, BP 20722 44307 Nantes Cedex 3, France (rabah@emn.fr).

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

For finite-dimensional systems the natural concept of controllability is when $\mathcal{R}_T = X$ (Kalman). For infinite-dimensional systems, as has been pointed out by several authors (Fattorini, Triggiani, Russel, Balakrishnan, and others), this concept is not realistic. It is easy to show that $\mathcal{R}_{T_1} \subset \mathcal{R}_{T_2}$ as $T_1 < T_2$. In general, there is no universal time T_0 such that $\mathcal{R}_{T_0} = \mathcal{R}_T$ for all $T > T_0$. However, for several classes of systems important for application this property holds (hyperbolic-type and neutral-type systems). In these cases, a natural way to formulate the controllability problem is the following setting:

(i) to find the maximal possible set $\mathcal{R}_T$ (depending on T),

(ii) to find the minimal T for which the set $\mathcal{R}_T$ becomes the maximal possible.

In order to obtain more profound and precise results by using this approach, it is important to consider a concrete class of systems and to use the specificity of this class. In this paper, we consider the problem of controllability for a general class of neutral systems with distributed delays given by the equation

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}[z(t) - Kz_t] = Lz_t + Bu(t), \quad t \geq 0, \\ z_0 = f, \end{cases} \tag{1.2}$$

where $z_t : [-1, 0] \to \mathbb{C}^n$ is the history of z defined by $z_t(s) = z(t + s)$. The difference and delay operators K and L, respectively, are defined by

$$Kf = A_{-1}f(-1) \quad \text{and} \quad Lf = \int_{-1}^{0} A_2(\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}f(\theta)\,\mathrm{d}\theta + \int_{-1}^{0} A_3(\theta)f(\theta)\,\mathrm{d}\theta$$

for $f \in H^1([-1, 0], \mathbb{C}^n)$, where A_{-1} is a constant $n \times n$ matrix, A_2, A_3 are $n \times n$ matrices whose elements belong to $L_2(-1, 0)$, and B is a constant $n \times r$ matrix.

We consider the operator model of the neutral-type system (1.2) introduced by Burns, Herdman, and Stech [3] in product spaces (see also [5]). The state space is $M_2(-1, 0; \mathbb{C}^n) = \mathbb{C}^n \times L_2(-1, 0; \mathbb{C}^n)$, shortly M_2, and (1.2) can be reformulated as

$$\dot{x}(t) = \mathcal{A}x(t) + \mathcal{B}u(t), \quad x(0) = \begin{pmatrix} y \\ f \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} B \\ 0 \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} 0 & L \\ 0 & \frac{\mathrm{d}}{\mathrm{d}\theta} \end{pmatrix}, \tag{1.3}$$

with $\mathcal{D}(\mathcal{A}) = \{(y, z(\cdot)) \in M_2 : z \in H^1([-1, 0]; \mathbb{C}^n), y = z(0) - A_{-1}z(-1)\}$.

In the particular case when $A_2(\theta) = A_3(\theta) = 0$, which corresponds to $L = 0$, we will use the notation $\widetilde{\mathcal{A}}$ for $\mathcal{A}$.

Suppose that the initial condition for the system (1.2) is $z(t) = z_0(t), t \in [-1, 0]$, and let us put $z_t(\theta) = z(t + \theta), \theta \in [-1, 0]$, and $y = z(0) - A_{-1}z(-1)$. The semigroup generated by $\mathcal{A}$ is given by

$$\mathrm{e}^{\mathcal{A}t}\begin{pmatrix} y \\ z_0(\cdot) \end{pmatrix} = \begin{pmatrix} z_t(0) - A_{-1}z_t(-1) \\ z_t(\cdot) \end{pmatrix} = \begin{pmatrix} z(t) - A_{-1}z(t-1) \\ z(t + \cdot) \end{pmatrix}.$$

It can be shown that the reachability set $\mathcal{R}_T$ is such that $\mathcal{R}_T \subset \mathcal{D}(\mathcal{A})$ for all $T > 0$. This is a consequence of the fact that for all $u(\cdot) \in L_2$ the corresponding solution of (1.2) is in H^1 and then the solution of (1.3) is in $\mathcal{D}(\mathcal{A})$ (see [5, Proposition 2.2] for the existence of the solution and [5, Corollary 2.7] for the property of the reachability subset). This naturally leads to the following definition of exact controllability.

Definition 1.1. *The system* (1.3) *is exactly null-controllable by controls from* L_2 *at the time* T *if* $\mathcal{R}_T = \mathcal{D}(\mathcal{A})$. *This means that the set of solutions of the system* (1.2), $\{z(t), t \in [T - 1, T]\}$, *coincides with* $H^1([T - 1, T]; \mathbb{C}^n)$.

This problem was the focus of attention of several authors in the 1970s and 1980s. The main results were devoted to systems with one or several discrete delays. This may be explained by the fact that the explicit, in this case, form of solutions is known and, as a result, the semigroup describing the solutions of (1.2) is known explicitly.

The main result for the system

$$\dot{z}(t) - A_{-1}\dot{z}(t-h) = A_0 z(t) + A_1 z(t-h) + Bu$$

is that the exact controllability holds if and only if (see [9, 12])

(i) $\operatorname{rank}(\,\Delta(\lambda) \quad B\,) = n$,

(ii) $\operatorname{rank}(\,B \quad A_{-1}B \quad \cdots \quad A_{-1}^{n-1}B\,) = n$,

where $\Delta(\lambda) = \lambda I - \lambda A_{-1}\mathrm{e}^{-\lambda h} - A_0 - A_1\mathrm{e}^{-\lambda h}$. For the particular case of scalar control (B is $n \times 1$ matrix) the time of exact controllability is given in [6]: $T > nh$, where h is the delay. For the general case, it is shown in [2] that the reachability set cannot increase for $T > nh$.

The case of noncommensurate delays with a distributed term was precisely studied in the paper by Yamamoto [16]. General conditions were given using the input-output technique. Conditions of approximate controllability (in [16], quasi reachability) in the time domain were explicitly given for a system without distributed delay (see also [8]).

In contrast to the above-mentioned works, we consider the model with distributed delays (1.2). In this case, we know only that the solution of (1.2) exists but the corresponding semigroup is not explicitly known. Then the technique using the explicit form of the solution, via an expression of the semigroup, cannot be used. So one needs another tool to analyze the controllability. In the similar situation of the controllability problems for hyperbolic systems, the powerful technique of the moment problem has been proved to be useful. It is caused by the fact that the operators corresponding to hyperbolic systems are as a rule skew-adjoint or close to skew-adjoint and then they possess a basis of eigenvectors. The expansion of the steering conditions in this basis allows the controllability problem for these systems to be reduced to a trigonometric problem with respect to some families of exponentials. Thus, the further analysis concerns the solvability of the non-Fourier trigonometric moment problem and is based on the profound theory of the Riesz bases of exponentials. This theory, originated by the famous Paley–Wiener theorem, has essentially been developed in the last decades (see monographs by Avdonin and Ivanov [1] and by Young [17] and the references therein).

The main idea of our work is to apply the moment problem method to the analysis of controllability of neutral-type systems. Note in this context that the case of neutral-type systems differs essentially from those mentioned above since the operator $\mathcal{A}$ of the system is not skew-adjoint and, moreover, may not have a basis of eigenvectors or even generalized eigenvectors. The first element of our consideration is the spectral analysis of the operator model (1.3) given in our previous works, together with Rezounenko [10, 11]. In these papers, it is shown that, under the condition that the matrix A_{-1} is not singular, the operator $\mathcal{A}$ (even if it does not verify the Riesz basis property) possesses a Riesz basis of finite-dimensional invariant subspaces. This allows the construction of a special Riesz basis in the space M_2 in which the steering conditions

$$\begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix} = \int_0^T \mathrm{e}^{\mathcal{A}(T-\tau)}\mathcal{B}u(\tau)\mathrm{d}\tau \tag{1.4}$$

take the form of a moment problem quadratically close to some special non-Fourier moment problems with respect to a family of quasi polynomials. These questions are

considered in sections 2 and 3. Let us notice that the detailed attention accorded to the construction of the needed Riesz basis is essentially motivated by the fact that, in the general case, the operator $\mathcal{A}$ may not possess a spectral Riesz basis. Otherwise, for example, if the eigenvalues of the matrix A_{-1} are simple, our construction would be much simpler. The main tool of the analysis of solvability of the obtained moment problem is based on the theory of families of exponentials [1, 17]. The basic elements of this approach used in our paper are given in section 4. Below we give a complete analysis of the controllability problem for neutral-type systems. In the course of the main part of the work, we consider the case when $\det A_{-1} \neq 0$. In this case, the controllability of system (1.2) is equivalent to the solvability of the moment problem obtained in section 3. We consider first the single input case in section 5 and give not only the conditions of exact null-controllability but also determine the time of controllability. These results are an extension of the result obtained in [6]. In section 6, we consider the solvability of the moment problem for the multivariable case ($\dim B = n \times r, r > 1$). We introduce some special indices m_1 and $\overline{m}$ which enable the moment problem to be characterized. We show that the exact null-controllability holds for time $T > \overline{m}$ and does not hold for $T < m_1$. Finally, we complete the results on controllability by getting rid of the assumption $\det A_{-1} \neq 0$ in section 7. We then obtain the precise time of controllability using the first controllability index of the pair (A_{-1}, B), say n_1 (cf., for example, [15, Chapter 5]). Our main result is the following.

THEOREM 1.2. *The system* (1.3) *is exactly null-controllable if and only if the following conditions are verified.*

(i) *There is no* $\lambda \in \mathbb{C}$ *and* $y \in \mathbb{C}^n, y \neq 0$, *such that* $\Delta^*_{\mathcal{A}}(\lambda)y = 0$ *and* $B^*y = 0$, *where*

$$\Delta^*_{\mathcal{A}}(\lambda) = \lambda I - \lambda \mathrm{e}^{-\lambda} A^*_{-1} - \lambda \int_{-1}^{0} \mathrm{e}^{\lambda s} A^*_2(s)\mathrm{d}s - \int_{-1}^{0} \mathrm{e}^{\lambda s} A^*_3(s)\mathrm{d}s,$$

or equivalently $\operatorname{rank}(\,\Delta_{\mathcal{A}}(\lambda) \quad B\,) = n$ *for all* $\lambda \in \mathbb{C}$.

(ii) *There is no* $\mu \in \sigma(A_{-1})$ *and* $y \in \mathbb{C}^n, y \neq 0$, *such that* $A^*_{-1}y = \bar{\mu}y$ *and* $B^*y = 0$, *or equivalently* $\operatorname{rank}(\,B \quad A_{-1}B \quad \cdots \quad A^{n-1}_{-1}B\,) = n$.

If conditions (i) *and* (ii) *hold, then the system is controllable at the time* $T > n_1$ *and not controllable at the time* $T \leq n_1$, *where* n_1 *is the controllability index of the pair* (A_{-1}, B).

If the delay is h instead of 1, the time of exact controllability is $T = n_1 h$.

2. The choice of the basis. In this section, we assume that the matrix A_{-1} is not singular, $\det A_{-1} \neq 0$.

Let us recall [10] that the spectrum of $\widetilde{\mathcal{A}}$ (the state operator corresponding to the case $A_2 = A_3 = 0$) consists of only the eigenvalues that are the roots of the equation $\det \Delta_{\widetilde{\mathcal{A}}}(\lambda) = \det(\lambda I - \lambda \mathrm{e}^{-\lambda} A_{-1}) = 0$, i.e.,

$$\sigma(\widetilde{\mathcal{A}}) = \{\lambda^{(k)}_m = \ln|\mu_m| + \mathrm{i}(\arg \mu_m + 2k\pi)\} \cup \{0\},$$

where $\{\mu_m, m = 1, \ldots, \ell\} = \sigma(A_{-1})$.

The operator $\widetilde{\mathcal{A}}$ possesses a Riesz basis of generalized eigenvectors which may be characterized as follows (see [10, 11]).

Let ν_m be the number of Jordan blocks corresponding to $\mu_m \in \sigma(A_{-1})$ and let $p_{m,j},\ j = 1, \ldots, \nu_m$, be the dimension of the corresponding blocks; then

1. to any $\lambda_m^{(k)} \neq 0$ and to any $j = 1, \dots, \nu_m$ there corresponds a Jordan chain of generalized eigenvectors of $\widetilde{\mathcal{A}}$, noted $\{\widetilde{\varphi}_{m,k}^{j,1}, \widetilde{\varphi}_{m,k}^{j,2}, \dots, \widetilde{\varphi}_{m,k}^{j,p_{m,j}}\}$ such that

$$\widetilde{\mathcal{A}}\widetilde{\varphi}_{m,k}^{j,1} = \lambda_m^{(k)}\widetilde{\varphi}_{m,k}^{j,1}, \quad (\widetilde{\mathcal{A}} - \lambda_m^{(k)} I)\widetilde{\varphi}_{m,k}^{j,s} = \widetilde{\varphi}_{m,k}^{j,s-1}, \qquad s = 2, \dots, p_{m,j};$$

2. the root space of $\widetilde{\mathcal{A}}$ corresponding to $0 \in \sigma(\widetilde{\mathcal{A}})$ is of dimension

$$n + \dim \operatorname{Ker}(A_{-1} - I)^n.$$

If $1 = \mu_g \in \sigma(A_{-1}),\ g \in \{1, \dots, \ell\}$, then for any $j \in \{1, \dots, \nu_g\}$ there exists a Jordan chain $\{\widetilde{\varphi}_0^{j,1}, \widetilde{\varphi}_0^{j,2}, \dots, \widetilde{\varphi}_0^{j,p_{g,j}+1}\}$ such that

$$\widetilde{\mathcal{A}}\widetilde{\varphi}_0^{j,1} = 0, \quad \widetilde{\mathcal{A}}\widetilde{\varphi}_0^{j,s} = \widetilde{\mathcal{A}}\widetilde{\varphi}_0^{j,s-1}, \qquad s = 2, \dots, p_{m,j} + 1,$$

and, besides, there exist $n - \nu_g$ linearly independent eigenvectors $\widetilde{\varphi}_0^j,\ j = \nu_g + 1, \dots, n$, such that $\widetilde{\mathcal{A}}\widetilde{\varphi}_0^j = 0$;

3. any collection $\{\widetilde{\varphi}_{m,k}^{j,s},\ s = 1, \dots, p_{m,j},\ j = 1, \dots, \nu_m\}$ forms a basis in the space $\operatorname{Ker}(\widetilde{\mathcal{A}} - \lambda_m^{(k)} I)^n,\ \lambda_m^{(k)} \neq 0$. The collection

$$\{\widetilde{\varphi}_0^{j,s},\ s = 1, \dots, p_{m,j} + 1,\ j = 1, \dots, \nu_m\} \bigcup \{\widetilde{\varphi}_0^j,\ j = \nu_g + 1, n\}$$

forms a basis in $\operatorname{Ker}\widetilde{\mathcal{A}}^n$.

In the following, we refer to this basis as $\{\widetilde{\varphi}\}$ omitting the indices when they are not necessary. The concrete form of interest to us is that which corresponds to the nonzero eigenvalues and then takes the form

$$\widetilde{\varphi}_{m,k}^{j,s} = \begin{pmatrix} 0 \\ \mathrm{e}^{\lambda_m^{(k)} t} P_m^{j,s}(\theta) \end{pmatrix} \tag{2.1}$$

with

$$P_m^{j,s}(\theta) = \sum_{r=1}^{s} D_{m,j}^r \sum_{i=0}^{s-r} \beta_r^{i,s} \frac{\theta^i}{i!}. \tag{2.2}$$

The constant vectors $D_{m,j}^i$ form a basis in $\mathbb{C}^n$ designed from the Jordan chains of the matrices A_{-1}. These vectors and the constants $\beta_q^{i,s}$ in the polynomials $P(\theta)$ do not depend on k.

This gives that, in particular,

$$\inf\{\|\widetilde{\varphi}_{m,k}^{j,s}\|, k \in \mathbb{Z}\} = \rho > 0, \qquad \sup\{\|\widetilde{\varphi}_{m,k}^{j,s}\|, k \in \mathbb{Z}\} = R < \infty. \tag{2.3}$$

The corresponding biorthogonal basis to $\{\widetilde{\varphi}\}$ will be denoted by $\{\widetilde{\psi}\}$.

PROPOSITION 2.1. *For any m, k the vectors of the biorthogonal basis $\{\widetilde{\psi}\}$ form the following Jordan chain with respect to the adjoint operator $\widetilde{\mathcal{A}}^*$:*

$$(\widetilde{\mathcal{A}}^* - \bar{\lambda}_m^{(k)} I)\widetilde{\psi}_{m,k}^{j,p_{m,j}} = 0, \qquad (\widetilde{\mathcal{A}}^* - \bar{\lambda}_m^{(k)} I)\widetilde{\psi}_{m,k}^{j,s} = \widetilde{\psi}_{m,k}^{j,s+1}, \qquad s = 0, \dots, p_{m,j} - 1,$$

where $\bar{\lambda}$ is the complex conjugate of λ.

Proof. To prove the statement we need to observe that

$$\left\langle (\widetilde{\mathcal{A}} - \lambda_m^{(k)} I)\widetilde{\varphi}, \widetilde{\psi}_{m,k}^{j,p_{m,j}} \right\rangle = 0 \quad \forall \widetilde{\varphi} \in \{\widetilde{\varphi}\}.$$

Hence $\widetilde{\psi}_{m,k}^{j,p_{m,j}} \in \mathcal{D}(\widetilde{\mathcal{A}}^*)$ and $(\widetilde{\mathcal{A}}^* - \bar{\lambda}_m^{(k)} I)\widetilde{\psi}_{m,k}^{j,p_{m,j}} = 0$. Next, for $s = 0, \dots, p_{m,j} - 1$ we have

$$\left\langle (\widetilde{\mathcal{A}} - \widetilde{\lambda}_m^{(k)} I)\widetilde{\varphi}, \widetilde{\psi}_{m,k}^{j,s} \right\rangle = 0 \quad \forall \widetilde{\varphi} \in \{\widetilde{\varphi}\} \backslash \{\widetilde{\varphi}_{m,k}^{j,s+1}\}$$

and

$$\left\langle (\widetilde{\mathcal{A}} - \lambda_m^{(k)} I)\widetilde{\varphi}_{m,k}^{j,s+1}, \widetilde{\psi}_{m,k}^{j,s} \right\rangle = 1.$$

This means that $\widetilde{\psi}_{m,k}^{j,s} \in \mathcal{D}(\widetilde{\mathcal{A}}^*)$ and $(\widetilde{\mathcal{A}}^* - \bar{\lambda}_m^{(k)} I)\widetilde{\psi}_{m,k}^{j,s} = \widetilde{\psi}_{m,k}^{j,s+1}$. □

Let us give the concrete form of elements $\widetilde{\psi}_{m,k}^{j,s}$ corresponding to the nonzero eigenvalues.

PROPOSITION 2.2. *Let $\{C_{m,j}^1, \dots, C_{m,j}^{p_{m,j}}\}$ be the jth Jordan chain of A_{-1}^* corresponding to $\bar{\mu}_m$:*

$$A_{-1}^* C_{m,j}^1 = \bar{\mu}_m C_{m,j}^1, \quad (A_{-1}^* - \bar{\mu}_m I) C_{m,j}^s = C_{m,j}^{s-1}, \qquad s = 2, \dots, p_{m,j}.$$

Then the vectors $\widetilde{\psi}_{m,k}^{j,p_{m,j}-r}$ are of the form

$$\widetilde{\psi}_{m,k}^{j,p_{m,j}-i} = \begin{pmatrix} \frac{1}{\bar{\lambda}_m^{(k)}} \sum_{s=0}^{i} q_{m,j}^{i,1+s} C_{m,j}^{1+s} \\ \mathrm{e}^{-\bar{\lambda}_m^{(k)}\theta} \sum_{s=0}^{i} \widetilde{q}_{m,j}^{i,1+s}(\theta) C_{m,j}^{1+s} \end{pmatrix}, \tag{2.4}$$

where $i = 0, \dots, p_{m,j} - 1$ and the coefficients q and $\widetilde{q}(\theta)$ do not depend on k.

In particular, the eigenvectors are given by

$$\widetilde{\psi}_{m,k}^{j,p_{m,j}} = \beta_{m,k}^j \begin{pmatrix} \frac{1}{\bar{\lambda}_m^{(k)}} C_{m,j}^1 \\ \mathrm{e}^{\bar{\lambda}_m^{(k)}\theta} C_{m,j}^1 \end{pmatrix}, \qquad \beta_{m,k}^j \in \mathbb{C}. \tag{2.5}$$

Proof. The proof may be obtained by a simple calculation. □

Let us now recall [10] that the space M_2 possesses a Riesz basis of $\mathcal{A}$-invariant finite-dimensional subspaces $\{V\} = \{V_m^{(k)},\ |k| > N,\ m = 1, \dots, \ell\} \cup \{\widehat{V}_N\}$, where

$$V_m^{(k)} = P_m^{(k)} M_2, \qquad P_m^{(k)} = \frac{1}{2\pi \mathrm{i}} \int_{L_m^{(k)}} R(\lambda, \mathcal{A}) \mathrm{d}\lambda,$$

where $L_m^{(k)}$ are circles of fixed radius $r < r_0 = \frac{1}{3} \min\{|\lambda_m^k - \lambda_i^j|,\ (m,k) \neq (i,j)\}$, centered at λ_m^k, and $\widehat{V}_N$ is the subspace spanned on all the generalized eigenvectors of $\mathcal{A}$ whose eigenvalues are located outside the circles $L_m^{(k)}$, $|k| > N$, $m = 1, \dots, \ell$, with $\dim \widehat{V}_N = 2(N+1)n$. Let us remark that this Riesz basis property is valid for all sufficiently large $N \geq N_0$. Moreover, it is shown in [10] that

$$\sum_{k \in \mathbb{Z}} \sum_{m=1}^{\ell} \left\| P_m^{(k)} - \widetilde{P}_m^{(k)} \right\|^2 < \infty, \tag{2.6}$$

where

$$\widetilde{P}_m^{(k)} = \frac{1}{2\pi \mathrm{i}} \int_{L_m^{(k)}} R(\lambda, \widetilde{\mathcal{A}}) \mathrm{d}\lambda, \qquad \mathrm{Im} \widetilde{P}_m^{(k)} = \mathrm{Ker}(\widetilde{\mathcal{A}} - \lambda_m^{(k)} I)^n.$$

So, in this sense, the basis $\{V\}$ is asymptotically quadratic close to the spectral basis of $\widetilde{\mathcal{A}}$. Consider now the biorthogonal to $\{V\}$ basis of subspaces

$$\{W\} = \{W_m^{(k)},\ |k| > N,\ m = 1, \dots, \ell\} \cup \{\widehat{W}_N\},$$

i.e., the basis that may be defined by

$$W_m^{(k)} = \left(\sum_{\substack{(i,j)\neq(m,k)\\ |i|>N, j=1,\dots,\ell}} V_j^{(i)} + \widehat{V}_N \right)^{\perp}, \quad m = 1, \dots, \ell,\ |k| > N,$$

and

$$\widehat{W}_N = \left(\sum_{\substack{|i|>N\\ j=1,\dots,\ell}} V_j^{(i)} \right)^{\perp}.$$

One can easily check that the basis $\{W\}$ consists of $\mathcal{A}^*$-invariant subspaces and, besides,

$$W_m^{(k)} = P_m^{(k)*} M_2, \qquad P_m^{(k)*} = \frac{1}{2\pi \mathrm{i}} \int_{\bar{L}_m^{(k)}} R(\lambda, \mathcal{A}^*) \mathrm{d}\lambda, \qquad |k| > N,\ m = 1, \dots, \ell,$$

where $\bar{L}_m^{(k)}$ are the complex conjugate circles to $L_m^{(k)}$.

The finite-dimensional operator $\mathcal{A}^*_{\big|\widehat{W}_N}$ has the spectrum $\sigma(\mathcal{A}^*_{\big|\widehat{W}_N})$ which is the complex conjugate of the spectrum $\sigma(\mathcal{A}_{\big|\widehat{V}_N})$. Let us consider this relation in more detail. Let $\widehat{\lambda}_m,\ m = 1, \dots, \ell_N$, be the eigenvalues of $\mathcal{A}_{\big|\widehat{V}_N}$, and $\widehat{\nu}_m$ the number of Jordan blocks corresponding to $\widehat{\lambda}_m \in \sigma(\mathcal{A}_{\big|\widehat{V}_N})$ with the dimension of blocks $\widehat{p}_{m,j},\ j = 1, \dots, \widehat{\nu}_m$. Let $\widehat{\varphi}_m^{j,s},\ j = 1, \dots, \widehat{\nu}_m;\ s = 1, \dots, \widehat{p}_{m,j}$, be a Jordan basis of generalized vectors of $\mathcal{A}_{\big|\widehat{V}_N}$, i.e.,

$$\mathcal{A}\widehat{\varphi}_m^{j,1} = \widehat{\lambda}_m \widehat{\varphi}_m^{j,1}, \qquad (\mathcal{A} - \widehat{\lambda}_m I)\widehat{\varphi}_m^{j,s} = \widehat{\varphi}_m^{j,s-1}, \qquad s = 2, \dots, \widehat{p}_{m,j}, \tag{2.7}$$

for the subspace $\widehat{V}_N$.

We can now formulate the following statement.

Proposition 2.3. *The family*

$$\{\widehat{\psi}_m^{j,s},\ m = 1, \dots, \ell_N;\ j = 1, \dots, \widehat{\nu}_m;\ s = 1, \dots, \widehat{p}_{m,j}\} \subset \widehat{W}_N$$

biorthogonal to $\{\widehat{\varphi}_m^{j,s}\}$, *i.e.,* $\langle \widehat{\varphi}_m^{j,s}, \widehat{\psi}_r^{i,k} \rangle = \delta_{\{(m,j,s),(r,i,k)\}}$, *forms a Jordan basis of generalized eigenvectors of* $\mathcal{A}^*_{\big|\widehat{V}_N}$:

$$\mathcal{A}^* \widehat{\psi}_m^{j,\widehat{p}_{m,j}} = \bar{\widehat{\lambda}}_m \widehat{\psi}_m^{j,\widehat{p}_{m,j}}, \qquad (\mathcal{A}^* - \bar{\widehat{\lambda}}_m I)\widehat{\psi}_m^{j,s} = \widehat{\psi}_m^{j,s+1}, \qquad s = 0, \dots, \widehat{p}_{m,j} - 1,$$

for the subspace $\widehat{W}_N$.

Proof. The proof is analogous to the proof of Proposition 2.1. □

Now we have all the elements to define the Riesz basis that we will use for the analysis of the steering condition (1.4).

We consider the spectral, with respect to the operator $\widetilde{\mathcal{A}}$, basis $\{\widetilde{\varphi}\}$ described above and the corresponding biorthogonal basis $\{\widetilde{\psi}\}$. For a given $N > N_0$ we put

$$\psi^{j,s}_{m,k} = P^{(k)*}_m \widetilde{\psi}^{j,s}_{m,k},\ |k| > N,\ m = 1.\ldots,\ell;\quad j = 1,\ldots,\nu_m;\ s = 1,\ldots,p_{m,j}. \tag{2.8}$$

Then we complete the collection (2.8) by the set

$$\{\widehat{\psi}^{j,s}_m;\ m = 1,\ldots,\ell_N;\ j = 1,\ldots,\widehat{\nu}_m; s = 1,\ldots,\widehat{p}_{m,j}\},$$

which contains $2(N+1)n$ vectors forming a Jordan basis in $\widehat{W}_N$ (Proposition 2.3).

THEOREM 2.4. *Let the condition* (2.3) *be satisfied and let* N *be sufficiently large. Then the collection*

$$\{\psi\} = \{\psi^{j,s}_{m,k}\} \cup \{\widehat{\psi}^{j,s}_m\},$$

where $\psi^{j,s}_{m,k}$ *are given by* (2.8) *and* $\psi^{j,s}_m$ *are defined in Proposition* 2.3*, constitutes a Riesz basis of* M_2*.*

Proof. We start with the fact that under condition (2.3) the collection $\{\widetilde{\psi}\}$ forms a Riesz basis in M_2. In particular, this implies that the collection $\{\widetilde{\psi}^{j,s}_{m,k},\ |k| > N\}$ forms a Riesz basis in the closure of its linear span

$$\operatorname{Cl}\operatorname{Lin}\{\widetilde{\psi}^{j,s}_{m,k},\ |k| > N\} = \operatorname{Cl}\sum_{|k|>N}\sum_{m=1}^{\ell}\operatorname{Ker}\left(\widetilde{\mathcal{A}}^* - \bar{\lambda}^{(k)}_m I\right)^{\max_j p_{m,j}}.$$

On the other hand, from (2.6) and (2.8) we have

$$\sum_{|k|>N}\sum_{m,j,s}\left\|\psi^{j,s}_{m,k} - \widetilde{\psi}^{j,s}_{m,k}\right\|^2 = \sum_{|k|>N}\sum_{m,j,s}\left\|P^{(k)*}_m - \widetilde{P}^{(k)*}_m\right\|^2\left\|\widetilde{\psi}^{j,s}_{m,k}\right\|^2.$$

This implies that for any $\varepsilon > 0$ there exists a large N_1 such that if $N > N_1$, then

$$\sum_{|k|>N}\sum_{m,j,s}\left\|\psi^{j,s}_{m,k} - \widetilde{\psi}^{j,s}_{m,k}\right\|^2 < \varepsilon. \tag{2.9}$$

Hence for this N, the family $\{\psi^{j,s}_{m,k}\}_{|k|>N}$ is quadratically close to $\{\widetilde{\psi}^{j,s}_{m,k}\}_{|k|>N}$ and, therefore, forms a Riesz basis in the closure of its linear span

$$\operatorname{Cl}\operatorname{Lin}\{\psi^{j,s}_{m,k},\ |k| > N\} = \operatorname{Cl}\sum_{|k|>N}\sum_{m=1}^{\ell} W^{(k)}_m.$$

Since, due to Proposition 2.3, the set $\{\widehat{\psi}^{j,s}_m\}$ is a basis in $\widehat{W}_N$ and

$$\operatorname{Cl}\sum_{|k|>N}\sum_{m}^{\ell} W^{(k)}_m + \widehat{W}_N = M_2,$$

the union

$$\{\psi^{j,s}_{m,k},\ |k| > N\} \cup \{\widehat{\psi}^{j,s}_m\}$$

is a Riesz basis in M_2. This ends the proof. □

By $\{\varphi\}$ we denote the Riesz basis biorthogonal to the basis $\{\psi\}$ in Theorem 2.4.

Remark 2.5. The elements of the basis $\{\varphi\}$ which correspond to the part $\{\widehat{\psi}_m^{j,s}\}$ are the generalized eigenvectors $\{\widehat{\varphi}_m^{j,s}\}$ of the operator $\mathcal{A}$ (see (2.7)). So,

$$\{\varphi\} = \{\varphi_{m,k}^{j,s},\ |k| > N\} \cup \{\widehat{\varphi}_m^{j,s}\}.$$

Moreover, it is easy to show that there exists N_1 such that for any given $N > N_1$ and $m = 1, \dots, \ell$ *the collection* $\{\varphi_{m,k}^{j,s},\ j = 1, \dots, \nu_m; s = 1, \dots, p_{m,j}\}$ *is a basis of* $V_m^{(k)}$. The chosen basis $\{\varphi\}$ will be used in our further analysis of the steering conditions by the moment problem method. In this context, we notice that the construction of a proper basis becomes rather complicated only in the case when the spectrum of the matrix A_{-1} is not simple and, as a consequence, the operator $\mathcal{A}$ may not possess a spectral Riesz basis. If all eigenvalues of A_{-1} are simple, the basis $\{\varphi\}$ constructed in this section coincides with a spectral basis of $\mathcal{A}$.

3. Expansion of the steering condition in the Riesz basis. In order to use the results of section 2, we assume that the matrix A_{-1} is not singular.

Let us expand the steering condition (1.4) with respect to the basis $\{\varphi\}$ and to the biorthogonal basis $\{\psi\}$. Consider a state $x = \binom{y}{z(\cdot)} \in M_2$; this state is reachable at time T if and only if

$$\sum_{\varphi\in\{\varphi\}} \langle x, \psi\rangle\, \varphi = \sum_{\varphi\in\{\varphi\}} \int_0^T \left\langle \mathrm{e}^{\mathcal{A}t}\mathcal{B}u(t), \psi\right\rangle \mathrm{d}t\varphi, \quad u(\cdot) \in L_2(-1, 0; \mathbb{C}^r).$$

Then the steering condition (1.4) can be substituted by the following system of equalities:

$$\langle x, \psi\rangle = \int_0^T \left\langle \mathrm{e}^{\mathcal{A}t}\mathcal{B}u(t), \psi\right\rangle \mathrm{d}t, \quad \psi \in \{\psi\},\ u(\cdot) \in L_2(-1, 0; \mathbb{C}^r). \tag{3.1}$$

Let $\{b_1, \dots, b_r\}$ be an arbitrary basis in $\mathrm{Im}B$, the image of the matrix B and $\mathbf{b}_i = \binom{b_i}{0} \in M_2,\ i = 1, \dots, r$ (more precision on the choice of this basis will be given in section 6). Then the right-hand side of (3.1) takes the form

$$\int_0^T \left\langle \mathrm{e}^{\mathcal{A}t}\mathcal{B}u(t), \psi\right\rangle \mathrm{d}t = \sum_{i=1}^r \int_0^T \left\langle \mathrm{e}^{\mathcal{A}t}\mathbf{b}_i, \psi\right\rangle u_i(t)\mathrm{d}t. \tag{3.2}$$

Let us omit the index i for $\mathbf{b}_i$ and for any $\mathbf{b} \in \{\mathbf{b}_1, \dots, \mathbf{b}_r\}$ transform the term $\left\langle \mathrm{e}^{\mathcal{A}t}\mathbf{b}u(t), \psi\right\rangle,\ \psi \in \{\psi_{m,k}^{j,s},\ |k| > N\}$ as follows:

$$\begin{aligned} \left\langle \mathrm{e}^{\mathcal{A}t}\mathbf{b}, \psi_{m,k}^{j,s}\right\rangle &= \left\langle \mathrm{e}^{\mathcal{A}t}\mathbf{b}, P_m^{(k)*}\widetilde{\psi}_{m,k}^{j,s}\right\rangle \\ &= \left\langle P_m^{(k)}\mathrm{e}^{\mathcal{A}t}\mathbf{b}, \widetilde{\psi}_{m,k}^{j,s}\right\rangle \\ &= \left\langle \widetilde{P}_m^{(k)}\mathrm{e}^{\widetilde{\mathcal{A}}t}\mathbf{b}, \widetilde{\psi}_{m,k}^{j,s}\right\rangle + \left\langle (P_m^{(k)}\mathrm{e}^{\mathcal{A}t} - \widetilde{P}_m^{(k)}\mathrm{e}^{\widetilde{\mathcal{A}}t})\mathbf{b}, \widetilde{\psi}_{m,k}^{j,s}\right\rangle. \end{aligned} \tag{3.3}$$

LEMMA 3.1. *There exists a sequence* $\{\alpha_k\}, \sum_{|k|>N}\alpha_k^2 < \infty$, *such that for all* $m = 1, \dots, \ell;\ j = 1, \dots, \nu_m$, *and* $s = 1, \dots, p_{m,j}$ *the following estimates hold:*

$$\left|\left\langle (P_m^{(k)}\mathrm{e}^{\mathcal{A}t} - \widetilde{P}_m^{(k)}\mathrm{e}^{\widetilde{\mathcal{A}}t})\mathbf{b}, \widetilde{\psi}_{m,k}^{j,s}\right\rangle\right| \le \frac{\alpha_k}{|k|}, \quad |k| > N, \quad t \in [0, T]. \tag{3.4}$$

Proof. Let us denote by $R(\lambda,\mathcal{A})$ and $R(\lambda,\widetilde{\mathcal{A}})$ the resolvents of the operators $\mathcal{A}$ and $\widetilde{\mathcal{A}}$. Taking into account the relation (2.3) we can write

$$\begin{aligned}
&\left|\left\langle (P_m^{(k)}\mathrm{e}^{\mathcal{A}t}-\widetilde{P}_m^{(k)}\mathrm{e}^{\widetilde{\mathcal{A}}t})\mathbf{b},\widetilde{\psi}_{m,k}^{j,s}\right\rangle\right| \\
&\quad=\left|\left\langle \tfrac{1}{2\pi\mathrm{i}}\int_{L_m^k}\mathrm{e}^{\lambda t}\left(R(\lambda,\mathcal{A})-R(\lambda,\widetilde{\mathcal{A}})\right)\mathrm{d}\lambda\cdot\mathbf{b},\widetilde{\psi}_{m,k}^{j,s}\right\rangle\right| \\
&\quad=\tfrac{1}{2\pi}\left|\int_{L_m^k}\mathrm{e}^{\lambda t}\left\langle\left(R(\lambda,\mathcal{A})-R(\lambda,\widetilde{\mathcal{A}})\right)\mathbf{b},\widetilde{\psi}_{m,k}^{j,s}\right\rangle\mathrm{d}\lambda\right| \\
&\quad\le\tfrac{1}{2\pi}\int_{L_m^k}\left|\mathrm{e}^{\lambda t}\right|\left\|\left(R(\lambda,\mathcal{A})-R(\lambda,\widetilde{\mathcal{A}})\right)\mathbf{b}\right\|\left\|\widetilde{\psi}_{m,k}^{j,s}\right\||\mathrm{d}\lambda| \\
&\quad\le C\int_{L_m^k}\left|\mathrm{e}^{\lambda t}\right|\left\|\left(R(\lambda,\mathcal{A})-R(\lambda,\widetilde{\mathcal{A}})\right)\mathbf{b}\right\||\mathrm{d}\lambda|
\end{aligned}\tag{3.5}$$

with $C>0$. Now we need to estimate $\|(R(\lambda,\mathcal{A})-R(\lambda,\widetilde{\mathcal{A}}))\mathbf{b}\|$. In order to do that, we need to use an explicit expression for the resolvents of the operators $\mathcal{A}$ and $\widetilde{\mathcal{A}}$ given in [10, Proposition 2.2] (for the proof see [11] and also [5]). We obtain

$$\left(R(\lambda,\mathcal{A})-R(\lambda,\widetilde{\mathcal{A}})\right)\mathbf{b}=\begin{pmatrix}(I-A_{-1}\mathrm{e}^{-\lambda})\left(\Delta_{\mathcal{A}}^{-1}(\lambda)-\Delta_{\widetilde{\mathcal{A}}}^{-1}(\lambda)\right)b\\ \mathrm{e}^{\lambda\theta}\left(\Delta_{\mathcal{A}}^{-1}(\lambda)-\Delta_{\widetilde{\mathcal{A}}}^{-1}(\lambda)\right)b\end{pmatrix},$$

where

$$\Delta_{\mathcal{A}}(\lambda)=\lambda I-\lambda\mathrm{e}^{-\lambda}A_{-1}+\lambda\int_{-1}^{0}\mathrm{e}^{\lambda s}A_2(s)\mathrm{d}s-\int_{-1}^{0}\mathrm{e}^{\lambda s}A_3(s)\mathrm{d}s,$$

and, from that, $\Delta_{\widetilde{\mathcal{A}}}(\lambda)=\lambda I-\lambda\mathrm{e}^{-\lambda}A_{-1}$. From the relation

$$\Delta_{\widetilde{\mathcal{A}}}^{-1}(\lambda)-\Delta_{\mathcal{A}}^{-1}(\lambda)=\Delta_{\mathcal{A}}^{-1}(\lambda)\left(\lambda\int_{-1}^{0}\mathrm{e}^{\lambda s}A_2(s)\mathrm{d}s+\int_{-1}^{0}\mathrm{e}^{\lambda s}A_3(s)\mathrm{d}s\right)\Delta_{\widetilde{\mathcal{A}}}^{-1}(\lambda),$$

and using the estimates [11, formulas (25), (26)],

$$\left\|\Delta^{-1}_{\mathcal{A}\text{ or }\widetilde{\mathcal{A}}}(\lambda)\right\|\le C_1|\lambda|^{-1},\quad C_1>0,\quad \lambda\in L_m^{(k)},\quad |k|>N,\ m=1,\dots,\ell,$$

we get the inequality

$$\left\|\left(R(\lambda,\mathcal{A})-R(\widetilde{\lambda,\mathcal{A}})\right)\mathbf{b}\right\|\le\frac{C_2}{|\lambda|^2}\left(|\lambda|\left\|\int_{-1}^{0}\mathrm{e}^{\lambda s}A_2(s)\mathrm{d}s\right\|+\left\|\int_{-1}^{0}\mathrm{e}^{\lambda s}A_3(s)\mathrm{d}s\right\|\right).\tag{3.6}$$

For $\lambda\in L_m^{(k)}$ we can put $\lambda=\widetilde{\lambda}+2k\pi\mathrm{i}$, where $\widetilde{\lambda}\in L_m^{(0)}=\{\xi:|\xi-\ln|\mu_m|+\mathrm{i}\arg\mu_m|=r\}$. This yields

$$\begin{aligned}
\left\|\int_{-1}^{0}\mathrm{e}^{\lambda s}A_2(s)\mathrm{d}s\right\| &= \left\|\int_{-1}^{0}\mathrm{e}^{\widetilde{\lambda}s}A_2(s)\mathrm{e}^{2\pi\mathrm{i}ks}\mathrm{d}s\right\| = C_2^{(k)}(\widetilde{\lambda}),\\
\left\|\int_{-1}^{0}\mathrm{e}^{\lambda s}A_3(s)\mathrm{d}s\right\| &= \left\|\int_{-1}^{0}\mathrm{e}^{\widetilde{\lambda}s}A_3(s)\mathrm{e}^{2\pi\mathrm{i}ks}\mathrm{d}s\right\| = C_3^{(k)}(\widetilde{\lambda}),
\end{aligned}\tag{3.7}$$

where

$$\sum_{|k|>N}\left[C_j^{(k)}(\widetilde{\lambda})\right]^2 \le \int_{-1}^{0}|\mathrm{e}^{\widetilde{\lambda}s}|^2\|A_j(s)\|^2\mathrm{d}s,\quad j=2,3.$$

Since for all values of the parameter

$$\widetilde{\lambda}\in\bigcup_{m=1}^{\ell}L_m^{(0)}$$

the L_2-norm of the matrix functions $\mathrm{e}^{\widetilde{\lambda}s}A_j(s)$, $j=2,3$, on the interval $[-1,0]$ are uniformly bounded, then there exists $\delta>0$ such that

$$\sum_{|k|>N}\left(C_j^{(k)}(\widetilde{\lambda})\right)^2\le\delta<\infty,\quad j=2,3,\quad \widetilde{\lambda}\in\bigcup_{m=1}^{\ell}L_m^{(0)}. \tag{3.8}$$

Finally, from (3.6) and (3.7), for $\widetilde{\lambda}\in L_m^{(k)}$ we obtain

$$\left\|\left(R(\lambda,\mathcal{A})-R(\lambda,\widetilde{\mathcal{A}})\right)\mathbf{b}\right\|\le\frac{C_2}{|\widetilde{\lambda}+2k\pi\mathrm{i}|}\left(|\widetilde{\lambda}+2k\pi\mathrm{i}|C_2^{(k)}(\widetilde{\lambda})+C_3^{(k)}(\widetilde{\lambda})\right),\quad \widetilde{\lambda}\in L_m^{(0)}. \tag{3.9}$$

Then the inequalities (3.5), (3.6), and (3.9) give the validity of (3.4). The proof is complete. □

Let us consider the first term on the right-hand side of (3.3). Since $\widetilde{\psi}_{m,k}^{j,s}\in W_m^{(k)}$ and due to Proposition 2.1, we have

$$\begin{aligned}\left\langle\widetilde{P}_m^{(k)}\mathrm{e}^{\widetilde{\mathcal{A}}t}\mathbf{b},\widetilde{\psi}_{m,k}^{j,s}\right\rangle &= \left\langle\mathbf{b},\mathrm{e}^{\widetilde{\mathcal{A}}^*t}\widetilde{P}_m^{(k)*}\widetilde{\psi}_{m,k}^{j,s}\right\rangle\\ &= \left\langle\mathbf{b},\mathrm{e}^{\widetilde{\mathcal{A}}^*t}\widetilde{\psi}_{m,k}^{j,s}\right\rangle\\ &= \left(\langle\mathbf{b},\widetilde{\psi}_{m,k}^{j,p_{m,j}}\rangle\tfrac{t^{p_{m,j}-s}}{(p_{m,j}-s)!}+\cdots+\langle\mathbf{b},\widetilde{\psi}_{m,k}^{j,s}\rangle\right)\mathrm{e}^{\lambda_m^{(k)}t}.\end{aligned} \tag{3.10}$$

Lemma 3.2. *There exists a constant δ_1 such that*

$$\left|\left\langle\mathbf{b},\widetilde{\psi}_{m,k}^{j,s}\right\rangle\right|\le\frac{\delta_1}{|k|} \tag{3.11}$$

for all $|k|>N$, $m=1,\dots,\ell, j=1,\dots,\nu_m, s=1,\dots,p_{m.j}$. Moreover, we have

$$\left\langle\mathbf{b},\widetilde{\psi}_{m,k}^{j,p_{m.j}}\right\rangle=\frac{1}{\bar{\lambda}_m^{(k)}}\left\langle b,C_{m,j}^1\right\rangle,$$

where $C_{m,j}^1$ are the eigenvectors of A_{-1} corresponding to $\bar{\mu}_m$ and $\widetilde{\psi}_{m,k}^{j,p_{m.j}}$ are as in the formula (2.5) *in Proposition* 2.2 *with $\beta_{m,k}^j=1$.*

Proof. Taking into account (2.3), we get

$$\begin{aligned}\left|\langle\mathbf{b},\widetilde{\psi}_{m,k}^{j,s}\rangle\right| &= \left|\langle\mathbf{b},\widetilde{P}_m^{(k)*}\widetilde{\psi}_{m,k}^{j,s}\rangle\right|\\ &= \left|\langle\widetilde{P}_m^{(k)}\mathbf{b},\widetilde{\psi}_{m,k}^{j,s}\rangle\right|\\ &= \tfrac{1}{2\pi}\left|\int_{L_m^{(k)}}\langle R(\lambda,\widetilde{\mathcal{A}})\mathbf{b},\widetilde{\psi}_{m,k}^{j,s}\rangle\mathrm{d}\lambda\right|\\ &\le C\int_{L_m^{(k)}}\left\|R(\lambda,\widetilde{\mathcal{A}})\mathbf{b}\right\|\mathrm{d}\lambda,\qquad C>0,\end{aligned} \tag{3.12}$$

where C is a constant. The explicit expression of the resolvent $R(\lambda, \widetilde{\mathcal{A}})$ is given by (see [10, 11])

$$R(\lambda, \widetilde{\mathcal{A}})\mathbf{b} = \begin{pmatrix} (I - A_{-1}\mathrm{e}^{-\lambda})\Delta_{\widetilde{\mathcal{A}}}^{-1}(\lambda)b \\ \mathrm{e}^{\lambda\theta}\Delta_{\widetilde{\mathcal{A}}}^{-1}(\lambda)b \end{pmatrix}.$$

Since $\|\Delta_{\widetilde{\mathcal{A}}}^{-1}\| \leq C_1|\lambda|^{-1}$ with $C_1 > 0$ and for $\lambda \in L_m^{(k)},\ |k| > N,\ m = 1, \ldots, \ell$, we obtain the estimate

$$\left\| R(\lambda, \widetilde{\mathcal{A}})\mathbf{b} \right\| \leq \frac{C_2}{|k|}, \qquad C_2 > 0, \quad \lambda \in L_m^{(k)}.$$

This, together with (3.12), leads to (3.11).

The second statement follows directly:

$$\left\langle \mathbf{b}, \widetilde{\psi}_{m,k}^{j,p_{m.j}} \right\rangle = \left\langle \begin{pmatrix} b \\ 0 \end{pmatrix}, \widetilde{\psi}_{m,k}^{j,p_{m.j}} \right\rangle = \frac{1}{\bar{\lambda}_m^{(k)}} \left\langle b, C_{m,j}^1 \right\rangle.$$

This completes the proof. □

LEMMA 3.3. *Assume that* $\left\langle b, C_{m,j}^s \right\rangle = 0,\ s = 1, \ldots, r,\ r < p_{m,j}$; *then*

$$\left\langle \mathbf{b}, \widetilde{\psi}_{m,k}^{j,p_{m,j}-r} \right\rangle = \frac{q_{m,j}^r}{\bar{\lambda}_m^{(k)}} \left\langle b, C_{m,j}^{r+1} \right\rangle, \tag{3.13}$$

where the coefficients $q_{m,j}^r$ *do not depend on* k.

Proof. This is a direct consequence of the relation (2.4). □

Let us denote by $q_{m,k}^{j,s,d}(t)$ the polynomials

$$q_{m,k}^{j,s,d}(t) = k \left(\frac{\left\langle \mathbf{b}_d, \widetilde{\psi}_{m,k}^{j,p_{m,j}} \right\rangle}{(p_{m,j} - s)!} t^{p_{m,j}-s} + \frac{\left\langle \mathbf{b}_d, \widetilde{\psi}_{m,k}^{j,p_{m,j}-1} \right\rangle}{(p_{m,j} - s - 1)!} t^{p_{m,j}-s-1} + \cdots + \left\langle \mathbf{b}_d, \widetilde{\psi}_{m,k}^{j,s} \right\rangle \right) \tag{3.14}$$

and by $f_{m,k}^{j,s,d}(t)$ the functions

$$f_{m,k}^{j,s,d}(t) = k \left\langle \left(P_m^{(k)} \mathrm{e}^{\mathcal{A}t} - \widetilde{P}_m^{(k)} \mathrm{e}^{\widetilde{\mathcal{A}}t} \right) \mathbf{b}_d, \widetilde{\psi}_{m,k}^{j,s} \right\rangle. \tag{3.15}$$

With these notations (and also due to (3.2), (3.3), and (3.10)), the infinite part of the system (3.1) corresponding to $\psi \in \{\psi_{m,k}^{j,s}\},\ |k| > N$, reads as

$$k \left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \psi_{m,k}^{j,s} \right\rangle = \sum_{d=1}^{r} \int_0^T \left(\mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j,s,d}(t) + f_{m,k}^{j,s,d}(t) \right) u_d(t) \mathrm{d}t. \tag{3.16}$$

Moreover, due to Lemmas 3.1, 3.2, and 3.3, the functions (3.14) and (3.15) verify the following properties:

(P1) the coefficients of the polynomials $\{q(t)\}$ are uniformly bounded as $|k| > N$;

(P2) the set of leading coefficients of the nontrivial polynomials $\{q(t)\}$ does not have a limit point at 0;

(P3) $\sum_{|k|>N} |f_{m,k}^{j,s,d}(t)|^2 < \alpha < \infty,\ t \in [0, T],\ \alpha > 0.$

Next we observe that if $\psi = \widehat{\psi}_m^{j,s},\ m = 1, \dots, \ell_N,\ j = 1, \dots, \widehat{\mu}_m,\ s = 1, \dots, \widehat{p}_{m,j}$, then

$$\begin{aligned}\left\langle \mathrm{e}^{\mathcal{A}t}\mathbf{b}_d, \psi \right\rangle, &= \left\langle \mathbf{b}_d, \mathrm{e}^{\mathcal{A}^* t}\psi \right\rangle \\ &= \widehat{q}_m^{j,s,d}(t)\mathrm{e}^{\widehat{\lambda}_m t},\end{aligned}$$

where

$$\widehat{q}_m^{j,s,d}(t) = \left(\frac{\left\langle \mathbf{b}_d, \widehat{\psi}_{m,k}^{j,p_{m,j}} \right\rangle}{(\widehat{p}_{m,j} - s)!} t^{\widehat{p}_{m,j}-s} + \frac{\left\langle \mathbf{b}_d, \widehat{\psi}_{m,k}^{j,\widehat{p}_{m,j}-1} \right\rangle}{(\widehat{p}_{m,j} - s - 1)!} t^{\widehat{p}_{m,j}-s-1} + \dots + \left\langle \mathbf{b}_d, \widehat{\psi}_{m,k}^{j,s} \right\rangle \right).$$

Therefore, the finite part of the system (3.1) corresponding to $\psi \in \{\widehat{\psi}_m^{j,s}\}$ reads as

$$\left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \widehat{\psi}_m^{j,s} \right\rangle = \sum_{d=1}^{r} \int_0^T \mathrm{e}^{\widehat{\lambda}_m t} \widehat{q}_{m,k}^{j,s,d}(t) u_d(t) \mathrm{d}t. \tag{3.17}$$

Thus, we observe that the state $\binom{y_T}{z_T(\cdot)} \in M_2$ is reachable from 0 at the time $T > 0$ if and only if the equalities (3.16) and (3.17) hold for some controls $u_d(\cdot) \in L_2(0, T),\ d = 1, \dots, r$. These equalities pose a kind of moment problem, which is the main object of our further analysis.

4. The problem of moments and the Riesz basis property. In this section, we recall the general properties of the problem of moments that will be applied to the analysis of the problem (3.16)–(3.17) given in section 3.

Consider a collection of functions $\{g_k(t),\ t \in [0, \infty[\}_{k\in\mathbb{N}}$ assuming that for any $k \in \mathbb{N},\ T > 0,\ g_k(\cdot) \in L_2(0, T)$, and consider the following problem of moments:

$$s_k = \int_0^T g_k(t) u(t) \mathrm{d}t, \qquad k \in \mathbb{N}. \tag{4.1}$$

We start with the following well-known fact, which is a consequence of Bari theorem [4, Chapter 6] and [17, Chapter 4] (see also [14] for a direct proof and the references therein).

PROPOSITION 4.1. *The following statements are equivalent:*

(i) *For the scalars $s_k,\ k \in \mathbb{N}$, the problem (4.1) has a solution $u(\cdot) \in L_2(0, T)$ if and only if $\{s_k\} \in \ell_2$, i.e., $\sum_{k\in\mathbb{N}} s_k^2 < \infty$;*

(ii) *the family $\{g_k(t)\}_{k\in\mathbb{N}},\ t \in [0, T]$, forms a Riesz basis in the closure of its linear span*

$$\mathrm{Cl}\,\mathrm{Lin}\{g_k(t),\ k \in \mathbb{N}\}.$$

Using this proposition, we can prove the following result.

PROPOSITION 4.2. *Let us denote by $\mathcal{L}(0, T)$ the closed subspace*

$$\mathrm{Cl}\,\mathrm{Lin}\{g_k(t),\ k \in \mathbb{N}\} \subset L_2(T_0, T).$$

Let us suppose that for some $T_1 > 0$ the functions $\{g_k(t)\}_{k\in\mathbb{N}},\ t \in [0, T_1]$, form a Riesz basis in $\mathcal{L}(0, T_1) \subset L_2(0, T_1)$ and $\mathrm{codim}\mathcal{L}(0, T_1) < \infty$. Then for any $0 < T < T_1$, there exists an infinite-dimensional subspace $\ell^T \subset \ell_2$ such that the problem of moments (4.1) is unsolvable for $\{s_k\} \in \ell^T$ if $\{s_k\} \neq \{0\}$.

Proof. We introduce for all $T > 0$ the operator $Q_T : L_2(0,T) \to \ell_2$ given by

$$Q_T u(\cdot) = \left\{ \int_0^T g_k(t) u(t) \mathrm{d}t \right\}_{k \in \mathbb{N}}. \tag{4.2}$$

This gives $Q_{T_1}(L_2(0,T_1)) = \ell_2$ by Proposition 4.1. The operator Q_{T_1} is bounded due to the closed graph theorem. It is easy to see that the adjoint operator $Q^*_{T_1}$ acts as

$$Q^*_{T_1}\{s_k\}_{k\in\mathbb{N}} = \sum_{k\in\mathbb{N}} s_k g_k(t) \in \mathcal{L}(0,T_1).$$

Let us denote now by $Q_1 : \mathcal{L}(0,T_1) \to \ell_2$ the one-to-one operator defined as follows:

$$Q_1 u(\cdot) = \{c_k\} \qquad \text{for} \qquad u(\cdot) = \sum_{k\in\mathbb{N}} c_k g_k(\cdot),\ \{c_k\}, \in \ell_2.$$

Now consider the decomposition $L_2(0,T_1) = X_1 \oplus X_2$, where

$$X_1 = \{u(\cdot) : u(t) \equiv 0,\ t \in [0,T]\}, \qquad X_2 = \{u(\cdot) : u(t) \equiv 0,\ t \in [T,T_1]\},$$

and observe that since $\operatorname{codim} \mathcal{L}(0,T_1) < \infty$, then the intersection $X = X_1 \cup \mathcal{L}(0,T_1)$ is an infinite-dimensional subspace in $L_2(0,T_1)$. Finally, let us denote $\ell^T = Q_1(X)$. The above considerations prove that this subspace is infinite dimensional. Taking $u(\cdot) \in X_2$ and $\{s_k\} \in \ell^T$, we obtain

$$\langle Q_{T_1} u(\cdot), \{s_k\}\rangle = \langle u(\cdot), Q^*_T\{s_k\}\rangle = \int_0^T u(t) \sum_{k\in\mathbb{N}} s_k g_k(t) \mathrm{d}t = 0,$$

because $\sum_{k\in\mathbb{N}} s_k g_k(t) \in X_1$. Thus $Q_{T_1}(X_2) \perp \ell^T$ and, therefore, (4.1) is unsolvable for $\{s_k\} \in \ell^T$ if $\{s_k\} \neq \{0\}$. □

PROPOSITION 4.3. *Let us consider the moment problem*

$$s_k = \sum_{d=1}^{r} \int_0^T g^d_k(t) u_d(t) \mathrm{d}t, \qquad k \in \mathbb{N}, \tag{4.3}$$

with the assumption

$$\sum_{k\in\mathbb{N}} \int_0^T |g^d_k(t)|^2 \mathrm{d}t < \infty, \qquad d = 1, \dots, r. \tag{4.4}$$

Then the set $S_{0,T}$ of sequences $\{s_k\}$ for which problem (4.3) *is solvable is a nontrivial submanifold of ℓ_2, i.e., $S_{0,T} \neq \ell_2$.*

Proof. Let us introduce the operator $Q^r_T : L^r_2(0,T) \to \ell^2$ defined by

$$Q^r_T u(\cdot) = \{s_k\}_{k\in\mathbb{N}} = \left\{ \sum_{d=1}^{r} \int_0^T g^d_k(t) u_d(t) \mathrm{d}t \right\}_{k\in\mathbb{N}}, \qquad u(\cdot) = (u_1(\cdot), \dots, u_r(\cdot)).$$

Then, if $\|u(\cdot)\| \le 1$, we obtain

$$\sum_{k=N}^{\infty} |s_k|^2 = \sum_{d=1}^{r} \sum_{k=N}^{\infty} \left| \int_0^T g_k^d(t) u_d(t) \right|^2 \mathrm{d}t \le \sum_{d=1}^{r} \sum_{k=N}^{\infty} \int_0^T |g_k^d(t)|^2 \mathrm{d}t,$$

and then $\sum_{k=N}^{\infty} |s_k|^2 \to 0$ as $N \to \infty$. This means that the set $\{Q_T^r u(\cdot),\ \|u(\cdot)| \le 1\}$ satisfies the criterion of compactness in ℓ^2 (see, for example, [7, Chapter 5]). Hence Q_T^r is a compact operator and therefore $\mathrm{Im} Q_T^r \ne \ell_2$. □

In the following, our analysis will be based on the theory of families of exponential developed by Avdonin and Ivanov in [1]. We are particularly interested in the basis properties of such families.

Let $\delta_1, \dots, \delta_\ell$ be different, modulus $2\pi\mathrm{i}$, complex numbers, and let $m_1, \dots, m_\ell$ and N be natural integers. Let us denote by $\widetilde{\mathcal{E}}_N$ the family

$$\widetilde{\mathcal{E}}_N = \left\{ \mathrm{e}^{(\delta_s + 2\pi\mathrm{i}k)t}, t\mathrm{e}^{(\delta_s+2\pi\mathrm{i}k)t}, \dots, t^{m_s-1}\mathrm{e}^{(\delta_s+2\pi\mathrm{i}k)t} \right\}_{\substack{|k|>N \\ s=1,\dots,\ell}}.$$

Next, let $\varepsilon_1, \dots, \varepsilon_r$ be another collection of different complex numbers such that $\varepsilon_j \ne \delta_s + 2\pi\mathrm{i}k,\ j = 1, \dots, r;\ s = 1, \dots, \ell;\ |k| > N$, and let $m_1', \dots, m_r'$ be positive integers. Let us denote by $\mathcal{E}_0$ the collection

$$\mathcal{E}_0 = \left\{ \mathrm{e}^{\varepsilon_j t}, t\mathrm{e}^{\varepsilon_j t}, \dots, t^{m_s'-1}\mathrm{e}^{\varepsilon_j t} \right\}_{j=1,\dots,r}.$$

The following theorem is the main tool of our further analysis.

THEOREM 4.4. (i) *If* $\sum_{j=1}^{r} m_j' = (2N+1)\sum_{s=1}^{\ell} m_s$, *then the family* $\mathcal{E} = \widetilde{\mathcal{E}}_N \cup \mathcal{E}_0$ *constitutes a Riesz basis in* $L_2(0, \sum_{s=1}^{\ell} m_s)$.

(ii) *If* $T > \sum_{s=1}^{\ell} m_s$, *then independently of the number of elements in* $\mathcal{E}_0$, *the family* $\mathcal{E}$ *forms a Riesz basis of the closure of its linear span in the space* $L_2(0,T)$.

Proof. (i) We make use of [1, Theorem II.4.23]. According to this theorem, let us consider the complex function

$$f(z) = \mathrm{e}^{\frac{\mathrm{i}z}{2}\sum_{s=1}^{\ell} m_s} \prod_{s=1}^{\ell} \left(\sin\left(\frac{z}{2} - \frac{\delta_s}{2} \right) \right)^{m_s} R(z),$$

where

$$R(z) = \prod_{j=1}^{r} (z - \varepsilon_j)^{m_j'} \left(\prod_{\substack{s=1,\dots,\ell \\ |k|\le N}} (z - \delta_s - 2\pi\mathrm{i}k)^{m_s} \right)^{-1} \to 1 \quad \text{as} \quad z \to \infty.$$

One can easily verify that $f(z)$ extended to the points $\delta_s + 2\pi\mathrm{i}k,\ s = 1, \dots, \ell,\ |k| \le N$, by continuity, is an entire function of the sine type (see [1, Definition II.1.27] and also [17, section 4.5]). Representing

$$\prod_{s=1}^{\ell} \left(\sin\left(\frac{z}{2} - \frac{\delta_s}{2} \right) \right)^{m_s} = \prod_{s=1}^{\ell} \left(\frac{\mathrm{e}^{\frac{\mathrm{i}z}{2}} \mathrm{e}^{\frac{-\mathrm{i}\delta_s}{2}} - \mathrm{e}^{\frac{-\mathrm{i}z}{2}} \mathrm{e}^{\frac{\mathrm{i}\delta_s}{2}}}{2\mathrm{i}} \right)^{m_s},$$

we get

$$\prod_{s=1}^{\ell}\left(\sin\left(\frac{z}{2}-\frac{\delta_s}{2}\right)\right)^{m_s}=C_0\mathrm{e}^{\frac{\mathrm{i}z}{2}\sum_{s=1}^{\ell}m_s}+C_1\mathrm{e}^{\frac{-\mathrm{i}z}{2}\sum_{s=1}^{\ell}m_s}+\sum_{j=2}^{N_0}C_j\mathrm{e}^{\mathrm{i}zq_j},$$

where C_j are constants and $q_j\in\{-\frac{1}{2}\sum_{s=1}^{\ell}m_s,\frac{1}{2}\sum_{s=1}^{\ell}m_s\}$, $j=2,\dots,N_0$. And then the growth indicator of the function f (see [1, Paragraph II.1.4.2]) is of the form

$$\begin{aligned}h_f(\phi)&=\limsup_{\rho\to\infty}\frac{1}{\rho}\ln|f(\rho\mathrm{e}^{\mathrm{i}\phi})|\\&=\limsup_{\rho\to\infty}\frac{1}{\rho}\ln\left|C_0\mathrm{e}^{\mathrm{i}z\sum_{s=1}^{\ell}m_s}+C_1+\sum_{j=2}^{N_0}C_j\mathrm{e}^{\mathrm{i}z(q_j+\frac{1}{2}\sum_{s=1}^{\ell}m_s)}\right|\\&=\begin{cases}0, & \phi\in[0,\pi],\\ -\sum_{s=1}^{\ell}m_s\sin\phi, & \phi\in[-\pi,0].\end{cases}\end{aligned}$$

Therefore, the indicator diagram is $G_f=\big[-\mathrm{i}\sum_{s=1}^{\ell}m_s,0\big]$. Finally, observe that the set of zeros of f is exactly

$$\{\delta_s+2\pi\mathrm{i}k\}_{\substack{s=1,\dots,\ell\\|k|>N}}\bigcup\{\varepsilon_j\}_{j=1,\dots,r},$$

the roots $\delta_s+2\pi\mathrm{i}k$ are of multiplicity m_s, and the roots ε_j are of multiplicity m_j'. To summarize, we conclude that $f(z)$ is a generating function (see [1, Definition II.4.21]) of the family $\mathcal{E}$ on the interval $[0,\sum_{s=1}^{\ell}m_s]$ and, therefore, this family is a Riesz basis of $L_2(0,\sum_{s=1}^{\ell}m_s)$. The statement is proved.

(ii) Let us denote $\gamma=T-\sum_{s=1}^{\ell}m_s>0$ and choose a complex number μ such that

$$\mu+\frac{2\pi\mathrm{i}m}{\gamma}\neq\delta_s+2\pi\mathrm{i}k\quad\text{and}\quad\mu+\frac{2\pi\mathrm{i}m}{\gamma}\neq\varepsilon_j$$

for all $m,k\in\mathbb{Z}$, $s=1,\dots,\ell$, $j=1,\dots,r$. Let us put $m'=\sum_{j=1}^{r}m_j'$ and

$$\begin{aligned}\mathcal{E}_1^{(m')}&=\left\{\mathrm{e}^{(\mu+\frac{2\pi\mathrm{i}m}{\gamma})t}\right\}_{m\in\mathbb{Z}\setminus\{1,\dots,m'\}},\\\widetilde{\mathcal{E}}_N&=\left\{\mathrm{e}^{(\delta_s+2\pi\mathrm{i}k)t},t\mathrm{e}^{(\delta_s+2\pi\mathrm{i}k)t},\dots,t^{m_s-1}\mathrm{e}^{(\delta_s+2\pi\mathrm{i}k)t}\right\}_{\substack{|k|>N\\s=1,\dots,\ell}}\end{aligned}$$

and consider the family

$$\widetilde{\mathcal{E}}_N\cup\mathcal{E}_0\cup\mathcal{E}_1^{(m')}.$$

Now let us introduce a complex function of the sine type given by

$$f_1(z)=\mathrm{e}^{\mathrm{i}(\frac{z}{2}\sum_{s=1}^{\ell}m_s+\gamma)}\prod_{s=1}^{\ell}\left(\sin\left(\frac{z}{2}-\frac{\delta_s}{2}\right)\right)^{m_s}R_1(z)\sin\gamma\left(\frac{z}{2}-\frac{\mu}{2}\right),$$

where

$$R_1(z)=\prod_{j=1}^{r}(z-\varepsilon_j)^{m_j'}\left(\prod_{m=1}^{m_j'}\left(z-\mu-\frac{2\pi\mathrm{i}m}{\gamma}\right)\right)^{-1}\to1\quad\text{as}\quad z\to\infty.$$

Then, by arguments analogous to those given in the proof of part (i), it follows that $f_1(z)$ is a generating function of $\widetilde{\mathcal{E}}_N \cup \mathcal{E}_0 \cup \mathcal{E}_1^{(m')}$ on the interval $[0,T]$. Therefore, this family forms a Riesz basis of $L_2(0,T)$. Now, since $\mathcal{E} \subset \widetilde{\mathcal{E}}_N \cup \mathcal{E}_0 \cup \mathcal{E}_1^{(m')}$, this means that $\mathcal{E}$ forms a Riesz basis in $\operatorname{Cl}\operatorname{Lin}\mathcal{E} \subset L_2(0,T)$. The proof of the theorem is complete. □

Now we apply Theorem 4.4 to the collection of functions appearing in (3.16). Let us fix $d \in \{1,\dots,r\}$ and choose an arbitrary subset $L \subset \{1,\dots,\ell\}$. Next, for any $m \in L$ we choose $j(m) \in \{1,\dots,\nu_m\}$ and denote $J(L) = \{j(m)\}_{m\in L}$. Finally, for any couple $(m, j(m))$, $m \in L$, we put $\pi_{m,j(m)} = \deg q_{m,k}^{j(m),1,d}(t) + 1$. Let us recall that from (3.14) and Lemmas 3.2 and 3.3 it follows that this degree does not depend on k.

THEOREM 4.5. *For any choice of $d, L, J(L)$, for any $p'_{m,j(m)}$, such that $1 \le p'_{m,j(m)} \le \pi_{m,j(m)}$, and for any $T \ge n' = \sum_{m\in L} p'_{m,j(m)}$ the collection of functions*

$$\Phi_1 = \left\{ \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,d}(t),\ |k| > N;\ m \in L;\ s = \pi_{m,j(m)} - p'_{m,j(m)} + 1, \dots, \pi_{m,j(m)} \right\}$$

constitutes a Riesz basis of $\operatorname{Cl}\operatorname{Lin}\Phi_1$ *in* $L_2(0,T)$.

If in addition N is large enough, then the family

$$\Phi_2 = \left\{ \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,d}(t) + f_{m,k}^{j(m),s,d}(t) \right\}_{|k|>N;\ m\in L;\ s=\pi_{m,j(m)}-p'_{m,j(m)}+1,\dots,\pi_{m,j(m)}}$$

also forms a Riesz basis of $\operatorname{Cl}\operatorname{Lin}\Phi_2$ *in* $L_2(0,T)$.

If $T = n'$, the subspaces $\operatorname{Cl}\operatorname{Lin}\Phi_1$ *and* $\operatorname{Cl}\operatorname{Lin}\Phi_2$ *are of finite codimension* $(2N+1)n'$ *in* $L_2(0,n')$.

Proof. Consider the linear operator $\mathcal{T} : \operatorname{Lin}\Phi_1 \to \operatorname{Lin}\Phi_1$ defined on the elements of Φ_1 by the equalities

$$\mathcal{T}(\mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,d}(t)) = \mathrm{e}^{\lambda_m^{(k)} t} t^{p'_{m,j(m)}-s}$$

for $|k| > N$; $m \in L$; $s = \pi_{m,j(m)} - p'_{m,j(m)} + 1, \dots, \pi_{m,j(m)}$. It follows from the properties (P1) and (P2) (see section 3) and Theorem 4.4 that the operator $\mathcal{T}$ is bounded in the sense of $L_2(0,T)$ and its extension to $L = \operatorname{Cl}\operatorname{Lin}\Phi_1$ is a bounded one-to-one operator from L to L. Hence, since the images of the elements of Φ_1 form a Riesz basis of L (Theorem 4.4), then Φ_1 is also a Riesz basis of this subspace of $L_2(0,T)$.

Next, let us introduce in $L_2(0,T)$ an equivalent norm $\|\cdot\|_1$ in which the system Φ_1 becomes orthonormal. Let Φ_1^c be an orthonormal complement of the basis Φ_1 to a basis of $L_2(0,T)$. Now using the property (P3), we choose the scalar N large enough so that

$$\sum_{\substack{|k|>N\\ m\in L\\ s\in I_m}} \|f_{m,k}^{j(m),s,d}\|_1^2 \le C \sum_{\substack{|k|>N\\ m\in L\\ s\in I_m}} \|f_{m,k}^{j(m),s,d}\|_{L_2}^2 < 1,$$

where $I_m = \pi_{m,j(m)} - p'_{m,j(m)} + 1, \dots, \pi_{m,j(m)}$. Then $\Phi_2 \cup \Phi_1^c$ is quadratically close in $\|\cdot\|_1$ to the orthonormal system $\Phi_1 \cup \Phi_1^c$ with a quadratic distance less than 1. This means that $\Phi_2 \cup \Phi_1^c$ forms also a Riesz basis in $L_2(0,T)$ (see Gohberg and Krein [4]). As a consequence, Φ_2 is a Riesz basis in $\operatorname{Cl}\operatorname{Lin}\Phi_2$.

Finally, let us observe that in the case $T = n'$ the space L, which is also presented as

$$L = \operatorname{Cl}\operatorname{Lin}\left\{ \mathrm{e}^{\lambda_m^{(k)} t} t^{p'_{m,j(m)}-s}, |k| > N; m \in L, s = \pi_{m,j(m)} - p'_{m,j(m)} + 1, \dots, \pi_{m,j(m)} \right\},$$

is of codimension $(2N+1)n'$ in $L_2(0,T)$ (see Theorem 4.4). Then Φ_1^c consists of exactly $(2N+1)n'$ elements. The proof is complete. □

5. Analysis of the controllability for a single control. Let us study the solvability of the systems of equalities (3.16) and (3.17). We assume again that the matrix A_{-1} is not singular, $\det A_{-1} \neq 0$.

Consider the sequence of functions

$$\left\{\int_0^T \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,d}(t) + f_{m,k}^{j(m),s,d}(t)\mathrm{d}t\right\} = \left\{\int_0^T \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,d}(t)\mathrm{d}t + \int_0^T f_{m,k}^{j(m),s,d}(t)\mathrm{d}t\right\} \tag{5.1}$$

for $|k| > N,\ s = 1, \dots, p_{m,j(m)}$, and any fixed d, m, j and $u(\cdot) \in L_2(0,T)$. It follows from Theorem 4.5 that all nonzero functions of the collection

$$\left\{\mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,d}(t),\ |k| > N;\ s = 1, \dots, p_{m,j}\right\}$$

form a Riesz basis of their linear span in $L_2(0,T')$ if T' is large enough. Therefore, by Proposition 4.1, the first term of (5.1) belongs to the class ℓ_2. On the other hand, the second term also belongs to ℓ_2 due to Proposition 4.3. This gives the following proposition.

PROPOSITION 5.1. *If the state* $\binom{y_T}{z_T(\cdot)}$ *is reachable from* 0 *by the system* (1.3), *then it satisfies the following equivalent conditions:*

(C1) $\sum_{\substack{|k|>N \\ m,j,s}} k^2 |\langle \binom{y_T}{z_T(\cdot)}, \psi_{m,k}^{j,s}\rangle|^2 < \infty$,

(C2) $\sum_{\substack{|k|>N \\ m=1,\dots,\ell}} k^2 \|P_m^{(k)} \binom{y_T}{z_T(\cdot)}\|^2 < \infty$,

(C3) $\binom{y_T}{z_T(\cdot)} \in \mathcal{D}(\mathcal{A})$.

Proof. The condition (C1) follows from the previous consideration. Note that actually the validity of (C1) does not depend on the choice of the basis $\{\psi\}$. In fact, we can observe that

$$P_m^{(k)} \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix} = \sum_{\substack{j=1,\dots,\nu_m \\ s=1,\dots,p_{m,j}}} \left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \psi_{m,k}^{j,s} \right\rangle \phi_{m,k}^{j,s}.$$

From here and since $\{\psi\}$ is a Riesz basis [14], we deduce that there exist two constants c and C (independently of m and k) such that

$$c^2 \sum_{j,s} \left| \left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \psi_{m,k}^{j,s} \right\rangle \right|^2 \le \|P_m^{(k)} \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}\|^2 \le C^2 \left| \left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \psi_{m,k}^{j,s} \right\rangle \right|^2 \tag{5.2}$$

and this gives the equivalence between (C1) and (C2). Let us show now that (C1) and (C2) are equivalent to (C3).

First of all, we notice that from the explicit form of the resolvent $R(\lambda, \mathcal{A})$ given in [11, Proposition 1] and by arguments and estimates given in the proof of [11, Theorem 2.9], it follows that there exists a constant C such that

$$\|R(\lambda, \mathcal{A})| \le C, \qquad \lambda \in L_m^{(k)}, \quad |k| > N,\ m = 1, \dots, \ell. \tag{5.3}$$

Let $\mathcal{A}_m^{(k)} : V_m^{(k)} \to V_m^{(k)}$ be the restriction of the operator $\mathcal{A}$ to its invariant subspace $V_m^{(k)}$. Then due to (5.3) we have

$$\begin{aligned} \|\mathcal{A}_m^{(k)} v\| &\le \int_{L_m^{(k)}} |\lambda| \|R(\lambda,\mathcal{A})\| \|v\| |\mathrm{d}\lambda| \le C_1 |k| \|v\|, \\ \left\| \left(\mathcal{A}_m^{(k)}\right)^{-1} v \right\| &\le \int_{L_m^{(k)}} \frac{1}{|\lambda|} \|R(\lambda,\mathcal{A})\| \|v\| |\mathrm{d}\lambda| \le \frac{C_1'}{|k|} \|v\|, \end{aligned}$$

where $v \in V_m^{(k)}$ and the constants C_1, C_1' do not depend on m, k. From this, one can obtain for $v \in V_m^{(k)}$ the inequality

$$\frac{1}{C_1'} \|v\| \le \frac{1}{k} \|\mathcal{A}_m^{(k)} v\| \le C_1 \|v|. \tag{5.4}$$

With our notations, the condition (C3) is obviously equivalent to

$$\sum_{\substack{|k|>N \\ m=1,\dots,\ell}} \left\| \mathcal{A}_m^{(k)} P_m^{(k)} \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix} \right\|^2 < \infty.$$

But, on the other hand, due to (5.4) this condition is equivalent to (C2). This completes the proof. □

From Proposition 5.1 it follows once more, as was pointed out in the introduction (see also [5]), that the set $\mathcal{R}_T$ of the states reachable from 0 by virtue of the system (1.3) and controls from $L_2(0,T)$ is always a subset of $\mathcal{D}(\mathcal{A})$. This justifies also Definition 1.1 given in the introduction: *the system* (1.3) *is said to be null-controllable at the time* T *if* $\mathcal{R}_T = \mathcal{D}(\mathcal{A})$. Next, we give the necessary conditions of null-controllability.

THEOREM 5.2. *Assume that the system* (1.3) *is null-controllable by controls from* $L_2(0,T)$ *for some* $T > 0$. *Then the following two conditions hold.*

(i) *There is no* $\lambda \in \mathbb{C}$ *and* $y \in \mathbb{C}^n, y \ne 0$, *such that* $\Delta_{\mathcal{A}}^*(\lambda) y = 0$ *and* $B^* y = 0$, *where*

$$\Delta_{\mathcal{A}}^*(\lambda) = \lambda I - \lambda \mathrm{e}^{-\lambda} A_{-1}^* - \lambda \int_{-1}^{0} \mathrm{e}^{\lambda s} A_2^*(s) \mathrm{d}s - \int_{-1}^{0} \mathrm{e}^{\lambda s} A_3^*(s) \mathrm{d}s,$$

or equivalently $\operatorname{rank} \begin{pmatrix} \Delta_{\mathcal{A}}(\lambda) & B \end{pmatrix} = n$ *for all* $\lambda \in \mathbb{C}$.

(ii) *There is no* $\mu \in \sigma(A_{-1})$ *and* $y \in \mathbb{C}^n, y \ne 0$, *such that* $A_{-1}^* y = \bar{\mu} y$ *and* $B^* y = 0$, *or equivalently* $\operatorname{rank} \begin{pmatrix} B & A_{-1}B & \cdots & A_{-1}^{n-1}B \end{pmatrix} = n$.

First we prove the following lemma.

LEMMA 5.3. *Condition* (i) *of Theorem* 5.2 *is equivalent to the following condition:*

(i′) *There is no eigenvector* g *of the adjoint operator* $\mathcal{A}^*$ *belonging to* $\operatorname{Ker} \mathcal{B}^*$.

Proof of Lemma 5.3. We make use of the following explicit form of $\mathcal{A}^*$:

$$\mathcal{A}^* \begin{pmatrix} y \\ z(\cdot) \end{pmatrix} = \begin{pmatrix} A_2^*(0) y + z(0) \\ -\frac{\mathrm{d}}{\mathrm{d}\theta} \left(z(\theta) + A_2^*(\theta) y \right), + A_3^*(\theta) y \end{pmatrix}$$

with the domain

$$\begin{aligned} \mathcal{D}(\mathcal{A}^*) = \big\{ \big(y, z(\cdot) \big) \in M_2 : z(\theta) + A_2^*(\theta) y \in H^1([-1,0], \mathbb{C}^n), \\ (A_{-1}^* A_2^*(0) - A_2^*(-1)) y = z(-1) - A_{-1}^* z(0) \big\}. \end{aligned}$$

From this expression of the adjoint operator, one can show that $\mathcal{A}^*g = \lambda g$ if and only if

$$g = \begin{pmatrix} y \\ \left(\lambda \mathrm{e}^{-\lambda\theta} I - A_2^*(\theta) + \lambda \mathrm{e}^{-\lambda\theta} \int_0^\theta \mathrm{e}^{\lambda s} A_2^*(s)\mathrm{d}s + \mathrm{e}^{-\lambda\theta} \int_0^\theta \mathrm{e}^{\lambda s} A_3^*(s)\mathrm{d}s\right) y \end{pmatrix},$$

where $y \in \operatorname{Ker} \Delta^*(\lambda)$. Since $\mathcal{B}^*g = B^*y$, the proof of the lemma is complete. □

Proof of Theorem 5.2. Let (i) be false. Then by Lemma 5.3 there exists a vector $g \neq 0$ such that $\mathcal{A}^*g = \lambda g$ and $g \in \operatorname{Ker} \mathcal{B}^*$. Consider an arbitrary state $\binom{y_T}{z_T(\cdot)} \in \mathcal{R}_T$, i.e., which is of the form (1.4). This gives

$$\left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, g \right\rangle = \int_0^T \left\langle u(t), \mathcal{B}^* \mathrm{e}^{\mathcal{A}^* t} g \right\rangle \mathrm{d}t = 0.$$

This means that $\mathcal{R}_T$ is not dense in M_2 and so cannot be equal to $\mathcal{D}(\mathcal{A})$ which is dense in M_2 because $\mathcal{A}$ is an infinitesimal generator. Hence null-controllability is impossible.

Now let condition (ii) not hold, i.e., there exists a nonzero vector $y \in \mathbb{C}^n$ such that

$$A_{-1}^* y = \bar{\mu}_m y \qquad \text{and} \qquad B^* y = 0. \tag{5.5}$$

With our notations, we can represent y as

$$y = \sum_{j=1}^{\nu_m} \alpha_j C_{m,j}^1,$$

where $C_{m,j}^1$ is a basis of the eigenspace of A_{-1}^* corresponding to the eigenvalue $\bar{\mu}_m$. Among the moment equalities (3.16) we can extract those corresponding to $s = p_{m,j}$ (for fixed m and $j = 1, \dots, \nu_m$), i.e.,

$$s_k^j = k \left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \psi_{m,k}^{j,p_{m,j}} \right\rangle = \sum_{d=1}^{r} \int_0^T \left(\mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j,p_{m,j},d}(t) + f_{m,k}^{j,p_{m,j},d}(t) \right) u_d(t) \mathrm{d}t \tag{5.6}$$

for $|k| > N,\ j = 1, \dots, \nu_m$. From (3.14) and Lemma 3.2 we have

$$q_{m,k}^{j,p_{m,j},d}(t) = k \left\langle \mathbf{b}_d, \bar{\psi}_{m,k}^{j,p_{m,j}} \right\rangle = \frac{k}{\bar{\lambda}_m^{(k)}} \left\langle b_d, C_{m,j}^1 \right\rangle. \tag{5.7}$$

Let us show that the moment problem (5.6) cannot be solved for all $\{s_k^j\} \in \ell_2$.

Assume the opposite; then the problem

$$\widetilde{s}_k^j = \sum_{j=1}^{\nu_m} \bar{\alpha}_j s_k^j = \sum_{d=1}^{r} \int_0^T \sum_{j=1}^{\nu_m} \bar{\alpha}_j \left(\mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j,p_{m,j},d}(t) + f_{m,k}^{j,p_{m,j},d}(t) \right) u_d(t) \mathrm{d}t$$

is also solvable for all $\{\widetilde{s}_k^j\} \in \ell_2$. On the other hand, (5.5) and (5.7) imply

$$\sum_{j=1}^{\nu_m} \bar{\alpha}_j \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j,p_{m,j},d}(t) = \mathrm{e}^{\lambda_m^{(k)} t} \frac{k}{\bar{\lambda}_m^{(k)}} \left\langle b_d, \sum_{j=1}^{\nu_m} \alpha_j C_{m,j}^1 \right\rangle = 0.$$

Hence the latter moment problem reads as

$$\widetilde{s}_k^j = \sum_{d=1}^{r} \int_0^T g_k^d(t) u_d(t) \mathrm{d}t, \tag{5.8}$$

where $g_k^d(t) = \sum_{j=1}^{\nu_m} \bar{\alpha}_j f_{m,k}^{j,p_{m,j},d}(t)$, $|k| > N$, and, due to the property (P3), these functions satisfy

$$\sum_{|k|>N} \int_0^T |g_k^d(t)|^2 \mathrm{d}t < \infty.$$

However, by Proposition 4.3 it follows that the set of solvability of (5.8) is a linear submanifold $\ell' \subset \ell_2$, $\ell' \neq \ell_2$. From the obtained contradiction, we conclude that there exist sequences $\{s_k^j\}_{\substack{|k|>N \\ j=1,\dots,\nu_m}}$ for which (5.6) is not solvable. This means that there exist states $\binom{y_T}{z_T(\cdot)}$ that satisfy (C1) but are not reachable from 0 by the system (1.3). Thus $\mathcal{R}_T \neq \mathcal{D}(\mathcal{A})$. □

The following results will be used to prove the main results on controllability. They are also of independent interest.

LEMMA 5.4. *Assume that for an abstract system* (1.1) *the following conditions hold:*

(a) $\mathcal{R}_T \subset \mathcal{D}(\mathcal{A})$ *for all* $T > 0$,

(b) *for some* $T_0 > 0$ *the set* $\mathcal{R}_{T_0}$ *is a closed subspace of finite codimension in the space* $X_{\mathcal{A}} = \mathcal{D}(\mathcal{A})$, *with the standard graph norm* $\|x\|_{\mathcal{A}} = \sqrt{\|x\|^2 + \|\mathcal{A}x\|^2}$.

Then for all $T \geq T_0$ *we have* $\mathcal{R}_T = L$, *where* L *is a subspace of* $\mathcal{D}(\mathcal{A})$ *invariant by the semigroup* e^{At} *and* $0 < \operatorname{codim} L \leq \operatorname{codim} \mathcal{R}_{T_0} < \infty$.

Proof. Taking into account the inclusion $\mathcal{R}_{T_1} \subset \mathcal{R}_{T_2}$ as $T_1 \subset T_2$ we infer from assumptions (a) and (b) that there exists $\varepsilon > 0$ such that

$$\mathcal{R}_T = L, \quad T \in]T_0, T_0 + \varepsilon], \tag{5.9}$$

where L is a subspace such that $0 \leq \operatorname{codim} L \leq \operatorname{codim} \mathcal{R}_{T_0}$. Let us show that the relation (5.9) holds also for all $T > T_0$. To do that it is enough to prove that

$$\mathcal{R}_{T_0 + \frac{3}{2}\varepsilon} = L. \tag{5.10}$$

Let us put

$$\mathcal{R}_{T_2}^{T_1} = \left\{ x : x = \int_{T_1}^{T_2} \mathrm{e}^{\mathcal{A}t} B u(t) \mathrm{d}t, u \in L_(T_1, T_2; U) \right\}$$

and $\mathcal{R}_T^0 = \mathcal{R}_T$. Let us prove first that

$$\mathcal{R}_{T_0 + \frac{3}{2}\varepsilon}^{T_0 + \varepsilon} \subset L. \tag{5.11}$$

In fact, it is easy to see that $\mathcal{R}_{T_0 + \frac{3}{2}\varepsilon}^{T_0 + \varepsilon} = \mathrm{e}^{\mathcal{A}\frac{\varepsilon}{2}} \mathcal{R}_{T_0 + \varepsilon}^{T_0 + \frac{\varepsilon}{2}}$. On the other hand, it follows from (5.9) that $L = \mathcal{R}_{T_0 + \varepsilon} = \mathcal{R}_{T_0 + \frac{\varepsilon}{2}}$ and hence $\mathcal{R}_{T_0 + \varepsilon}^{T_0 + \frac{\varepsilon}{2}} \subset \mathcal{R}_{T_0 + \varepsilon} = L$. Therefore

$$\mathrm{e}^{\mathcal{A}\frac{\varepsilon}{2}} \mathcal{R}_{T_0 + \varepsilon}^{T_0 + \frac{\varepsilon}{2}} \subset \mathrm{e}^{\mathcal{A}\frac{\varepsilon}{2}} L = \mathrm{e}^{\mathcal{A}\frac{\varepsilon}{2}} \mathcal{R}_{T_0 + \frac{\varepsilon}{2}} = \mathcal{R}_{T_0 + \varepsilon}^{\frac{\varepsilon}{2}} \subset \mathcal{R}_{T_0 + \varepsilon} = L.$$

Now from (5.11) and from the obvious relation

$$\mathcal{R}_{T_0+\frac{3}{2}\varepsilon} = \mathcal{R}_{T_0+\varepsilon} + \mathcal{R}^{T_0+\varepsilon}_{T_0+\frac{3}{2}\varepsilon},$$

we infer that

$$L \subset \mathcal{R}_{T_0+\frac{3}{2}\varepsilon} = \mathcal{R}_{T_0+\varepsilon} + \mathcal{R}^{T_0+\varepsilon}_{T_0+\frac{3}{2}\varepsilon} \subset L + L = L,$$

which proves (5.10).

Thus (5.9) is valid for all $T > T_0$. Then $L = \cup_{T>0}\mathcal{R}_T$, and, therefore, it is an invariant subspace for the semigroup $\{e^{\mathcal{A}t}\}_{t\geq 0}$. The lemma is proved. □

In the following, we denote by $X_{\mathcal{A}}$ the space $\mathcal{D}(\mathcal{A}) \subset M_2$ with the graph norm.

THEOREM 5.5. *For the system* (1.3) *let there exist a natural* N *and* $T_0 > 0$ *such that the moment problem* (3.16) *for* $T = T_0$ *and* $|k| > N$ *is solvable for all the vectors*

$$\left\{ k \left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \psi^{j,s}_{m,k} \right\rangle \right\}_{|k|>N}$$

satisfying the condition (C1). *Then, from the condition* (i′) *of Lemma* 5.3, *it follows that* $\mathcal{R}_T = \mathcal{D}(\mathcal{A})$ *as* $T > T_0$.

Proof. Let us denote by $L_N \subset \mathcal{D}(\mathcal{A})$ the subspace

$$L_N = \mathrm{Cl}_{X_{\mathcal{A}}} \sum_{\substack{|k|>N \\ m=1,\dots,\ell}} V^{(k)}_m$$

and by P_N a projector onto L_N in $X_{\mathcal{A}}$. From the assumption on solvability of the problem (3.16), it follows that $P_N\mathcal{R}_{T_0} = L_N$. This, in particular, means that $\mathcal{R}_{T_0}$ is a subspace of finite codimension: $\operatorname{codim} L_N = (2N+1)n$ in $\mathcal{D}(\mathcal{A})$. Then by Lemma 5.4 we conclude that $\mathcal{R}_T = L$ as $T > T_0$, where $L \subset \mathcal{D}(\mathcal{A})$ is invariant with respect to $\{e^{\mathcal{A}t}\}_{t\geq 0}$ and $\operatorname{codim} L \leq (2N+1)n$. Let us prove that under the condition (i′) we have in fact $L = \mathcal{D}(\mathcal{A})$.

Assume the contrary. Then let us consider the dual space $X^*_{\mathcal{A}}$ and denote by $L^\perp \subset X^*_{\mathcal{A}}$ the subspace of functionals on $X_{\mathcal{A}}$ which are 0 on L. Obviously $L^\perp$ is finite dimensional. Denote by $\mathcal{A}^*_1$ the infinitesimal extension of $\mathcal{A}^*$ to the space $X^*_{\mathcal{A}}$ generating the semigroup $e^{\mathcal{A}^*_1 t}$. Since, due to Lemma 5.4, L is invariant with respect to $\{e^{\mathcal{A}t}\}_{t\geq 0}$, and then $L^\perp$ is invariant with respect to $\{e^{\mathcal{A}^*_1 t}\}_{t\geq 0}$. Taking into account the finite dimensionality of $L^\perp$, we conclude that $L^\perp \subset \mathcal{D}(\mathcal{A}^*_1)$ and there exists an eigenvector g of the operator $\mathcal{A}^*_1$ that lies in $L^\perp$. Let us notice now that the collection of subspaces $\{V^{(k)}_m, m = 1,\dots,\ell; |k| > N\}$ is a Riesz basis also for the space $X_{\mathcal{A}}$ and all these subspaces are invariant for the operator $\mathcal{A}_1 = \mathcal{A}_{|_{\mathcal{D}(\mathcal{A})}}$. This implies that the collection $\{W^{(k)}_m, m = 1,\dots,\ell; |k| > N; \widehat{W}_N\}$ is a Riesz basis of invariant subspaces for $\mathcal{A}^*_1$ in the space $X^*_{\mathcal{A}}$. From this, we infer that all the eigenvectors of $\mathcal{A}^*_1$ lie in

$$\bigcup_{m,k} W^{(k)}_m \bigcup \widehat{W}_N \subset \mathcal{D}(\mathcal{A}^*)$$

and, therefore, g is also an eigenvector for $\mathcal{A}^*$. Since $g \in L^\perp$, then

$$\left\langle \int_0^T e^{\mathcal{A}t}\mathcal{B}u(t)\mathrm{d}t, g \right\rangle = 0, \quad u(\cdot) \in L_2(0,T;U).$$

If we put $u(t) \equiv u \in U,\ t \in [0,T]$, the latter relation brings

$$0 = \int_0^T \left\langle u, \mathcal{B}^* \mathrm{e}^{\mathcal{A}^* t} g \mathrm{d}t \right\rangle = \int_0^T \langle u, \mathcal{B}^* g \rangle \, \mathrm{e}^{\mu t} \mathrm{d}t = \langle u, \mathcal{B}^* g \rangle \int_0^T \mathrm{e}^{\mu t} \mathrm{d}t$$

for all $u \in U$, where $\mathcal{A}^* g = \bar{\mu} g$. This implies $g \in \operatorname{Ker} \mathcal{B}^*$, which contradicts (C1). That completes the proof. □

Now we are ready to prove the first important result of our work.

THEOREM 5.6. *Let conditions* (i) *and* (ii) *of Theorem* 5.2 *hold. Then*

(i) *the system* (1.3) *is null-controllable at the time T as $T > n$;*

(ii) *if the system* (1.3) *is of single control ($r = 1$), then the estimation of the time of controllability in* (i) *is exact, i.e., the system is not controllable at time $T = n$.*

If the delay is h instead of 1, the time of exact controllability is $T = nh$.

Proof. Here we prove (i) for the case of a single control. In the case of multivariable control we obtain a more precise estimate for the time of controllability in section 6.

First of all, let us observe that conditions (i) and (ii) of Theorem 5.2 imply, in the case of single control, that all the eigenspaces of $\mathcal{A}^*$ and $\widetilde{\mathcal{A}}^*$ are one dimensional. In fact, otherwise we will have that there exists an eigenvector g of $\mathcal{A}^*$ or $\widetilde{\mathcal{A}}^*$ such that $\langle \mathbf{b}, g \rangle = 0$. But we know that g has the form

$$g = \begin{pmatrix} y \\ z(\theta) \end{pmatrix},$$

where y is a nonzero eigenvector for the pencil $\Delta^*(\lambda)$ (that is, $\Delta^*(\lambda_0) y = 0$ for some λ_0) or of the matrix A_{-1}^*, respectively. Since $\langle \mathbf{b}, g \rangle = 0$ gives $\langle b, y \rangle = 0$ we arrive at a contradiction with the conditions of Theorem 5.2.

Thus, equalities (3.16) and (3.17) take, in our case, the form

$$k \left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \psi_{m,k}^{1,s} \right\rangle = \int_0^T \left(\mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{1,s}(t) + f_{m,k}^{1,s}(t) \right) u(t) \mathrm{d}t, \tag{5.12}$$

where $|k| > N,\ m = 1, \dots, \ell,\ s = 1, \dots, p_{m,1}$, and

$$\left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \widehat{\psi}_m^{1,s} \right\rangle = \int_0^T \mathrm{e}^{\widehat{\lambda}_m t} \widehat{q}_m^{1,s}(t) u(t) \mathrm{d}t, \tag{5.13}$$

where $m = 1, \dots, \ell_N,\ s = 1, \dots, \widehat{p}_{m,1}$. From Lemmas 3.2 and 5.3, it follows that all polynomials $\{q(t)\},\ \{\widehat{q}(t)\}$ are nontrivial and $\deg q_{m,k}^{1,s}(t) = p_{m,1} - s,\ \deg \widehat{q}_m^{1,s}(t) = \widehat{p}_{m,1} - s$. This gives

$$\sum_{m=1}^{\ell} p'_{m,1} = \sum_{m=1}^{\ell} \left(\deg q_{m,k}^{1,1} + 1 \right) = \sum_{m=1}^{\ell} p_{m,1} = n.$$

Applying Theorem 4.5, we find that for a large enough N, the collection

$$\Phi = \left\{ \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{1,s}(t) + f_{m,k}^{1,s}(t) \right\}_{\substack{|k|>N \\ m=1,\dots,\ell \\ s=1,\dots,p_{m,1}}} \bigcup \left\{ \mathrm{e}^{\widehat{\lambda}_m t} \widehat{q}_m^{1,s}(t) \right\}_{\substack{m=1,\dots,\ell_N \\ s=1,\dots,\widehat{p}_{m,1}}}$$

forms a Riesz basis in $\operatorname{Cl} \operatorname{Lin} \Phi \subset L_2(0,T)$. Then by Proposition 4.1 the moment problem (5.12) is solvable if and only if (C1) holds. Due to Theorem 5.5, this yields $\mathcal{R}_T = \mathcal{D}(\mathcal{A})$ for $T > n$.

To prove (ii) we first recall that the total number of elements of the family

$$\widehat{\Phi} = \left\{ \mathrm{e}^{\widehat{\lambda}_m t} \widehat{q}_m^{1,s}(t),\ m = 1, \dots, \ell_N;\ s = 1, \dots, \widehat{p}_{m,1} \right\}$$

equals $\sum_{m=1}^{\ell} \widehat{p}_{m,1} = (2N+2)n$. Since $\sum_{m=1}^{\ell} p_{m,1} = n$, we have

$$\sum_{m=1}^{\ell} \widehat{p}_{m,1} = (2N+1) \sum_{m=1}^{\ell} p_{m,1} + n.$$

On the other hand, it follows from Theorem 4.5 that in $L_2(0,n)$ we have

$$\operatorname{codim} \operatorname{Cl} \operatorname{Lin} \Phi_1 = (2N+1)n = (2N+1) \sum_{m}^{\ell} p_{m,1},$$

where

$$\Phi_1 = \left\{ \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{1,s}(t) + f_{m,k}^{1,s}(t),\ |k| > N,\ m = 1, \dots, \ell,\ s = 1, \dots, p_{m,1} \right\}.$$

This means that the family $\Phi = \Phi_1 \cup \widehat{\Phi}$ contains at least n functions, which are presented as linear combinations of the others. As a consequence, the set of reachability $\mathcal{R}_T$ for $T = n$ cannot be equal to $\mathcal{D}(\mathcal{A})$. More precisely, the codimension of $\mathcal{R}_T$ in $\mathcal{D}(\mathcal{A})$ satisfies the estimation $n \le \operatorname{codim} \mathcal{R}_T < \infty$. The theorem is proved. □

Remark 5.7. It is clear that the system (1.3) is also uncontrollable at time $T < n$. Moreover, it follows from Proposition 4.2 that, in this case, the set $\operatorname{Cl} \mathcal{R}_T$ is of infinite codimension in $X_{\mathcal{A}}$.

6. Controllability in the multivariable case. Let us now consider the multivariable case: $\dim B = r$ with also the assumption that the matrix A_{-1} is not singular, $\det A_{-1} \neq 0$.

Let $\{b_1, \dots, b_r\}$ be an arbitrary basis noted β. Let us introduce a set of integers. We denote $B_i = (\, b_{i+1}, \dots, b_r \,)$, $i = 0, 1, \dots, r-1$, which gives in particular $B_0 = B$ and $B_{r-1} = (\, b_r \,)$ and we put formally $B_r = 0$. We need in the following the integers

$$m_i^{\beta} = \operatorname{rank} (\, B_{i-1} \quad A_{-1}B_{i-1} \quad \cdots \quad A_{-1}^{n-1}B_{i-1} \,) - \operatorname{rank} (\, B_i \quad A_{-1}B_i \quad \cdots \quad A_{-1}^{n-1}B_i \,) \tag{6.1}$$

corresponding to the basis β. Let us denote

$$m_1 = \max_{\beta} m_1^{\beta}, \qquad \overline{m} = \min_{\beta} \max_{i} m_i^{\beta}, \tag{6.2}$$

for all possible choices of a basis β. It is easy to show that for all β, there exists i such that $m_i^{\beta} \ge m_1$ and then $\overline{m} \ge m_1$. Indeed, assume that m_1 is realized on the basis $\beta = \{b_1, \dots, b_r\}$, and consider an arbitrary basis $\beta_0 = \{b_1^0, \dots, b_r^0\}$. Then there exists i such that $\operatorname{Lin} \{b_i^0, \dots, b_r^0\} \not\subset \operatorname{Lin} \{b_2, \dots, b_r\}$ but $\operatorname{Lin} \{b_{i+1}^0, \dots, b_r^0\} \subset \operatorname{Lin} \{b_2, \dots, b_r\}$. For this integer i we have $m_i^{\beta_0} \ge m_1$.

Now we can formulate the main result of this section.

THEOREM 6.1. *Let conditions* (i) *and* (ii) *of Theorem* 5.2 *hold. Then*

(i) *the system* (1.3) *is null-controllable at the time* $T > \overline{m}$*;*

(ii) *the system* (1.3) *is not controllable at the time* $T < m_1$.

If the delay is h instead of 1, then in (i) and (ii) $\overline{m}$ and m_1 must be replaced by $\overline{m}h$ and $m_1 h$, respectively.

Proof. We show first that the system is not controllable at the time $T < m_1$.

Assume that the system is controllable. Let the basis where m_1 is realized be $\{b_1, \dots, b_r\}$. Consider now the relation (3.16) together with (3.14) and (3.15) given by the controllability problem. The basis $\{\widetilde{\psi}\}$ arising in (3.14) and (3.15) is given by (2.4) and expressed via the rootvectors $C^s_{m,j}$ of the matrix A^*_{-1}.

Let us choose the vectors $C^s_{m,j}$. Consider the subspace

$$\operatorname{Im}\left(\begin{array}{cccc} B_1 & A_{-1}B_1 & \cdots & A_{-1}^{n-1}B_1 \end{array}\right),$$

where $B_1 = \left(\begin{array}{ccc} b_2 & b_3 & \cdots \quad b_r \end{array}\right)$. Then the subspace

$$\mathcal{N}_1 = \{\operatorname{Im}\left(\begin{array}{cccc} B_1 & A_{-1}B_1 & \cdots & A_{-1}^{n-1}B_1 \end{array}\right)\}^{\perp} = \bigcap_{i=0}^{n-1} \operatorname{Ker} B_1^* {A^*_{-1}}^i$$

is invariant by A^*_{-1}. The condition of controllability gives that $A^*_{-1|_{\mathcal{N}_1}}$ has, for each eigenvalue $\overline{\mu}$, only one Jordan chain. Indeed, on the contrary, if there are two chains in $\mathcal{N}_1$, then there are two independent eigenvectors corresponding to the same eigenvalue in $\mathcal{N}_1 \subset \operatorname{Ker} B_1^*$, i.e., both these vectors are orthogonal to $b_2, \dots, b_r$. Then there exists a linear combination of these eigenvectors, which is orthogonal also to b_1 and, as a consequence, belongs to $\operatorname{Ker} B^*$. This contradicts the controllability condition.

Note also that $\dim \mathcal{N}_1 = n - \dim \operatorname{Im}\left(\begin{array}{cccc} B_1 & A_{-1}B_1 & \cdots & A_{-1}^{n-1}B_1 \end{array}\right) = m_1$. We take the corresponding vectors $C^s_{m,j}$ in $\mathcal{N}_1$. This implies

$$\left\langle C^s_{m,j}, b_i \right\rangle = 0, \quad i = 2, \dots, r.$$

This means that m is chosen such that $\overline{\mu}_m$ is an eigenvalue of $A^*_{-1|_{\mathcal{N}_1}}$, j is the number of the unique Jordan chain in $\mathcal{N}_1$, and s is the index of the vectors in the Jordan chain.

Let $\Omega_{\mathcal{N}_1}$ be the set of indices of the eigenvalues $\overline{\mu}_m \in \sigma(A^*_{-1|_{\mathcal{N}_1}})$. For each $m \in \Omega_{\mathcal{N}_1}$ we have a Jordan chain, say with the number $j(m)$ in $\mathcal{N}_1$. The indices of the corresponding generalized eigenvectors are

$$s \in \left\{p_{m,j(m)}, p_{m,j(m)} - 1, \dots, p_{m,j(m)} - p'_{m,j(m)} + 1\right\} = I_m,$$

the length of the Jordan chain is $p'_{m,j(m)}$, and $C^{p_{m,j(m)}}_{m,j(m)}$ is an eigenvector. For $m \in \Omega_{\mathcal{N}_1}$, $j(m)$ and $s \in I_m$ we consider the relation (3.16) which is the expression of the controllability condition. As $C^s_{m,j(m)} \in \mathcal{N}_1$, we get

$$k \left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \psi^{j(m),s}_{m,k} \right\rangle = \int_0^T \mathrm{e}^{\lambda^{(k)}_m t} q^{j(m),s,1}_{m,k}(t) u_1(t) \mathrm{d}t + \sum_{d=1}^{r} \int_0^T f^{j,s,d}_{m,k}(t) u_d(t) \mathrm{d}t \tag{6.3}$$

for $|k| > N$. From the hypothesis of controllability at the time T, it follows that the left-hand side gives an arbitrary element of ℓ_2, and then the relation (6.3) may be represented by the expression

$$x = [Q_1 + F]\, u(\cdot), \quad x \in \ell_2, \tag{6.4}$$

where the operators Q_1 and F are linear bounded operators from $L_2(0,T)$ to ℓ_2 defined by

$$Q_1 u(\cdot) = \left\{ \int_0^T \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,1}(t) u_1(t) \mathrm{d}t;\ m \in \Omega_{\mathcal{N}_1},\ |k| > N,\ s \in I_m \right\},$$

$$Fu(\cdot) = \left\{ \sum_{d=1}^{r} \int_0^T f_{m,k}^{j(m),s,d}(t) u_d(t) \mathrm{d}t;\ m \in \Omega_{\mathcal{N}_1},\ |k| > N, s \in I_m \right\}.$$

We now need the following lemmas.

LEMMA 6.2. *The operator F is compact.*

Proof. By (P3) the operator F is compact in the same way as Q_T^r is compact in Proposition 4.3. □

LEMMA 6.3. *The image of the operator Q_1 is of infinite codimension.*

Proof. Let us recall that $m_1 = \dim \mathcal{N}_1$. Then for each k, the sum of the length of Jordan chains of the operator $\widetilde{\mathcal{A}}$ corresponding to the Jordan chains in $\mathcal{N}_1$ is m_1:

$$\sum_{m \in \Omega_{N_1}} p'_{m,j(m)} = m_1.$$

Let us first show that the family

$$\left\{ \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,1}(t),\ m \in \Omega_{\mathcal{N}_1},\ |k| > N,\ s \in I_m \right\} \tag{6.5}$$

forms a Riesz basis of the closure of its linear span in the space $L_2(0, m_1)$. In order to do that, we have to consider the family

$$\left\{ \mathrm{e}^{\lambda_m^{(k)} t}, \mathrm{e}^{\lambda_m^{(k)} t} t, \dots, \mathrm{e}^{\lambda_m^{(k)} t} t^{p_{m,j(m)}-1};\ m \in \Omega_{\mathcal{N}_1},\ |k| > N, \right\}. \tag{6.6}$$

This family forms a Riesz basis of the closure of its linear span in $L_2(0, m_1)$ by Theorem 4.4. Moreover, the closure of its linear span is of finite codimension $(2N+1)m_1$ since it may be completed by a family of $(2N+1)m_1$ functions to get a Riesz basis of $L_2(0, m_1)$. The relation between the families (6.5) and (6.6) may be written as

$$\mathcal{T}(\mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,1}(t)) = \mathrm{e}^{\lambda_m^{(k)} t} t^{p_{m,j(m)}-s},$$

where $\mathcal{T}$ is a linear bounded invertible operator in the closure of the linear span of the family (6.5). This implies that this family forms a Riesz basis in the closure of its linear span, which is of finite codimension. Then, from Proposition 4.2, the problem of moments

$$s_{m,k}^{j(m),s} = \int_0^T \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,1}(t) u_1(t) \mathrm{d}t$$

is not solvable in a subspace of infinite codimension and this implies that $\operatorname{Im} Q_1$ is of infinite codimension. Lemma 6.3 is proved. □

Let us now show that, from the fact that $\operatorname{Im} Q_1$ is of infinite codimension and F is a compact operator, we have $\operatorname{Im}[Q_1 + F] \neq \ell_2$.

The necessary and sufficient condition of the equality $\operatorname{Im}[Q_1 + F] = \ell_2$ is (cf., for example, [13, Theorem 4.13])

$$\exists \gamma > 0 \quad \forall x \in \ell_2, \quad \|[Q_1 + F]^* x\| \geq \gamma \|x\|.$$

We know that $\operatorname{Ker} Q_1^* = (\operatorname{Im} Q_1)^\perp$ is an infinite-dimensional closed subspace. This implies that

$$\forall x \in \operatorname{Ker} Q_1^*, \quad \|F^* x\| \geq \gamma \|x\|$$

for the same scalar γ, and this is impossible because F is a compact operator. Hence $\operatorname{Im}[Q_1 + F] \neq \ell_2$ and this implies that (6.4) is not possible for all $x \in \ell_2$.

Then the relation (6.3) is not possible for all $(y_T, z_T(\cdot)) \in \mathcal{D}(\mathcal{A})$ if $T < m_1$. Part (ii) of the theorem is proved.

Let us now prove part (i) of the theorem.

First we choose a basis for the relation (3.16). Let $\beta = \{b_1, \dots, b_r\}$ be an arbitrary basis of $\operatorname{Im} B$ and $T > \max\{m_i^\beta, i = 1, \dots, r\}$. Consider now the subspaces

$$\mathcal{N}_i = \bigcap_{j=0}^{n-1} \operatorname{Ker} B_i^* {A_{-1}^*}^j,$$

where $B_i = (\, b_{i+1} \quad \cdots \quad b_r \,)$, $i = 1, \dots, r-1$, $B_0 = B$, and $B_r = 0$. We have $\mathcal{N}_0 = 0$, $\mathcal{N}_r = \mathbb{C}^n$, and $\mathcal{N}_i \subset \mathcal{N}_{i+1}$ for $i = 0, \dots, r-1$. The subspaces $\mathcal{N}_i$ are invariant by A_{-1}^*. In order to construct the basis $\{\psi\}$ corresponding to $|k| > N$, we first choose a basis of generalized eigenvectors of A_{-1}^* in the following way. Let us take a basis in $\mathcal{N}_1$ as in the first part of the proof. Then we complete this basis up to a basis of $\mathcal{N}_2$ by extending some Jordan chains from $\mathcal{N}_1$ and by adding Jordan chains corresponding to some other eigenvalues. In the same way, we extend our basis up to the basis of $\mathcal{N}_3, \dots, \mathcal{N}_r = \mathbb{C}^n$.

Remark 6.4. The part of the obtained basis of $\mathcal{N}_i$ not belonging to $\mathcal{N}_{i-1}$, $i = 1, \dots, r$, does not contain two chains corresponding to the same eigenvalue. We have already proved that in $\mathcal{N}_1$ there do not exist two chains with the same eigenvalue. Suppose now that $\mathcal{N}_2$ contains the end of the first chain and the beginning of the second chain corresponding to the same eigenvalue. Let $y_1^{\mathrm{o}} \in \mathcal{N}_2, y_1^{\mathrm{o}} \notin \mathcal{N}_1$ be the continuation of the first chain from $\mathcal{N}_1$. If the maximal order of the rootvectors in $\mathcal{N}_1$ is p, then the order of y_1^{o} is $p+1$. Let y_2^{n} be a new eigenvector in the second chain of $\mathcal{N}_2$ corresponding to the same eigenvalue. Let us consider the vector $y = \alpha y_1^{\mathrm{o}} + \beta y_2^{\mathrm{n}}$. We know that y_1^{o} is not orthogonal to b_2, because if $y_1^{\mathrm{o}} \perp b_2$, then $y_1^{\mathrm{o}} \in \mathcal{N}_1$. Then one can choose α, β such that $y \perp b_2$. Then, for this choice of α and β, $y \in \mathcal{N}_1$ and it is a rootvector of higher order than p in $\mathcal{N}_1$, which contradicts the construction of $\mathcal{N}_1$ (the maximal order in $\mathcal{N}_1$ is $p-1$). This proves the remark for $i = 2$. For $i > 2$, the proof is the same.

We have then a basis in $\mathbb{C}^n$ of Jordan chains of A_{-1}^* formed by successive bases of

$$\mathcal{N}_1 \subset \mathcal{N}_2 \subset \cdots \subset \mathcal{N}_r = \mathbb{C}^n.$$

Let us denote by $\Omega_{\mathcal{N}_i/\mathcal{N}_{i-1}}$, $i = 1, \dots, r$ ($\Omega_{\mathcal{N}_1/\mathcal{N}_0} = \Omega_{\mathcal{N}_1}$), the set of the indices m of the eigenvalues of the matrix A_{-1}^* for which there exists chains in $\mathcal{N}_i$ not belonging to N_{i-1}. Since for any $m \in \Omega_{\mathcal{N}_i/\mathcal{N}_{i-1}}$ such a chain is unique we can denote its number by $j(m)$.

Using the constructed basis, we obtain a basis $\{\psi\}$ by the relation (2.4). In this basis, the relations (3.16) may be written as follows, and noted as $(\mathbf{R}_i, i = 1, \dots, r)$.

The first family $(\mathbf{R}_1)$ is

$$(\mathbf{R}_1)\qquad k\left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \psi_{m,k}^{j(m),s} \right\rangle = \int_0^T \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,1}(t) u_1(t)\mathrm{d}t + \sum_{d=1}^{r} \int_0^T f_{m,k}^{j(m),s,d}(t) u_d(t)\mathrm{d}t \tag{6.7}$$

for $m \in \Omega_{\mathcal{N}_1},\ s = p_{m,j(m)} - p_{m,j(m)}^{1\prime} + 1, \dots, \pi_{m,j(m)}^1, p_{m,j(m)} - 1,\ |k| > N,\ \pi_{m,j(m)}^1 = p_{m,j(m)}$.

The second family $(\mathbf{R}_2)$ is

$$(\mathbf{R}_2)\qquad \begin{aligned} k\left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \psi_{m,k}^{j(m),s} \right\rangle &= \int_0^T \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,1}(t) u_1(t)\mathrm{d}t \\ &\quad + \int_0^T \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,2}(t) u_2(t)\mathrm{d}t \\ &\quad + \sum_{d=1}^{r} \int_0^T f_{m,k}^{j(m),s,d}(t) u_d(t)\mathrm{d}t \end{aligned} \tag{6.8}$$

for $m \in \Omega_{\mathcal{N}_2/\mathcal{N}_1},\ s = \pi_{m,j(m)}^2 - p_{m,j(m)}^{2\prime} + 1, \dots, \pi_{m,j(m)}^2,\ |k| > N$, where $\pi_{m,j(m)}^2$ and $p_{m,j(m)}^{2\prime}$ are some integer.

The last one $(\mathbf{R}_r)$ being

$$(\mathbf{R}_r)\qquad k\left\langle \begin{pmatrix} y_T \\ z_T(\cdot) \end{pmatrix}, \psi_{m,k}^{j(m),s} \right\rangle = \sum_{d=1}^{r} \int_0^T \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,d}(t) u_d(t)\mathrm{d}t + \sum_{d=1}^{r} \int_0^T f_{m,k}^{j,s,d}(t) u_d(t)\mathrm{d}t \tag{6.9}$$

for $m \in \Omega_{\mathcal{N}_r/\mathcal{N}_{r-1}},\ j = j(m),\ s = \pi_{m,j(m)}^r - p_{m,j(m)}^{r\prime} + 1, \dots, \pi_{m,j(m)}^r,\ |k| > N$, with some integer $\pi_{m,j(m)}^r$ and $p_{m,j(m)}^{r\prime}$.

For each $|k| > N$ the number of equalities $(\mathbf{R}_i, i = 1, \dots, r)$ is exactly m_i^β (see the definition of this number in (6.1)).

Remark 6.5. Let us specify that in the collections $(\mathbf{R}_i, i = 1, \dots, r)$, for each k there exists only one group of quasi polynomials corresponding to the given exponent. Moreover, for each $i, \dots, r$, the quasi polynomials corresponding to $d = i$ have degrees $0, 1, \dots, p^{i\prime}$, as follows from (2.4), (3.14), and Lemma 3.3.

Before solving the problems $(\mathbf{R}_i, i = 1, \dots, r)$ we solve first the same problems with $f_{m,k}^{j,s,d} = 0$, noted $(\mathbf{R}_i^0, i = 1, \dots, r)$.

LEMMA 6.6. *The problems* $(\mathbf{R}_i^0, i = 1, \dots, r)$ *obtained from* $(\mathbf{R}_i, i = 1, \dots, r)$ *by the assumption that* $f_{m,k}^{j,s,d} = 0$ *with* $T > \max\{m_i^\beta, i = 1, \dots, r\}$ *are solvable if and only if the left-hand side is an element of* ℓ_2.

Proof. Consider the problem $(\mathbf{R}_1^0)$ obtained from (6.7) with the assumption $f_{m,k}^{j,s,d} = 0$. This problem is solvable if and only if the left-hand side is ℓ_2 by Theorem 4.5. *If this problem is not solvable, then the problems* $(\mathbf{R}_i^0)$ *are not solvable.* If $(\mathbf{R}_1^0)$ is solvable, then we can find a solution $u_1(t)$. Then, in the problem $(\mathbf{R}_2^0)$ the term

$$\int_0^T \mathrm{e}^{\lambda_m^{(k)} t} q_{m,k}^{j(m),s,1}(t) u_1(t)\mathrm{d}t$$

on the right-hand side is determined and may be moved to the left-hand side. Hence $(\mathbf{R}_2^0)$ is a new moment problem with unknown function $u_2(t)$. It is solvable if and only if the left-hand side is in ℓ_2.

Repeating this argumentation up to $(\mathbf{R}_r^0)$, we obtain that the global problem $(\mathbf{R}_i^0, i = 1, \dots, r)$ is solvable if and only if the right-hand side in (6.7)–(6.9) is in ℓ_2. The proof of Lemma 6.6 is complete. □

Let us now return to the general problem $(\mathbf{R}_i, i = 1, \dots, r)$ given in (6.7)–(6.9). One can represent the equalities $(\mathbf{R}_i, i = 1, \dots, r)$ in the following operator form:

$$x = Q_N u(\cdot) + F_N u(\cdot), \quad x \in \ell_2, \quad u(\cdot) \in L_2(0, T; \mathbb{C}^r),$$

with N the integer for which the problem is considered ($|k| > N$). We shall prove that there exists N sufficiently large such that

$$\operatorname{Im} Q_N = \ell_2 \Longrightarrow \operatorname{Im} [Q_N + F_N] = \ell_2,$$

and the last equality means that $(\mathbf{R}_i, i = 1, \dots, r)$ is solvable if $(\mathbf{R}_i^0, i = 1, \dots, r)$ is solvable, i.e., if the left-hand sides in (6.7)–(6.9) are in ℓ_2.

Suppose that $\operatorname{Im} Q_N = \ell_2$; then there exists a constant $\gamma_N > 0$ such that $\|Q_N^* x\| \geq \gamma_N \|x\|$ for all $x \in \ell_2$ (see, for example, [13, Theorem 4.13]). Let $N > N_0$ and let us denote by ℓ_2^N the Hilbert space $\ell_2(N) = \{s_k, |k| > N : \sum_{|k|N} |s_k|^2 < \infty\}$; then $Q_N = P Q_{N_0}$ where $P : \ell_2^{N_0} \to \ell_2^N$ is the projector defined by

$$P(\{s_k, |k| > N_0\}) = \{s_k, |k| > N\}.$$

Then $Q_N^* = Q_{N_0}^* P^*$ and $\|P^* x\| = \|x\|$. This gives

$$\|Q_N^* x\| = \|Q_{N_0}^* P^* x\| \geq \gamma_{N_0} \|x\|.$$

This means that for all $N > N_0$, $\|Q_N^* x\| \geq \gamma \|x\|$ for all $x \in \ell_2$, where $\gamma = \gamma_0$.

Consider now the operator F_N. By the property (P3) (section 3) we have $\|F_N\| \to 0$ when $N \to \infty$. Hence the norm $\|Q_N - Q_N - F_N\| = \|F_N\|$ can be made arbitrarily small, say $\|F_N\| \leq \frac{\gamma}{2}$. This gives that the operator $Q_N + F_N$ is also surjective because

$$\|[Q_N^* + F_N^*]x\| \geq \|Q_N^* x\| - \|F_N^* x\| \geq \gamma \|x\| - \frac{\gamma}{2} \|x\| = \frac{\gamma}{2} \|x\|.$$

Then from Lemma 6.6 it follows that if $T > \max\{m_i^\beta, i = 1, \dots, r\}$, the moment problem $(\mathbf{R}_i, i = 1, \dots, r), |k| > N$, is solvable for all left-hand sides in ℓ_2.

Applying now Theorem 5.5, we conclude that $\mathcal{R}_T = \mathcal{D}(\mathcal{A})$. The proof of the theorem is complete. □

7. Controllability in the general case. In the previous section, we use the assumption that the system (1.2) is a pure neutral-type system ($\det A_{-1} \neq 0$). However, this condition is in fact a technical assumption that allows the use of the Riesz basis of eigenspaces of the operator $\mathcal{A}$ in M_2 and the moment problem approach.

In this section, we show that conditions (i) and (ii) of Theorem 5.2 are necessary and sufficient for exact controllability for the general neutral systems (A_{-1} may be a singular matrix). We obtain also the precise time of controllability. From Theorem 6.1 it is not clear what happens if the time T is such that $m_1 \leq T \leq \overline{m}$ even if the conditions of controllability are satisfied. In this section, the exact time of controllability is given. In order to do that, we need the classical concept of the controllability indices.

Recall that the first index n_1 may be defined as the minimal integer ν such that (see, for example, [15, Chapter 5])

$$\operatorname{rank}\left(B, A_{-1}B, \ldots, {A_{-1}}^{\nu-1}B\right) = n.$$

LEMMA 7.1. *Assume that the pair $(\mathcal{A}_{-1}, B)$ is controllable. Let n_1 be the index of controllability of the couple (A_{-1}, B) and let $\overline{m}, m_1$ be defined by* (6.2). *Then $m_1 \le n_1 \le \overline{m}$.*

Proof. Let $\beta = \{b_1, \ldots, b_r\}$ be an arbitrary basis of $\operatorname{Im} B$. Then

$$A^{n_1}b_1 \in \operatorname{Im}\left(B \quad AB \quad \cdots \quad A^{n_1-1}B\right) = \operatorname{Im}\left(B \quad AB \quad \cdots \quad A^{n-1}B\right).$$

This may be written as

$$A^{n_1}b_1 \in \operatorname{Lin}\{b_1, Ab_1, \ldots, A^{n_1-1}b_1\} + \operatorname{Im}\left(B_1 \quad AB_1 \quad \cdots \quad A^{n_1-1}B_1\right).$$

This gives that $m_1^\beta \le n_1$ for all β. Hence $m_1 \le n_1$.

Let us now consider the indices $m_1^\beta, \cdots, m_r^\beta$. By the definition of the integers m_i^β we get that

$$\left\{b_1, \ldots, A^{m_1{}^\beta-1}b_1, b_2, \ldots, A^{m_2{}^\beta-1}b_2, \ldots, b_r, \ldots, A^{m_r{}^\beta-1}b_r\right\}$$

is a basis in $\mathbb{C}^n$. This may be verified remarking first that, by definition, the vectors $b_r, \ldots, A^{m_r^\beta-1}b_r$ are linearly independent and

$$\operatorname{Lin}\{b_r, \ldots, A^{m_r^\beta-1}b_r\} = \operatorname{Lin}\{b_r, \ldots, A^{n-1}b_r\}.$$

Then we have to consider the previous step, i.e., m_{r-1}^β, and state that

$$\left\{b_{r-1}, \ldots, A^{m_{r-1}^\beta}b_{r-1}, b_r, \ldots, A^{m_r-1}b_r\right\}$$

are also linearly independent and

$$\operatorname{Lin}\left\{b_{r-1}, \ldots, A^{m_{r-1}^\beta-1}b_{r-1}, b_r, \ldots, A^{m_r{}^\beta-1}b_r\right\} = \operatorname{Lin}\left\{b_{r-1}, \ldots, A^{n-1}b_{r-1}, b_r, \ldots, A^{n-1}b_r\right\}$$

and so on.

We have then $m_1^\beta + \cdots + m_r^\beta = n$ and then $\operatorname{rank}\left(B \quad AB \quad \cdots \quad A^{\overline{m}^\beta-1}B\right) = n$, where $\max\{m_i^\beta, i = 1, \ldots, r\}$. This gives $n_1 \le \overline{m}^\beta$ for all β and hence $n_1 \le \overline{m}$. This completes the proof of the lemma. □

It is well known that in contrast to indices $m_1, \overline{m}$, the controllability index n_1 is invariant under feedback. This means that n_1 is the same for all couples $(A_{-1} + BP, B)$, where P is an $r \times n$ matrix. Then one can choose a feedback matrix P and a basis in $\mathbb{C}^n$ such that $A_{-1} + BP$ take the following form (see [15, Theorem 5.10 and Corollary 5.3]):

$$F = \operatorname{diag}\{F_1, \ldots, F_r\},$$

where

$$F_i = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_1^i & a_2^i & a_3^i & \cdots & a_{n_i}^i \end{pmatrix}$$

and B becomes

$$G = \operatorname{diag}\{g_1, \dots, g_r\},$$

where $g_i = (\,0 \quad 0 \quad \cdots \quad 1\,)^{\mathrm{T}}$, the dimension being $n_i \times 1$. It is easy to check that $\overline{m}(F,G) = m_1(F,G) = n_1$. Moreover, the spectrum of F may be chosen arbitrarily by means of an appropriate choice of P.

Let us now return to the controllability problem for the system (1.2) (or equivalently (1.3)). We first give a preliminary result.

LEMMA 7.2. *The system* (1.2) *is exactly null-controllable at the time T if and only if the perturbed system*

(7.1)
$$\dot z(t) = (A_{-1} + BP)\dot z(t-1) + \int_{-1}^{0} A_2(\theta)\dot z(t+\theta)\mathrm{d}\theta + \int_{-1}^{0} A_3(\theta) z(t+\theta)\mathrm{d}\theta + Bu$$

is exactly null-controllable at the same time T.

Proof. Obviously it is enough to prove one implication only. Assume that the system (1.2) is controllable at the time T. It means that for any function $f(t) \in H^1(T-1, T; \mathbb{C}^n)$ there exists a control $u(t) \in L_2(0, T; \mathbb{C}^n)$ such that the solution of the equation

$$\dot z(t) = A_{-1}\dot z(t-1) + \int_{-1}^{0} A_2(\theta)\dot z(t+\theta)\mathrm{d}\theta + \int_{-1}^{0} A_3(\theta) z(t+\theta)\mathrm{d}\theta + Bu(t), \tag{7.2}$$

with the initial condition $z(t) = 0, t \in [-1, 0]$, verifies $z(t) = f(t), t \in [T-1, T]$. Let us rewrite (7.2) in the form

$$\dot z(t) = (A_{-1} + BP)\dot z(t-1) + \int_{-1}^{0} A_2(\theta)\dot z(t+\theta)\mathrm{d}\theta + \int_{-1}^{0} A_3(\theta) z(t+\theta)\mathrm{d}\theta + Bv(t),$$

where $v(t) = u(t) - P\dot z(t-1), t \in [0, T]$. Since $z(t-1) \in H^1([0, T]; \mathbb{C}^n)$, then $v(t) \in L_2(0, T; \mathbb{C}^n)$. Thus, the control $v(t)$ transfers the state $z(t) = 0, t \in [-1, 0]$, to the state $z(t) = f(t), t \in [T-1, T]$, by virtue of the perturbed system. This means that it is also controllable at the time T. ∎

We have the following result, which concludes our considerations.

THEOREM 7.3. *Let the neutral-type system* (1.2) *be in the general form, i.e., without the assumption* $\det A_1 \neq 0$. *Conditions* (i) *and* (ii) *of Theorem* 5.2 *are necessary and sufficient for the exact controllability of the system. Under these conditions, the precise time of controllability is $T = n_1$. This means that the system is not controllable for $T \leq n_1$ and is controllable for $T > n_1$.*

If the delay is h instead of 1, then the exact time of controllability is $n_1 h$.

Proof. According to Theorem 5.2, the proof of necessity is needed for the case when $\det A_1 = 0$. Let us first show that condition (ii) holds. Assume that (ii) is not verified. Then there exist vectors $z_0 \neq 0$ such that $A_{-1}^* z_0 = \lambda_0 z_0$ and $B^* z_0 = 0$. If for all such vectors $\lambda_0 \neq 0$, then one can find P_0 such that $A_{-1} + BP_0$ is not singular. Then, according to Lemma 7.2, the perturbed system (7.1) with $P = P_0$ is exactly null-controllable. This gives that the pair $(A_{-1} + BP_0, B)$ is controllable, which contradicts the existence of such vectors z_0.

Suppose that for some vector $z_0 \neq 0$, we have $A_{-1}^* z_0 = 0$ and $B^* z_0 = 0$. Then, multiplying (1.2) by z_0 we get

$$\langle \dot{z}(t), z_0 \rangle = \left\langle A_{-1}\dot{z}(t-1) + \int_{-1}^{0} A_2(\theta)\dot{z}(t+\theta)\mathrm{d}\theta + \int_{-1}^{0} A_3(\theta) z(t+\theta)\mathrm{d}\theta + Bu(t), z_0 \right\rangle$$

and the exact null-controllability definition means that this relation holds for an arbitrary function $\langle \dot{z}(t), z_0 \rangle \in L_2(T-1, T)$. As $\langle A_{-1}\dot{z}(t-1), z_0\rangle = 0$ and $\langle Bu(t), z_0\rangle = 0$, this gives, after a change of variables,

$$\begin{aligned}
\langle \dot{z}(t), z_0 \rangle &= \left\langle \int_{-1}^{0} A_2(\theta)\dot{z}(t+\theta)\mathrm{d}\theta + \int_{-1}^{0} A_3(\theta) z(t+\theta)\mathrm{d}\theta, z_0 \right\rangle \\
&= \left\langle \int_{t-1}^{t} A_2(s-t)\dot{z}(s)\mathrm{d}s + \int_{t-1}^{t} A_3(s-t) z(s)\mathrm{d}s, z_0 \right\rangle \\
&= \begin{pmatrix} K_2(z_0) & K_3(z_0) \end{pmatrix} \begin{pmatrix} \dot{z}(\cdot) \\ z(\cdot) \end{pmatrix},
\end{aligned}$$

where $K_j : L_2((T-2, T); \mathbb{C}^n) \longrightarrow L_2(T-1, T), j = 2, 3$, are linear operators defined by

$$\left(K_j(z_0)w\right)(t) = \int_{t-1}^{t} A_j(s-t)w(s)\mathrm{d}s = \int_{T-1}^{T} \widehat{A}_j(s-t)w(s)\mathrm{d}s,$$

and

$$\widehat{A}_j(s) = \begin{cases} A_j(s), & s \in [-1, 0], \\ 0, & s \notin [-1, 0]. \end{cases}$$

The operators $K_j, j = 2, 3$, are clearly compact operators because

$$\int_{T-1}^{T} \left(\int_{T-2}^{T} \|\widehat{A}_j(s-t)\|^2 \mathrm{d}s \right) \mathrm{d}t = \int_{T-1}^{T} \left(\int_{-1}^{0} \|A_j(\theta)\|^2 \mathrm{d}\theta \right) \mathrm{d}t < \infty;$$

see, for example, [7, Chapter 6]. Then the image of the operator $\begin{pmatrix} K_2(z_0) & K_3(z_0) \end{pmatrix}$ cannot coincide with $L_2(T-1, T)$. Thus, such a vector z_0 does not exist. This gives that condition (ii) is necessary.

Let us now prove the necessity of condition (i). If A_{-1} is nonsingular, it is proved in Theorem 5.2. Assume now that A_{-1} is singular. Then (since we have proved (ii)) we can choose a matrix P such that $A_{-1} + BP$ is not singular. According to Lemma 7.2, the perturbed system (7.1) is still exactly null-controllable. Using Theorem 5.2, we have the following statement: there do not exist $\lambda \in \mathbb{C}$ and $y \in \mathbb{C}^n$ such that

$$\left[\lambda I - \lambda \mathrm{e}^{-\lambda} \left(A_{-1}^* + P^* B^* \right) - \lambda \int_{-1}^{0} \mathrm{e}^{\lambda s} A_2^*(s)\mathrm{d}s - \int_{-1}^{0} \mathrm{e}^{\lambda s} A_3^*(s)\mathrm{d}s \right] y = 0$$

and $B^* y = 0$. This gives condition (i).

Now we prove the sufficiency. Assume that conditions (i) and (ii) are verified. Then they are also verified for the perturbed system. From condition (ii) we can choose a matrix P such that $A_{-1} + BP$ is nonsingular and $\overline{m}(A_{-1} + BP, B) = m_1(A_{-1} + BP, B) = n_1$. And this gives that the perturbed system is exactly null-controllable at the time $T > n_1$ and is not controllable at the time $T < n_1$. By Lemma 7.2 we infer that our system (1.2) satisfies the same condition.

Moreover, it is easy to prove, arguing as in the proof of Theorem 5.6, that the system (1.2) is also not controllable at the time $T = n_1$. More precisely, the codimension of $\mathcal{R}_{n_1}$ in X_A is finite and not less than n_1. For $T < n_1$, the codimension of $\mathcal{R}_T$ is infinite. □

8. Conclusion and perspectives. The main goal of this paper is to demonstrate how the moment problem approach can be used in the controllability problem for delay systems of neutral type. To this end, we chose a quite general model (1.2) with distributed delays in the function and its derivative, a pointwise neutral term determined by a matrix A_{-1}, and the control term by a matrix B. Using our approach, we have given a complete analysis of the exact null-controllability for this model. Namely,

(i) we showed that the maximal possible set of the states reachable from 0 by the system at some time $T > 0$ is the space H^1;
(ii) we found the conditions of the parameters of the system under which this set of reachability can be maximally possible (the conditions of exact controllability);
(iii) we proved that, under the above conditions, the system is exactly controllable at the time T if and only if $T > n_1$, where n_1 is the first controllability index of the couple (A_{-1}, B) (the time of exact controllability).

As a perspective, we consider the extension of our approach to systems with several pointwise neutral terms and to the general case of distributed neutral-type delay,

$$Kf = \int_{-1}^{0} \mathrm{d}\mu(\theta) f(\theta), \quad f \in C([-1, 0], \mathbb{C}^n),$$

where μ is a matrix-valued function of bounded variation and continuous at zero. One can prove that, for this class of systems, the generalized Riesz basis property of the model operator $\mathcal{A}$ is preserved. However, the immediate spectral analysis of this operator is more complex. In the case when the delays in the neutral terms are commensurable, the results on exact controllability are expected to be similar to those obtained in the present paper. In the general case, the formulation and the proofs may be much more complicated. This problem is to be considered in our forthcoming works.

REFERENCES

[1] S. A. Avdonin and S. A. Ivanov, *Families of Exponentials. The Method of Moments in Controllability Problems for Distributed Parameter Systems*, Cambridge University Press, Cambridge, UK, 1995.
[2] H. T. Banks, M. Q. Jacobs, and C. E. Langenhop, *Characterization of the controlled states in $W_2^{(1)}$ of linear hereditary systems*, SIAM J. Control, 13 (1975), pp. 611–649.
[3] J. A. Burns, T. L. Herdman, and H. W. Stech, *Linear functional differential equations as semigroups on product spaces,* SIAM J. Math. Anal., 14 (1983), pp. 98–116.
[4] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monogr. 18, AMS, Providence, RI, 1969.
[5] K. Ito and T. J. Tarn, *A linear quadratic optimal control for neutral systems*, Nonlinear Anal., 9 (1985), pp. 699–727.
[6] M. Q. Jacobs and C. E. Langenhop, *Criteria for function space controllability of linear neutral systems*, SIAM J. Control Optim., 14 (1976), pp. 1009–1048.
[7] L. A. Liusternik and V. J. Sobolev, *Elements of Functional Analysis*, Russian Monographs and Texts on Advanced Mathematics and Physics, 5, Hindustan Publishing, Delhi, Gordon and Breach Publishers, New York, 1961.

[8] A. Manitius and R. Triggiani, *Function space controllability of linear retarded systems: A derivation from abstract operator conditions*, SIAM J. Control Optim., 16 (1978), pp. 599–645.
[9] D. A. O'Connor and T. J. Tarn, *On the function space controllability of linear neutral systems*, SIAM J. Control Optim., 21 (1983), pp. 306–329.
[10] R. Rabah, G. M. Sklyar, and A. V. Rezounenko, *Generalized Riesz basis property in the analysis of neutral type systems*, C. R. Math. Acad. Sci. Paris, 337 (2003), pp. 19–24.
[11] R. Rabah, G. M. Sklyar, and A. V. Rezounenko, *Stability analysis of neutral type systems in Hilbert space*, J. Differential Equations, 214 (2005), pp. 391–428.
[12] H. Rivera Rodas and C. E. Langenhop, *A sufficient condition for function space controllability of a linear neutral system*, SIAM J. Control Optim., 16 (1978), pp. 429–435.
[13] W. Rudin, *Functional Analysis*, 2nd ed., McGraw–Hill, New York, 1991.
[14] D. Ullrich, *Divided differences and systems of nonharmonic Fourier series*, Proc. Amer. Math. Soc., 80 (1980), pp. 47–57.
[15] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, Springer, New York, 1985.
[16] Y. Yamamoto, *Reachability of a class of infinite-dimensional systems: An external approach with applications to general neutral systems*, SIAM J. Control Optim., 27 (1989), pp. 217–234.
[17] R. Y. Young, *An Introduction to Nonharmonic Analysis*, Academic Press, New York, 1980.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2182–2198

ON THE GENERICITY OF THE DIFFERENTIAL OBSERVABILITY OF CONTROLLED DISCRETE-TIME SYSTEMS*

SABEUR AMMAR†, MOHAMED MABROUK‡, AND JEAN-CLAUDE VIVALDA§

Abstract. In this paper, we prove the genericity of the differential observability for discrete-time systems with more outputs than inputs.

Key words. observability, nonlinear systems, discrete-time systems, transversality theory

AMS subject classifications. 93B07, 93B29, 93C55

DOI. 10.1137/060677938

1. Introduction.

In this paper, we study the genericity of the differential observability for discrete-time controlled nonlinear systems such that

$$\begin{cases} x_{k+1} = f(x_k, u_k), \\ y_k = h(x_k, u_k), \\ x_k \in X,\ u_k \in U,\ y_k \in \mathbb{R}^p, \end{cases} \tag{1}$$

where

(i) X and U are C^∞ compact connected second-countable manifold with dimensions n and m, respectively;

(ii) $f : X \times U \to X$ is a parametrized diffeomorphism: that is to say, for every $u \in U$, the mapping $f(\cdot, u)$ is a C^∞ diffeomorphism; we denote by $\mathrm{Diff}_{\mathrm{U}}(X)$ the set of all parametrized diffeomorphisms;

(iii) $h : X \times U \to \mathbb{R}^p$ is a C^∞ mapping.

To be more specific, we shall introduce some notations. Given $f \in \mathrm{Diff}_{\mathrm{U}}(X)$ and $h \in C^\infty(X \times U, \mathbb{R}^p)$, we denote by $\underline{u_N}$ the finite sequence $(u_0, \dots, u_{N-1})$ of elements of U, and we define recursively $f^k(x, \underline{u_k})$ by

$$f^1(x, \underline{u_1}) = f(x, u_0),$$

$$f^{k+1}(x, \underline{u_{k+1}}) = f(f^k(x, \underline{u_k}), u_k) \qquad \text{for } k \geq 1.$$

Let us recall the notion of observability investigated in this paper.

DEFINITION 1. *Two initial conditions x_0 and $\bar{x}_0$ and an input u (i.e., a sequence $(u_k)_{k\geq 0}$ of elements of U) being given, x_k and $\bar{x}_k$ denote the points $x_k = f^k(x_0, \underline{u_k})$ and $\bar{x}_k = f^k(\bar{x}_0, \underline{u_k})$.*

System (1) *is said to be observable for input u if for any initial conditions $x_0 \neq \bar{x}_0$, there exists an index k (possibly depending on the initial conditions) such that $x_k \neq \bar{x}_k$.*

*Received by the editors December 18, 2006; accepted for publication (in revised form) June 9, 2007; published electronically December 21, 2007. All of the authors are involved in the PAI Utique-CMCU 05G1507.

http://www.siam.org/journals/sicon/46-6/67793.html

†ISIT. Com de Hammam Sousse, Route principale Numéro 1, 4011 Hammam Sousse, 4002 Sousse, France (ammar_3021@yahoo.fr).

‡Faculté des sciences de Gabès, Cité Riadh, Zirig 6072 Gabès, Tunisie, France (Mohamed.Mabrouk@loria.fr).

§Inria-Lorraine (projet CORIDA) et LMAM (UMR 7122), Université de Metz, Ile du Saulcy, 57 045 Metz Cedex 01, France (vivalda@loria.fr).

System (1) *is said to be observable if it is observable for each input.*

Below, we are introducing a stronger notion of observability. We consider the mapping $\Theta_{2n+1}^{f,h}$ from $X \times U^{2n+1}$ to $\mathbb{R}^{(2n+1)p} \times U^{2n+1}$ defined by

$$\Theta_{2n+1}^{f,h}(x, \underline{u_{2n+1}}) = (h(x,u_0), h(f^1(x,\underline{u_1}),u_1), \ldots, h(f^{2n}(x,\underline{u_{2n}}),u_{2n}), \underline{u_{2n+1}}).$$

Notice that this mapping is the discrete-time analogous of the mapping $S\Phi_k^\Sigma$ defined in [4].

DEFINITION 2. *We shall say that system* (1) *is* strongly observable *if the related mapping* $\Theta_{2n+1}^{f,h}$ *defined above is one-to-one.*

In [3], we proved that system (1) is generically strongly observable as long as $p > \dim U$; more precisely, we proved that the set of pairs (f,h) which make the mapping $\Theta_{2n+1}^{f,h}$ one-to-one is a residual.

In this article, we deal with a stronger notion of observability.

DEFINITION 3. *We shall say that system* (1) *is* strongly differentially observable *if, for every fixed sequence* $\underline{u_{2n+1}}$, *the mapping* $\bar{\Theta}_{2n+1}^{f,h}$ *from* X *to* $\mathbb{R}^{(2n+1)p}$ *defined by*

$$\bar{\Theta}_{2n+1}^{f,h}(x) = (h(x,u_0), h(f^1(x,\underline{u_1}),u_1), \ldots, h(f^{2n}(x,\underline{u_{2n}}),u_{2n}))$$

is an embedding.

In the continuation of [3], the goal of this paper is to prove that system (1) is strongly differentially observable as long as $p > \dim U$.

On this subject, one has to mention first the important work from Gauthier and Kupka. In a first paper, also with Hammouri (see [5]), the authors investigated the genericity of observability for uncontrolled continuous-time systems. This work was generalized by Gauthier and Kupka in [6, 4], where the authors proved the genericity of differential observability for systems with more outputs than inputs. To be more precise, in their paper, the authors consider the set, denoted by $\mathscr{O}$, of systems Σ such that the mapping $S\Phi_\Sigma^N$ (analogous, for continuous time systems, to mapping $\Theta_{2n+1}^{f,h}$) is an embedding; they show that, provided that a bound on the derivative of the control is given, this set is $\mathscr{O}$ open and dense. When this condition is not assumed, the authors prove that this set is residual (and therefore dense) but the openness property remains an open problem. In our case, we assume that the controls belong to a compact manifold, so this difficulty disappears. Also, we do not have to consider the case of nonsmooth controls. Nevertheless, there are some other difficulties; for example, we have to pay special attention to periodic points of f_u. As far as we are concerned by discrete-time systems, we have to cite several papers on the subject of the genericity of the observability: first, a paper written by Aeyels [2] in which the author considers uncontrolled continuous-time systems and the discrete-time systems obtained by discretizing the continuous ones. In [2], the author introduced the notion of P-observability. The system

$$\begin{cases} \dot{x} &= f(x), \\ y &= h(x) \end{cases} \tag{2}$$

is said to be P-observable if, given a time $T > 0$ and a finite subset P of $[0,T]$, for every pair (x,y) of distinct elements in X^2, there exists a $t_i \in P$ such that $h \circ \Phi_{t_i}(x) \neq h \circ \Phi_{t_i}(y)$, where Φ denotes the flow of f. One of the results in this paper is the proof of the existence of an open and dense set of vector fields such that (a vector field f in this set being fixed) the subset of functions h belonging to $C^r(X,\mathbb{R})$

such that the system (f,h) is P-observable is open and dense in $C^r(X,\mathbb{R})$. This is true for almost any finite subset P of $(2\dim X+1)$ points in $[0,T]$.

To an uncontrolled discrete-time system such that

$$\begin{cases} x_{k+1} &= f(x_k), \\ y_k &= h(x_k), \\ x_k \in M, \text{ compact manifold, } y_k \in \mathbb{R}\,, \end{cases} \tag{3}$$

is attached a map analogous to the map $\Theta^{f,h}_{2n+1}$ defined above: consider

$$\begin{aligned} \Phi: \quad M &\longrightarrow \mathbb{R}^{2n+1} \\ x &\longmapsto (h(x), h\circ f(x), \ldots, h\circ f^{2n}(x)), \end{aligned}$$

where n is the dimension of manifold M. In [10], the proof that, generically, Φ is an embedding is sketched, while in [8] and [11] the same result is proved in greater detail (see also the concluding remarks of [2]). In the case of controlled discrete-time systems, in [9], the authors investigate controlled discrete-time systems and obtained some results which are similar (but not identical) to the one presented here; namely, they present a result of genericity of the observability, but it is not a result about observability for every input. As in the present paper, the tools used in the work of these authors belong to the transversality theory.

Before going straight to the point, we want to add some words about the fact that the observation function h depends on u. This situation is not common in automatic control theory, but the opposite assumption leads to clumsy statements. Nevertheless, as explained in the conclusion of [3], the result of genericity can also be proved for systems where h does not depend on u. The paper is organized as follows: in the next section, some facts from transversality theory are recalled; in section 3, the main result is stated together with some definitions and lemmas; in section 4, our result is proved through the demonstration of five lemmas.

2. Some facts from transversality theory. In this section we recall some theorems from differential topology which will be intensively used in the proof of the main result of this paper. For details on the C^∞ Whitney topology, the reader is referred to the book "Stable Mappings and their Singularities" [7].

If X and Y are two smooth manifolds, then $J^k(X,Y)$ will denote, as usual, the set of k-jets from X to Y, $\alpha: J^k(X,Y)\to X$ is the source map, and $\beta: J^k(X,Y)\to Y$ is the target map; moreover, we denote by $C^r(X,Y)$ $(1\le r\le+\infty)$ the set of C^r maps from X to Y. If f is in $C^\infty(X,Y)$, then j^kf denotes the k-jet of f. Recall that the set $C^\infty(X,Y)$ endowed with the Whitney topology is a Baire space and so every residual set of $C^\infty(X,Y)$ (i.e., every countable intersection of open dense subsets) is dense.

The notion of transversality is of paramount importance for our purpose, and we recall its definition below.

Definition 4. *Let f be a smooth mapping between two smooth manifolds X and Y, W a submanifold of Y, and x a point in X. We shall say that f intersects W* transversely *at x if either*

(i) $f(x)\notin W$ *or*

(ii) $f(x)\in W$ *and* $T_{f(x)}Y = T_{f(x)}W + df_x(T_xX)$,

with T_xX denoting the tangent space to X at x and df_x the Jacobian of f at x. We shall say that f intersects W transversely if it intersects W transversely at x for all x in W. We shall use the symbol $\pitchfork$ to denote the transversality.

The following theorem states a result of genericity [7].

THEOREM 1 (Thom transversality theorem). *Let X and Y be smooth manifold, W a submanifold of $J^k(X,Y)$, and let*

$$T_W = \{f \in C^\infty(X,Y) \mid j^k f \pitchfork W\}.$$

Then T_W is a residual subset of $C^\infty(X,Y)$ in the C^∞ topology. Moreover, if W is closed, then T_W is open.

The following result generalizes the above theorem to multijet spaces. We first define the set $X^{(s)} = \{\,(x_1,\dots,x_s) \in X^s \mid x_i \neq x_j \text{ for } 1 \leq i < j \leq s\,\}$ and the mapping

$$\begin{aligned} \alpha^s : \big(J^k(X,Y)\big)^s &\longrightarrow X^s \\ (\sigma_1,\dots,\sigma_s) &\longmapsto \big(\alpha(\sigma_1),\dots,\alpha(\sigma_s)\big), \end{aligned}$$

and we let $J^k_s(X,Y) = (\alpha^s)^{-1}(X^{(s)})$, notice that $J^k_s(X,Y)$ is a submanifold of $(J^k(X,Y))^s$.

For $f \in C^\infty(X,Y)$, we can define

$$\begin{aligned} j^k_s f : \quad X^{(s)} &\longrightarrow J^k_s(X,Y) \\ (x_1,\dots,x_s) &\longmapsto \big(j^k f(x_1),\dots,j^k f(x_s)\big). \end{aligned}$$

THEOREM 2 (multijet transversality theorem). *Let W be a submanifold of $J^k_s(X,Y)$, and let*

$$T_W = \{f \in C^\infty(X,Y) \mid j^k_s f \pitchfork W\}.$$

Then T_W is a residual subset of $C^\infty(X,Y)$ in the C^∞ topology. Moreover, if W is compact, then T_W is open.

We shall use also a transversality theorem due to Abraham (see [1]). Let $\mathcal{A}, X$, and Y be C^r manifolds and ρ a map from $\mathcal{A}$ to $C^r(X,Y)$.

For $a \in \mathcal{A}$, we write ρ_a, the C^r map

$$\begin{aligned} \rho_a : X &\longrightarrow Y \\ x &\longmapsto \rho_a(x) = \rho(a)(x), \end{aligned}$$

and we say that ρ is a C^r representation if the evaluation map

$$\begin{aligned} \mathrm{ev}_\rho : \mathcal{A} \times X &\longrightarrow Y \\ (a,x) &\longmapsto \rho_a(x) = \rho(a)(x) \end{aligned}$$

is a C^r map from $\mathcal{A} \times X$ to Y.

THEOREM 3 (Abraham transversal density theorem). *Let $\mathcal{A}, X, Y$ be C^r manifolds, $\rho : \mathcal{A} \to C^r(X,Y)$ a C^r representation, $W \subset Y$ a submanifold (not necessarily closed), and $\mathrm{ev}_\rho : \mathcal{A} \times X \to Y$ the evaluation map. Define $\mathcal{A}_W \subset \mathcal{A}$ by*

$$\mathcal{A}_W = \{a \in \mathcal{A} \mid \rho_a \pitchfork W\}.$$

Assume that

1. *X has a finite dimension n and W has a finite codimension q in Y;*
2. *$\mathcal{A}$ and X are second countable;*

3. $r > \max(0, n-q)$;
4. $\mathrm{ev}_\rho \pitchfork W$.

Then $\mathcal{A}_W$ is residual in $\mathcal{A}$.

Notice that manifold $\mathcal{A}$ is not necessarily finite dimensional; it may be a Banach space or an open subset of a Banach space.

Finally, we shall need the following theorem that can also be found in [1].

THEOREM 4 (openness of transversal intersection). *Let $\mathcal{A}$, X, and Y be C^r manifolds with X finite dimensional, $W \subset Y$ a closed C^r submanifold, K a compact subset of X, and $\rho : \mathcal{A} \to C^r(X, Y)$ a C^r representation. Then the subset $\mathcal{A}_{KW} \subset \mathcal{A}$ defined by*

$$\mathcal{A}_{KW} = \{a \in \mathcal{A} \mid \rho_a \pitchfork_x W \text{ for } x \in K\}$$

is open.

3. Main result. We state here our main result and some lemmas used in the proof of our theorem. Our framework is the set $\mathrm{Diff}_{\mathrm{U}}(X) \times C^\infty(X \times U, \mathbb{R}^p)$ equipped with the Whitney topology; obviously, $\mathrm{Diff}_{\mathrm{U}}(X)$ is open in $C^\infty(X \times U, X)$ for this topology. In the theorem below, we assume that $\dim U < p$.

THEOREM 5. *The set of mappings $(f, h) \in \mathrm{Diff}_{\mathrm{U}}(X) \times C^\infty(X \times U, \mathbb{R}^p)$ such that the mapping $\bar{\Theta}^{f,h}_{2n+1}$ is an embedding is open and dense in $\mathrm{Diff}_{\mathrm{U}}(X) \times C^\infty(X \times U, \mathbb{R}^p)$ equipped with the Whitney topology.*

We begin by proving the easiest part of this result: the openness of the set of mappings (f, h) such that $\Theta^{f,h}_{2n+1}$ is an embedding.

Proof. Consider the mapping

$$\begin{aligned} \Phi : \quad X \times U^{2n+1} &\longrightarrow C^\infty(X \times U^{2n+1}, (\mathbb{R}^p)^{2n+1} \times U^{2n+1}) \\ (f, \underline{u_{2n+1}}) &\longmapsto (\Theta^{f,h}_{2n+1}, \underline{u_{2n+1}}), \end{aligned}$$

which is obviously continuous for the Whitney topology. Clearly, $\Phi(f, \underline{u_{2n+1}})$ is an embedding iff the mapping $\Theta^{f,h}_{2n+1}(\cdot, \underline{u_{2n+1}})$ is an embedding for every finite sequence $\underline{u_{2n+1}} \in U^{2n+1}$. Now, since X and U are compact manifolds, the set of embeddings from $X \times U^{2n+1}$ to $(\mathbb{R}^p)^{2n+1} \times U^{2n+1}$ is open for the Whitney topology, so, due to the continuity of Φ, the set of mappings $\Theta^{f,h}_{2n+1}(\cdot, \underline{u_{2n+1}})$, which are embeddings for every $\underline{u_{2n+1}}$, is open. ∎

We shall now prove the density part of the theorem. Notice that in the continuous-time case, the set of pairs (f, h) (with f a parametrized vector field) is a Banach space for the C^r topology ($r < +\infty$), but this is not the case for the set of pairs (f, h), where f is a parametrized diffeomorphism. So, it is not possible to copy directly the reasoning of [6]. The proof of this theorem will be somewhat awkward and will be based on several technical lemmas. Before stating these lemmas, we describe below our global strategy.

Suppose that $\mathscr{P}_1(f, h)$ and $\mathscr{P}_2(f, h)$ are two properties depending on $(f, h) \in \mathrm{Diff}_{\mathrm{U}}(X) \times C^\infty(X \times U, \mathbb{R}^p)$ whose conjunction is equivalent to the fact that $\bar{\Theta}^{f,h}_{2n+1}$ is an immersion. In Proposition 1, we shall prove that, for a given $f \in \mathrm{Diff}_{\mathrm{U}}(X)$, a given integer $r \geq 1$, and for every integer l there exists a subset $U^r_l(f)$ of $C^\infty(X \times U, \mathbb{R}^p)$, open and dense for the C^r topology, such that if h belongs to the intersection $\bigcap_{l \geq 0} U^r_l(f)$, then the pair (f, h) satisfies property $\mathscr{P}_1$. Moreover, we shall prove that, for every integer l, the set

$$\mathscr{U}^r_l = \bigcup_{f \in \mathcal{D}_U} \{f\} \times U^r_l(f)$$

($\mathcal{D}_U$ open dense set of $\mathrm{Diff}_{\mathrm{U}}$) is open dense in $\mathrm{Diff}_{\mathrm{U}}(X) \times C^\infty(X \times U, \mathbb{R}^p)$ equipped with the C^r topology. In Proposition 2, we shall prove that the set

$$E = \{\, (f,h) \in \mathrm{Diff}_{\mathrm{U}}(X) \times C^\infty(X \times U, \mathbb{R}^p) \mid \mathscr{P}_2(f,h) \text{ is true} \,\}$$

contains a residual set of $\mathrm{Diff}_{\mathrm{U}}(X) \times C^\infty(X \times U, \mathbb{R}^p)$. Hence, clearly, the set $E \cap (\bigcap_{k \geq 0 r \geq 1} \mathscr{U}_k^r)$ contains a residual set for the C^∞ topology and a pair (f,h) belonging to this set satisfies both properties $\mathscr{P}_1$ and $\mathscr{P}_2$.

We shall give the definition of periodic points before stating our propositions.

Definition 5. *Let $f \in \mathrm{Diff}_{\mathrm{U}}(X)$. We shall say that the point $(x, \underline{u_{2n+1}}) \in X \times U^{2n+1}$ is periodic for f if there exist two different integers $s' < s$ in $\{0, \ldots, 2n\}$ such that $f^{s'}(x, \underline{u_{s'}}) = f^s(x, \underline{u_s})$. If $(x, \underline{u_{2n+1}})$ is a periodic point, then its period is the smallest integer s such that the above equality is satisfied.*

Notations. We denote by $\mathscr{P}_f$ the set of all periodic points of f; obviously, $\mathscr{P}_f$ is a closed subset of $X \times U^{2n+1}$. We denote also by $\mathscr{P}_f^c$ the set complement of $\mathscr{P}_f$: $\mathscr{P}_f^c = X \times U^{2n+1} \setminus \mathscr{P}_f$.

First, we want to state a lemma about a property of continuity of the sets of periodic points; before that, we recall the definition of the Hausdorff distance between sets.

Definition 6. *Let (E,d) be a metric space, if A and B are subsets of E, then the Hausdorff distance between A and B is defined by*

$$\delta(A,B) = \sup_{x \in A} d(x,B) + \sup_{y \in B} d(y,A).$$

We suppose that X and U are equipped with distances which are compatible with their topologies, so we can speak of the Hausdorff distance on $X \times U^{2n+1}$, and we state the following lemma.

Lemma 1. *There exists an open and dense set in $\mathrm{Diff}_{\mathrm{U}}(X)$, denoted by $\mathcal{D}_{\mathrm{U}}$, such that for each $f \in \mathcal{D}_{\mathrm{U}}$,*

(i) *if $\mathscr{P}_f = \varnothing$, then $\mathscr{P}_g = \varnothing$ for every g in some neighborhood of f;*

(ii) *if $\mathscr{P}_f \neq \varnothing$, then $\delta(\mathscr{P}_f, \mathscr{P}_g)$ tends to 0 as g tends to f for the C^∞ topology.*

Property $\mathscr{P}_1(f,h)$ is related to the periodic points of f and is the object of the following proposition.

Proposition 1. *Let $f \in \mathcal{D}_{\mathrm{U}}$ be given. For each $r > 0$ there exists a sequence $(U_l^r(f))_{l \geq 1}$ of open and dense sets for the C^r topology included in $C^\infty(X \times U, \mathbb{R}^p)$ such that for every mapping h in $\cap_{l \geq 1} U_l^r(f)$, the mapping $\bar{\Theta}_{2n+1}^{f,h}$ is an immersion at each point of $\mathscr{P}_f^c$.*

Moreover, for every nonzero integer l, the set

$$\mathscr{U}_l^r = \bigcup_{f \in \mathcal{D}_{\mathrm{U}}} \{f\} \times U_l^r(f)$$

is open and dense in $\mathrm{Diff}_{\mathrm{U}}(X) \times C^\infty(X \times U, \mathbb{R}^p)$ for the C^r topology.

The second proposition is concerned with property $\mathscr{P}_2(f,h)$; before stating it, we introduce some sets of covectors. We denote by π the canonical projection from T^*X, the cotangent bundle of X to X, and, given an integer $k > n$, we define the set $(T^*X)^{\otimes k}$ by

$$(T^*X)^{\otimes k} = \{\, (p_1, \ldots, p_k) \in (T^*X)^k \mid \pi(p_1) = \cdots = \pi(p_k) \,\}$$

and the set $V(k, T^*X)$ by

$$V(k, T^*X) = \{\, (p_1, \ldots, p_k) \in (T^*X)^{\otimes k} \mid \operatorname{rank}(p_1, \ldots, p_k) < n \,\}.$$

Clearly, $(T^*X)^{\otimes k}$ is a submanifold of $(T^*X)^k$ and $V(k, T^*X)$ is a finite union of submanifolds of $(T^*X)^{\otimes k}$ whose codimension (the codimension of the highest dimensional submanifold of the union) is equal to $k - n + 1$ (see [5]).

We state now our second proposition.

PROPOSITION 2. *The set of pairs* $(f, h) \in \mathrm{Diff}_{\mathrm{U}}(X) \times C^\infty(X \times U, \mathbb{R}^p)$ *such that the mapping* $\bar{\Theta}^{f,h}_{2n+1}$ *is an immersion at each point of* $\mathscr{P}_f$ *is residual.*

Notation. If p is a point of a manifold M, then, hereafter, we shall denote by T_pM the tangent space to M at p.

4. Proof of the main result.

4.1. Proof of Lemma 1. For the proof of this result, we need the following lemma.

LEMMA 2. *There exists a residual subset, denoted by* $\mathcal{R}$, *of* $\mathrm{Diff}_{\mathrm{U}}(X)$ *such that if* f *is in this subset, then* $\mathscr{P}_f$ *is either the empty set or a finite union of submanifolds of* $X \times U^{2n+1}$ *of codimension greater than or equal to* n.

Proof. Let f be in $\mathrm{Diff}_{\mathrm{U}}(X)$ and s be a positive integer less than or equal to $2n$. Consider the following mapping:

$$\begin{aligned} j^0_s f : \quad (X \times U)^{(s)} \quad &\longrightarrow \quad J^0_s(X \times U, X) \\ ((x_0, u_0), \ldots, (x_{s-1}, u_{s-1})) \quad &\longmapsto \quad \big((x_0, u_0, f^1(x_0, \underline{u_1})), \ldots, \\ &\qquad\qquad (x_{s-1}, u_{s-1}, f^1(x_0, \underline{u_s}))\big). \end{aligned}$$

Let s' be a nonnegative integer less than s, and, in $(X \times U \times X)^s$, consider the submanifold $W_{s',s}$ defined as

$$W_{s',s} = \{\, ((x_0, u_0, z_0), \ldots, (x_{s-1}, u_{s-1}, z_{s-1})) \mid x_i = z_{i-1} \text{ for } i = 1, \ldots, s-1 \text{ and } z_{s-1} = z_{s'-1} \,\}.$$

Notice that the codimension of $W_{s',s}$ is equal to sn. By applying the multijet transversality theorem, we can assert that the set of diffeomorphisms f in $\mathrm{Diff}_{\mathrm{U}}(X)$ transverse to $W_{s',s}$ is a residual, so, generically, $V_{s',s} = (j^0_s f)^{-1}(W_{s',s})$ is either empty or a submanifold of $(X \times U)^{(s)}$ (and also of $(X \times U)^s$) of codimension $s\,n$. Denoting by π the projection from $(X \times U)^s$ to $X \times U^s$, it follows that $\pi(V_{s',s})$ is either empty or a submanifold of codimension greater or equal to $s\,n - (s-1)n = n$ from which we deduce that generically the set of periodic points of f with period equal to s is either empty or a finite union of submanifold of $X \times U^{2n+1}$ of codimensions greater than or equal to n. □

We have now proved that, generically, $\mathscr{P}_f$ is not equal to the whole space $X \times U^{2n+1}$. We are now ready to prove Lemma 1; hereafter, we shall denote by f_u the diffeomorphism $x \mapsto f(x, u)$, the control u being fixed.

For the existence of $\mathcal{D}_U$, we shall prove that, in fact, the set $\mathscr{R}$ of the above lemma is open. Take f in $\mathcal{R}$ and suppose that $\mathscr{P}_f$ is empty. Let d be a distance on X which is compatible with the topology of X. For every pair (s', s) of integers such that $0 \le s' < s \le 2n$, we define $\beta_{s',s}$ as

$$\beta_{s',s} = \inf\{\, d(f^{s'}(x, \underline{u_{s'}}), f^s(x, \underline{u_s})) \mid (x, \underline{u_{2n+1}}) \in X \times U^{2n+1} \,\}.$$

Due to the compactness of $X \times U^{2n+1}$, the real numbers $\beta_{s',s}$ are all positive; otherwise there would exist $(x, \underline{u_{2n+1}})$ such that $f^{s'}(x, \underline{u_{s'}}) = f^s(x, \underline{u_s})$, which contradicts the emptiness of $\mathscr{P}_f$. Take now $\varepsilon > 0$ and consider a neighborhood $\mathscr{V}$ of f such that if g is in this neighborhood, then $d(f^k(x, \underline{u_k}), g^k(x, \underline{u_k})) < \varepsilon$ for all $(x, \underline{u_{2n+1}}) \in X \times U^{2n+1}$ and all $k = 1, \ldots, 2n$. If g is in $\mathscr{V}$, then we have

$$\begin{aligned} d(f^{s'}(x, \underline{u_{s'}}), f^s(x, \underline{u_s})) &\leq d(f^{s'}(x, \underline{u_{s'}}), g^{s'}(x, \underline{u_{s'}})) + d(g^{s'}(x, \underline{u_{s'}}), g^s(x, \underline{u_s})) \\ &\quad + d(g^s(x, \underline{u_s}), f^s(x, \underline{u_s}) \\ &\leq 2\varepsilon + d(g^{s'}(x, \underline{u_{s'}}), g^s(x, \underline{u_s})). \end{aligned}$$

This inequality implies that

$$\beta_{s',s} \leq 2\varepsilon + d(g^{s'}(x, \underline{u_{s'}}), g^s(x, \underline{u_s})) \tag{4}$$

for all $(x, \underline{u_{2n+1}})$ in $X \times U^{2n+1}$ and all pair (s', s). So if $(x, \underline{u_{2n+1}})$ is a periodic point for g, there exists a pair (s', s) such that $d(g^{s'}(x, \underline{u_{s'}}), g^s(x, \underline{u_s})) = 0$, and from (4), we then have

$$\beta_{s',s} \leq 2\varepsilon.$$

If ε is chosen small enough, then this last inequality is not possible and so a diffeomorphism g belonging to $\mathscr{V}$ cannot have periodic points.

Suppose now that f is such that $\mathscr{P}_f$ is nonempty. Then we shall prove the existence of a neighborhood $\mathscr{V}$ of f such that if g is in $\mathscr{V}$, then $j^0_s g$ is transverse to the submanifolds W'_s, s defined in the proof of Lemma 2. We reason by contradiction; for every positive integer k, take the neighborhood $\mathscr{V}_k$ of f constituted by the mappings g such that $d(f(x,u), g(x,u)) < 1/k$ and $d'(j^1 f(x,u), j^1 g(x,u)) < 1/k$ (d' denotes a distance on $J^1(X \times U, X)$). Suppose the existence of a sequence of diffeomorphisms $g_k \in \mathscr{V}_k$ and elements $v_k \in X \times U^{2n+1}$ such that $v_k = (x_k, \underline{u^k_{2n+1}})$ is a periodic point (with period s_k) of g_k at which the mapping $j^0_{s_k} g_k$ is not transverse to the submanifold $W_{s'_k, s_k}$ defined in the proof of the above lemma. As $X \times U^{2n+1}$ is compact, we can assume that the sequence $(v_k)_{k \geq 1}$ is convergent with limit v and can also assume that s'_k and s_k do not depend on k. First, we shall see that $v = (x, \underline{u_{2n+1}})$ is a periodic point for f; we have

$$\begin{aligned} d(f^{s'}(x, \underline{u_{s'}}), f^s(x, \underline{u_s})) &\leq d(f^{s'}(x, \underline{u_{s'}}), f^{s'}(x_k, \underline{u^k_{s'}})) + d(f^{s'}(x_k, \underline{u^k_{s'}}), g^{s'}_k(x_k, \underline{u^k_{s'}})) \\ &\quad + d(g^{s'}_k(x_k, \underline{u^k_{s'}}), g^s_k(x_k, \underline{u^k_s})) + d(g^s_k(x_k, \underline{u^k_s}), f^s(x_k, \underline{u^k_s})) \\ &\quad + d(f^s(x_k, \underline{u^k_s}), f^s(x, \underline{u_s})) \\ &\leq 2/k + d(f^{s'}(x, \underline{u_{s'}}), f^{s'}(x_k, \underline{u^k_{s'}})) + d(f^s(x_k, \underline{u^k_s}), f^s(x, \underline{u_s})). \end{aligned}$$

So, due to the continuity of f, for every $\varepsilon > 0$, we have

$$d(f^{s'}(x, \underline{u_{s'}}), f^s(x, \underline{u_s})) \leq 2/k + \varepsilon,$$

provided that k is chosen large enough. This proves that $f^{s'}(x\underline{u_{s'}}) = f^s(x, \underline{u_s})$. Now at point v, $j^0 f$ is transverse to $W_{s'_n s}$ and, if k is large enough, this is still true for g_k

at point v_k: a contradiction. We proved the existence of an open and dense set $\mathcal{D}_U$ such that if f belongs to this set, $\mathscr{P}_f$ is either empty or a finite union of submanifolds of codimensions at least n.

Now take f in $\mathcal{D}_U$ and assume that $\mathscr{P}_f$ is nonempty; for every $\varepsilon > 0$, due to the compactness of set $\mathscr{P}_f$, we can cover it with a finite union of open balls: $\mathscr{P}_f \subset \cup_{i=1}^{N} B(v_i, \varepsilon)$, the v_i being elements of $\mathscr{P}_f$. Let K be the complement in $X \times U^{2n+1}$ of this union of open balls, as $\mathscr{P}_f$ is a finite union of n-dimensional submanifolds, and K is nonempty if ε is small enough; in this case, we define the numbers $\beta_{s',s}$ as follows:

$$\beta_{s',s} = \inf\{\, d(f^{s'}(x, \underline{u_{s'}}), f^{s}(x, \underline{u_s})) \mid (x, \underline{u_{2n+1}}) \in K \,\}.$$

Obviously, due to the compactness of K, all the $\beta_{s',s}$'s are positive, and, if g belongs to a sufficiently small neighborhood of f, by reasoning as above, we can obtain the inequality (4) for all $(x, \underline{u_{2n+1}}) \in K$. Consequently, if g belongs to a sufficiently small neighborhood of f, then the periodic points of g cannot belong to K and are all located in the union $\cup B(v_i, \varepsilon)$, which implies that the distance $d(v, \mathscr{P}_f)$ is less than ε for every v in $\mathscr{P}_g$.

The proof that, for every $w \in \mathscr{P}_f$, the distance $d(w, \mathscr{P}_g)$ can be made arbitrarily small for g in a sufficiently small neighborhood of f is a little harder. Take f in $\mathcal{D}_U$ and let $(x_0, \underline{u^0_{2n+1}})$ be a periodic point of f such that $f^s(x_0, \underline{u^0_s}) = f^{s'}(x_0, \underline{u^0_{s'}})$; then we can regard $f^{s'}(x_0, \underline{u^0_{s'}})$ as a fixed point of the map $f_{u^0_{s-1}} \circ \cdots \circ f_{u^0_{s'}}$ and, for $i = s', \ldots, s-1$, we shall prove the existence of functions $x \mapsto u_i(x)$, defined in a neighborhood of $f^{s'}(x_0, \underline{u^0_{s'}})$, such that the function $f^{s',s}(x)$ defined as

$$f^{s',s}(x) = f(f \ldots (f(x, u_{s'}(x)), u_{s'+1}(x)), \ldots, u_{s-1}(x))$$

has a fixed point at $f^{s'}(x_0, \underline{u^0_{s'}})$ and is transverse to $\Delta X \triangleq \{\, (x, x) \mid x \in X \,\}$ at this fixed point. To this end, we exploit the fact that the mapping $j^0_s f$ is transverse to $W_{s',s}$ at $w = ((x_0, u^0_0), \ldots, (x^0_{s-1}, u^0_{s-1}))$ (where $x^0_i = f(x^0_{i-1}, u^0_{i-1}) = f^i(x_0, \underline{u_i})$ for $i = 1, \ldots, s-1$). Given a tangent vector of $(X \times U \times X)^s$ at $j^0_s f(w)$ represented by the finite sequence $(l_0, m_0, l'_0), \ldots, (l_{s-1}, m_{s-1}, l'_{s-1})$, where $l_i \in \mathrm{T}_{x_i} X$, $m_i \in \mathrm{T}_{u^0_i} U$, and $l'_i \in \mathrm{T}_{x_{i+1}} X$, consider the two following sequences:

(i) $(t_0, \mu_0), \ldots, (t_{s-1}, \mu_{s-1})$ with $t_i \in \mathrm{T}_{x_i} X$ $\mu_i \in \mathrm{T}_{u_i} U$, which represents a tangent vector of $(X \times U)^{(s)}$ at w;

(ii) and $(\bar{t}_0, \bar{\mu}_0, \bar{t}_1), (\bar{t}_1, \bar{\mu}_1, \bar{t}_2) \ldots, (\bar{t}_{s-1}, \bar{\mu}_{s-1}, \bar{t}_s)$ with $\bar{t}_i \in \mathrm{T}_{x_i} X$, $\bar{\mu}_i \in \mathrm{T}_{u_i} U$, and $\bar{t}_s = \bar{t}_{s'}$, which represents a tangent vector of $W_{s',s}$ at $j^0_s f(w)$. The union of these two sequences gives a solution of the following system:

$$\left\{\begin{array}{rl} t_0 + \bar{t}_0 = l_0, \\ \mu_0 + \bar{\mu}_0 = m_0, \\ A_0 t_0 + B_0 \mu_0 + \bar{t}_1 = l'_0, \\ \vdots \qquad\qquad \vdots \\ t_{s-1} + \bar{t}_{s-1} = l_{s-1}, \\ \mu_{s-1} + \bar{\mu}_{s-1} = m_{s-1}, \\ A_{s-1} t_{s-1} + B_{s-1} \mu_{s-1} + \bar{t}_s = l'_{s-1}, \end{array}\right. \tag{5}$$

where $\bar{t}_s = \bar{t}_{s'}$, and A_i denotes the partial derivative with respect to x of f at point

(x_i^0, u_i^0) and B_i the partial derivative with respect to u of f at the same point. Manipulating the equalities of system (5), we get the formula

$$A_{s-1}A_{s-2}\cdots A_{s'}\cdot t_{s'} + \sum_{i=1}^{s-s'}\Big(\prod_{j=1}^{i-1} A_{s-j}\Big)B_{s-i}\cdot\mu_{s-i}$$

$$= l'_{s-1} + \sum_{i=2}^{s-s'}\Big(\prod_{j=1}^{i-1} A_{s-j}\Big)\cdot l'_{s-i} - \sum_{i=1}^{s-s'-1}\Big(\prod_{j=1}^{i} A_{s-j}\Big)\cdot l_{s-i} - \bar{t}_s.$$

Taking into account that $\bar{t}_s = \bar{t}_{s'} = l_{s'} - t_{s'}$, we get

$$A_{s-1}A_{s-2}\cdots A_{s'}\cdot t_{s'} - t_{s'} + \sum_{i=1}^{s-s'}\Big(\prod_{j=1}^{i-1} A_{s-j}\Big)B_{s-i}\cdot\mu_{s-i}$$

$$= l'_{s-1} + \sum_{i=2}^{s-s'}\Big(\prod_{j=1}^{i-1} A_{s-j}\Big)\cdot l'_{s-i} - \sum_{i=1}^{s-s'-1}\Big(\prod_{j=1}^{i} A_{s-j}\Big)\cdot l_{s-i} - l_{s'}. \tag{6}$$

Consider now the mapping $f^{s',s}(x)$ with the smooth mappings $x \mapsto u_i(x)$ chosen in such a way that $u_i(x_{s'}^0) = u_i^0$, the derivative of this mapping at $x_{s'}^0$, is given by the formula

$$df^{s',s}(x_{s'}^0) = A_{s-1}A_{s-2}\cdots A_{s'} + \sum_{i=1}^{s-s'}\Big(\prod_{j=1}^{s-s'-i} A_{s-j}\Big)B_{s'+i-1}K_{s'+i-1},$$

where the K_j's denote the derivatives of the mappings $u_j(x)$ at point $x_{s'}^0$. Now, as is well known [7], the mapping $f^{s',s}$ is transverse to ΔX at point $x_{s'}^0$ iff the following equation of unknown t

$$df^{s',s}(x_{s'}^0).t - t = l \tag{7}$$

has a solution for all vectors l belonging to the tangent space to X at $x_{s'}^0$. If the linear mapping $A_{s-1}\cdots A_{s'}$ does not admit 1 as an eigenvalue, it suffices to choose the u_i's such that their derivatives vanish at $x_{s'}^0$ and, in this case, (7) will have a solution for every t.

In the case where $A_{s-1}\cdots A_{s'}$ does not have this property, we can suppose that $t_{s'}$ is nonzero. As a matter of fact take $\tau_{s'} \neq 0$ such that $A_{s-1}\cdots A_{s'}\cdot\tau_{s'} = \tau_{s'}$ and define recursively the τ_i's by $\tau_{i+1} = A_i\tau_i$ for $i = s', \ldots, s-1$ and $\tau_i = A_i^{-1}\tau_{i+1}$ for $i = 0, \ldots, s'-1$; obviously, we have $\bar{t}_s - \tau_s = \bar{t}_{s'} - \tau_{s'}$ and, if we replace t_i by $t_i + \tau_i$ $(i = 0, \ldots, s-1)$ and $\bar{t}_i$ by $\bar{t}_i - \tau_i$ $(i = 0, \ldots, s)$, then equalities (5) remain satisfied and, obviously, we have $t_{s'} \neq 0$ or $t_{s'} + \tau_{s'} \neq 0$. Since $t_{s'}$ can be assumed to be nonzero, it is possible to design the mappings $x \mapsto u_i(x)$ in such a way that $K_j\cdot t_{s'} = \mu_j$, as the right-hand member of (6) can be chosen arbitrarily; we can see that (7) admits always a solution. We can reformulate this property of transversality of $f^{s',s}$ by saying that the submanifold V_f, constituted by the points $(x, f^{s',s}(x))$ and locally defined around the point $x_{s'}^0$, intersects ΔX at $(x_{s'}^0, x_{s'}^0)$ and is transverse to ΔX at this point. Now, if we take a mapping g close to f, then $g^{s',s}$ will be close to $f^{s',s}$ and the submanifold V_g constituted by the points $(x, g^{s',s}(x))$ will be close from V_f, and since V_f, V_g, and ΔX are n-dimensional submanifolds of the manifold $X \times X$

of dimension $2n$ since the intersection between V_f and ΔX is nonempty, the same is true for the intersection $V_g \cap \Delta X$ provided that V_g is close enough to V_f; moreover, the intersection points in $V_g \cap \Delta X$ will be close to $x^0_{s'}$. Now let $\bar{x}^0_{s'}$ be a point in the intersection $V_g \cap \Delta X$, a fixed point of $g^{s',s}$, and be regarded as the image of a point $(\bar{x}_0, \underline{\bar{u}_{s'}})$ by the mapping $g^{s'}$. Now this last point is closed to $(x_0, \underline{u_{s'}})$ if g is close to f and is a periodic point of g.

4.2. Proof of Proposition 1. We denote by h_u the mapping $x \mapsto h(x,u)$ and by $df_u(x)$ (resp., $dh_u(x)$) the derivative of f_u (resp., h_u) at x; subsequently, we shall regard the p components of $dh_u(x)$ as covectors of $T^*_x X$.

Consider the representation ρ from $C^\infty(X \times U, \mathbb{R}^p)$ to $C^\infty(\mathscr{P}^c_f, (T^*X)^{\otimes(2n+1)p})$ defined through the following evaluation map ev_ρ:

$$\begin{aligned} \mathrm{ev}_\rho : \quad C^\infty(X \times U, \mathbb{R}^p) \times \mathscr{P}^c_f \quad &\longrightarrow \quad (T^*X)^{\otimes(2n+1)p} \\ (h, x, \underline{u_{2n+1}}) \qquad &\longmapsto \quad \big(dh_{u_0}(x), d(h_{u_1} \circ f_{u_0})(x), \dots, \\ &\qquad\qquad d(h_{u_{2n}} \circ f_{u_{2n}} \circ \dots \circ f_{u_0})(x)\big). \end{aligned}$$

We shall prove the existence of a residual set in $C^\infty(X \times U, \mathbb{R}^p)$ such that if h belongs to this set, then ρ_h is transverse to $V((2n+1)p, T^*X)$; we shall do this thanks to the Abraham transversal density theorem with $\mathcal{A} = C^\infty(X \times U, \mathbb{R}^p)$, endowed with the C^r topology ($\mathcal{A}$ is a Banach space), $X = \mathscr{P}^c_f$, $Y = (T^*X)^{\otimes(2n+1)p}$, and $W = V((2n+1)p, T^*X)$ (notice that W is closed).

Clearly, the first three hypotheses of this theorem are satisfied for every r large enough and to prove that $\mathrm{ev}_\rho \pitchfork W$, it is sufficient to prove that the evaluation map is a submersion. The point $(x_0, \underline{u_{2n+1}})$ being given, the mapping $\mathrm{ev}_\rho(\cdot, x_0, \underline{u_{2n+1}})$ from $C^\infty(X \times U, \mathbb{R}^p)$ to $(T^*M)^{\otimes(2n+1)p}$ is linear and so is equal to its derivative; hence, in order to prove that ev_ρ is a submersion it is sufficient to prove that for every $(p_0, \dots, p_{2n}) \in (T^*M)^{\otimes(2n+1)p}$ there exists $h \in C^\infty(X \times U, \mathbb{R}^p)$ such that

$$\text{(8)} \qquad \begin{cases} p_0 = dh_{u_0}(x_0), \\ p_i = d(h_{u_i} \circ f_{u_{i-1}} \circ \dots \circ f_{u_0})(x_0) \qquad \text{for } i = 1, \dots, 2n. \end{cases}$$

But, letting $x_i = f^i(x_0, \underline{u_i})$, $d(h_{u_i} \circ f_{u_{i-1}} \circ \dots \circ f_{u_0})(x_0) = dh_{u_i}(x_i) \circ d(f_{u_{i-1}} \circ \dots \circ f_{u_0})(x_0)$, the pairs (x_i, u_i) being mutually different, it is always possible to find $h \in C^\infty(X \times U, \mathbb{R}^p)$ such that the relations (8) are satisfied.

So we can apply the Abraham transversal density theorem: the set of mappings h in $C^\infty(X \times U, \mathbb{R}^p)$ such that the mapping ρ_h is transverse to $V((2n+1)p, T^*X)$ is a residual set denoted by $\mathscr{R}_r(f)$; now the intersection of the $\mathscr{R}_r(f)$'s gives a set $\mathscr{R}(f)$ which is residual for the C^∞ topology. Now notice that the codimension of $V((2n+1)p, T^*X)$ is greater than or equal to $(2n+1)p - n + 1$, which is greater than $n + (2n+1)m$, the dimension of $\mathscr{P}^c_f$, so saying that ρ_h is transverse to $V((2n+1)p, T^*X)$ is equivalent to saying that the range of ρ_h does not intersect $V((2n+1)p, T^*X)$, which is equivalent to saying that $\bar{\Theta}^{f,h}_{2n+1}$ is an immersion at each point of $\mathscr{P}^c_f$. At this stage, we have proven the existence of a residual set $\mathscr{R}(f)$ included in $C^\infty(X \times U, \mathbb{R}^p)$ such that for every h in $\mathscr{R}(f)$ the mapping $\bar{\Theta}^{f,h}_{2n+1}$ is an immersion at each point of $\mathscr{P}_f$.

For the sake of readability, we shall denote by the same letter, d, distances defined on X or on $X \times U$ which are compatible with the topologies of these spaces.

Given $f \in \mathcal{D}_U$, for every positive integer l consider the compact $K_l(f)$ defined by

$$K_l(f) = \begin{cases} X \times U^{2n+1} & \text{if } \mathscr{P}_f \text{ is empty,} \\ \{\, v \in X \times U^{2n+1} \mid d(v, \mathscr{P}_f) \geq 1/l \,\} & \text{if } \mathscr{P}_f \neq \varnothing. \end{cases}$$

Endow $C^\infty(X \times U, X)$ with the C^r topology and consider the sets $U_k^r(f)$ defined as

$$U_l^r(f) = \{\, h \in C^\infty(X \times U, \mathbb{R}^p) \mid \rho_h \pitchfork_x W \text{ for } x \in K_l(f) \,\}.$$

Using Theorem 4, we can see that $U_l^r(f)$ is open for the C^r topology, and, since $\mathscr{R}_r(f)$ is obviously included in $U_l^r(f)$, it is also dense. Proving that the set

$$\mathscr{U}_l^r = \bigcup_{f \in \mathcal{D}_U} \{f\} \times U_l^r(f)$$

is open is quite a delicate task.

We first prove that, given $\varepsilon > 0$ and $f_0 \in \mathcal{D}_U$, there exists a neighborhood $\mathscr{V}_{f_0}$ of f_0 such that if f belongs to this neighborhood, then for all v in $K_l(f)$, we have $d(v, K_l(f_0)) < \varepsilon$. To prove this point, we will reason by contradiction. We take neighborhoods (in the C^0 topology) of f_0 under the form

$$\mathscr{V}_n = \{\, f \in \mathcal{D}_U \mid d(f_0(x,u), f(x,u)) < 1/n \text{ for all } (x,u) \in X \times U \,\}.$$

Assume the existence of a positive real number ε_0 such that, for every positive integer n, there exist $f_n \in \mathscr{V}_n$ and $v_n \in K_l(f_n)$ such that $d(v_n, K_l(f_0)) > \varepsilon_0$. For each v_n let $\bar{v}_n \in \mathscr{P}_{f_0}$ such that $d(v_n, \bar{v}_n) = d(v_n, \mathscr{P}_{f_0})$. Due to the compactness of $X \times U$, we can suppose that all of these sequences are convergent, and consequently, the sequence $d_n \triangleq d(v_n, \mathscr{P}_{f_0})$ is convergent. Assume that $\lim_{n\to\infty} d_n < 1/l$; then there exists an $\alpha > 0$ such that $d_n < 1/l - \alpha$ for every positive n. In this case, there exist w_n in $\mathscr{P}_{f_n}$ such that $d(\bar{v}_n, w_n)$ tends to 0 as n tends to infinity and we have

$$\begin{aligned} d(v_n, w_n) &\leq d(v_n, \bar{v}_n) + d(\bar{v}_n, w_n) \\ &< 1/l - \alpha + d(\bar{v}_n, w_n) \\ &< 1/l - \alpha/2 \text{ for all n large enough.} \end{aligned}$$

This last inequality is in contradiction with the appurtenance of v_n to $K_l(f_n)$.

In the case where $\lim_{n\to\infty} d_n = 1/l$, let v be the limit of v_n. We then have $d(v, \mathscr{P}_{f_0}) = 1/l$ and so v is in $K_l(f_0)$; therefore, $d(v_n, K_l(f_0)) \leq d(v_n, v)$ and this last quantity becomes less than ε_0 if n is large enough: a contradiction.

Now, let (f_0, h_0) be in $\mathscr{U}_l^r$. Then take a neighborhood $\mathscr{V}_{f_0}$ of f_0 for the C^r topology such that if f belongs to this neighborhood, then the distance $d(v, K_l(f_0))$ is less than ε for every v in $K_l(f)$. If ε is chosen small enough, then there exists a neighborhood $\mathscr{W}_{h_0}$ of h_0 such that if the pair (f,h) belongs to $\mathscr{V}_{f_0} \times \mathscr{W}_{h_0}$, then the representation ρ_h^f related to the pair (f,h) is such that its evaluation map is a submersion at every point of $K_l(f)$.

4.3. Proof of Proposition 2. In order to prove this result, we need three lemmas. Let x_0 be a periodic point of order $s \leq 2n$, that is to say there exists $s' < s$ such that $f^{s'}(x_0, \underline{u_{s'}}) = f^s(x_0, \underline{u_s})$ and $f^i(x_0, \underline{u_i}) \neq f^j(x_0, \underline{u_j})$ if $i, j < s$; we denote

by x_i the iterated of x_0 by f, to be more precise, $x_i = f^i(x_0, \underline{u_i})$, and we also put $z_i, = f(x_i, u_i)$ and $y_i = h(x_i, u_i)$. We consider the list L,

$$(x_0, u_0, z_0, y_0), \ldots, (x_{2n}, u_{2n}, z_{2n}, y_{2n}),$$

and we say that two elements (x_i, u_i, z_i, y_i) and (x_j, u_j, z_j, y_j) are equivalent if $(x_i, u_i) = (x_j, u_j)$. In each equivalence class, we retain the term of the least index and obtain the following list L' extracted from L:

$$(x_{i_0}, u_{i_0}, z_{i_0}, y_{i_0}), \ldots, (x_{i_r}, u_{i_r}, z_{i_r}, y_{i_r})$$

with $i_0 < i_1 < \cdots < i_r$ (necessarily $i_0 = 0$). We then claim the following lemma.

LEMMA 3. *In the list L' above, we can find $r + 1$ equalities between the terms x_i and z_i.*

Proof. Let j be an index less than r. We consider two cases:

1. If $i_{j+1} = i_j + 1$, then we have $z_{i_j} = x_{i_j+1} = x_{i_{j+1}}$;
2. if $i_{j+1} > i_j + 1$, then the term of index $i_j + 1$ was removed from list L because there exists an index $k < i_j + 1$ such that $(x_k, u_k) = (x_{i_j+1}, u_{i_j+1})$; hence in the list L', there exists an index $i_l \leq i_j$ such that $x_{i_l} = x_{i_j+1} = z_{i_j}$.

To sum up, for each index $i_0, \ldots, i_r$, we can write $z_{i_j} = x_{i_l}$ with $i_l = i_j + 1$ or $i_l \leq i_j$. Thus, we can write r equalities between the x_i's and the z_i's.

Now, to show the existence of a $(r + 1)$th equality, we consider two cases:

1. The term (x_s, u_s, z_s, u_s) belongs to the list L' and in this case we have the additional equality $x_s = x_{s'}$ (notice that the term $(x_{s'}, u_{s'}, z_{s'}, y_{s'})$ cannot be removed from list L because there does not exist an index $i < s$ such that $x_{s'} = x_i$).

2. The term (x_s, u_s, z_s, u_s) does not belong to the list L' and in this case we have $(x_s, u_s) = (x_{s'}, u_{s'})$ because there exists an index $i < s$ such that $(x_s, u_s) = (x_i, u_i)$ and this index i is necessarily equal to s'.

If $i_r < s$, then $i_r + 1 \leq 2n$ and there exists an index $k \leq r$ such that $x_{i_r+1} = x_{i_k}$ but $x_{i_r+1} = z_{i_r}$.

If $s < i_r$, then let k_0 be the least index such that $s < i_{k_0}$. We also have

$$\begin{aligned} x_{i_{k_0}} &= z_{i_{k_0}-1} \\ &= f(x_{i_{k_0}-1}, u_{i_{k_0}-1}), \end{aligned}$$

but the term of index $i_{k_0} - 1$ does not belong to list L' and so there exists i_j such that $i_j < i_{k_0} - 1$ and $(x_{i_{k_0}-1}, u_{i_{k_0}-1}) = (x_{i_j}, u_{i_j})$, so $x_{i_{k_0}} = z_{i_j}$. This last equality has not been taken into account above because $i_{k_0} \geq s + 1$ and $s \geq i_j + 1$ and so $i_{k_0} \geq i_j + 2$. We then have $i_{k_0} > i_j$ and cannot have $i_{k_0} = i_j + 1$. □

The next two lemmas are concerned with the derivatives of the components of $\Theta^{f,h}_{2n+1}$.

LEMMA 4. *Let r be a given nonnegative integer and $(i_0, \ldots, i_{n-1})$ be a given sequence of indices in $\{0, \ldots, r\}$. Given $r + 1$ matrices $(A_0, \ldots, A_r)$ in $\mathrm{GL}(n, \mathbb{R})$, we consider the related sequence of matrices $(\tilde{A}_0, \ldots, \tilde{A}_{n-1})$, where*

(i) $\tilde{A}_0 = A_0$;
(ii) *for $j \geq 1$, $\tilde{A}_j = A_{i_j} \tilde{A}_{j-1}$.*

Let $1 \leq k \leq n - 1$ and consider the subset W_k of $\mathrm{GL}(n, \mathbb{R})^{r+1} \times \mathbb{P}^{n-1}$ ($\mathbb{P}^{n-1}$ is the projective space of dimension $n - 1$) constituted by the elements $(A_0, \ldots, A_r, l)$ such that, with $(\tilde{A}_0, \ldots, \tilde{A}_{n-1})$ being the sequence related to $(A_0, \ldots, A_r)$,

(i) *the family $(l, \tilde{A}_0 l, \ldots, \tilde{A}_{k-2} l)$ is linearly independent (this family reduces to (l) if $k = 1$);*

(ii) *the family* $(l, \tilde{A}_0 l, \dots, \tilde{A}_{k-1} l)$ *is linearly dependent.*

The set W_k *is a submanifold of* $\mathrm{GL}(n,\mathbb{R})^{r+1} \times \mathscr{P}_f^{n-1}$ *with codimension equal to* $n-k$.

Proof. Denote by V_k the set of sequences $(v_0, \dots, v_{n-1})$ of n elements of $\mathbb{R}^n$ such that

(i) the family $(v_0, \dots, v_{k-1})$ is linearly independent;

(ii) the family $(v_0, \dots, v_k)$ is linearly dependent;

then V_k is a submanifold of $\mathbb{R}^{n^2}$ of codimension $n-k$. This claim is a particular case of a more general proposition concerned with codimension of sets of linear homomorphisms; its proof can be found in [7, Prop. 5.3 p. 60].

We consider the domains of charts U_i, $i = 1, \dots, n$ of $\mathbb{P}^{n-1}$, where U_i is the set of lines of $\mathbb{R}^n$ generated by $(x_1, \dots, x_n)$ with $x_i \neq 0$; we shall define mappings φ_i $(i = 1, \dots, n)$ from U_i to $\mathbb{R}^{n^2}$. In what follows, the computations will be made only with φ_1, the reasoning being the same for the other mappings φ_i with $i \geq 2$.

We define φ_1 by

$$\begin{array}{rccc} \varphi_1 : & \mathrm{GL}(n,\mathbb{R})^{r+1} \times \mathbb{P}^{n-1} & \longrightarrow & \mathbb{R}^{n^2} \\ & (A_0, \dots, A_r, l) & \longmapsto & (\bar{l}, \tilde{A}_0 \bar{l}, \dots, \tilde{A}_{n-1} \bar{l}), \end{array}$$

where $\bar{l}$ is the element of $\mathbb{R}^n$ which represents l and whose first component is equal to 1. Clearly $W_k \cap U_1 = \varphi_1^{-1}(V_k)$ and we shall show that φ_1 is transverse to V_k which will prove that the codimension of W_k is the same as the one of V_k: $n-k$.

To this end we begin by the characterization of the tangent space of V_k; let $(v_0, \dots, v_{n-1})$ be an element of V_k, and from the matrix M whose columns are the vectors v_i $(i = 0, \dots, n-1)$ we can extract a squared $k \times k$ nonsingular submatrix, without loss of generality. We can assume that this submatrix is constituted by the k first lines and columns of M, and we write the $n \times (k+1)$ matrix whose columns are the vectors $(v_0, \dots, v_k)$ as follows:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

with A a squared matrix of order k, C a $(n-k) \times k$ matrix, and the column $(BD)^{\mathrm{T}}$ the vector $\tilde{A}_{k-1} \bar{l}$. The sequence $(v_0, \dots, v_{n-1})$ belongs to V_k iff $D - CA^{-1}B = 0$ [7, p. 60], so a vector $(t_0, \dots, t_{n-1})$ is tangent to V_k at $(v_0, \dots, v_{n-1})$ iff

$$L' - H'A^{-1}B + CA^{-1}HA^{-1}B - CA^{-1}L = 0, \tag{9}$$

where H is a square matrix of order k, H' is a $(n-k) \times k$ matrix, L is a column matrix with k lines, and L' is a column matrix with $n-k$ lines such that the columns of the matrix

$$\begin{pmatrix} H & L \\ H' & L' \end{pmatrix}$$

are constituted by the vectors $(t_0, \dots, t_k)$; notice that this equality can be written as

$$\begin{pmatrix} -CA^{-1} & I_{n-k} \end{pmatrix} \left\{ \begin{pmatrix} L \\ L' \end{pmatrix} - \begin{pmatrix} H \\ H' \end{pmatrix} A^{-1}B \right\} = 0. \tag{10}$$

Let $e = (A_0, \dots, A_r, l)$ be an element of $\mathrm{GL}(n,\mathbb{R})^{r+1} \times \mathbb{P}^{n-1}$ such that $\varphi_1(e) \in V_k$. We shall consider only the derivative with respect to $A_{i_{k-1}}$. We then have

$$\begin{aligned} &\mathrm{d}\varphi_1(e).(0, \dots, 0, M, 0, \dots, 0) \\ &\quad = (0, \psi_0(M, \bar{l}_0), \dots, \psi_{k-2}(M, \bar{l}_0), M\tilde{A}_{k-2}\bar{l}_0 + A_{i_{k-1}}\psi_{k-2}(M, \bar{l}_0), \dots), \end{aligned}$$

where $\psi_i(M, \bar{l}_0)$ is the expression obtained by replacing successively each occurrence of $A_{i_{k-1}}$ in $\tilde{A}_i$ by M; obviously, if $\tilde{A}_i$ does not contain $A_{i_{k-1}}$, then $\psi_i(M, \bar{l}_0)$ is zero. Denote by $w_0, \dots, w_{k-2}$ the vectors $\bar{l}_0, \tilde{A}_0\bar{l}_0, \dots, \tilde{A}_{k-3}\bar{l}_0$. The expression $\psi_i(M, \bar{l}_0)$ is either zero or a sum of terms of the form $B_j M w_j$ with $0 \le j \le k-2$, the vectors $w_0, \dots, w_{k-2}$ being linearly indépendent, and we can find a matrix M such that $Mw_j = 0$ for $j = 0, \dots, k-2$; moreover, the term $M\tilde{A}_{k-2}\bar{l}_0$ can be made equal to an arbitrary vector. By replacing in the right-hand member of (10), $(H, H')^{\mathrm{T}}$ and $(L, L')^{\mathrm{T}}$ by the corresponding components of $\mathrm{d}\varphi_1(e)$, we find the expression

$$\begin{pmatrix} -CA^{-1} & I_{n-k} \end{pmatrix} M\tilde{A}_{k-2}\bar{l}_0,$$

which can be made nonzero. In conclusion, we can find $n-k$ (the codimension of V_k) independent matrices M such that $\mathrm{d}\varphi_1(e)(0, \dots, M, \dots, 0)$ does not belong to the tangent space of V_k. □

We first introduce a definition.

DEFINITION 7. *Consider $r+1$ matrices $A_0, \dots, A_r$ in $\mathrm{GL}(n, \mathbb{R})$ and $r+1$ matrices $C_0, \dots, C_r$ in $\mathscr{M}_{p,n}(\mathbb{R})$. We say that the finite sequence of n matrices $D_0, \dots, D_{n-1}$ is differentially related to the family $(A_0, C_0, A_1, C_1, \dots, A_r, C_r)$ if*

(i) *$D_0 = C_0$ and for $j \ge 1$, each D_j is equal to $C_{i_j}\tilde{A}_{j-1}$, where $i_j \in \{0, \dots, r\}$, and $\tilde{A}_{j-1}$ is a product of j matrices taken in the set $\{A_0, \dots, A_r\}$;*

(ii) *if $D_j = C_{i_j}\tilde{A}_{j-1}$, then D_{j+1} has the form $D_{j+1} = C_{i_{j+1}}A_{i_j}\tilde{A}_{j-1}$;*

(iii) *each matrix $C_0, \dots, C_r$ is involved at least one time in the D_i's; to be more precise, there exist indices $0 \le i_0 < \cdots < i_{r-1}$ such that $C_{i_j} = C_j$.*

For the sake of readability, hereafter, we shall denote by $\mathscr{M}$ the set $\mathrm{GL}(n, \mathbb{R}) \times \mathscr{M}_{p,n}(\mathbb{R})$.

LEMMA 5. *Take $r+1$ matrices $A_0, \dots, A_r$ in $\mathrm{GL}(n, \mathbb{R})$ and $r+1$ matrices $C_0, \dots, C_r$ in $\mathscr{M}_{p,n}(\mathbb{R})$ and $D_0, \dots, D_{r-1}$ a sequence differentially related to the family $(A_0, C_0, \dots, A_r, C_r)$. Consider the set*

$$W = \{ (A_0, C_0, A_1, C_1, \dots, A_r, C_r) \in \mathscr{M}^{r+1} \mid \exists x \in \mathbb{R}^n, D_0x = D_1x = \cdots = D_{\bar{r}}x = 0 \},$$

where $\bar{r} = \max(r, n-1)$. We claim that W is a submanifold of $\mathscr{M}^{r+1}$ with codimension greater than $(r+1)m$.

Proof. We denote by $\mathbb{P}^{n-1}$ the $n-1$-dimensional real projective space, and for $k = 0, \dots, n-1$ we consider the sets M_k of elements

$$(A_0, C_0, \dots, A_r, C_r, \ell) \in \mathscr{M}^{r+1} \times \mathbb{P}^{n-1}$$

such that the family $(\ell, \tilde{A}_0\ell, \dots, \tilde{A}_{k-1}\ell)$ is linearly independent (if $k = 0$, this family reduces to (ℓ)). Obviously, the sets M_k are open in $\mathscr{M}^{r+1} \times \mathbb{P}^{n-1}$; remark also that if E is the subset of matrices of $\{C_0, \dots, C_r\}$ involved in the terms $D_0, \dots, D_k$ and if $\operatorname{card} E < r+1$, then there exist r' matrices $D_{j_1}, \dots, D_{j_{r'}}$ whose indices $j_1, \dots, j_{r'}$ are greater than k and such that $\{C_{i_{j_1}}, \dots, C_{i_{j_{r'}}}\}$ is the set complement of E in $\{C_0, \dots, C_r\}$.

We also define the sets N_k as the subsets of elements $(A_0, C_0, \dots, A_r, C_r, \ell)$ of M_k such that

(i) if $k < n-1$, then the family $(\ell, \tilde{A}_0\ell, \dots, \tilde{A}_{k-1}\ell, \tilde{A}_k\ell)$ is linearly dependent;

(ii) $D_0\ell = D_1\ell = \cdots = D_k\ell = D_{j_1}\ell = \cdots = D_{j_{r'}}\ell = 0$.

If we denote by π the projection from $\mathscr{M}^{r+1} \times \mathbb{P}^{n-1}$ to $\mathscr{M}^{r+1}$, then obviously $W \subset \cup_{k=0}^{n-1}\pi(N_k)$. Clearly, the codimension of N_k is equal to $n-(k+1)+(k+1+r')p$

and so the codimension of $\pi(N_k)$ is greater than or equal to $n-(k+1)+(k+1+r')$ $p-(n-1)=(k+1+r')p-k$. As W is included in the union of the projections of the N_k's, its codimension is greater than or equal to $\min_{0\le k\le n-1}\operatorname{codim}(N_k)$ but

$$\begin{aligned}(k+1+r')p-k &\ge (k+1+r')(m+1)-k \qquad \text{since } p>m\\ &= (k+1+r')m+r'+1\\ &\ge (r+1)m+r'+1\\ &> (r+1)m. \qquad \square\end{aligned}$$

The proof will result from an application of the multijet transversality theorem. To fix $s\le 2n$, we shall prove that the set of pairs (f,h) such that the mapping $\Theta^{f,h}_{2n+1}$ is an immersion at each periodic point x of f is residual as a finite intersection of residual sets.

If x_0 is a s-periodic point for $f\in \mathrm{Diff}_{\mathrm{U}}(X)$, as shown in the proof of Lemma 3, then we can find a list of $r+1$ equalities between the x_i's and the z_i's.

We consider the mapping

$$\begin{array}{rccl} j^1_{r+1}(f,h): & (X\times U)^{(r+1)} & \longrightarrow & J^1_{r+1}(X\times U, X\times\mathbb{R}^p)\\ & (\xi_0,\upsilon_0,\dots,\xi_r,\upsilon_r) & \longmapsto & (j^1(f,h)(\xi_0,\upsilon_0),\dots,j^1(f,h)(\xi_r,\upsilon_r)),\end{array}$$

and we shall define a submanifold W of $J^1_{r+1}(X\times U, X\times\mathbb{R}^p)$. Let $(\mathscr{O}_i\times\mathscr{U}_i,(\varphi_i,\psi_i))$ be a chart of $X\times U$ at (x_i,u_i), and the local expression of the above mapping is given by

$$\begin{array}{rccl} \bar{j}: & \varphi_0(\mathscr{O}_0)\times\psi_0(\mathscr{U}_0)\times\cdots\times\varphi_r(\mathscr{O}_r)\times\psi_r(\mathscr{U}_r) & \longrightarrow & E_0\times\cdots\times E_r\\ & (\bar{\xi}_0,\bar{\upsilon}_0\dots,\bar{\xi}_r,\bar{\upsilon}_r) & \longmapsto & (\beta_0,\dots,\beta_r)\end{array}$$

with

$$\begin{aligned} Ei &= \varphi_i(\mathscr{O}_i)\times\psi_i(\mathscr{U}_i)\times \mathrm{GL}(n,\mathbb{R})\times\mathscr{M}_{p,n}(\mathbb{R}),\\ \beta_i &= (\bar{\xi}_i,\bar{\upsilon}_i,\bar{f}(\bar{\xi}_i,\bar{\upsilon}_i),\bar{h}(\bar{\xi}_i,\bar{\upsilon}_i),d\bar{f}(\bar{\xi}_i,\bar{\upsilon}_i),d\bar{h}(\bar{\xi}_i,\bar{\upsilon}_i)),\end{aligned}$$

and $\bar{\xi}_i$, $\bar{\upsilon}_i$, $\bar{f}$, and $\bar{h}$ the local expressions of ξ_i, υ_i, f, and h, respectively. We put $A_i=d\bar{f}(\bar{\xi}_i,\bar{\upsilon}_i)$ and $C_i=d\bar{h}(\bar{\xi}_i,\bar{\upsilon}_i)$. We define locally the submanifold W defined by one of the sets of the $r+1$ equalities shown in the proof of Lemma 3 and by the relations between matrices A_i and C_i as shown in Lemma 5. We shall give an estimate of $\operatorname{codim}(W)$.

To this end, we denote by W_1 the submanifold containing W which is defined only by the equalities bearing on matrices A_i and C_i and by π the projection from the Cartesian product $E\triangleq E_0\times\cdots\times E_r$ to $(\mathrm{GL}(n,\mathbb{R})\times\mathscr{M}_{p,n}(\mathbb{R}))^{r+1}$; the codimension of W in E is equal to the codimension of W as a submanifold of W_1 plus the codimension of W_1 in E. The codimension of W as a submanifold of W_1 is equal to $(r+1)\,n$ and the codimension of W_1 in E is greater than or equal to the codimension of its projection $W_2\triangleq\pi(W_1)$. Applying Lemma 5, we can say that the codimension of a set of matrices as W_2 is greater than $(r+1)m$, so the codimension of W is greater than $(r+1)(n+m)$, which is the dimension of $(X\times U)^{(r+1)}$. Thus $\bar{j}$ is transverse to W iff $\bar{j}(\bar{\xi}_0,\bar{\upsilon}_0,\dots,\bar{\xi}_r,\bar{\upsilon}_r)\notin W$ for all $(\bar{\xi}_0,\bar{\upsilon}_0,\dots,\bar{\xi}_r,\bar{\upsilon}_r)$.

For each r from 0 to $2n$, we consider a countable family $\mathscr{F}$ of charts covering $(X \times U)^{r+1}$ and consider the application of the Thom transversality theorem to each chart of $\mathscr{F}$ as explained previously. If $(f,h) \in \mathrm{Diff_U}(X) \times C^\infty(X, \mathbb{R}^p)$ and $(x_0, \underline{u_{2n+1}})$ is a periodic point of f with period no greater than $2n$, then starting from x_0 we consider the list L' as in Lemma 3; the element $(x_0, u_0, \ldots, x_{i_r}, u_{i_r})$ constituting L' belongs to one of the charts of the family $\mathscr{F}$, and together with the z_i's satisfy $r+1$ equalities as explained in Lemma 3. Moreover, the above reasoning shows that the set of pairs (f,h), such that $\Theta^{f,h}_{2n+1}$ is an immersion at each periodic point of f lying in one of the charts of family $\mathscr{F}$, is residual by considering the (countable) intersection of all the residual sets related to the charts of $\mathscr{F}$; Proposition 2 is then proven.

5. Conclusion. Let us denote by $\mathscr{O}_1$ the set of pairs $(f,h) \in \mathrm{Diff_U}(X) \times C^\infty(X \times U, \mathbb{R}^p)$ such that the mapping $\Theta^{f,h}_{2n+1}$ is an immersion; the conjunction of Propositions 1 and 2 proves that $\mathscr{O}_1$ is residual. Let us denote by $\mathscr{O}_2$ the set of pairs $(f,h) \in \mathrm{Diff_U}(X) \times C^\infty(X \times U, \mathbb{R}^p)$ such that the mapping $\Theta^{f,h}_{2n+1}$ is one-to-one; in [3], under the assumption that X and U are compact, we proved that $\mathscr{O}_2$ is also residual. Let us denote by $\mathscr{O}_3$ the set of pairs $(f,h) \in \mathrm{Diff_U}(X) \times C^\infty(X \times U, \mathbb{R}^p)$ such that the mapping $\Theta^{f,h}_{2n+1}$ is an embedding; we can conclude that $\mathscr{O}_3$ is residual as the intersection of $\mathscr{O}_1$ and $\mathscr{O}_2$. Moreover, we also proved the openness of $\mathscr{O}_3$; therefore we can assert that $\mathscr{O}_3$ is open and dense.

Acknowledgment. The authors thank the second reviewer for his fruitful remarks and suggestions.

REFERENCES

[1] R. Abraham and J. Robbin, *Transversal Mappings and Flows*, W. A. Benjamin, New York, 1967.

[2] D. Aeyels, *Generic observability of differentiable systems*, SIAM J. Control Optim., 19 (1981), pp. 595–603.

[3] S. Ammar and J.-C. Vivalda, *On the genericity of the observability of controlled discrete-time systems*, ESAIM Control Optim. Calc. Var., 11 (2005), pp. 161–179.

[4] J. Gauthier and I. Kupka, *Deterministic Observation Theory and Applications*, Cambridge University Press, Cambridge, UK, 2002.

[5] J.-P. Gauthier, H. Hammouri, and I. Kupka, *Observers for nonlinear systems*, in Proceedings of the IEEE 30th CDC, Brighton, England, 1991, pp. 1483–1489.

[6] J.-P. Gauthier and I. Kupka, *Observability for systems with more outputs than inputs and asymptotic observers*, Math. Z., 223 (1996), pp. 47–78.

[7] M. Golubitsky and V. Guillemin, *Stable Mappings and Their Singularities*, Springer-Verlag, New York, 1973.

[8] J. Stark, *Delay embedding for forced systems.* I. *Deterministic forcing*, J. Nonlinear Sci., 9 (1999), pp. 255–332.

[9] J. Stark, D. Broomhead, M. Davies, and J. Huke, *Delay embedding for forced systems.* II. *Stochastic forcing*, J. Nonlinear Sci., 13 (2003), pp. 519–577.

[10] F. Takens, *Detecting strange attractors in turbulence*, in Dynamical Systems and Turbulence, Lecture Notes in Math. 898, Springer-Verlag, Berlin, 1981, pp. 366–381.

[11] J.-C. Vivalda, *On the genericity of the observability of uncontrolled discrete nonlinear systems*, SIAM J. Control Optim., 42 (2003), pp. 1509–1522.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2199–2241

AN INTRODUCTION TO QUANTUM FILTERING*

LUC BOUTEN†, RAMON VAN HANDEL†, AND MATTHEW R. JAMES‡

Abstract. This paper provides an introduction to quantum filtering theory. An introduction to quantum probability theory is given, focusing on the spectral theorem and the conditional expectation as a least squares estimate, and culminating in the construction of Wiener and Poisson processes on the Fock space. We describe the quantum Itô calculus and its use in the modeling of physical systems. We use both reference probability and innovations methods to obtain quantum filtering equations for system-probe models from quantum optics.

Key words. quantum filtering, quantum probability, quantum stochastic processes

AMS subject classifications. 93E11, 81P15, 81S25, 81Q10, 81R15, 34F05

DOI. 10.1137/060651239

1. Introduction.

Since even before the industrial revolution, feedback control has played a major role in the development of technology. Nowadays, many machines and devices that make up our everyday lives use feedback to provide efficient and reliable performance despite ever increasing complexity and miniaturization, and a rich control theory has been developed to aid in the design of feedback controllers based on device models from classical physics. As microtechnology is making way for nanotechnology, however, we are now rapidly approaching the boundary of the classical world past which the effects of quantum mechanics cannot be neglected.

The laws of quantum mechanics tell us that any description of the phenomena at small scales is inherently nondeterministic in nature. This opens new areas of application for stochastic control theory, which could play a role in a future generation of technology. In particular, as observations of quantum systems are inherently noisy, the theory of filtering—the extraction of information from a noisy signal—forms an integral part of quantum feedback control theory.

Quantum filtering was already implicit in early work on quantum measurement theory by Davies in the 1960s [24, 25]. In its modern form, the study of quantum filtering and control was pioneered by Belavkin in a series of articles dating back to the early 1980s [9, 10, 11, 12, 13]. The theory developed by Belavkin provides an essential foundation for statistical inference in, e.g., quantum optical systems, and much of what we will discuss in the second half of this article is based on his work. The theory gained popularity in the physics community after it was independently developed on a more heuristic level by Carmichael in the early 1990s [22] under the name "quantum trajectory theory" and has since been widely applied in the description of quantum optical experiments and as a computational tool.

*Received by the editors January 30, 2006; accepted for publication (in revised form) June 20, 2007; published electronically December 21, 2007.
http://www.siam.org/journals/sicon/46-6/65123.html

†Physical Measurement and Control 266-33, California Institute of Technology, Pasadena, CA 91125 (bouten@its.caltech.edu, ramon@its.caltech.edu). The research of these authors are supported by Army Research Office grant DAAD19-03-1-0073. The first author is additionally supported by National Science Foundation grant PHY-0456720.

‡Department of Engineering, Australian National University, Canberra, ACT 0200, Australia (matthew.james@anu.edu.au). The research of this author is supported by the Australian Research Council.

Based on the foundations of quantum filtering theory, methods from classical nonlinear and stochastic control can be developed and applied to design feedback control laws for quantum systems. These methods may be optimal in some sense or otherwise designed with relevant considerations in mind (e.g., stability). The resulting controllers are intended to be implemented with some classical technology (e.g., digital or analog electronics). Recent experiments implementing quantum feedback controls [3, 35, 58, 21] have led to renewed interest in the field, which is now rapidly expanding [10, 64, 28, 27, 17, 43, 15, 61, 16, 44, 37, 30, 60, 62, 18]. We believe that a fruitful interaction between stochastic control and theoretical and experimental physics will be essential in paving the way towards the engineering of quantum technologies.

This paper provides an introduction to quantum filtering theory. There are three key ingredients that are required for the development of the theory. First, we need to capture both classical probability and quantum mechanics within the framework of a generalized probability theory, called noncommutative or quantum probability theory. The central object in this theory, the spectral theorem, provides a link between quantum systems and the associated probabilistic measurement outcomes. Second, we need a noncommutative generalization of the concept of conditional expectations. As in classical probability, we will find that a suitably restricted definition of the quantum conditional expectation is none other than a least squares estimator, which elucidates its role in quantum filtering theory. Finally, we need a noncommutative analog of stochastic calculus and quantum stochastic differential equations (QSDEs). This provides a broad class of models for which we can obtain filtering equations.

A typical physical scenario, to which the theory that we will develop can be applied, is illustrated schematically in Figure 5.1. A cloud of (usually cold, trapped) atoms interacts with the electromagnetic field in free space; this can be coherent light from a laser, or even the vacuum. Depending on their internal state the atoms can, for example, emit radiation into the field. If we detect this radiation using an optical detection setup, we can try to infer some information on the internal state of the atoms—this is precisely the goal of quantum filtering theory. If we wanted to control the state of the atoms, we could then feed back some function of the state estimates through a suitable actuator. Recent laboratory experiments (e.g., [58]) implement precisely such a setup and provide a motivating example for the theory.

We begin in section 2 by providing some background for quantum filtering. This includes a discussion of the quantum mechanics and quantum probability in the simplest, finite-dimensional context. In section 3 quantum probability is developed in detail. Then in section 4 we show how Wiener and Poisson processes emerge in a particular quantum probabilistic model based on the Fock space, and how these can be used to develop a noncommutative stochastic calculus. In section 5 we introduce a class of system-observation models that describe typical experiments in quantum optics. Section 6 deals with the derivation of quantum filtering equations using the reference probability approach, while section 7 gives an alternative derivation using the innovations or martingale method.

Scope. It has been our aim to make quantum probability and filtering theory accessible, modulo a set of technicalities, to readers with a minimal number of prerequisites. We (only) presume some familiarity with probability theory and elementary functional analysis. We have put an emphasis on introducing the mathematical structures of quantum probability theory and on demonstrating their significance and their use. As a consequence we do not everywhere achieve the highest level of rigor; we are particularly lax in the use of unbounded operators and their domains. It is

our hope that skimming over these technicalities has enabled us to paint a clearer picture of the pillars of the theory and of the essential techniques involved. That being said, we should point out that many of the tools described in this paper are applied regularly and successfully by physicists without paying any attention to the technical issues involved; the reader should not hesitate to get his feet wet!

It is an ambitious project to introduce an unfamiliar probability theory, a new stochastic calculus, and to even solve a nontrivial problem (filtering) within the confines of about 40 pages. Though we have tried to give a pedagogical treatment, the explanations are sometimes necessarily terse; we hope that the reader will be sufficiently compelled to work his way through the paper. Needless to say there are many omissions; one that particularly deserves mention is the linear case: indeed, the quantum Kalman filter, and the corresponding theory of quantum LQG control, can be developed along similar lines to the filters we will discuss. We have chosen to omit this topic in order to avoid the technicalities of QSDEs with unbounded coefficients, but refer instead to [30] and the references therein.

Notation. The sets of natural, real, and complex numbers are denoted $\mathbf{N}$, $\mathbf{R}$, and $\mathbf{C}$, respectively. In general, script symbols (e.g., $\mathscr{Y}$) are used for von Neumann algebras, while calligraphic symbols (e.g., $\mathcal{Y}$) stand for σ-algebras. $\mathcal{B}$ is the Borel σ-algebra on $\mathbf{R}$. Classical probability spaces are denoted as $(\Omega, \mathcal{F}, \mathbf{P})$, and $E_{\mathbf{P}}$ denotes the expectation with respect to the measure $\mathbf{P}$. Blackboard symbols (e.g., $\mathbb{P}$) denote states on von Neumann algebras. Sans serif symbols (e.g., H) are used for Hilbert spaces. Hilbert space adjoints, as well as the scalar complex conjugate, are indicated by *, and the Hilbert space inner product is denoted by $\langle \cdot, \cdot \rangle$. The commutator of two bounded operators is denoted by $[X, Y] = XY - YX$. I is the identity operator.

2. Background and motivation. In this article we adopt a modern quantum probability formulation of quantum mechanics. *Quantum probability* is the noncommutative counterpart of Kolmogorov's axiomatic characterization of classical probability theory. In addition to the natural interpretation and mathematical tools provided by Kolmogorov's formalism, one of its major successes is that conditioning is a derived concept rather than an additional axiom. The situation is much the same in quantum probability; in particular, the conditioning axiom or "projection postulate," as it is traditionally posed in quantum mechanics, can emerge as a consequence of conditional expectation and the physical idea that in a single experiment one only has direct access to information contained in a commutative subalgebra of observables.

Considering the success of the classical (Kolmogorov) theory, it should come as no surprise that the mathematical abstraction provided by the framework of quantum probability pays off significantly (as we will see throughout the article). Introductory physics textbooks on quantum mechanics rarely use such a description, however. In this section we introduce the basic concepts of quantum probability in their simplest form and attempt to provide contact with ideas about quantum mechanics that readers may be familiar with. This is intended to provide a reference point for interpreting the quantum probabilistic framework used in this paper.

2.1. Some textbook quantum mechanics. According to the textbook by Merzbacher [51, p. 1], "Quantum mechanics is the theoretical framework within which it has been found possible to describe, correlate, and predict the behavior of a vast range of physical systems, from particles through nuclei, atoms, and radiation to molecules and condensed matter." Central to quantum mechanics are the notions of *observables*, which are mathematical representations of physical quantities that can (in principle) be measured, and *states*, which summarize the status of physical systems

and permit the calculation of statistical quantities (such as probabilities, expectations, correlations) of observables.

Indeed, the reader may be familiar with the *Schrödinger wavefunction* $\psi(q,t)$ for a particle of mass m moving in a force field $V(q)$ (dependent on position q, in one dimension for simplicity). If Q is the observable representing position (defined in Example 3.9), the expected position of the particle when in a state described by $\psi(q,t)$ at time t is defined to be

$$\langle Q\rangle = \int q|\psi(q,t)|^2 dq. \tag{2.1}$$

The wavefunctions are normalized to one $\int |\psi(q,t)|^2 dq = 1$, so that $|\psi(q,t)|^2$ could be interpreted as the probability density of the position of the particle. The dynamics of the particle are described by the famous *Schrödinger wave equation*

$$i\hbar\frac{\partial\psi(q,t)}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi(q,t)}{\partial q^2} + V(q)\psi(q,t), \tag{2.2}$$

where $\hbar = h/2\pi$, h is *Planck's constant*, and $i^2 = -1$.

The key distinction between classical (i.e., nonquantum) and quantum mechanics is that quantum mechanics is *noncommutative*, meaning that there exist observables that do not commute, a fact which has deep implications. The momentum observable P (defined in Example 3.9) does not commute with the position observable Q; in fact $[Q,P] = QP - PQ = i\hbar I$. The most famous implication of this failure of commutativity is *Heisenberg's uncertainty relation*, which asserts that

$$\Delta Q\Delta P \geq \frac{1}{2}|\langle i[Q,P]\rangle| = \frac{\hbar}{2}, \tag{2.3}$$

where the variances are defined by $\Delta Q = (\langle Q^2\rangle - \langle Q\rangle^2)^{1/2}$, $\Delta P = (\langle P^2\rangle - \langle P\rangle^2)^{1/2}$. Naive interpretation of the Heisenberg uncertainty relation can be misleading; we will discuss its precise meaning in the following section. Nonetheless, it evidently implies that there is a fundamental irreducible randomness in quantum mechanics. This is in contrast to classical randomness, which in principle can be eliminated with enough effort and information. Experimental evidence has repeatedly confirmed the irreducible randomness of quantum mechanical observations.

Let us make this somewhat vague discussion a little more precise. For simplicity, we will work in this section only in a finite-dimensional setting (in which observations can only take a finite number of values; i.e., they are finite-state random variables). First, recall that if $A = A^*$ is a self-adjoint operator on a finite-dimensional Hilbert space $\mathsf{H} = \mathbf{C}^n$, it has at most n (distinct) real eigenvalues. The set $\mathrm{spec}(A) = \{a_j\}$ of eigenvalues of A is called the spectrum of A, and A can be written as

$$A = \sum_{a\in\mathrm{spec}(A)} a\,P_a, \tag{2.4}$$

where P_a is the projection operator onto the subspace of H spanned by vectors with eigenvalue a. The projections resolve the identity $\sum_{a\in\mathrm{spec}(A)} P_a = I$.

In this finite-dimensional setting, the following operational characterization of quantum mechanical models (often referred to as the "postulates" of quantum mechanics) can be found in most introductory textbooks.

Observables. Physical quantities like position, momentum, spin, etc. are represented by self-adjoint operators on the Hilbert space H and are called *observables*. These are the noncommutative counterparts of random variables.

States. A state is meant to provide a summary of the status of a physical system that enables the calculation of statistical quantities associated with observables. A generic state is specified by a *density matrix* ρ, which is a self-adjoint operator on H that is positive $\rho \geq 0$ and normalized $\mathrm{Tr}[\rho] = 1$. This is the noncommutative counterpart of a probability density.

Measurement. A *measurement* is a physical procedure or experiment that produces numerical results related to observables. In any given measurement, the allowable results take values in the spectrum $\mathrm{spec}(A)$ of a chosen observable A. Given the state ρ, the value $a \in \mathrm{spec}(A)$ is observed with probability $\mathrm{Tr}[\rho P_a]$. Consequently, the expectation of an observable A is given by $\langle A \rangle = \mathrm{Tr}[\rho A]$.

Conditioning. Suppose that a measurement of A gives rise to the observation $a \in \mathrm{spec}(A)$. Then we must condition the state in order to predict the outcomes of subsequent measurements by updating the density matrix ρ using

$$\rho \mapsto \rho'[a] = \frac{P_a \rho P_a}{\mathrm{Tr}[\rho P_a]}. \tag{2.5}$$

This is known as the "projection postulate."

Evolution. A *closed* (i.e., isolated) quantum system evolves in a *unitary* fashion: a physical quantity that is described at time $t = 0$ by an observable A is described at time $t > 0$ by $A(t) = U(t)^* A U(t)$, where $U(t)$ is a unitary operator for each time t. The unitary is generated by the *Schrödinger equation*

$$i\hbar \frac{d}{dt} U(t) = H(t) U(t), \tag{2.6}$$

where the (time-dependent) Hamiltonian $H(t)$ is a self-adjoint operator for each t.

Before continuing, we make the following remarks.

Remark 2.1. *Pure states.* The set of density matrices ρ is convex; we can thus wonder what are the extremal points in this set, i.e., those that correspond to the most informative states. It is not difficult to show that the set of extremal density matrices is the set of projections onto one-dimensional subspaces. Thus we can specify any extremal state uniquely (up to a phase factor $e^{i\varphi}$) by a single unit vector $\psi \in \mathsf{H}$ in the corresponding subspace, and $\mathrm{Tr}[\rho X] = \langle \psi, X\psi \rangle$ for any operator X. In classical probability theory, the set of probability measures is also convex and the extremal measures are deterministic (Dirac) measures. In the quantum mechanical setting, on the other hand, the Heisenberg uncertainty relation implies that even extremal states do not give deterministic measurement outcomes for all observables.

Historically, and in most textbooks, quantum mechanics is first formulated in terms of the extremal states (called *pure states*) and the description is later generalized to density matrices (*mixed states*). The Schrödinger wavefunction $\psi(q, t)$ is an example of a pure state vector in an infinite-dimensional Hilbert space setting. □

Remark 2.2. *Heisenberg vs. Schrödinger picture.* In the above description of time evolution we work with a fixed state while the observables change in time. This conforms to the usual treatment in classical probability theory, where the underlying probability measure is fixed at the outset and the random variables are time dependent (stochastic processes). In quantum mechanics this is known as the *Heisenberg picture*; equally (or perhaps more) popular is the *Schrödinger picture*, in which the observables

are considered fixed and the density matrix evolves as $\rho(t) = U(t)\rho U(t)^*$. The two pictures are essentially equivalent as $\mathrm{Tr}[\rho A(t)] = \mathrm{Tr}[\rho(t)A]$ for any observable A.

Note that if we start in a pure state, then unitary evolution preserves this property; in terms of the state vector, $\psi(t) = U(t)\psi$. Intuitively, this enforces the physical idea that no information is lost from an isolated system. Together with (2.6) we obtain the traditional Schrödinger equation for $\psi(t)$, of which (2.2) is a special case (for a specific choice of H, in infinite dimensions). *We will always work in the Heisenberg picture*, however, as we will be dealing with (quantum) stochastic processes. □

As a basic illustration we discuss the following simple example.

Example 2.3. One of the classic experimental demonstrations of the necessity of quantum mechanics was performed in 1922 by Stern and Gerlach. A silver atom is subjected to an inhomogeneous magnetic field. The atom possesses an intrinsic magnetic moment and hence experiences a force that is proportional to the component of its magnetic moment in the direction of the field gradient. As Stern and Gerlach did not prepare the atom in a particular orientation, they expected it to be deflected randomly in a continuous range of directions corresponding to a random orientation of the magnetic moment. Repeated runs of the experiment showed, however, that the atom is randomly deflected into two discrete directions only—the reason being that in quantum mechanics the intrinsic magnetic moment (or spin) observable is discrete, rather than continuous. Atoms deflected in the upper direction are said to have "spin up," while those in the lower direction have "spin down."

A simple model of a spin is as follows. Let $\mathsf{H} = \mathbf{C}^2$, and consider the observable

$$\sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{2.7}$$

representing spin in the z direction. We have $\mathrm{spec}(\sigma_z) = \{-1, 1\}$, which correspond to spin down and spin up, respectively. In terms of the eigenprojections

$$P_{z,1} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad P_{z,-1} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

we can write $\sigma_z = P_{z,1} - P_{z,-1}$. The next step is to introduce a state. Consider a pure state, given by the vector $\psi = (c_1 \; c_{-1})^T$ with $|c_1|^2 + |c_{-1}|^2 = 1$. If we observe σ_z, we obtain the outcome 1 (spin up) with probability $\langle \psi, P_{z,1}\psi\rangle = |c_1|^2$, or the outcome -1 with probability $\langle \psi, P_{z,-1}\psi\rangle = |c_{-1}|^2$. □

2.2. A first look at quantum probability. The description of quantum mechanics in the previous section contains the rudiments of a viable probability theory. We will now formalize these ideas, once again restricting ourselves to the finite-dimensional case for simplicity (the general theory, which will be discussed in section 3, is conceptually very similar). Two key ideas, which we elaborate on below, form the essence of the formalism: the first is that a set of measurements made in a single realization[1] of a quantum experiment corresponds to a particular choice of a commutative algebra of observables; and the second is that any such commutative algebra is entirely equivalent to a classical (Kolmogorov) probability model.

A classical probability model is described by a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Here Ω, the sample space, is not of essential importance; the basic ingredients of the theory

[1]By a *realization* or an *experiment* we mean that random variables are assigned a definite value, as is the case if we perform measurements on a single physical system. In classical probability this corresponds to the choice of a sample point $\omega \in \Omega$; the quantum case is a little more subtle.

are the events that can occur, contained in the σ-algebra $\mathcal{F}$, and their probabilities, which are determined by the measure $\mathbf{P}$. Equivalently, we could describe an event $F \in \mathcal{F}$ by a random variable χ_F which takes the value 1 if F occurs and 0 otherwise (the indicator function on F), and the probability of the event is simply the expectation of χ_F. We have already encountered such objects in the previous section: events are precisely those observables that are projection operators ($P = P^* = P^2$), and the probability of an event P is given by $\mathbb{P}(P) = \mathrm{Tr}[\rho P]$. Thus the set of projections, together with the linear map $\mathbb{P}$, play much the same role as the classical pair $\mathcal{F}, \mathbf{P}$.

We run into trouble in the quantum case when we try to ascribe joint probabilities to certain events. This is always possible in classical probability theory: the joint probability of the events A and B is $\mathbf{P}(A \cap B) = E_{\mathbf{P}}(\chi_A \chi_B)$. But given two projection operators P, Q, the operator PQ is not guaranteed to be a projection or even an observable ($(PQ)^* = QP$), *unless* P and Q commute. This simple observation is no coincidence; it has the following physical interpretation: in a single realization of a quantum probability model, we can only verify the truth of a set of commuting events. This is in contrast with classical probability where in every realization any event is either true or false, whether we choose to observe it or not. In quantum probability we can a priori choose to verify the truth of an arbitrary event, but subsequently some of the other events (those that do not commute with the observed event, said to be *incompatible*) become meaningless within the same realization.

The incompatibility of events is a significant conceptual departure from classical probability and requires a little getting used to. In many ways, however, this is the only essential departure from classical probability theory. We now begin to construct the mathematical formalism of quantum probability, and we will show that it is indeed very close to Kolmogorov's theory.

Consider the following idea. Suppose we decide to measure an observable A and obtain a particular outcome $a \in \mathrm{spec}(A)$. Then we do not need to perform another measurement to know that any function $f(A)$ would give the outcome $f(a)$; in essence, this is merely a relabeling of the measurement outcomes of A. Indeed,

$$A = \sum_{a \in \mathrm{spec}(A)} a\, P_a \quad \Longrightarrow \quad f(A) = \sum_{a \in \mathrm{spec}(A)} f(a)\, P_a, \tag{2.8}$$

and all such operators commute with each other. Thus measuring A "automatically" measures all functions $f(A)$. The set of operators $\mathscr{A} = \{X : X = f(A),\ f : \mathbf{R} \to \mathbf{C}\}$ forms a *commutative $*$-algebra*, i.e., arbitrary (complex) linear combinations, products, and adjoints of operators in $\mathscr{A}$ are still in $\mathscr{A}$, $I \in \mathscr{A}$, and all elements of $\mathscr{A}$ commute. We will call $\mathscr{A}$ the $*$-algebra *generated*[2] *by* A. A linear map $\mathbb{P} : \mathscr{A} \to \mathbf{C}$ that is positive ($\mathbb{P}(A) \geq 0$ if $A \geq 0$) and normalized ($\mathbb{P}(I) = 1$) is called a *state* on $\mathscr{A}$ (clearly we can always write such a state as $A \mapsto \mathrm{Tr}[\rho A]$ for some density matrix ρ). Note that the projections $P \in \mathscr{A}$ are precisely those events that we can distinguish by measuring A, and $\mathbb{P}(P)$ gives their probabilities. We can similarly generate the commutative $*$-algebra of functions of an arbitrary set of commuting observables.

The algebraic structure we have introduced is of fundamental importance as it provides us with a direct connection to the classical theory, as follows.

THEOREM 2.4 (spectral theorem, finite-dimensional case). *Let $\mathscr{A}$ be a commutative $*$-algebra of operators on a finite-dimensional Hilbert space, and let $\mathbb{P}$ be a state*

[2]In fact, it is the smallest $*$-algebra of operators that contains A.

on $\mathscr{A}$. Then there are a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a map ι from $\mathscr{A}$ onto the set of measurable functions on Ω that is a $*$-isomorphism, i.e., a linear bijection with $\iota(AB) = \iota(A)\iota(B)$ (pointwise) and $\iota(A^*) = \iota(A)^*$, and moreover $\mathbb{P}(A) = E_{\mathbf{P}}(\iota(A))$.

Proof. The proof is an elementary exercise in linear algebra. As the Hilbert space H has dimension $n < \infty$, we can, without loss of generality, suppose that $\mathsf{H} = \mathbf{C}^n$ and that $\mathscr{A}$ is a commutative $*$-algebra of complex $n \times n$ matrices. As all the elements of $\mathscr{A}$ commute, we can find a unitary matrix U such that U^*AU is a diagonal matrix for every $A \in \mathscr{A}$. Let $\Omega = \{1, \ldots, n\}$. Define $\iota(A) : \Omega \to \mathbf{C}$ by $\iota(A)(i) = (U^*AU)_{ii}$ for every $A \in \mathscr{A}$. Next, define $\mathcal{F} = \sigma\{\iota(A) : A \in \mathscr{A}\}$. Finally, define $\mathbf{P}(S) = \mathbb{P}(\iota^{-1}(\chi_S))$ for every $S \in \mathcal{F}$. We have now explicitly constructed $(\Omega, \mathcal{F}, \mathbf{P})$ and ι. □

Evidently the commutative $*$-algebra structure is completely equivalent to classical probability theory; by simultaneously diagonalizing all the operators in the algebra, we obtain an explicit representation of measurable random variables as the functions on the diagonals. We also note the following. Suppose we are given some (large) commutative $*$-algebra $\mathscr{A}$, and consider a subalgebra $\mathscr{B} \subset \mathscr{A}$ generated by a single element $B \in \mathscr{A}$. If we apply the map ι to $\mathscr{B}$, we obtain precisely the subset of functions on Ω that are measurable with respect to $\sigma\{\iota(B)\}$. Thus subalgebras play the same role in quantum probability as sub-σ-algebras in classical probability; they allow us to keep track of particular subsets of information.

We do not a priori have a basis for specifying a particular commutative $*$-algebra; given a quantum system, we could decide to measure any of a large set of incompatible observables. The discussion up to this point motivates the following definition.

Definition 2.5 (quantum probability space, finite-dimensional case). *A pair $(\mathscr{N}, \mathbb{P})$, where $\mathscr{N}$ is a (not necessarily commutative) $*$-algebra of operators on a finite-dimensional Hilbert space and $\mathbb{P}$ is a state on $\mathscr{N}$, is called a (finite-dimensional) quantum probability space.*

Usually we will choose $\mathscr{N}$ to be the set of all (bounded) operators $\mathscr{B}(\mathsf{H})$ on some underlying Hilbert space H. The principles of quantum probability now amount to the following. In each realization, we must make a choice of commutative $*$-subalgebra $\mathscr{A} \subset \mathscr{N}$ which fixes the observations. Every statistic that pertains to these observations (e.g., the statistics compiled by repeating the experiment many times with the same choice of $\mathscr{A}$) is now described by the classical probability model obtained through the spectral theorem. The reader should convince himself that the operational description given in the previous section fits neatly within this model (with the exception of conditioning, which we discuss in section 2.4).

Notice that in contrast to a classical probability space $(\Omega, \mathcal{F}, \mathbf{P})$, there are no sample points $\omega \in \Omega$ in a quantum probability space. The sample points emerge through the spectral theorem after the choice of a commutative $*$-subalgebra.

Example 2.6. Let us reformulate Example 2.3. Set $\mathsf{H} = \mathbf{C}^2$ and choose $\mathscr{N} = \mathscr{B}(\mathsf{H}) = M_2$, the $*$-algebra of 2×2 complex matrices. The pure state is defined by $\mathbb{P}(A) = \langle \psi, A\psi \rangle = \psi^* A \psi$ (recall that $\psi = (c_1 \ c_{-1})^T$ with $|c_1|^2 + |c_{-1}|^2 = 1$).

The observable σ_z, used to represent spin measurement in the z direction, generates a commutative $*$-subalgebra $\mathscr{A}_z \subset \mathscr{N}$. It is not difficult to see that $\mathscr{A}_z$ is simply the linear span of the events $P_{z,1}$ and $P_{z,-1}$. Let us now apply the spectral theorem; we obtain the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ where $\Omega = \{1, 2\}$, $\mathcal{F} = \{\varnothing, \{1\}, \{2\}, \Omega\}$, $\mathbf{P}(\{1\}) = |c_1|^2$, etc., and $\iota(P_{z,1}) = \chi_{\{1\}}$, $\iota(P_{z,-1}) = \chi_{\{2\}}$. In particular, the random variable $\iota(\sigma_z) : (1, 2) \mapsto (1, -1)$ has precisely the right properties.

Now suppose we do not wish to measure the intrinsic angular momentum (spin)

in the z direction, but in the x direction. This corresponds to the observable

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{2.9}$$

which has the spectral decomposition $\sigma_x = P_{x,1} - P_{x,-1}$ with

$$P_{x,1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad P_{x,-1} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

The observable σ_x also generates a commutative $*$-subalgebra $\mathscr{A}_x = \mathrm{span}\{P_{x,1}, P_{x,-1}\}$ to which we can apply the spectral theorem. However, as σ_x and σ_z do not commute, they cannot be jointly represented on a classical probability space through the spectral theorem. In other words, σ_x and σ_z are incompatible and their joint statistics are undefined; hence they cannot both be observed in the same realization. □

To conclude this section, let us say a few words about the interpretation of the Heisenberg uncertainty relation. The relation says that the product of the variances of two noncommuting observables is bounded from below by a positive constant. It is important to realize, however, that the two observables cannot be measured in the same realization as they are incompatible—in particular, the *co*variance of the observables is undefined. Rather, the uncertainty relation is a statement about the properties of quantum states: for any state, the statistics of the two observables, compiled in the course of separate realizations in each of which only one of the observables is measured, must obey the Heisenberg inequality.[3]

2.3. Composite systems. We will often wish to form a composite probability model from two separate probability spaces. In classical probability theory, two probability spaces $(\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbf{P}_2)$ can be merged into a single probability space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, \mathbf{P}_1 \times \mathbf{P}_2)$ where $\mathbf{P}_1 \times \mathbf{P}_2$ is the product measure. We now briefly describe the noncommutative counterpart.

Consider a composite system constructed from two quantum probability spaces $(\mathscr{N}_1, \mathbb{P}_1)$, $(\mathscr{N}_2, \mathbb{P}_2)$ of operators on the Hilbert spaces H_1 and H_2, respectively. The composite quantum probability space consists of operators on the tensor product Hilbert space $\mathsf{H}_1 \otimes \mathsf{H}_2$; for vectors $\psi_1, \phi_1 \in \mathsf{H}_1$ and $\psi_2, \phi_2 \in \mathsf{H}_2$, the inner product on $\mathsf{H}_1 \otimes \mathsf{H}_2$ is given by

$$\langle \psi_1 \otimes \psi_2, \phi_1 \otimes \phi_2 \rangle = \langle \psi_1, \phi_1 \rangle \langle \psi_2, \phi_2 \rangle,$$

which is extended by linearity to any vector in the tensor product space. The algebra $\mathscr{N}_1 \otimes \mathscr{N}_2$ is generated by elements of the form

$$(A_1 \otimes A_2)(\psi_1 \otimes \psi_2) = A_1\psi_1 \otimes A_2\psi_2,$$

where $A_1 \in \mathscr{N}_1$ and $A_2 \in \mathscr{N}_2$. Finally, the product state is defined by

$$(\mathbb{P}_1 \otimes \mathbb{P}_2)(A_1 \otimes A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$$

[3]In the physics literature one often find statements to the effect that the Heisenberg uncertainty relation limits the precision with which we can "imperfectly" observe two noncommuting observables simultaneously, i.e., within the same realization. This is a misconception. Though the idea of an imperfect measurement can be implemented rigorously (see, e.g., [38]), this gives rise to an uncertainty relation which is different from Heisenberg's uncertainty relation [4].

and is extended by linearity. The quantum probability space $(\mathscr{N}_1 \otimes \mathscr{N}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$ of operators on the Hilbert space $\mathsf{H}_1 \otimes \mathsf{H}_2$ describes the composite system. The reader should verify that if $\mathscr{N}_1$ and $\mathscr{N}_2$ are commutative, then applying the spectral theorem to the composite system is equivalent to applying the spectral theorem to the individual subsystems, then forming the composite classical probability space.

2.4. Conditional expectations. Let us recall for a moment the Stern–Gerlach experiment of Examples 2.3 and 2.6. We have introduced the observables σ_z and σ_x, corresponding to spin in the z and x directions. These observables are incompatible, so we cannot measure them in the same realization. Recall that in order to measure σ_z, Stern and Gerlach apply a field gradient in the z direction; the atom then acquires momentum in that direction proportional to σ_z, and we can determine the value of σ_z in that realization by observing whether the atom is deflected up (1) or down (-1). Similarly, σ_x is measured by orienting the field gradient along the x axis.

We would not be measuring both σ_z and σ_x by applying both field gradients simultaneously, but rather as magnetic fields add vectorially, this would measure the spin in some other direction in the x-z plane whose observable commutes with neither σ_z nor σ_x. On the other hand, we could first apply the field gradient in the z direction until we can resolve σ_z, then turn this field off and switch on a field in the x direction to resolve σ_x. It is a characteristic feature of quantum mechanics that the measurement outcomes in such a procedure can differ drastically depending on what order we apply the fields. It is thus of crucial importance to specify precisely how such measurements are performed by including in the quantum probability space a model of the measurement apparatus (or *probe*).

We defer the discussion of the Stern–Gerlach measurement with magnetic fields until we have developed the necessary machinery in section 3. For the sake of example, we develop in this section a simpler probe model which shows the main features of the procedure. We will see that this probe model, together with the concept of conditional expectations, reproduces precisely the traditional projection postulate of section 2.1.

Let us begin by discussing *conditional expectations* in the noncommutative context. The key observation we need is the following. The conditional probability of an event B given an event A is the probability that B is true given that A is true in the same realization. Hence the concept of conditioning inherently makes sense only in the context of quantities that can be observed in the same realization of an experiment. This means that we can only define conditional expectations in commutative subalgebras of a quantum probability space; but as long as we are restricted to the commutative case, the spectral theorem allows us to define any probabilistic operation directly in terms of the associated classical probability space (see [18]).

To be more precise, let $(\mathscr{N}, \mathbb{P})$ be a quantum probability space, $\mathscr{A} \subset \mathscr{N}$ be a commutative subalgebra, and $B \in \mathscr{N}$ be a self-adjoint element that commutes with every $A \in \mathscr{A}$. Then B and $\mathscr{A}$ generate a larger commutative subalgebra $\mathscr{C} \subset \mathscr{N}$, to which we can apply the spectral theorem to obtain a $*$-isomorphism ι. The conditional expectation is now simply inherited from the classical space as $\mathbb{P}(B|\mathscr{A}) = \iota^{-1}(E_{\mathbf{P}}(\iota(B)|\sigma\{\iota(\mathscr{A})\}))$. Note, however, that if B, C are two self-adjoint operators that commute with every $A \in \mathscr{A}$, this does not necessarily imply that B and C commute. The set $\mathscr{A}' = \{B \in \mathscr{N} : AB = BA \ \forall A \in \mathscr{A}\}$, the *commutant* of $\mathscr{A}$ (in $\mathscr{N}$), is the largest $*$-subalgebra of operators that can be conditioned on $\mathscr{A}$. The conditional expectation is defined as above for its self-adjoint elements, and extends to all of $\mathscr{A}'$ by linearity.

From this discussion and the definition of the classical conditional expectation, we extract the following definition directly in terms of the quantum probability space.

DEFINITION 2.7 (conditional expectation, finite-dimensional case). *Let $(\mathscr{N}, \mathbb{P})$ be a finite-dimensional quantum probability space and let $\mathscr{A} \subset \mathscr{N}$ be a commutative $*$-subalgebra. Then $\mathbb{P}(\cdot|\mathscr{A}) : \mathscr{A}' \to \mathscr{A}$ is called (a version of) the conditional expectation from $\mathscr{A}'$ onto $\mathscr{A}$ if $\mathbb{P}(\mathbb{P}(B|\mathscr{A})A) = \mathbb{P}(BA)$ for all $A \in \mathscr{A}$, $B \in \mathscr{A}'$.*

As we will see in section 3, the discussion above generalizes directly to the infinite-dimensional case. In finite dimensions it is convenient to give an explicit expression for the conditional expectation. Note that a finite-dimensional $*$-algebra is a finite-dimensional linear space. Then $\langle A, B\rangle_{\mathbb{P}} = \mathbb{P}(A^*B)$ turns the algebra into a pre-Hilbert space; i.e., it is a Hilbert space except that $A \mapsto \langle A, A\rangle_{\mathbb{P}} = \|A\|_{\mathbb{P}}^2$ may have a nontrivial null space. In particular, the fundamental property $\mathbb{P}(\mathbb{P}(B|\mathscr{A})A) = \mathbb{P}(BA)$ for all $A \in \mathscr{A}$ is precisely that of orthogonal projection from $\mathscr{A}'$ onto the linear subspace $\mathscr{A}$, which in a pre-Hilbert space is uniquely determined up to an event of zero probability. Note that the classical characterization of $\mathbb{P}(B|\mathscr{A})$ as the least mean square estimate of B in $\mathscr{A}$ follows immediately. We will elaborate on this point in section 3.

An explicit expression for $\mathbb{P}(B|\mathscr{A})$ is easily obtained if we find an orthogonal basis for $\mathscr{A}$. Any commutative $*$-algebra in finite dimensions is spanned by a set of projections that resolve the identity. This is easily seen: in n dimensions any self-adjoint operator is a linear combination of at most n projections that resolve the identity, and as all the operators in the $*$-algebra commute they must be expressible as linear combinations of the same projections. Let $\mathscr{A} = \operatorname{span}\{P_a\}$ for some set of orthogonal projections P_a resolving the identity. Then a version of the conditional expectation is given by

$$\mathbb{P}(B|\mathscr{A}) = \sum_{P \in \{P_a\} : \mathbb{P}(P) \neq 0} \frac{P}{\|P\|_{\mathbb{P}}} \left\langle \frac{P}{\|P\|_{\mathbb{P}}}, B \right\rangle_{\mathbb{P}} = \sum_{P \in \{P_a\} : \mathbb{P}(P) \neq 0} \frac{\mathbb{P}(PB)}{\mathbb{P}(P)} P. \tag{2.10}$$

Note what could happen if we naively fill in some $B \notin \mathscr{A}'$. Then $\langle P, B\rangle_{\mathbb{P}} \neq \langle B, P\rangle_{\mathbb{P}}$ for some $P \in \{P_a\}$, which implies that we obtain complex coefficients in the sum even if B is an observable. Hence the expression does not make sense unless $B \in \mathscr{A}'$.

Example 2.8. This example serves to illustrate conditional expectations; it is not meant to represent a particular physical scenario. Consider $\mathsf{H} = \mathbf{C}^3$, $\mathscr{N} = M_3$, and $\mathbb{P}(X) = \langle \psi, X\psi\rangle$ with $\psi = (1\ 1\ 1)^T/\sqrt{3}$. Define $A, B \in \mathscr{N}$ by

$$A = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 5 \end{pmatrix} = 4 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + 5 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Let $\mathscr{A}$ be the $*$-algebra generated by A. Then

$$\mathscr{A}' = \left\{ \begin{pmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & x \end{pmatrix} \ : \ a, b, c, d, x \in \mathbf{C} \right\}.$$

Note that $\mathscr{A}'$ is not a commutative algebra, despite that every element of $\mathscr{A}'$ commutes with every element of $\mathscr{A}$. As $B \in \mathscr{A}'$, we can use (2.10) to calculate

$$\mathbb{P}(B|\mathscr{A}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} = 1 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + 2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \in \mathscr{A}.$$

The observable $\mathbb{P}(B|\mathscr{A})$ is the orthogonal projection of B onto $\mathscr{A}$ with respect to the inner product $\langle A, B\rangle_{\mathbb{P}} = \mathbb{P}(A^*B)$. By the projection theorem, $\mathbb{P}(B|\mathscr{A})$ is an element of $\mathscr{A}$ that minimizes the mean square error $\|B - \mathbb{P}(B|\mathscr{A})\|_{\mathbb{P}}$. □

We now proceed to develop a simple probe model that reproduces the projection postulate. Recall that the conditional *probability* of an event P given a commuting event Q is simply given by $\mathbb{P}(PQ)/\mathbb{P}(Q)$. This is equivalent to $\mathbf{P}(A \cap B)/\mathbf{P}(B)$ by the spectral theorem, where A and B are the sets corresponding to P and Q.

Example 2.9. *Simple probe model.* We will work in a generic n-dimensional setting, $n < \infty$. Let $\mathsf{H} = \mathbf{C}^n$, $\mathscr{N} = M_n$ (the set of $n \times n$ complex matrices), and let $\mathbb{P}(X) = \mathrm{Tr}[\rho X]$ be some state on $\mathscr{N}$. Let A, B be two observables in $\mathscr{N}$ that do not commute. Hence we cannot measure A and B directly in the same realization. However, we can have the system interact with an external probe system, in such a way that the observable A is copied to some probe observable A' after the interaction. If A' commutes with B, we interpret this procedure (like in the Stern–Gerlach example) as an (indirect) measurement of A followed by a (direct) measurement of B.

The strategy is simple. First, we describe the probe system by a separate probe quantum probability space $(\mathscr{N}_p, \mathbb{P}_p)$ and form the composite space $(\mathscr{N} \otimes \mathscr{N}_p, \mathbb{P} \otimes \mathbb{P}_p)$. Next, we introduce an interaction. Recall from section 2.1 that the evolution of an isolated system is described by a unitary transformation. Hence, we will choose a probe observable $I \otimes A'$ and construct a suitable unitary operator U so that the probe observable $U^*(I \otimes A')U$ after the interaction gives the same outcome as $A \otimes I$ would have before the interaction. Note that, by construction, the system observable $B \otimes I$ commutes with $I \otimes A'$ after the interaction, $[U^*(I \otimes A')U, U^*(B \otimes I)U] = 0$. Hence we can measure them within the same realization.

We now fill out the details of this model. Let $A = \sum_{a \in \mathrm{spec}(A)} a\, P_a$, and we denote by m the number of elements in $\mathrm{spec}(A)$ (the number of possible measurement outcomes). For the probe algebra, we choose $\mathsf{H}_p = \mathbf{C}^m$, $\mathscr{N}_p = M_m$. Now fix an observable $A' \in \mathscr{N}_p$ that has the same spectrum as A. $A' = \sum_{a \in \mathrm{spec}(A)} a\, P'_a$ and that P'_a are projections onto one-dimensional subspaces of H_p; hence we can fix an orthonormal basis of vectors $\psi_a \in \mathsf{H}_p$ such that $P'_a = \psi_a \psi_a^*$. Now define the operator $X'_{ab} = \psi_b \psi_a^* + \psi_a \psi_b^* + \sum_{c \neq a,b} \psi_c \psi_c^* \in \mathscr{N}_p$ for $a \neq b$, and $X'_{aa} = I$; these operators switch the events P'_a and P'_b in the sense $X'_{ab} P'_a X'_{ab} = P'_b$, $X'_{ab} P'_b X'_{ab} = P'_a$, and $X'_{ab} P'_c X'_{ab} = P'_c$ for $c \neq a, b$. Finally, set $\mathbb{P}_p(X) = \mathrm{Tr}[X P'_p]$ where we have fixed some $p \in \mathrm{spec}(A)$ at the outset.

Now consider the operator $U \in \mathscr{N} \otimes \mathscr{N}_p$ defined by $U = \sum_{a \in \mathrm{spec}(A)} P_a \otimes X'_{ap}$. As $(X'_{ap})^2 = I$ it follows that $U^*U = UU^* = U^2 = I$; i.e., U is unitary. Note that $U^*(I \otimes P'_c)U = P_c \otimes P'_p + (1 - P_c) \otimes P'_c$ if $c \neq p$, $U^*(I \otimes P'_p)U = \sum_a P_a \otimes P'_a$. We calculate $(\mathbb{P} \otimes \mathbb{P}_p)(U^*(I \otimes P'_c)U(P_c \otimes I))/(\mathbb{P} \otimes \mathbb{P}_p)(P_c \otimes I) = 1$ for every c; i.e., the conditional probability that $U^*(I \otimes A')U$ gives the outcome c, given that we have observed $A \otimes I$ with outcome c, is one. Thus the unitary interaction U precisely copies the system observable A onto the probe observable A'.

We can now measure the system observable B after interaction with the probe. In particular, let us calculate the expectation of B conditioned on the probe measurement. Define $\mathscr{A}$ as the commutative $*$-algebra generated by $U^*(I \otimes A')U$, and note that $U^*(B \otimes I)U \in \mathscr{A}'$. Thus we can use (2.10) to calculate

$$(\mathbb{P} \otimes \mathbb{P}_p)(U^*(B \otimes I)U|\mathscr{A}) = \sum_c \frac{(\mathbb{P} \otimes \mathbb{P}_p)(U^*(B \otimes P'_c)U)}{(\mathbb{P} \otimes \mathbb{P}_p)(U^*(I \otimes P'_c)U)}\, U^*(I \otimes P'_c)U$$

$$= \sum_c \frac{\mathbb{P}(P_c B P_c)}{\mathbb{P}(P_c)}\, U^*(I \otimes P_c')U = \sum_c \mathrm{Tr}[\rho_c B]\, U^*(I \otimes P_c')U,$$

where $\rho_c = P_c \rho P_c / \mathrm{Tr}[\rho P_c]$. This is precisely the projection postulate of section 2.1.

This example may be somewhat bewildering, and we encourage the reader to work through the procedure for a particular model (e.g., that of Example 2.8), paying particular attention to which operators do and do not commute. The reader should convince himself that different answers are obtained if one first measures B then A.

Finally, we note that though we have here measured A through a probe and B directly, there is no reason to stop here. If, in addition to A and B, we want to measure an observable C that does not commute with B, we would introduce a second probe to measure B as well. Now suppose that $C = A$. If we first measure A through the probe, then measure A again, we would (obviously) obtain the same outcome. However, if we first probe A, then probe B, and then measure A, we obtain a different outcome from that of the first measurement of A! The reader is encouraged to work out also this case. The reason for this phenomenon is that the interaction with the probe that is used for the observation of B disturbs the system in such a way that its value of A is changed. This effect is known as "measurement back action." □

The previous example, in particular the construction of the probe and the corresponding interaction, may seem rather ad hoc, and indeed we have only chosen this rather artificial example to reproduce the projection postulate. This is not a shortcoming of the theory we have outlined, however, but rather highlights the importance of including a reasonable model of the probe in the quantum probability space. Indeed, most realistic measurement setups are not of this type and the projection postulate of section 2.1 cannot be used to describe such systems. For example, we will see in section 3 that the Stern–Gerlach measurement is only approximately described by the projection postulate. Later we will describe even more complicated optical measurements in which we wish to condition system observables based on the observation of stochastic processes in continuous time (the signal from a photodetector). It is the latter, most practically useful case where we need quantum filtering theory.

Remark 2.10. It is important to realize that statements like the projection postulate do not really implement the notion of conditioning; they consist of a pure conditioning component and of a particular physical probe model which has no statistical significance. One also finds in the literature generalizations of the projection postulate, called instruments, which implement different types of probes [25, 40]. In the quantum probability context of this paper it is most natural to separate the two parts; we will take existing probe models from physics and concentrate on the calculation of the associated conditional expectations (filtering). □

3. Noncommutative probability theory. In the finite-dimensional case, we have seen in section 2 that quantum mechanics can be modeled as a noncommutative probability theory. In this section we present a general formulation for quantum probability that has wide applicability. We give a general definition of quantum probability space, prove the existence and uniqueness of conditional expectations, and prove a quantum version of Bayes' rule that is very helpful for quantum filtering.

Almost all of the features of the full theory can already be seen in the finite-dimensional case discussed in section 2; the main difficulties in the general case are the technicalities involved in the theory of infinite-dimensional Hilbert spaces. This parallels the difficulties in classical probability theory—though finite-state random

variables can be treated by almost trivial (counting, combinatoric) methods, the description of continuous random variables requires us to upgrade our machinery using methods of real analysis. Similarly, the elementary linear algebra that underlies finite-dimensional quantum probability must be upgraded to functional analysis if we wish to treat the infinite-dimensional case. Conceptually, however, the two cases are very similar, and the reader is encouraged to develop an intuitive understanding of the finite-dimensional case before tackling the full formalism. For a thorough introduction to functional analysis we refer the reader to the excellent textbook [56].

3.1. Quantum probability spaces. Let H be a complex Hilbert space, and denote by $\mathscr{B}(\mathsf{H})$ the set of all bounded (linear) operators on H. We restrict ourselves (for the time being) to bounded operators as we wish to construct $*$-algebras of such operators: attempting to do this with unbounded operators would get us into no end of trouble, as we would surely run into domain problems. Recall that for $A \in \mathscr{B}(\mathsf{H})$, the usual Hilbert space adjoint $A^* \in \mathscr{B}(\mathsf{H})$ is defined by $\langle \psi, A\phi\rangle = \langle A^*\psi, \phi\rangle \; \forall \psi, \phi \in \mathsf{H}$. With this involution $\mathscr{B}(\mathsf{H})$ is a $*$-algebra in the sense of section 2.

We wish to introduce a structure that plays the same role as a $*$-algebra in the finite-dimensional case. It turns out, however, that the $*$-algebra structure in itself is not sufficient in the infinite-dimensional case; we need to impose an additional technical condition in order to be able to prove an infinite-dimensional version of the spectral theorem (Theorem 2.4). The additional condition has a natural interpretation which we will discuss below; however, the reader should not be too worried about this technicality, particularly if he is not familiar with nets or locally convex topologies. In practice we will rarely need to verify this property directly.

DEFINITION 3.1. *A positive linear functional $\mu : \mathscr{B}(\mathsf{H}) \to \mathbf{C}$ is said to be* normal *if $\mu(\sup_\alpha A_\alpha) = \sup_\alpha \mu(A_\alpha)$ for any upper bounded increasing net $\{A_\alpha\}$ of positive elements in $\mathscr{B}(\mathsf{H})$. The locally convex topology on $\mathscr{B}(\mathsf{H})$ defined by the family of seminorms $\{A \mapsto |\mu(A)| : \mu \text{ normal}\}$ is called the* normal topology.

For a detailed discussion of nets, locally convex topologies, etc., see [56].

DEFINITION 3.2 (von Neumann algebra). *A von Neumann algebra $\mathscr{N}$ is a $*$-subalgebra of $\mathscr{B}(\mathsf{H})$ that is closed in the normal topology. A state $\mathbb{P}$ on $\mathscr{N}$ is normal if it is the restriction to $\mathscr{N}$ of a normal state on $\mathscr{B}(\mathsf{H})$.*

We can now extend the spectral theorem to the infinite-dimensional case, essentially showing that commutative von Neumann algebras with normal states are equivalent to classical probability spaces. See, e.g., [57, Proposition 1.18.1] for a proof. Conceptually, we are guided by the finite-dimensional case; Theorem 3.3 extends the idea of simultaneous diagonalization to infinite-dimensional operators. Though technically much more involved, the flavor of the procedure remains the same.[4]

THEOREM 3.3 (spectral theorem). *Let $\mathscr{C}$ be a commutative von Neumann algebra. Then there is a measure space $(\Omega, \mathcal{F}, \mu)$ and a $*$-isomorphism ι from $\mathscr{C}$ to $L^\infty(\Omega, \mathcal{F}, \mu)$, the algebra of bounded measurable complex functions on Ω up to μ-a.s. equivalence. Moreover, a normal state $\mathbb{P}$ on $\mathscr{C}$ defines a probability measure $\mathbf{P}$, which is absolutely continuous with respect to μ such that $\mathbb{P}(C) = E_{\mathbf{P}}(\iota(C))$ for all $C \in \mathscr{C}$.*

Before we continue, let us demonstrate the significance of the additional technical conditions on a von Neumann algebra. First, we give an example of a $*$-subalgebra

[4]The additional measure μ that shows up in the theorem has no direct physical significance; its job is to identify "enough" null sets in $L^\infty(\Omega)$ so we can construct the $*$-isomorphism ι. We can generally not use $\mathbf{P}$ for this purpose as there may be projections $P \in \mathscr{C}$ with $\mathbb{P}(P) = 0$; if ι were to map to $L^\infty(\Omega, \mathcal{F}, \mathbf{P})$, then necessarily $\iota(P) = 0$ and hence ι would not be invertible. The precise details of the construction are never an issue, as we will never use μ and only prove results $\mathbf{P}$-a.s.

of $\mathscr{B}(\mathsf{H})$ that is not a von Neumann algebra.

Example 3.4. Let $\mathsf{H} = L^2([0,1])$ and $\mathscr{A} = C([0,1])$, the commutative algebra of continuous functions on the unit interval. We can consider $A \in \mathscr{A}$ as an operator on H under pointwise multiplication; i.e., $(A\psi)(x) = A(x)\psi(x)$ for every $\psi \in \mathsf{H}$. Then $\mathscr{A}$ satisfies all the requirements of a von Neumann algebra except that it is not closed in the normal topology. Indeed, one can construct, for example, an increasing sequence of continuous functions that converges to $\chi_{[0,1/2]}$, which is discontinuous.

The problem is that the only indicator functions in $\mathscr{A}$ are $\chi_\varnothing$ and $\chi_{[0,1]}$: all other indicator functions on $[0,1]$ are discontinuous. Hence from a probabilistic point of view $\mathscr{A}$ defines a trivial theory, as the only events in $\mathscr{A}$ are the trivial ones. Nonetheless $\mathscr{A}$ is much larger than the algebra $\mathbf{C}$ that is generated by $\chi_\varnothing$ and $\chi_{[0,1]}$. Hence $\mathscr{A}$ cannot be $*$-isomorphic to the set of measurable functions on some measure space. The role of normal closure is to avoid this complication. Indeed, this property guarantees that any von Neumann algebra is generated by its projections [45]. □

Like normal closure, normality of the state is also required in order for the spectral theorem to hold. Note that for normal states the expectation of an increasing set of observables converges to the expectation of their least upper bound; i.e., the monotone convergence property holds. This corresponds to the more basic property of countable additivity. In the following example we construct a state which is not normal.

Example 3.5. Let $\mathsf{H} = \ell^2(\mathbf{N})$ and $\mathscr{A} = \ell^\infty(\mathbf{N})$, acting on H by pointwise multiplication. $\mathscr{A}$ is closed in the normal topology; i.e., it is a commutative von Neumann algebra. Now introduce a state on $\mathscr{A}$ which is given by the expression[5]

$$\mathbb{P}(A) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} A(n), \qquad A \in \mathscr{D} \subset \mathscr{A}, \tag{3.1}$$

on a suitably chosen linear subspace $\mathscr{D}$. $\mathbb{P}$ is not a normal state; to see this, let us introduce the events $P_n \in \mathscr{A}$ defined by $(P_n\psi)(k) = \psi(k)$ if $k \le n$, and zero otherwise. $\{P_n\}$ is an increasing sequence of projections in $\mathscr{A}$ whose least upper bound is the identity $P_\infty = I$. However, straightforward calculation shows that $\mathbb{P}(P_n) = 0$ for any finite n, whereas $\mathbb{P}(I) = 1$. We conclude that the state $\mathbb{P}$ is not normal.

Note that what we have constructed is precisely the classical model of a uniform distribution over the natural numbers $\mathbf{N}$. This does not give rise to a well-defined probability model in the sense of Kolmogorov, however, as the uniform distribution on $\mathbf{N}$ does not obey the property that the probability of a countable union of disjoint events is the sum of the probabilities of these events (which is exactly what went wrong above). Requiring that the state be normal is equivalent to requiring that it gives rise to a countably additive measure [46], which rules out our example. □

Remark 3.6. Definition 3.2 is one of many equivalent definitions of a von Neumann algebra. We have emphasized normality as it is close to the probabilistic notion of monotone convergence. Normal closure turns out to be equivalent to closure in several other topologies, notably the weak and strong operator topologies on $\mathscr{B}(\mathsf{H})$. We will not concern ourselves with topological issues in this article; see, e.g., [20, section 2.4].

The following definition should come as no surprise.

Definition 3.7 (quantum probability space). *A quantum probability space is a pair $(\mathscr{N}, \mathbb{P})$, where $\mathscr{N}$ is a von Neumann algebra and $\mathbb{P}$ is a normal state.*

[5] Equation (3.1) does not by itself define a state, as there are many $A \in \mathscr{A}$ for which the limit does not exist. However, note that (3.1) is well defined on a linear subspace, e.g., $\mathscr{D} = \{A \in \mathscr{A} : \exists c \in \mathbf{C} \text{ s.t. } \lim_{n\to\infty} A(n) = c\}$. Now $\mathbb{P}$ can be extended from $\mathscr{D}$ to $\mathscr{A}$ using the Hahn–Banach theorem.

The structure has precisely the same interpretation as in section 2, of which we briefly remind the reader. In each realization we must choose a commutative von Neumann subalgebra $\mathscr{A} \subset \mathscr{N}$ which fixes the observations. Every statistic that pertains to these observations is then described by the classical probability model obtained by applying the spectral theorem to $(\mathscr{A}, \mathbb{P})$. The equivalence between commutative quantum probability spaces and classical probability spaces is the foundation of the theory; a commutative quantum probability model *is* a classical probabilistic model, and we will often implicitly identify these two pictures.

In this article we will only use three types of von Neumann algebras. We list these below; they will be used throughout without comment.

(i) $\mathscr{A} = \mathscr{B}(\mathsf{H})$ is a von Neumann algebra. Moreover, any *vector state* on $\mathscr{A}$ ($\mathbb{P}(A) = \langle \psi, A\psi \rangle$ for fixed $\psi \in \mathsf{H}$), or any convex combination of vector states, is a normal state. Many models from quantum mechanics are described by such a model.

(ii) $\mathscr{A} = L^\infty(\Omega, \mathcal{F}, \mathbf{P})$, acting on $\mathsf{H} = L^2(\Omega, \mathcal{F}, \mathbf{P})$ by pointwise multiplication, is a commutative von Neumann algebra. Moreover, any state of the form $\mathbb{P}(X) = E_{\mathbf{P}}(X)$ is a normal state. This is a *classical probability model*.

(iii) Given $\mathscr{S} \subset \mathscr{B}(\mathsf{H})$, recall that $\mathscr{S}' = \{X \in \mathscr{B}(\mathsf{H}) : XS = SX \ \forall S \in \mathscr{S}\}$ is called the *commutant* of $\mathscr{S}$ in $\mathscr{B}(\mathsf{H})$. The following theorem (see [45, Theorem 5.3.1] for a proof) allows us to construct von Neumann subalgebras of $\mathscr{B}(\mathsf{H})$.

THEOREM 3.8 (double commutant theorem). *Let $\mathscr{S} \subset \mathscr{B}(\mathsf{H})$ be any self-adjoint set, i.e., $S \in \mathscr{S} \Rightarrow S^* \in \mathscr{S}$. Then $\mathscr{A} = \mathscr{S}''$ is the smallest von Neumann subalgebra of $\mathscr{B}(\mathsf{H})$ that contains $\mathscr{S}$. In particular, $\mathscr{S}$ is a von Neumann algebra iff $\mathscr{S} = \mathscr{S}''$.*

Given any $\mathscr{S} \subset \mathscr{B}(\mathsf{H})$, we call $\mathrm{vN}(\mathscr{S}) = (\mathscr{S} \cup \mathscr{S}^*)''$ the von Neumann algebra generated by $\mathscr{S}$. We will repeatedly use this construction in the following. For example, suppose that we decide to measure in one realization some commuting set of observables $A_1, \ldots, A_n$. Then $\mathscr{A} = \mathrm{vN}(A_1, \ldots, A_n)$ is a commutative von Neumann algebra which, through the spectral theorem, describes the associated classical probability model. $\mathscr{A}$ is the quantum probability equivalent of the σ-algebra generated by a set of random variables.

3.2. Random variables. Now that we have a general definition of a quantum probability space, we can develop some tools to deal with random variables. Recall from section 2 that any self-adjoint element of a quantum probability space can be decomposed into events using (2.4), which gives its interpretation as an observable (random variable). Let us show how to do this in the infinite-dimensional case.

Let $(\mathscr{N}, \mathbb{P})$ be a quantum probability space and consider an element $A \in \mathscr{N}$ which is self-adjoint $A = A^*$. Then $\mathscr{A} = \mathrm{vN}(A) \subset \mathscr{N}$ is a commutative von Neumann algebra. By the spectral theorem, there is a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a $*$-isomorphism ι that maps A to some (measurable) random variable $a : \Omega \to \mathbf{R}$. We can now apply classical probability theory; in particular, for any Borel set $B \in \mathcal{B}$ we have the event $[a \in B] = \{\omega \in \Omega : a(\omega) \in B\} = a^{-1}(B) \in \mathcal{F}$. To map this event back to $\mathscr{A}$ we simply invert ι; the projection corresponding to $[a \in B]$ is denoted by $P_A(B) = \iota^{-1}(\chi_{[a \in B]})$, and we call the map P_A from $\mathcal{B}$ to the projections in $\mathscr{N}$ the *spectral measure* of A. But this object is a familiar one from functional analysis [56]; in fact, it is well known that we can express A in terms of its spectral measure by

$$A = \int_{\mathbf{R}} \lambda \, P_A(d\lambda), \tag{3.2}$$

where the integral is defined in a suitable sense [56, sections 7.3 and 8.3]. Equation (3.2) is precisely the infinite-dimensional counterpart of (2.4). We emphasize the

physical interpretation of $P_A(B)$: it is the event [A takes a value in B], which occurs with probability $\mathbb{P}(P_A(B))$.

This would be all there is to it, were it not for the fact that our algebras contain only bounded operators (recall that unbounded operators cannot be defined on the entire Hilbert space, and hence cannot be added or multiplied at will). Evidently we did not lose much by this choice, as the probabilistic model is already contained in an algebra of bounded operators by the spectral theorem. An unfortunate side effect, however, is that self-adjoint operators in the algebra can only represent *bounded* random variables, whereas many observations of interest are quite naturally unbounded (think of a Gaussian random variable). This means that we need to deal with unbounded observables separately. We briefly discuss one way of doing this.

Consider a von Neumann algebra $\mathscr{N} \subset \mathscr{B}(\mathsf{H})$. In general, an observable is defined by a (not necessarily bounded) self-adjoint operator A on some dense domain in H. We need to relate the unbounded operator A to $\mathscr{N}$. The trick we use is remarkably simple: we compute a bounded function of A. Define $T_A = (A+iI)^{-1}$. By elementary spectral theory [56], any self-adjoint A has a real spectrum, and hence $A+iI$ is invertible with bounded inverse. We say that A is *affiliated* to $\mathscr{N}$ if $T_A \in \mathscr{N}$. This is the equivalent of the classical notion of a random variable that is measurable with respect to some σ-algebra $\mathcal{G}$. Note that every self-adjoint A is affiliated to $\mathscr{B}(\mathsf{H})$, and if A is also bounded, then A is affiliated to $\mathscr{N}$ iff $A \in \mathscr{N}$.

We wish to represent A as a classical (unbounded) random variable. To this end, define the von Neumann algebra generated by A as $\mathrm{vN}(A) = \mathrm{vN}(T_A)$. Now note that T_A commutes with its adjoint, and hence $\mathrm{vN}(A)$ is a commutative von Neumann algebra to which we can apply the spectral theorem. All we need to do is to "package" A into T_A, apply ι, and "unpack" it on the other end; in other words, we define $\iota(A) = \iota(T_A)^{-1} - i$. Once we have done this, we can define a spectral measure P_A for A in the usual way, and indeed (3.2) still holds even for unbounded A [56]. We remark that A being affiliated to $\mathscr{N}$ corresponds to the fact that $P_A(B) \in \mathscr{N}$ for every $B \in \mathcal{B}$; this is precisely the classical notion of measurability.

Unbounded operators are a nuisance, but unfortunately they are a fact of life in mathematical physics. In this article, particularly in the later sections, we will occasionally add and multiply unbounded operators without justification; a detailed analysis of the operator domains is beyond our scope. Though this does not often cause trouble, the reader should keep in mind that a fully rigorous treatment must verify that any addition or multiplication of unbounded operators is indeed well defined. We quote one useful result: operators affiliated to a *commutative* von Neumann algebra can be added and multiplied at will (see [45, Theorem 5.6.15] and [54]).

Example 3.9. We take $\mathsf{H} = L^2(\mathbf{R})$ and $\mathscr{N} = \mathscr{B}(\mathsf{H})$. The vector

$$\psi \in \mathsf{H}, \qquad \psi(x) = (2\pi)^{-1/4}\sigma^{-1/2}\exp\left(-\frac{(x-\mu)^2}{4\sigma^2}\right)$$

defines the (pure) state $\mathbb{P}(X) = \langle \psi, X\psi \rangle$. Now consider the self-adjoint operators

$$(Q\psi)(x) = x\psi(x), \qquad (P\psi)(x) = -i\hbar\frac{d}{dx}\psi(x),$$

which are prototypical observables for the position Q and momentum P of a quantum particle. Both are unbounded observables, but their domains include at least the set of smooth functions with compact support which is dense in $L^2(\mathbf{R})$.

What random variables do these represent? We can read off from the definition that Q is a Gaussian random variable with mean μ and variance σ^2—as Q is already in "diagonal" form (Q is affiliated to $L^\infty(\mathbf{R}) \subset \mathscr{N}$), its spectral measure is given by

$$(P_Q(B)\psi)(x) = \chi_B(x)\psi(x)$$

and it is evident that $\mathbb{P}(P_Q(B))$ is a Gaussian measure with mean μ and variance σ^2. Alternatively, consider the characteristic function $q(k) = \mathbb{P}(e^{ikQ})$ of Q. Unlike Q, e^{ikQ} is a bounded operator, and we can directly compute

$$q(k) = \langle \psi, e^{ikQ}\psi\rangle = (2\pi)^{-1/2}\sigma^{-1}\int_{-\infty}^{\infty} e^{ikx}\, e^{-(x-\mu)^2/2\sigma^2}\, dx = e^{ik\mu - k^2\sigma^2/2}$$

which is the characteristic function of a Gaussian random variable with mean μ and variance σ^2. Similarly, e^{ikP} is a bounded operator, and we compute

$$p(k) = \mathbb{P}(e^{ikP}) = \langle \psi, e^{ikP}\psi\rangle = \int_{-\infty}^{\infty} \psi(x)\,\psi(x+\hbar k)\, dx = e^{-\hbar^2 k^2/8\sigma^2}$$

which is the characteristic function of a Gaussian random variable with mean zero and variance $\hbar^2/4\sigma^2$. Thus both Q and P are Gaussian random variables, but their joint distribution is undefined as they do not commute. Note that we cannot choose σ so that both Q and P have arbitrarily small variance: this is a manifestation of the Heisenberg uncertainty relation (compare (2.3)). □

The following example plays a central role in the physics of harmonic oscillators; we will encounter a very similar construction later for continuous-time quantum stochastic processes. We will need the following classic result (see, e.g., [45] for a proof).

THEOREM 3.10 (Stone's theorem). *Let $\mathscr{N}$ be a von Neumann algebra and let $\{U_t\}_{t\in\mathbf{R}} \subset \mathscr{N}$ be a group of unitary operators that is strongly continuous. Then there is a unique self-adjoint A affiliated to $\mathscr{N}$, the* Stone generator*, such that $U_t = e^{itA}$.*

Example 3.11. Let $\mathsf{H} = \ell^2(\mathbf{N})$ and $\mathscr{N} = \mathscr{B}(\mathsf{H})$. Define the complete orthonormal basis $\{\psi_n,\ n = 0, 1, \ldots\} \subset \mathsf{H}$, where $\psi_n(k) = 1$ if $k = n$ and $\psi_n(k) = 0$ otherwise. Moreover, we define for every $\alpha \in \mathbf{C}$ the *exponential vector* $e(\alpha) \in \mathsf{H}$ by $e(\alpha)(k) = \alpha^k/\sqrt{k!}$, and we remark that the linear span D of all exponential vectors is dense in H. The normalized exponential vectors $e(\alpha)e^{-|\alpha|^2/2}$ are called coherent vectors and can be used to define the coherent states $\mathbb{P}_\alpha(X) = \langle e(\alpha), Xe(\alpha)\rangle\, e^{-|\alpha|^2}$.

The simplest random variable we can investigate is defined by $(\lambda\psi)(k) = k\psi(k)$—i.e., this is the natural diagonal operator affiliated to $\ell^\infty(\mathbf{N}) \subset \mathcal{N}$. The spectral measure of λ is given by $(P_\lambda(B)\psi)(k) = \chi_B(k)\psi(k)$, from which we obtain directly

$$\mathbb{P}_\alpha(P_\lambda(B)) = \langle e(\alpha), P_\lambda(B)e(\alpha)\rangle\, e^{-|\alpha|^2} = \sum_{k\in B} \frac{e^{-|\alpha|^2}(|\alpha|^2)^k}{k!}.$$

Thus, evidently, λ is a Poisson-distributed random variable with intensity $|\alpha|^2$.

Can we find other interesting observables affiliated to $\mathscr{N}$? In many cases, physically relevant observables are found to be the Stone generators of particular unitary symmetry groups; see, e.g., [38] for a lucid discussion. Let us try to implement this procedure with the two-dimensional translation group. As a first attempt, let us define a translation operator by $D_\gamma e(\alpha) = e(\alpha+\gamma)\, e^{|\alpha|^2/2 - |\alpha+\gamma|^2/2}$ for $\gamma \in \mathbf{C}$; the constant

factor ensures that $\|D_\gamma e(\alpha)\| = \|e(\alpha)\|$, as must be the case for any unitary operator. Unfortunately, D_γ is not in fact unitary; a straightforward calculation shows

$$\langle e(\beta), D_\gamma^* D_\gamma e(\alpha)\rangle = \langle D_\gamma e(\beta), D_\gamma e(\alpha)\rangle = e^{\beta^*\alpha} e^{i\,\mathrm{Im}(\beta^*\gamma) - i\,\mathrm{Im}(\alpha^*\gamma)}$$

which contradicts unitarity $D_\gamma^* D_\gamma = I$, i.e., $\langle e(\beta), D_\gamma^* D_\gamma e(\alpha)\rangle = \langle e(\beta), e(\alpha)\rangle = e^{\beta^*\alpha}$. To fix this, define the *Weyl operator*

$$W_\gamma e(\alpha) = e(\alpha+\gamma)\, e^{|\alpha|^2/2 - |\alpha+\gamma|^2/2} e^{i\,\mathrm{Im}(\alpha^*\gamma)} = e(\alpha+\gamma)\, e^{-\gamma^*\alpha - |\gamma|^2/2}.$$

The Weyl operator is unitary and provides a projective unitary representation [38] in the sense that $W_\alpha W_\beta = W_{\alpha+\beta} e^{i\,\mathrm{Im}(\beta^*\alpha)}$. Note that it is sufficient to define the action of W_α only on exponential vectors; we can then extend to D by linearity, and as D is dense and W_α is bounded the Weyl operators are uniquely extended to all of H.

Now fix $\beta \in \mathbf{C}$ and consider the unitary group $\{W_{t\beta}\}_{t\in\mathbf{R}}$. This group is continuous ($W_{t\beta} e(\gamma) \to e(\gamma)$ as $t \to 0$) and hence by Stone's theorem, there exists a self-adjoint operator B_β such that $W_{t\beta} = e^{itB_\beta}$. Finding the distribution of the observable B_β is straightforward, as the characteristic function of B_β is given by

$$b_\beta(k) = \mathbb{P}_\alpha(W_{k\beta}) = \langle e(\alpha), e(\alpha+k\beta)\rangle\, e^{-k\beta^*\alpha - k^2|\beta|^2/2 - |\alpha|^2} = e^{2ik\,\mathrm{Im}(\alpha^*\beta) - k^2|\beta|^2/2}.$$

Hence B_β is a Gaussian random variable with mean $2\,\mathrm{Im}(\alpha^*\beta)$ and variance $|\beta|^2$.

Our next task is to obtain an explicit representation of B_β. We proceed as follows:

$$B_\beta e(\alpha) = \frac{1}{i}\frac{d}{dt} W_{t\beta} e(\alpha)\bigg|_{t=0} = i\beta^*\alpha\, e(\alpha) - i\,\frac{d}{dt} e(\alpha + t\beta)\bigg|_{t=0}.$$

One can verify explicitly that this expression makes sense, i.e., $B_\beta e(\alpha) \in \mathsf{H}$. Note that we cannot extend B_β to all of H, as B_β is unbounded. However, we see that the domain of B_β contains at least the exponential domain D.

Let us introduce the following notation. Define $q = B_i$, $p = B_{-1}$, and $a = (q+ip)/2$. Note that q and p are self-adjoint by Stone's theorem, whereas a has the adjoint $a^* = (q-ip)/2$. Moreover, we find that $a\, e(\alpha) = \alpha\, e(\alpha)$. But then

$$(a\, e(\alpha))(k) = \alpha\,\frac{\alpha^k}{\sqrt{k!}} = \sqrt{k+1}\,\frac{\alpha^{k+1}}{\sqrt{(k+1)!}} = \sqrt{k+1}\, e(\alpha)(k+1).$$

This implies that we can extend the domain of a to include also the $\{\psi_n\}$ by defining $a\,\psi_{k+1} = \sqrt{k+1}\,\psi_k$ (where $a\,\psi_0 = 0$). Furthermore, from

$$\langle \psi_m, a^*\psi_k\rangle = \langle a\psi_m, \psi_k\rangle = \sqrt{m}\,\delta_{(m-1)k} = \sqrt{k+1}\,\delta_{m(k+1)}$$

we can read off $a^*\psi_k = \sqrt{k+1}\,\psi_{k+1}$. a^* is known as the creation (or raising) operator and a as the annihilation (or lowering) operator.

Finally, note that $\lambda = a^*a$. From a classical probability point of view this is very remarkable indeed. Not only do both Poisson and Gaussian random variables emerge from the same state $\mathbb{P}_\alpha$, but there is even a *continuous* map $q, p \mapsto (q-ip)(q+ip)/4 = \lambda$ that transforms two Gaussian random variables into a Poisson random variable. One could never continuously transform a continuous classical random variable into a discrete classical random variable; however, we get away with it here because p, q, and λ do not commute with one another. Thus in each realization we can choose to measure either a discrete or a continuous random variable, but not both. □

Remark 3.12. Though presented rather differently, the last two examples are in fact $*$-isomorphic in the case that $\sigma^2 = \frac{1}{2}$ in the first example. For example, if $\alpha \in \mathbf{R}$, we can map $p \mapsto 2^{1/2}\hbar^{-1}P$, $q \mapsto 2^{1/2}Q$, and $\mathbb{P}_\alpha \mapsto \mathbb{P}_{\mu=2^{1/2}\alpha,\sigma=2^{-1/2}}$. From the expression for $b_\beta(k)$ we see that in a coherent state both p and q must have the same variance. In the first example we allowed for the variance of Q to shrink, though this necessarily increases the variance of P. This results in a "squeezed state" which can also be introduced in the context of the second example. We will not construct such states here; in the following, we will only use coherent states. □

3.3. Conditional expectation. We now consider conditional expectations, following the treatment of [18]. The following definition is identical to the one in section 2.

Definition 3.13 (conditional expectation). *Let $(\mathscr{N}, \mathbb{P})$ be a quantum probability space and let $\mathscr{A} \subset \mathscr{N}$ be a commutative von Neumann subalgebra. Then the map $\mathbb{P}(\cdot|\mathscr{A}) : \mathscr{A}' \to \mathscr{A}$ is called (a version of) the conditional expectation from $\mathscr{A}'$ onto $\mathscr{A}$ if $\mathbb{P}(\mathbb{P}(B|\mathscr{A})A) = \mathbb{P}(BA)$ for all $A \in \mathscr{A}$, $B \in \mathscr{A}'$.*

We briefly recall the significance of $\mathscr{A}'$. $\mathscr{A}$ is the algebra generated by our observations: it must be commutative, as we cannot observe incompatible events in a single experiment. We now wish to find the conditional statistics of an observable B that is not affiliated to $\mathscr{A}$. However, as we have already observed $\mathscr{A}$, this is only sensible if B commutes with every element in $\mathscr{A}$—there would be no physical way to test our predictions if we could not subsequently measure B in the same realization.

Remark 3.14. Recall that if $B = B^*$, we can use the spectral theorem to obtain explicitly $\mathbb{P}(B|\mathscr{A}) = \iota^{-1}(E_{\mathbf{P}}(\iota(B)|\sigma\{\iota(\mathscr{A})\}))$. This representation extends even to the case that B is an unbounded self-adjoint operator that is affiliated to $\mathscr{A}'$. For simplicity we will discuss below the properties of $\mathbb{P}(B|\mathscr{A})$ assuming that B is bounded, but with suitable care the treatment extends also to the unbounded case. □

Remark 3.15. A more general definition (see, e.g., [59]), of which Definition 3.13 is a special case, is often used in quantum probability. Unlike our definition, which is motivated by statistical inference and filtering, the more general "conditional expectation" allows for conditioning on noncommutative algebras and does not have a direct statistical interpretation. The more general definition is used, e.g., in the theory of noncommutative Markov processes [47]. We will not dwell on this further. □

Theorem 3.16. *The conditional expectation of Definition* 3.13 *exists and is unique with probability one* (*any two versions P and Q of $\mathbb{P}(B|\mathscr{A})$ satisfy $\|P-Q\|_{\mathbb{P}} = 0$, where $\|X\|_{\mathbb{P}}^2 = \mathbb{P}(X^*X)$*). *Moreover, $\mathbb{P}(B|\mathscr{A})$ is the least mean square estimate of B given $\mathscr{A}$ in the sense that $\|B - \mathbb{P}(B|\mathscr{A})\|_{\mathbb{P}} \le \|B - A\|_{\mathbb{P}}$ for all $A \in \mathscr{A}$.*

Proof. (i) *Existence.* We have already established that for self-adjoint $B \in \mathscr{A}'$, we can explicitly define a $\mathbb{P}(B|\mathscr{A})$ that satisfies the conditions of Definition 3.13 using the spectral theorem. The classical conditional expectation exists, and moreover the conditional expectation of a bounded random variable is bounded. Hence $\mathbb{P}(B|\mathscr{A})$ exists in $\mathscr{A}$ for self-adjoint $B \in \mathscr{A}'$. But any $B \in \mathscr{A}'$ can be written as $B = B_1 + iB_2$ with self-adjoint $B_1 = (B + B^*)/2$ and $B_2 = i(B^* - B)/2$. As $\mathbb{P}(B_1|\mathscr{A})$ and $\mathbb{P}(B_2|\mathscr{A})$ exist and $\mathbb{P}(B|\mathscr{A}) = \mathbb{P}(B_1|\mathscr{A}) + i\mathbb{P}(B_2|\mathscr{A})$ satisfies the conditions of Definition 3.13, existence is proved.

(ii) *Uniqueness with probability one.* Define the pre-inner product $\langle X, Y\rangle = \mathbb{P}(X^*Y)$ on $\mathscr{A}'$ (it might have nontrivial kernel). Then $\langle A, B - \mathbb{P}(B|\mathscr{A})\rangle = \mathbb{P}(A^*B) - \mathbb{P}(A^*\mathbb{P}(B|\mathscr{A})) = 0$ for all $A \in \mathscr{A}$ and $B \in \mathscr{A}'$, i.e., $B - \mathbb{P}(B|\mathscr{A})$ is orthogonal to $\mathscr{A}$. Now let P and Q be two versions of $\mathbb{P}(B|\mathscr{A})$. It follows that $\langle A, P - Q\rangle = 0$ for all $A \in \mathscr{A}$. But $P - Q \in \mathscr{A}$, so $\langle P - Q, P - Q\rangle = \|P - Q\|_{\mathbb{P}}^2 = 0$.

(iii) *Least squares.* Let P be a version of $\mathbb{P}(B|\mathscr{A})$. Then for all $K \in \mathscr{A}$

$$\|B-K\|_{\mathbb{P}}^2 = \|B-P+P-K\|_{\mathbb{P}}^2 = \|B-P\|_{\mathbb{P}}^2 + \|P-K\|_{\mathbb{P}}^2 \geq \|B-P\|_{\mathbb{P}}^2,$$

where, in the second step, we used that $(B-\mathbb{P}(B|\mathscr{A})) \perp (\mathbb{P}(B|\mathscr{A})-K) \in \mathscr{A}$. □

Remark 3.17. The usual elementary properties of classical conditional expectations and their proofs [63] carry over directly. In particular, we have linearity, positivity, invariance of the state $\mathbb{P}(\mathbb{P}(B|\mathscr{A})) = \mathbb{P}(B)$, invariance of $\mathscr{A}$ ($\mathbb{P}(B|\mathscr{A}) = B$ if $B \in \mathscr{A}$), the tower property $\mathbb{P}(\mathbb{P}(B|\mathscr{A})|\mathscr{C}) = \mathbb{P}(B|\mathscr{C})$ if $\mathscr{C} \subset \mathscr{A}$, the module property $\mathbb{P}(AB|\mathscr{C}) = B\,\mathbb{P}(A|\mathscr{C})$ for $B \in \mathscr{C}$, etc. As an example, let us prove linearity. It suffices to show that $Z = \alpha\,\mathbb{P}(A|\mathscr{C}) + \beta\,\mathbb{P}(B|\mathscr{C})$ satisfies $\mathbb{P}(ZC) = \mathbb{P}((\alpha A+\beta B)C)$ for all $C \in \mathscr{C}$. But this is immediate from the linearity of $\mathbb{P}$ and Definition 3.13. □

3.4. The Bayes formula. In section 2 we were able to calculate conditional expectations explicitly as all algebras were finite-dimensional. In most physical situations, however, at least the probe (and often the system as well) admits continuous observables and therefore we must deal with infinite-dimensional algebras. In this case it is usually not so simple to calculate the conditional expectations directly; however, the following Bayes-type formula will be of considerable assistance.

LEMMA 3.18 (Bayes formula [18]). *Let $\mathscr{C}$ be a commutative von Neumann subalgebra and let $\mathscr{C}'$ be equipped with a normal state $\mathbb{P}$. Choose $V \in \mathscr{C}'$ such that $V^*V > 0$ and $\mathbb{P}(V^*V) = 1$. Then we can define a new state on $\mathscr{C}'$ by $\mathbb{Q}(A) = \mathbb{P}(V^*AV)$ and*

$$\mathbb{Q}(X|\mathscr{C}) = \frac{\mathbb{P}(V^*XV|\mathscr{C})}{\mathbb{P}(V^*V|\mathscr{C})}, \qquad X \in \mathscr{C}'.$$

Proof. Let K be an element of $\mathscr{C}$. For all $X \in \mathscr{C}'$, we can write

$$\begin{aligned}\mathbb{P}(\mathbb{P}(V^*XV|\mathscr{C})K) &= \mathbb{P}(V^*XKV) = \mathbb{Q}(XK) = \mathbb{Q}(\mathbb{Q}(X|\mathscr{C})K)\\ &= \mathbb{P}(V^*V\mathbb{Q}(X|\mathscr{C})K) = \mathbb{P}(\mathbb{P}(V^*V\mathbb{Q}(X|\mathscr{C})K|\mathscr{C})) = \mathbb{P}(\mathbb{P}(V^*V|\mathscr{C})\mathbb{Q}(X|\mathscr{C})K).\end{aligned}$$

As this holds for all $K \in \mathscr{C}$, and as by construction the conditional expectations are elements of $\mathscr{C}$, we conclude that $\|\mathbb{P}(V^*XV|\mathscr{C}) - \mathbb{P}(V^*V|\mathscr{C})\mathbb{Q}(X|\mathscr{C})\|_{\mathbb{P}} = 0$, or equivalently, $\mathbb{P}(V^*XV|\mathscr{C}) = \mathbb{P}(V^*V|\mathscr{C})\mathbb{Q}(X|\mathscr{C})$ $\mathbb{P}$-a.s. □

We now have sufficient tools to deal with the Stern–Gerlach experiment described in section 2. Though the following example is not of much practical importance, it demonstrates the use of the Bayes theorem in a concrete setting. We will use a very similar "reference probability method" to obtain filtering equations later on.

Example 3.19. *Stern–Gerlach experiment.* Consider an atom with two degrees of freedom: a spin degree of freedom $\mathscr{N}_\mu = \mathscr{B}(\mathbf{C}^2)$ carrying the observables σ_x, σ_z etc., and a single spatial degree of freedom $\mathscr{N}_x = \mathscr{B}(\ell^2(\mathbf{N}))$ with the affiliated position q and momentum p observables defined[6] in Example 3.11 (we use the notations of that example). The total algebra describing the atom is then $\mathscr{N} = \mathscr{N}_\mu \otimes \mathscr{N}_x$. Initially the spin and position/momentum of the atom are uncorrelated; hence we work with the state $\mathbb{P} = \mathbb{P}_\mu \otimes \mathbb{P}_0$, where $\mathbb{P}_\mu$ is an arbitrary spin state and $\mathbb{P}_0(X) = \langle\psi_0, X\psi_0\rangle = \langle e(0), Xe(0)\rangle$. The latter implies that initially $I\otimes q$ and $I\otimes p$ (which we will interpret as position and momentum in the z direction) have zero mean and unit variance.

[6] We saw in Remark 3.12 that this description is $*$-isomorphic to the usual definition of position and momentum up to some numerical constants. These are not of essence, however, as they just correspond to a change of units in which we measure position and momentum. A little more care must be taken if we wish to make quantitative predictions on the outcomes of actual experiments; we will not worry about this, however, and work in arbitrary units.

To measure the spin, we apply a magnetic field gradient that is linear in q for some fixed period of time. The resulting force on the particle will cause its momentum to change; an observation of the momentum of the particle after the interaction should thus provide a measurement of its spin σ_z. In other words, the atomic spatial degree of freedom acts as a probe for the atomic spin degree of freedom. The action of the magnetic field is described by the unitary[7]

$$U = \exp(i\kappa\,\sigma_z \otimes q) = P_{z,1} \otimes e^{i\kappa q} + P_{z,-1} \otimes e^{-i\kappa q} = P_{z,1} \otimes W_{i\kappa} + P_{z,-1} \otimes W_{-i\kappa},$$

where $\kappa \in \mathbf{R}$ is the field gradient. Let us thus begin by calculating the characteristic function of $U^*(I \otimes p)U$, the momentum of the atom after the interaction:

$$\begin{aligned}\mathbb{P}(e^{ik\,U^*(I\otimes p)U}) &= \mathbb{P}(U^*(I \otimes W_{-k})U) = \mathbb{P}_\mu(P_{z,1})\,\mathbb{P}_0(W_{-i\kappa}W_{-k}W_{i\kappa}) \\ &\quad + \mathbb{P}_\mu(P_{z,-1})\,\mathbb{P}_0(W_{i\kappa}W_{-k}W_{-i\kappa}) = \mathbb{P}_\mu(P_{z,1})\,e^{2i\kappa k - k^2/2} + \mathbb{P}_\mu(P_{z,-1})\,e^{-2i\kappa k - k^2/2}.\end{aligned}$$

Hence the momentum of the atom after the interaction is distributed as a sum of two Gaussians of unit variance and means 2κ and -2κ, which are weighted, respectively, by $\mathbb{P}_\mu(P_{z,1})$ and $\mathbb{P}_\mu(P_{z,-1})$. Note that we cannot perfectly resolve the spin-up and down states using a Stern–Gerlach measurement; as the tails of the two Gaussians overlap, there is always a nonzero probability that we assign the wrong spin to the atom by looking, e.g., at the sign of the observed momentum. However, the error probability becomes very small when the gradient κ is large.

After the interaction, we may want to measure a spin observable $\sigma \in \mathscr{N}_\mu$ that does not necessarily commute with σ_z (e.g., σ_x). To describe this, let us calculate $\mathbb{P}(U^*(\sigma \otimes I)U|\mathrm{vN}(U^*(I \otimes p)U))$, the conditional expectation of the spin observable σ after the interaction given our observation of the momentum of the atom.

We begin by using the following elementary property: if U is a unitary operator and we define the state $\mathbb{Q}(X) = \mathbb{P}(U^*XU)$, then $\mathbb{P}(U^*XU|U^*\mathscr{C}U) = U^*\mathbb{Q}(X|\mathscr{C})U$ (this can be verified directly using Definition 3.13). Thus we obtain

$$\mathbb{P}(U^*(\sigma \otimes I)U|\mathrm{vN}(U^*(I \otimes p)U)) = U^*\mathbb{Q}(\sigma \otimes I|\mathrm{vN}(I \otimes p))U.$$

We would like to apply the Bayes rule to $\mathbb{Q}(\sigma \otimes I|\mathrm{vN}(I \otimes p))$. As U does not commute with $I \otimes p$, however, the Bayes rule does not apply in this form.

Fortunately, we can circumvent this problem using the following trick. Using the Baker–Campbell–Hausdorff formula, we can rewrite $e^{i\kappa q}$ as

$$e^{i\kappa q} = e^{i\kappa(a+a^*)} = e^{-\kappa^2/2}e^{i\kappa a^*}e^{i\kappa a}.$$

Beware that the Baker–Campbell–Hausdorff formula technically only holds for exponentials of bounded operators; thus here and below there will be domain issues, but these can be resolved with suitable care. As $a\,\psi_0 = 0$, we can write

$$e^{i\kappa q}\psi_0 = e^{-\kappa^2/2}e^{i\kappa a^*}e^{i\kappa a}\psi_0 = e^{-\kappa^2/2}e^{i\kappa a^*}\psi_0 = e^{-\kappa^2/2}e^{i\kappa a^*}e^{-i\kappa a}\psi_0 = e^{-\kappa^2}e^{\kappa p}\psi_0.$$

We obtain

$$\mathbb{P}_0(e^{-i\kappa q}Xe^{i\kappa q}) = \langle e^{i\kappa q}\psi_0, Xe^{i\kappa q}\psi_0\rangle = e^{-2\kappa^2}\langle e^{\kappa p}\psi_0, Xe^{\kappa p}\psi_0\rangle = e^{-2\kappa^2}\mathbb{P}_0(e^{\kappa p}Xe^{\kappa p}).$$

[7]This is the solution of (2.6) at some fixed time t for a suitable interaction Hamiltonian H.

It follows that we can equivalently replace U by V:

$$\mathbb{Q}(X) = \mathbb{P}(U^*XU) = \mathbb{P}(V^*XV), \quad V = e^{-\kappa^2} e^{\kappa\,\sigma_z \otimes p} = e^{-\kappa^2}(P_{z,1} \otimes e^{\kappa p} + P_{z,-1} \otimes e^{-\kappa p}).$$

V is not unitary, but it does commute with $I \otimes p$. Hence the Bayes rule gives

$$\mathbb{P}(U^*(\sigma \otimes I)U | \mathrm{vN}(U^*(I \otimes p)U)) = \frac{U^*\mathbb{P}(V^*(\sigma \otimes I)V | \mathrm{vN}(I \otimes p))U}{U^*\mathbb{P}(V^*V | \mathrm{vN}(I \otimes p))U}.$$

We can now use the module property and independence of $\sigma \otimes I$ and $I \otimes p$ under $\mathbb{P}$ to calculate explicitly the numerator and denominator; elementary manipulations give

$$\begin{aligned}
&\mathbb{P}[U^*(\sigma \otimes I)U | \mathrm{vN}(U^*(I \otimes p)U)] \\
&= \frac{\mathbb{P}_\mu(P_{z,1}\sigma P_{z,1}) e^{2\kappa\, U^*(I\otimes p)U} + \mathbb{P}_\mu(P_{z,-1}\sigma P_{z,-1}) e^{-2\kappa\, U^*(I\otimes p)U} + 2\,\mathrm{Re}\,\mathbb{P}_\mu(P_{z,-1}\sigma P_{z,1})}{\mathbb{P}_\mu(P_{z,1}) e^{2\kappa\, U^*(I\otimes p)U} + \mathbb{P}_\mu(P_{z,-1}) e^{-2\kappa\, U^*(I\otimes p)U}}.
\end{aligned}$$

By definition, $\mathbb{P}(U^*(\sigma \otimes I)U | \mathrm{vN}(U^*(I \otimes p)U))$ is affiliated to $\mathrm{vN}(U^*(I \otimes p)U)$, and indeed the expression above is simply a function of $U^*(I \otimes p)U$. If we observe $U^*(I \otimes p)U$ and obtain the value $\tilde{p}$, then the spectral theorem tells us that the conditional expectation takes the value given by the expression above if we simply substitute $\tilde{p}$ for $U^*(I \otimes p)U$. Note that the formula is not equivalent to the one given by the projection postulate for a measurement of σ_z. For large κ, however, we obtain approximately the projection postulate expression, and this becomes exact as $\kappa \to \infty$. □

4. Stochastic processes and quantum Itô calculus. After a general introduction to quantum probability, we now turn to one particular quantum probability space which we will use throughout the remainder of the article. In section 5 we shall argue that this model appropriately describes the quantum electromagnetic field and its interaction with matter. In the laboratory, the electromagnetic field can be measured by devices like photodetectors which can produce an electric current or even a discrete photocount. The statistics of data records from such experiments are well approximated by the model considered here. The model is rich and we will discover that it contains many interesting classical stochastic processes, i.e., a whole family of Poisson and Wiener processes. However, these processes do not commute with each other. An extension of the Itô calculus, due to Hudson and Parthasarathy [42], unites all these processes in one noncommutative stochastic calculus.

4.1. Poisson processes on Fock space. The theory we are about to discuss can be approached from many sides; here we have chosen to get started by finding a quantum probability space that naturally admits a Poisson process, and build the theory from there. As we have a particular classical process in mind, the general theory gives a hint as to how we could proceed. First, we define the process on a classical space $(\Omega, \mathcal{F}, \mathbf{P})$; equivalently, we can form the algebra $\mathscr{A} = L^\infty(\Omega, \mathcal{F}, \mathbf{P})$ acting on $\mathsf{H} = L^2(\Omega, \mathcal{F}, \mathbf{P})$ by pointwise multiplication, with a suitable state $\mathbb{P}$, and represent the process as a family of observables affiliated to $\mathscr{A}$. To create a noncommutative model, we could now broaden our horizon and consider $\mathscr{N} = (\mathscr{B}(\mathsf{H}), \mathbb{P})$ rather than just $\mathscr{A}$. Obviously such a construction does not necessarily carry a physical interpretation; this must be considered separately; see section 5. For the time being, however, we will use this convenient construction to provide us with a rich quantum stochastic model. The following discussion is heavily inspired by the work of Maassen [50].

Consider a classical Poisson process on a finite time interval $[0, T]$. We wish to describe the space of paths Ω. This is not difficult; a Poisson process on a finite time

interval has (a.s.) finitely many jumps n. Hence we can specify every relevant path by specifying its jump times. Let us thus introduce

$$\Omega = \bigcup_{n=0}^{\infty} \Omega_n, \qquad \Omega_0 = \{\varnothing\}, \quad \Omega_n = \{\{t_1, \ldots, t_n\} : t_1 < t_2 < \cdots < t_n \in [0,T]\}. \tag{4.1}$$

In other words, Ω is the set of ordered sequences in $[0,T]$ with a finite number of elements. We still need to introduce a σ-algebra $\mathcal{F}$ and a measure $\mathbf{P}$. To this end, consider Ω_n as a subset of the cube $([0,T]^n, e^{-T}\mu_n)$ where μ_n is the Lebesgue measure, so that Ω_n inherits a σ-algebra $\mathcal{F}_n$ and a measure $\mathbf{P}_n$ from the cube. Under $\mathbf{P}_n$ the jump times $t_1, \ldots, t_n$ are uniformly distributed (as must be the case for a Poisson process with fixed rate) and $\mathbf{P}_n(\Omega_n) = T^n e^{-T}/n!$. The measure $\mathbf{P}$ induced on Ω is precisely the probability measure of a Poisson process with unit rate.

We now introduce the Hilbert space $\mathsf{F} = L^2(\Omega, \mathcal{F}, \mathbf{P})$. It is called the *symmetric* or *Boson Fock space* and plays a central role in the following. We will also need the spaces $\mathsf{F}_{t]}$, $\mathsf{F}_{[t}$, and $\mathsf{F}_{[s,t]}$, defined identically to F except that the interval $[0,T]$ is replaced by $[0,t]$, $[t,T]$, and $[s,t]$, respectively. It is not difficult to see that for any $0 < s < t < T$ we have[8] $\Omega = \Omega_{s]} \times \Omega_{[s,t]} \times \Omega_{[t}$, and as the Poisson process has independent increments the measure splits up similarly. It follows that

$$\mathsf{F} = \mathsf{F}_{s]} \otimes \mathsf{F}_{[s,t]} \otimes \mathsf{F}_{[t} \qquad \forall\, 0 < s < t < T. \tag{4.2}$$

This important property is known as a continuous tensor product structure; it will play a key role in the definition of quantum stochastic integrals, as it gives a natural notion of adaptedness. Indeed, the algebra $\mathscr{W} = \mathscr{B}(\mathsf{F})$ splits up accordingly,

$$\mathscr{W} = \mathscr{W}_{s]} \otimes \mathscr{W}_{[s,t]} \otimes \mathscr{W}_{[t} = \mathscr{B}(\mathsf{F}_{s]}) \otimes \mathscr{B}(\mathsf{F}_{[s,t]}) \otimes \mathscr{B}(\mathsf{F}_{[t}). \tag{4.3}$$

A process of operators $\{X_t\}$ affiliated to $\mathscr{W}$ is said to be *adapted* if X_t is affiliated to $\mathscr{W}_{t]}$ for every t; equivalently, X_t is of the form $X_{t]} \otimes I$ as an operator on $\mathsf{F}_{t]} \otimes \mathsf{F}_{[t}$.

Next, let us introduce a set of interesting vectors. The reader should keep in mind Example 3.11 which is conceptually quite similar. Let $f \in L^\infty([0,T])$ be a complex Lebesgue measurable function. Then we can define the *exponential vector*

$$e(f)(\varnothing) = 1, \quad e(f)(\tau) = \prod_{t \in \tau} f(t), \qquad f \in L^\infty([0,T]),\ \tau \in \Omega. \tag{4.4}$$

It is not difficult to verify that $e(f) \in \mathsf{F}$, as

$$\langle e(g), e(f)\rangle = \sum_{n=0}^{\infty} \frac{e^{-T}}{n!} \left(\int_0^T g^*(t) f(t)\, dt \right)^n = \exp\left[\int_0^T (g^*(t) f(t) - 1)\, dt \right],$$

hence $\langle e(f), e(f)\rangle = e^{\|f\|_2^2 - T} < \infty$ for any $f \in L^\infty([0,T])$. We define D, the exponential domain, as the linear span of all $e(f)$, $f \in L^\infty([0,T])$, and we note that D is dense in F. The exponential vectors have the important property that they factorize over the continuous tensor product structure (4.2): indeed, it is evident from (4.4) that $e(f) = e(f_{s]}) \otimes e(f_{[s,t]}) \otimes e(f_{[t})$ where $f_{t]}$ is the restriction of f to $[0,t]$, etc.

[8] A more precise statement would be something like $\Omega = \Omega_{s]} \times \Omega_{(s,t]} \times \Omega_{(t}$; however, the only paths for which this makes a difference are those that have jumps exactly at times s or t, which is a set of $\mathbf{P}$-measure zero. For notational simplicity, we are free to always use closed time intervals $[s,t]$.

We are now ready to define a Poisson process. Let us first define it as a random variable on Ω; we simply write $N_t(\tau) = |\tau \cap [0,t]|$, where $|\tau|$ denotes the number of elements in the set $\tau \in \Omega$. The random variable N_t counts the number of jumps up to time t, and hence $\{N_t\}$ is by construction a Poisson process with unit rate under the measure $\mathbf{P}$. We now turn this into an operator process by pointwise multiplication:

$$(\Lambda_t\psi)(\tau) = N_t(\tau)\,\psi(\tau) = |\tau \cap [0,t]|\,\psi(\tau), \qquad \psi \in \mathsf{F},\ \tau \in \Omega,\ t \in [0,T]. \tag{4.5}$$

$\{\Lambda_t\}$ is called the *gauge process*; it is not difficult to see that though Λ_t is an unbounded operator,[9] it is affiliated to $\mathscr{W}_{t]}$ and hence the gauge process is adapted; in fact, the increments $\Lambda_t - \Lambda_s$ are even affiliated to $\mathscr{W}_{[s,t]}$. Furthermore, Λ_s and Λ_t commute for all $s,t \in [0,T]$, and indeed $\mathrm{vN}(\Lambda_t,\ t \in [0,T]) = L^\infty(\Omega,\mathcal{F},\mathbf{P}) \subset \mathscr{W}$ is commutative. Hence we could use the spectral theorem to map Λ_t back to a classical stochastic process. It is somewhat futile to diagonalize the operators using the spectral theorem, however, as we have already constructed them in diagonal form.

We have yet to introduce a state; a particularly interesting class of states are the *coherent states* $\mathbb{P}_f(X) = \langle e(f), X\,e(f)\rangle\, e^{T-\|f\|_2^2}$. Because of the continuous tensor product property, the coherent states split up as follows:

$$X = X_{s]} \otimes X_{[s,t]} \otimes X_{[t}, \qquad \mathbb{P}_f(X) = \mathbb{P}_{f_{s]}}(X_{s]})\ \mathbb{P}_{f_{[s,t]}}(X_{[s,t]})\ \mathbb{P}_{f_{[t}}(X_{[t}). \tag{4.6}$$

But as $\Lambda_t - \Lambda_s$ is affiliated to $\mathscr{W}_{[s,t]}$, it follows that under the state $\mathbb{P}_f$ the gauge process has independent increments. Furthermore, if we denote by $P_{\Lambda_t-\Lambda_s}(B)$ the spectral measure of $\Lambda_t - \Lambda_s$, then we have

$$\mathbb{P}_f(P_{\Lambda_t-\Lambda_s}(B)) = \mathbb{P}_{f_{[s,t]}}(\chi_B(|\tau \cap [s,t]|)) = \sum_{n\in B} \frac{e^{-\int_s^t |f(r)|^2\,dr}}{n!} \left(\int_s^t |f(r)|^2\,dr\right)^n.$$

Evidently, Λ_t is an inhomogeneous Poisson process with rate $|f(t)|^2$ under the state $\mathbb{P}_f$. Note in particular that as $e(1)(\tau) = 1$, we have for any $X \in L^\infty(\Omega,\mathcal{F},\mathbf{P})$ the relation $\mathbb{P}_1(X) = \langle 1, X\,1\rangle = E_{\mathbf{P}}(X)$; hence the fact that under $\mathbb{P}_1$ the gauge process is a Poisson process with unit rate is exactly what we expect from the definition of $\mathbf{P}$. Under $\mathbb{P}_0$, on the other hand, the gauge process does not register any counts; $\mathbb{P}_0 = \phi$ is called the *vacuum state*, and $e(0) = \Phi$ is called the *vacuum vector*.

4.2. Weyl operators and Wiener processes. We have now exhausted the diagonal observables affiliated to the space $(L^\infty(\Omega,\mathcal{F},\mathbf{P}),\mathbb{P}_f)$: every such observable is some functional of the Poisson process Λ_t with rate $|f|^2$. Let us thus explore whether we can find interesting observables affiliated to $\mathscr{W}$ that do not commute with Λ_t. To this end, we follow again essentially Example 3.11. Given $f,g \in L^\infty([0,T])$ we look for a unitary operator $W(f)$ that implements the translation group $W(f)e(g) \propto e(f+g)$. A calculation identical to the one in Example 3.11 shows that we should define

$$W(f)e(g) = e^{-\int_0^T \left(f^*(t)g(t) + \frac{1}{2}f^*(t)f(t)\right)dt} e(f+g) = e^{-\langle f,g\rangle_2 - \|f\|_2^2/2}\,e(f+g). \tag{4.7}$$

The unitary operator $W(f)$ is called a *Weyl operator* and provides a projective unitary representation in the sense that $W(f)W(g) = W(f+g)\,e^{i\,\mathrm{Im}\langle g,f\rangle_2}$. Note that it is

[9] As can be verified by explicit computation, the domain of Λ_t contains at least D, the exponential domain. The reader may ask himself why we have only defined exponential vectors $e(f)$ for $f \in L^\infty([0,T])$ rather than $f \in L^2([0,T])$: this is because the latter may not be in the domain of Λ_t. Our domain D is sometimes called the *restricted* exponential domain in the literature.

sufficient to define the action of $W(f)$ only on exponential vectors; we can extend to D by linearity, and as D is dense and $W(f)$ is bounded the Weyl operators are uniquely extended to all of F. An important property, which follows immediately from the definition of $W(f)$ and the continuous tensor product property, is that

$$W(f)e(g) = W(f_{s]})e(g_{s]}) \otimes W(f_{[s,t]})e(g_{[s,t]}) \otimes W(f_{[t})e(g_{[t}). \tag{4.8}$$

In particular, we see that $W(f\chi_{[0,t]})$ is an adapted operator process.

Now fix $f \in L^\infty([0,T])$ and consider the unitary group $\{W(kf)\}_{k\in\mathbf{R}}$; this group is in fact continuous [55], and hence by Stone's theorem (Theorem 3.10) there exists a self-adjoint $B(f)$ such that $W(kf) = e^{ikB(f)}$. The operators $B(f)$, $f \in L^\infty([0,T])$, are called *field operators*. Finding the distribution of the observable $B(f)$ is straightforward, as the characteristic function of $B(f)$ (under the coherent state $\mathbb{P}_g$) is given by

$$b_f(k) = \mathbb{P}_g(W(kf)) = \langle e(g), e(g+kf)\rangle e^{T-\|g\|_2^2 - k\langle f,g\rangle_2 - k^2\|f\|_2^2/2} = e^{2ik\,\mathrm{Im}\langle g,f\rangle_2 - k^2\|f\|_2^2/2}.$$

Hence $B(f)$ is a Gaussian random variable with mean $2\,\mathrm{Im}\langle g,f\rangle_2$ and variance $\|f\|_2^2$. In the vacuum, i.e., $g=0$, the mean vanishes; for simplicity, we will restrict ourselves to the vacuum case in the following.

Consider the operator process $\{B_t^\varphi = B(e^{i\varphi}\chi_{[0,t]}) : t \in [0,T]\}$ for some fixed, real function $\varphi \in L^\infty([0,T])$. B_t^φ is adapted, as we have already established that $W(f\chi_{[0,t]})$ is adapted for any f; moreover, $B(e^{i\varphi}\chi_{[s,t]}) = B_t^\varphi - B_s^\varphi$ is affiliated to $\mathscr{W}_{[s,t]}$ due to (4.8). This immediately tells us two important things. First, B_t^φ and B_s^φ commute for all $s,t \in [0,T]$; indeed, $B_t^\varphi - B_s^\varphi$ must commute with $B_s^\varphi - B_0^\varphi$, and commutativity follows from $B_0^\varphi = 0$. This means that $\mathrm{vN}(B_t^\varphi,\ t \in [0,T])$ is a commutative algebra and hence we can represent B_t^φ for every t as a classical random variable on the same probability space $(\Omega^\varphi, \mathcal{F}^\varphi, \mathbf{P}^\varphi)$; in particular, $\iota(B_t^\varphi)$ is a classical stochastic process. Second, (4.6) implies that the process B_t^φ has independent increments. But we have established $B_t^\varphi - B_s^\varphi$ is (in the vacuum) a mean zero Gaussian random variable with variance $t-s$, and as B_t^φ has independent increments we have established that $\iota(B_t^\varphi)$ is precisely a Wiener process on $(\Omega^\varphi, \mathcal{F}^\varphi, \mathbf{P}^\varphi)$.

Let us introduce the following notation. Define $Q_t = B(i\chi_{[0,t]})$, $P_t = B(-\chi_{[0,t]})$, and $A_t = (Q_t + iP_t)/2$. Note that Q_t and P_t are self-adjoint by Stone's theorem, whereas A_t has the adjoint $A_t^* = (Q_t - iP_t)/2$. We now compute

$$B(f)e(g) = \frac{1}{i}\frac{d}{dk}W(kf)e(g)\bigg|_{k=0} = i\langle f,g\rangle_2\, e(g) - i\,\frac{d}{dk}e(g+kf)\bigg|_{k=0}.$$

Evidently $A_t e(g) = \langle \chi_{[0,t]}, g\rangle_2\, e(g) = \int_0^t g(s)ds\, e(g)$. But then we can write

$$(A_t e(g))(\tau) = \int_0^t g(s)\,ds \prod_{r\in\tau} g(r) = \int_0^t g(s) \prod_{r\in\tau} g(r)\,ds = \int_0^t e(g)(\tau\cup\{s\})\,ds.$$

In particular, this formula extends to any $\psi \in \mathsf{F}$ for which the integral on the right-hand side (with $e(g)$ replaced by ψ) defines a normalizable vector. A_t is called the Fock space *annihilation operator*, as it generalizes the corresponding notion introduced in Example 3.11. The reader should verify that its adjoint can be expressed as

$$(A_t^*\psi)(\tau) = \sum_{s\in\tau\cap[0,t]} \psi(\tau\backslash\{s\})$$

on a sufficiently large domain. Not surprisingly, A_t^* is called the *creation operator*. It is conventional in quantum stochastic calculus to use A_t and its adjoint rather than Q_t and P_t; we shall conform to this standard.

In summary, we have constructed a quantum probability space $(\mathscr{W}, \phi)$ that admits an entire family (indexed by φ) of Wiener processes. Note, however, that these processes do not necessarily commute for different φ; in fact, it is not difficult to establish that $[B(f), B(g)]\psi = 2i\,\mathrm{Im}\langle f, g\rangle_2\,\psi$ on a suitably large domain (e.g., $\psi \in \mathsf{D}$). Therefore, even though every B_t^φ defines a Wiener process, these cannot be represented on the same classical probability space for different $\varphi_{1,2}$ unless $\mathrm{Im}(e^{i(\varphi_1-\varphi_2)}) = 0$.

We have also defined a Poisson process Λ_t, but unfortunately it vanishes in the vacuum. Consider, however, the process $\Lambda_t(f) = W(f)^*\Lambda_t W(f)$; for any Borel function b we can write $\phi(b(\Lambda_{t_1}(f), \ldots, \Lambda_{t_n}(f))) = \mathbb{P}_f(b(\Lambda_{t_1}, \ldots, \Lambda_{t_n}))$. Evidently $\Lambda_t(f)$ has the same statistics in the vacuum as does Λ_t under the coherent state $\mathbb{P}_f$. This shows that we can define even a whole family of Poisson processes in the vacuum. We do not lose much by restricting ourselves to the vacuum as an underlying state (as we will do in the remainder of the article), as we can always transform to a coherent state by "sandwiching" with Weyl operators. Note that like the family B_t^φ, the processes $\Lambda_t(f)$ do not commute amongst each other. We see that the quantum probability space $(\mathscr{W}, \phi)$ gives rise to a rich family of incompatible stochastic processes.

4.3. Quantum stochastic calculus. Now that we have obtained Wiener and Poisson processes, we can try to develop stochastic integrals with respect to these processes and an associated stochastic calculus. Note that if we were only interested in, e.g., integrating with respect to Q_t an adapted process which commutes with Q_t, then we could simply use the classical Itô integral definition through the spectral theorem. This will not suffice for our purposes, however, as we will want to consider stochastic differential equations that are driven simultaneously by the noncommuting noises Q_t and P_t (and even Λ_t). Moreover, we would like to have an Itô rule that tells us how to multiply stochastic integrals with respect to Q_t and P_t.

Our motivation for developing generalized quantum stochastic calculus is that this allows us to rigorously define and manipulate Schrödinger equations, as in (2.6), with a white-noise Hamiltonian formally defined by $H(t) = H_0 + H_1\,\dot{Q}_t + H_2\,\dot{P}_t$. In section 5 we will see that such models emerge naturally in applications. In this section we sketch the development of quantum stochastic calculus as it was introduced in a seminal paper by Hudson and Parthasarathy [42]. For a full development of this calculus we refer the reader to [42, 41, 55]. The Hudson–Parthasarathy approach has some technical issues, not surprisingly involving the unboundedness of operators, the full extent of which is still being explored. Though we cannot go into detail here, we will attempt to sketch some of the issues and give references to recent literature.

We work in the following setting. We wish to integrate processes against the three noises A_t, A_i^*, and Λ_t (the *fundamental noises*); i.e., we want to define $\int_0^t L_s\,dM_s$ where M_t is one of the fundamental noises. The noises are defined on the quantum probability space $(\mathscr{W}, \phi)$, but we will want to couple these noises to an external quantum system, the *initial system*,[10] with which they interact. To this end, let us introduce the initial Hilbert space h, $\mathscr{B} = \mathscr{B}(\mathsf{h})$, and the associated initial quantum

[10]This name has the following origin. Recall from section 2 that observables X evolve in time as $X_t = U_t^* X U_t$ (we will define a unitary evolution U_t in section 5). We would like to think of $X \otimes I \in \mathscr{B} \otimes \mathscr{W}$ as describing the external system; however, $U_t^*(X \otimes I)U_t$ will not be of the form $Y \otimes I$ except at $t = 0$. Hence the initial system observable $X \otimes I$ describes the external system at the initial time $t = 0$.

probability space $(\mathscr{B},\rho)$. We will choose our integrands L_t to be adapted processes on $(\mathscr{B}\otimes\mathscr{W},\rho\otimes\phi)$; i.e., each L_t is affiliated to $\mathscr{B}\otimes\mathscr{W}_{t]}$ and acts as I on $\mathscr{W}_{[t}$.

As usual, we begin with simple processes. Given $s<t$, recall that for the fundamental processes M_t-M_s is affiliated to $\mathscr{W}_{[s,t]}$, whereas for adapted processes L_s is affiliated to $\mathscr{B}\otimes\mathscr{W}_{s]}$; hence we can naturally write $L_s(M_t-M_s)=L_s\otimes(M_t-M_s)$. In particular, the increment M_t-M_s commutes with L_s, and we have no problems with operator multiplication of these unbounded operators. Let $\{t_i : i=0,\ldots,n,\ t_i<t_{i+1}\}$ be a sequence of times with $t_0=0$ and $t_n=T$. By definition, we set

$$L_t=\sum_{i=0}^{n-1}L_{t_i}\chi_{[t_i,t_{i+1})}(t)\quad\Longrightarrow\quad\int_0^t L_s\,dM_s=\sum_{i=0}^{n-1}L_{t_i}\otimes(M_{t_{i+1}\wedge t}-M_{t_i\wedge t}).$$

This definition makes sense as long as the operators L_t and M_t have a sufficiently large common dense domain that the sum is well defined. To enforce this, we will require that the domain of every L_t contains at least the exponential domain D.

Now comes the hard part in any integration theory: given a quadruple of suitably restricted adapted processes (E,F,G,H), such that these admit simple approximations (E^n,F^n,G^n,H^n), we wish to define the integral

$$I_t=\int_0^t(E_t\,d\Lambda_t+F_t\,dA_t+G_t\,dA_t^*+H_t\,dt)\tag{4.9}$$

as a limit, in some sense, of the corresponding integrals I_t^n over the simple processes. Recall that in the classical theory, the Itô isometry allows us to define the stochastic integral as a mean square limit of simple processes, and a little more work shows that every square-integrable process admits a mean square approximation by simple processes. Things are not quite so "simple" in the noncommutative case, however.

To see what goes wrong, consider for simplicity the case $\mathsf{h}=\mathbf{C}$ so that we can forget about the initial state ρ. We already encountered the noncommutative L^2 (semi)norm $\|X\|_\phi^2=\phi(X^*X)$ when we discussed conditional expectations. We are thus looking for a suitable unbounded operator I_t such that we have mean square convergence, $\|I_t-I_t^n\|_\phi^2=\langle(I_t-I_t^n)\Phi,(I_t-I_t^n)\Phi\rangle\to0$ as $n\to\infty$. But this is a very ill-defined problem, as it only depends on the action of I_t on the vacuum vector Φ; in particular, what do we choose as the domain of I_t, and how do we define I_t on vectors orthogonal to Φ? There could be a large number of inequivalent ways of doing this, giving rise to limiting operators with very different properties.[11]

The solution of Hudson and Parthasarathy works as follows. First of all, we fix the domain of I_t at the outset: every stochastic integral will have $\mathsf{h}\otimes\mathsf{D}$ as its domain (one could choose a dense domain in h as well; we will not worry about this). To specify I_t as a limit of simple integrals I_t^n, we choose I_t as the unique operator on $\mathsf{h}\otimes\mathsf{D}$ such that $\langle(I_t-I_t^n)\,v\otimes\psi,(I_t-I_t^n)\,v\otimes\psi\rangle\to0$ for every $\psi\in\mathsf{D}$, $v\in\mathsf{h}$ (it is sufficient to verify this for $\psi=e(f)$, $f\in L^\infty([0,T]))$. In essence this is like a mean square limit, but simultaneously for every coherent state. A suitable estimate replaces the Itô isometry [42, Corollary 1] and shows that this limit exists as long as $\int_0^T\|(E_s-E_s^n)\,v\otimes\psi\|^2ds\to0$ as $n\to\infty$ for every $\psi\in\mathsf{D}$, $v\in\mathsf{h}$ (and similarly for F,G,H), independent of the approximation. Finally, [42, Proposition 3.2] shows

[11] This was not a problem for the definition of conditional expectations; as all versions of the conditional expectation are affiliated to a single commutative algebra, they are a.s. equivalent by the spectral theorem. On the other hand, various "versions" of I_t that satisfy $\|I_t-I_t^n\|_\phi\to0$ need not even commute, and such operators are fundamentally inequivalent.

that every square-integrable process, i.e., $\int_0^T \|E_s\, v \otimes \psi\|^2 ds < \infty$ for all $\psi \in \mathsf{D}$, $v \in \mathsf{h}$, admits a suitable approximation by simple processes. We thus arrive at the following.

DEFINITION 4.1 (quantum Itô integral). *An operator process $\{X_t\}$ is stochastically integrable if it is adapted and square integrable. Given a quadruple (E, F, G, H) of such processes, the stochastic integral* (4.9) *is uniquely defined as the limit of simple approximations on the domain $\mathsf{h} \otimes \mathsf{D}$.*

Remark 4.2. It is often convenient to denote an expression of the form

$$X_t = X + \int_0^t (E_s\, d\Lambda_s + F_s\, dA_s + G_s\, dA_s^* + H_s\, ds),$$

symbolically as

$$dX_t = E_t\, d\Lambda_t + F_t\, dA_t + G_t\, dA_t^* + H_t\, dt, \qquad X_0 = X.$$

Both notations are used interchangeably in the literature.

A property that we will exploit in future is $\Lambda_t\Phi = A_t\Phi = 0$. It is immediate from the definition that stochastic integrals with respect to A_t and Λ_t acting on Φ vanish. Hence the vacuum expectations of stochastic integrals with respect to A_t and Λ_t vanish as well. Furthermore, as $\langle \Phi, A_t^*\Phi\rangle = \langle A_t\Phi, \Phi\rangle = 0$, we see that at least for simple processes (and indeed this holds for any integrand) the vacuum expectation of stochastic integrals with respect to A_t^* vanish. Note, however, that $A_t^*\Phi \neq 0$.

Our next task is to develop a stochastic calculus; the integrals defined above are not of much use, unless we have an Itô product rule with which they can be manipulated. Once again we run into unpleasant problems. If I_t and J_t are integrals of the form (4.9), there is no reason to expect that their product I_tJ_t is a well-defined operator on the domain $\mathsf{h} \otimes \mathsf{D}$. The idea of Hudson and Parthasarathy is inspired by the identity $\langle \psi', X^*Y\psi\rangle = \langle X\psi', Y\psi\rangle$ for bounded operators; rather than finding an expression for I_tJ_t, they calculate $\langle I_t\, v' \otimes \psi', J_t\, v \otimes \psi\rangle$ for every $v, v' \in \mathsf{h}$, $\psi, \psi' \in \mathsf{D}$, which is always well defined. One finds explicitly a lengthy expression [42, Theorems 4.3 and 4.4], which is essentially the quantum Itô rule expressed in terms of $\mathsf{h} \otimes \mathsf{D}$-matrix elements.

In practice, however, we are mostly interested in calculating actual operator products I_tJ_t. We will need the concept of an *adjoint pair*; two operators X and $X^\dagger$ are said to be an adjoint pair if $\langle v' \otimes \psi', X\, v \otimes \psi\rangle = \langle X^\dagger\, v' \otimes \psi', v \otimes \psi\rangle$ for every $v, v' \in \mathsf{h}$, $\psi, \psi' \in \mathsf{D}$. It is not difficult to verify that if (E, F, G, H) and $(E^\dagger, F^\dagger, G^\dagger, H^\dagger)$ are adjoint pairs, then I_t and $I_t^\dagger$ form an adjoint pair, where

$$I_t^\dagger = \int_0^t (E_t^\dagger\, d\Lambda_t + F_t^\dagger\, dA_t^* + G_t^\dagger\, dA_t + H_t^\dagger\, dt). \tag{4.10}$$

In essence, the adjoint $\dagger$ replaces the Hilbert space adjoint $*$ on the domain $\mathsf{h} \otimes \mathsf{D}$. Now suppose that we can verify explicitly that the product I_tJ_t is well defined; then we can read off an expression for I_tJ_t from the matrix elements $\langle I_t^\dagger\, v' \otimes \psi', J_t\, v \otimes \psi\rangle$. This gives the following explicit form of the quantum Itô rule.

THEOREM 4.3 (quantum Itô rule [55, Proposition 25.26]). *Let (F, G, H, I), (B, C, D, E), and $(B^\dagger, C^\dagger, D^\dagger, E^\dagger)$ be quadruples of stochastically integrable processes such that the latter two quadruples are adjoint pairs. Define the stochastic integrals*

$$dX_t = B_t\, d\Lambda_t + C_t\, dA_t + D_t\, dA_t^* + E_t\, dt,$$
$$dY_t = F_t\, d\Lambda_t + G_t\, dA_t + H_t\, dA_t^* + I_t\, dt,$$

and suppose that we have verified that the product X_tY_t is well defined and that $X_tF_t, \dots, X_tI_t$, $B_tY_t, \dots, E_tY_t$, and $B_tF_t, B_tG_t, \dots, E_tI_t$ are well defined and stochastically integrable. Then the process X_tY_t satisfies the relation

$$d(X_tY_t) = X_t\,dY_t + (dX_t)\,Y_t + dX_t\,dY_t,$$

where $X_t\,dY_t = X_tF_t\,d\Lambda_t + X_tG_t\,dA_t + X_tH_t\,dA_t^* + X_tI_t\,dt$, $(dX_t)\,Y_t = B_tY_t\,d\Lambda_t + C_tY_t\,dA_t + D_tY_t\,dA_t^* + E_tY_t\,dt$, and $dX_t\,dY_t = B_tF_t\,d\Lambda_t + C_tF_t\,dA_t + B_tH_t\,dA_t^* + C_tH_t\,dt$ are evaluated according to the following quantum Itô table.

$dX \setminus dY$	dA_t	$d\Lambda_t$	dA_t^*	dt
dA_t	0	dA_t	dt	0
$d\Lambda_t$	0	$d\Lambda_t$	dA_t^*	0
dA_t^*	0	0	0	0
dt	0	0	0	0

In particular, the theorem holds if B_t, C_t, D_t, E_t, and X_t are bounded processes [42]*, in which case the adjoints $B^\dagger$, $C^\dagger$, etc. are simply taken to be the Hilbert space adjoints B^*, C^*, etc., and X_t extends uniquely to a bounded operator in $\mathscr{W}_{t]}$.*

Remark 4.4. The choice to restrict attention to a fixed domain $\mathsf{h} \otimes \mathsf{D}$ allowed Hudson and Parthasarathy [42] to develop a viable quantum stochastic calculus. This choice, however, has quite a few drawbacks; we highlight one of the problems. Suppose X is self-adjoint; implicit in this statement is that the domains of X and X^* coincide. It can happen that if we restrict the domain of X, then the restricted operator admits many inequivalent self-adjoint extensions; see [56, pages 257–259] for an example. Hence the restriction to a fixed domain can become a real, physical problem, that prevents us from uniquely interpreting unbounded operators on $\mathsf{h} \otimes \mathsf{D}$ as observables.

Such problems have prompted the development of alternative approaches to quantum stochastic integration, and the topic is still under active investigation. In a significant recent achievement Attal and Lindsay [5], building on several earlier approaches (see, e.g., [52, 14] and the references therein), develop a theory in which the integrals achieve their maximal domains. Unfortunately, the theory is very technical and a little daunting for everyday use. A different approach that even precedes Hudson and Parthasarathy is that of Barnett, Streater, and Wilde [8]. Their theory is attractive as it is completely algebraic in nature (the Hilbert space and its domains do not play a fundamental role), but lacks a satisfactory Itô rule.

Despite these issues, the Hudson–Parthasarathy approach works quite well. In practice one usually works with a "noisy Schrödinger equation" (5.2), the solution of which is unitary and thus bounded. As long as the integrals and integrands are bounded, they are uniquely defined by their specification on a dense domain. In this article, in keeping with our attitude towards unbounded operators, we will not worry about such issues and assume that we can apply the quantum Itô rules. □

Example 4.5. In section 5 we will encounter quantum stochastic differential equations (QSDEs), the treatment of which proceeds along the same lines as the classical theory. We claim that the Weyl operator $W(f_{t]})$ is the solution of the QSDE

$$dW(f_{t]}) = \Big\{ f(t)\,dA_t^* - f(t)^*\,dA_t - \frac{1}{2}|f(t)|^2\,dt \Big\} W(f_{t]}). \tag{4.11}$$

In particular, one can verify the Weyl relation $W(f)W(g) = W(f+g)\,e^{i\,\mathrm{Im}\langle g,f\rangle_2}$

directly using the quantum Itô rule. From (4.11) and $W(kf) = e^{ikB(f)}$ we obtain

$$B(f) = \int_0^T (if(t)^*\, dA_t - if(t)\, dA_t^*).$$

Hence $dB_t^\varphi = ie^{-i\varphi(t)}\, dA_t - ie^{i\varphi(t)}\, dA_t^*$, and the quantum Itô rules reduce to the classical Itô rule $(dB_t^\varphi)^2 = dt$. Finally, recall that we defined Poisson processes $\Lambda_t(f) = W(f)^*\Lambda_t W(f) = W(f_{t]})^*\Lambda_t W(f_{t]})$ (the latter equality is due to $W(f) = W(f_{t]}) \otimes W(f_{[t})$ and the fact that $W(f_{[t}) \in \mathscr{W}_{[t}$ is unitary and commutes with the adapted process Λ_t). Using the quantum Itô rule we obtain the explicit representation

$$d\Lambda_t(f) = d\Lambda_t + f(t)^*\, dA_t + f(t)\, dA_t^* + |f(t)|^2\, dt, \tag{4.12}$$

for which the quantum Itô rules reduce to the classical product rule $(d\Lambda_t(f))^2 = d\Lambda_t(f)$ for a Poisson process. □

5. The filtering problem in quantum optics. Many realistic physical scenarios are very well described by quantum stochastic differential equations driven by the processes A_t, A_t^*, and Λ_t discussed in the previous section. Of course, as in the classical theory, white-noise systems are only an idealization of physical interactions; a Markov limit of wide-band noise in the spirit of Wong and Zakai (see [36] for details) gives stochastic models in the Itô form. For a large class of quantum systems, particularly those arising in the field of quantum optics, such approximations are extremely good and describe laboratory experiments essentially to experimental precision. Though a detailed discussion of the physics involved in the modelling of such systems is beyond the scope of this article, we here very briefly describe the physical origin of the equations that are widely used in the physics community [34], describe the measurements that are made, and set up the quantum filtering problem to be solved.

5.1. The quantum optics model. The basic model of quantum optics consists of some fixed physical system, e.g., a collection of atoms, in interaction with the electromagnetic field. The atomic observables are self-adjoint operators on a Hilbert space h. The description of the electromagnetic field and its interaction with the atoms follows from basic physical arguments (see the excellent monograph [23] for a thorough treatment of this theory, known as *quantum electrodynamics*). It turns out that the free electromagnetic field, i.e., an optical field in empty space, is described by a stationary Gaussian (noncommutative) wide-band noise $\tilde{a}(t, \mathbf{r})$ that propagates through space at the speed of light c; i.e., if we restrict ourselves to a single spatial dimension, $\tilde{a}(t + \tau, z) = \tilde{a}(t, z - c\tau)$. If we now place the atoms at the origin $z = 0$, then the quantum dynamics is given by a Schrödinger equation of the form

$$\frac{d}{dt}\tilde{U}(t) = [-iH + L\,\tilde{a}^*(t,0) - L^*\,\tilde{a}(t,0)]\,\tilde{U}(t), \qquad \tilde{U}(0) = I, \tag{5.1}$$

where $L \in \mathscr{B}$ is an atomic (dipole) operator and $H \in \mathscr{B}$ is an atomic Hamiltonian, H being self-adjoint. This equation, which follows directly from the physical model, has wide-band right-hand side. Note that we have set $\hbar = 1$ for convenience, a convention ubiquitous in physics (the only consequence is a change of units).

We now want to approximate the wide-band noise by white noise. This can be done in a rigorous way [1, 2, 36], but we will not detail the procedure here (a brief sketch can be found in [62]). Suffice it to say that one arrives at the following QSDE:

$$dU_t = \Big\{L\, dA_t^* - L^*\, dA_t - \frac{1}{2}L^*L\, dt - iH\, dt\Big\}U_t, \qquad U_0 = I, \tag{5.2}$$

which is driven by the noncommuting white-noise processes A_t and A_t^*. Note that this is almost precisely of the same form as (5.1), except that we have added the Itô correction term $-\frac{1}{2}L^*LU_t\,dt$. A Picard iteration argument [42, 55] ensures existence and uniqueness of the solution. The adjoint U_t^* satisfies

$$dU_t^* = U_t^*\Big\{L^*\,dA_t - L\,dA_t^* - \frac{1}{2}L^*L\,dt + iH\,dt\Big\}, \qquad U_0^* = I.$$

Using the quantum Itô rule we can calculate $d(U_t^*U_t) = d(U_tU_t^*) = 0$; i.e., the solution U_t is unitary for all t (as the solution of a Schrödinger equation should be).

Henceforth we will take (5.2) as our physical model. U_t defines the time evolution or *flow* $j_t : X \mapsto U_t^*(X \otimes I)U_t$ of every atomic observable $X \in \mathscr{B}$ (recall the time evolution in section 2.1); i.e., an observation of $X \in \mathscr{B}$ at time t is described by the observable $X_t = j_t(X)$. Using the Itô rules, we find an explicit dynamical equation

$$dj_t(X) = j_t(\mathcal{L}_{L,H}(X))\,dt + j_t([L^*, X])\,dA_t + j_t([X, L])\,dA_t^*, \quad X \in \mathscr{B}, \tag{5.3}$$

where the so-called Lindblad generator [48] is given by

$$\mathcal{L}_{L,H}(X) = i[H, X] + L^*XL - \frac{1}{2}(L^*LX + XL^*L), \qquad X \in \mathscr{B}.$$

In quantum probability, this object plays the same role as the infinitesimal generator of a Markov diffusion in classical probability theory.

Remark 5.1. Though it is unusual, one could use a very similar notation in classical stochastic models. Suppose some system is described by an underlying configuration x_t that obeys $dx_t = b(x_t)\,dt + \sigma(x_t)\,dW_t$. Then the "observables" in the theory, i.e., things we could try to measure, are functions f of the configuration of the system. The observable f at time t is described by the random variable $j_t(f) = f(x_t)$. Using the classical Itô rules, we get $dj_t(f) = j_t(\mathcal{L}f)\,dt + j_t(\Sigma f)\,dW_t$ where $\mathcal{L}f(x) = \sum_i b^i(x)\partial_i f(x) + \frac{1}{2}\sum_{ij}\sigma^i(x)\sigma^j(x)\partial_i\partial_j f(x)$ is the generator of the Markov diffusion x_t, and $\Sigma f(x) = \sum_i \sigma^i(x)\partial_i f(x)$. This expression is the classical analog of (5.3); the sample paths x_t do not have a quantum counterpart, however. □

5.2. Measurements. Having described the system and its interaction with the field, let us now turn to the observations that we can perform. Unlike in classical models, where one observes the system directly (with the addition of some corrupting noise), in quantum models an observation is generally performed in the field. From the system's perspective, the interaction with the field looks like an (albeit noncommutative) noisy driving force. Similarly, however, the field is perturbed by its interaction with the atoms and carries off information as it propagates away after the interaction. By performing a measurement in the field, then, we can attempt to perform statistical inference of the atomic observables. The entire setup is depicted in Figure 5.1.

To calculate the perturbation of the field by the atoms we once again calculate $U_t^*YU_t$, where now, however, Y is a field observable. The field observable of interest depends on the type of measurement we choose to perform. Without entering into the details, we mention two types of measurement that are extremely common in quantum optics: direct photodetection (photon counting), for which the observation at time t is given by $Y_t^\Lambda = U_t^*\Lambda_tU_t$, and homodyne detection, for which $Y_t^W = U_t^*(A_t + A_t^*)U_t$ (more generally $Y_t^W = U_t^*(e^{-i\varphi}A_t + e^{i\varphi}A_t^*)U_t$). We refer the reader to [6, 7] for a detailed treatment of quantum optical measurements. Using the Itô rules we obtain

$$dY_t^\Lambda = d\Lambda_t + j_t(L)\,dA_t^* + j_t(L^*)\,dA_t + j_t(L^*L)\,dt, \tag{5.4}$$

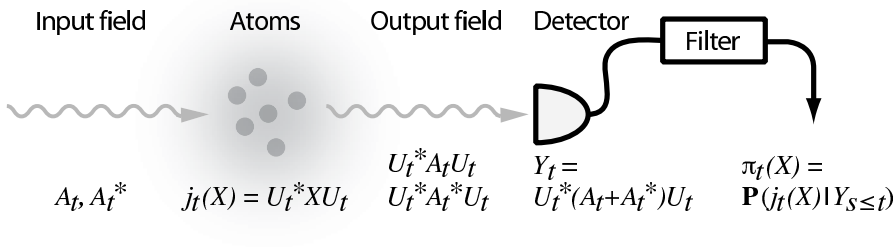


FIG. 5.1. *Diagram of the quantum filtering setup in quantum optics. An optical field, described by the field operators A_t, A_t^*, interacts with a system, e.g., a cloud of atoms. After the atom-field interaction the field operators, as well as system operator, X, are rotated by the unitary U_t. The field is detected, giving rise to the observation Y_t. Finally, the quantum filter (implemented on a classical signal processor) estimates atomic observables based on the field observations.*

$$dY_t^W = j_t(L + L^*)\,dt + dA_t + dA_t^*. \tag{5.5}$$

Intuitively, it would appear that Y_t^Λ is like a Poisson process whose intensity is controlled by $j_t(L^*L)$ (recall Example 4.5), whereas Y_t^W looks like a noisy observation of $j_t(L + L^*)$. One should be careful with this conclusion, however, as $j_t(L)$ need not commute with A_t or A_t^*, nor with itself at different times.

It is essential, however, that the observation process commutes with itself at different times and is hence equivalent to a classical stochastic process through the spectral theorem. An observation process that does not obey this property cannot be observed in a single realization of an experiment and is physically meaningless. Let us show that the observations processes we have defined above do obey this property, which is called the *self-nondemolition* property. Let Z be any operator of the form $I \otimes Z_{s]} \otimes I$ on $\mathsf{h} \otimes \mathsf{F}_{s]} \otimes \mathsf{F}_{[s}$ and let $t \ge s$. Then the Itô rules give directly

$$U_t^* Z U_t = U_s^* Z U_s + \int_s^t U_\tau^* \mathcal{L}_{L,H}(Z) U_\tau \, d\tau + \int_s^t U_\tau^* [L^*, Z] U_\tau \, dA_\tau + \int_s^t U_\tau^* [Z, L] U_\tau \, dA_\tau^*.$$

Now let $Z = A_s + A_s^*$ or $Z = \Lambda_s$. In both cases $\mathcal{L}_{L,H}(Z) = [Z, L] = 0$ as L and H are system observables and Z is a field observable. Hence $Y_s^W = U_t^*(A_s + A_s^*)U_t$ and $Y_s^\Lambda = U_t^* \Lambda_s U_t$ for all $t \ge s$. It is now easily verified, using the unitarity of U_t and the fact that $A_s + A_s^*$ and Λ_s are commutative processes, that $[Y_t^W, Y_s^W] = [Y_t^\Lambda, Y_s^\Lambda] = 0$ for all t, s. We denote by $\mathscr{Y}_t^W$ and $\mathscr{Y}_t^\Lambda$ the commutative von Neumann algebras generated by the observation processes Y_s^W and Y_s^Λ, $s \le t$, respectively. Do note, however, that Y_t^W and Y_t^Λ do not commute with each other; in any experiment, we can choose to perform only one of these measurements. Once we have made this choice, however, we can use the spectral theorem to represent the observations Y_t as a classical stochastic process $\iota(Y_t)$ on a probability space.

5.3. Statement of the filtering problem. Moving on to the next step in our program, we now wish to use the information gained from the measurement process to infer something about the system. To find a least mean square estimate of a system observable $X \in \mathscr{B}$ at time t, given the observations Y_t up to this time, we must calculate the conditional expectation

$$\pi_t(X) = \mathbb{P}(j_t(X) | \mathscr{Y}_t), \tag{5.6}$$

where $\mathscr{Y}_t = \mathrm{vN}(Y_s : 0 \le s \le t)$. The remainder of this article is devoted to finding a recursive equation for $\pi_t(X)$ (the *filtering equation*). Recall, however, that the conditional expectation is only defined if $j_t(X)$ is in the commutant of $\mathscr{Y}_t$, the interpretation being that statistical inference of an observable is only physically meaningful if the conditional statistics could possibly be tested through a compatible experiment. Through an entirely identical procedure to the one used to show the self-nondemolition property, we can show that $j_t(X)$ is in the commutant of $\mathscr{Y}_t$ for any $X \in \mathscr{B}$. This is known as the *nondemolition property*, which can be written as

$$[j_t(X), Y_s] = 0 \quad \forall s \le t,\ X \in \mathscr{B}. \tag{5.7}$$

We note that we have now obtained a system-theoretic model of our system and observations, defined on the quantum probability space $(\mathscr{B} \otimes \mathscr{W}, \mathbb{P} = \rho \otimes \phi)$ by

$$dj_t(X) = j_t(\mathcal{L}_{L,H}(X))\,dt + j_t([L^*, X])\,dA_t + j_t([X, L])\,dA_t^*, \tag{5.8}$$

$$dY_t = j_t(L + L^*)\,dt + dA_t + dA_t^* \tag{5.9}$$

in the case of homodyne detection, or by (5.8), and

$$dY_t = d\Lambda_t + j_t(L)\,dA_t^* + j_t(L^*)\,dA_t + j_t(L^*L)\,dt \tag{5.10}$$

in the case of counting observations. These equations define a system-observation model in direct analogy to such models used throughout classical nonlinear filtering and stochastic control theory.

Remark 5.2. Unlike in a classical filtering scenario, we have not added any independent corrupting noise to the observations. Nonetheless, the filtering problem does not reduce to a problem with complete observations because the system is driven by noise that does not commute with the observations. Hence the problem of partial observations is intrinsic to quantum measurement theory. The quantum filtering problem considered here is the simplest possible one; one could add additional corrupting noise as in the classical case, have the system interact with multiple fields (some of which are observed, others unobserved), etc. These are not essential complications, however, and filters for such models are obtained much in the same way. □

6. The reference probability method. The goal of this section is to derive the quantum filtering equation, a recursive equation for $\pi_t(X)$, using a method that is close to the classical reference probability method of Duncan [29], Mortensen [53], and Zakai [66]. We consider first the homodyne detection case, then the photon counting case. In section 7 we will rederive the filtering equation for the homodyne detection case using martingale methods; the chief advantage of the reference probability method is that it is somewhat simpler to apply. The following approach is based on [18].

6.1. Homodyne detection. Let us briefly recall the classical reference probability procedure; for an introduction see, e.g., [32]. In order to simplify the filtering problem, one starts by introducing a new probability measure, using a Girsanov transformation [49, section 6.3], under which the measurement record is a Wiener process. Then various (elementary) properties of the conditional expectation allow the filtering problem to be expressed, and solved, with respect to the new measure. We now apply this logic to the quantum filtering problem. Note that we have already applied the method in Example 3.19; the following is essentially a continuous time version of that example.

We consider the homodyne detection setup given by (5.8) and (5.9). We could try to find a new state under which Y_t is a Wiener process; however, it will be more convenient to work not in terms of Y_t but in terms of $Z_t = A_t + A_t^*$, as it is very easy to manipulate Z_t using the methods of section 4. Thus before we really start filtering, let us transform the problem in terms of Z_t. Introduce the state $\mathbb{Q}^t$ defined by

$$\mathbb{Q}^t(X) = \mathbb{P}(U_t^* X U_t), \tag{6.1}$$

with U_t as in section 5, and we fix from now on $\mathbb{P} = \rho \otimes \phi$. Now recall from Example 3.19 that $\mathbb{Q}(X) = \mathbb{P}(U^*XU)$ implies $\mathbb{P}(U^*XU|U^*\mathscr{C}U) = U^*\mathbb{Q}(X|\mathscr{C})U$ (this is easily checked using the definition of the conditional expectation). Thus we have

$$\mathbb{P}(j_t(X)|\mathscr{Y}_t) = U_t^* \mathbb{Q}^t(X|\mathscr{C}_t) U_t, \qquad X \in \mathscr{B}, \tag{6.2}$$

where $\mathscr{C}_t = \mathrm{vN}(Z_s : 0 \le s \le t)$. Note that $\mathscr{Y}_t = U_t^* \mathscr{C}_t U_t$ follows from the fact that $U_s^* Z_s U_s = U_t^* Z_s U_t$ for $t \ge s$, the property we used in section 5.2 to prove self-nondemolition of Y_t. The ease with which we will now be able to manipulate $\mathbb{Q}^t(X|\mathscr{C}_t)$ highlights the usefulness of the transformation (6.2).

Our strategy will be as follows. We wish to calculate $\mathbb{Q}^t(X|\mathscr{C}_t)$; however, the state $\mathbb{P}$ has the nice property that $Z_{s\le t}$, which generates $\mathscr{C}_t$, is a $\mathbb{P}$-Wiener process. We want to use the Bayes formula, Lemma 3.18, in order to express $\mathbb{Q}^t(X|\mathscr{C}_t)$ in terms of $\mathbb{P}$-conditional expectations. We run into a problem, however, as the "change of measure" operator U_t that relates $\mathbb{P}$ with $\mathbb{Q}^t$ does not satisfy the requirement of Lemma 3.18 that[12] $U_t \in \mathscr{C}_t'$. To solve this problem, we will replace U_t by a different operator V_t which is affiliated to $\mathscr{C}_t'$, but which still defines the same state in the sense that $\mathbb{P}(U_t^* X U_t) = \mathbb{P}(V_t^* X V_t)$ for every X. The following technique, to our knowledge, first appeared in [39]; it replaces Girsanov's theorem in the quantum context.

LEMMA 6.1. *Let V_t be the solution of the QSDE*

$$dV_t = \Big\{ L\,(dA_t^* + dA_t) - \frac{1}{2} L^* L\,dt - iH\,dt \Big\} V_t. \tag{6.3}$$

Then V_t is affiliated to $\mathscr{C}_t'$ and $\mathbb{Q}^t(X) = \mathbb{P}(V_t^* X V_t)$ for all $X \in \mathscr{B} \otimes \mathscr{W}$.

Proof. Let us assume for simplicity that the state ρ on $\mathscr{B}$ is pure; we can always obtain a mixed state later by taking convex combinations. Then $\mathbb{P}(X) = \langle \psi \otimes \Phi, X\, \psi \otimes \Phi \rangle$ for some vector $\psi \in \mathsf{h}$ (and $\Phi \in \mathsf{F}$ is the vacuum vector). To show that $\mathbb{P}(U_t^* X U_t) = \mathbb{P}(V_t^* X V_t)$, it is thus sufficient to show that

$$U_t\, \psi \otimes \Phi = V_t\, \psi \otimes \Phi. \tag{6.4}$$

Note that $\big\langle (U_t - V_t)\psi \otimes \Phi,\, (U_t - V_t)\psi \otimes \Phi \big\rangle = \big\langle \psi \otimes \Phi,\, (U_t - V_t)^*(U_t - V_t)\psi \otimes \Phi \big\rangle$. A simple application of the quantum Itô rule and the fact that vacuum expectations of stochastic integrals vanish show that $\big\langle \psi \otimes \Phi,\, (U_t - V_t)^*(U_t - V_t)\psi \otimes \Phi \big\rangle = 0$ and (6.4) holds. ∎

Note that the only difference between the equation for U_t, (5.2), and the equation for V_t, (6.3), is that we have modified the coefficient in front of dA_t. In principle, we could change the integrand of the A_t-integral arbitrarily without affecting how the QSDE acts on the vacuum; essentially this is due to the fact that any integral with respect to A_t vanishes when acting on the vacuum, as remarked after Remark 4.2.

[12] If this were the case, then we could calculate $Y_t = U_t^* Z_t U_t = Z_t U_t^* U_t = Z_t$; i.e., the observations would carry no information about the system and the filtering problem would be trivial.

In Lemma 6.1 we exploit this fact to modify U_t precisely so that it is in the commutant of $\mathscr{C}_t$; indeed, (6.3) is driven only by the noise $Z_t = A_t + A_t^*$ and its coefficients are in $\mathscr{B} \subset \mathscr{C}_t'$. We are now ready to apply the Bayes formula, Lemma 3.18. Together with Lemma 6.1 and (6.2), we immediately obtain the following result.

THEOREM 6.2 (noncommutative Kallianpur–Striebel). *Define for any system operator $X \in \mathscr{B}$ the unnormalized conditional expectation*

$$\sigma_t(X) = U_t^* \, \mathbb{P}(V_t^* X V_t | \mathscr{C}_t) \, U_t \in \mathscr{Y}_t. \tag{6.5}$$

Then the conditional expectation (5.6) *is given by*

$$\pi_t(X) = \frac{\sigma_t(X)}{\sigma_t(I)} \qquad \forall \, X \in \mathscr{B}. \tag{6.6}$$

We now obtain an explicit expression for $\sigma_t(X)$.

THEOREM 6.3 (unnormalized quantum filtering equation). *The unnormalized conditional expectation $\sigma_t(X)$ satisfies the following linear QSDE:*

$$d\sigma_t(X) = \sigma_t(\mathcal{L}_{L,H}(X)) \, dt + \sigma_t(L^* X + X L) \, dY_t. \tag{6.7}$$

To obtain (6.7) we will need to take conditional expectations of quantum Itô integrals. Let us briefly show how to do this. First, we claim that if K_t is an adapted process with K_s affiliated to $\mathscr{C}_s'$, then $\mathbb{P}(K_s|\mathscr{C}_t) = \mathbb{P}(K_s|\mathscr{C}_s)$ for $s \le t$. This follows from the fact that $\mathscr{C}_t = \mathscr{C}_s \otimes \mathscr{C}_{[s,t]}$ and that K_s is independent from $\mathscr{C}_{[s,t]}$ by adaptedness. Second, conditional expectations and integrals can be exchanged as follows:

$$\mathbb{P}\left(\int_0^t K_s \, ds \middle| \mathscr{C}_t \right) = \int_0^t \mathbb{P}(K_s|\mathscr{C}_s) \, ds, \qquad \mathbb{P}\left(\int_0^t K_s \, dZ_s \middle| \mathscr{C}_t \right) = \int_0^t \mathbb{P}(K_s|\mathscr{C}_s) \, dZ_s.$$

These properties are immediate if K_t is a simple process, and a proof of the general case is not difficult.

Proof. Using the quantum Itô rules we have

$$V_t^* X V_t = X + \int_0^t V_s^* \mathcal{L}_{L,H}(X) V_s \, ds + \int_0^t V_s^* (L^* X + X L) V_s \, d(A_s + A_s^*).$$

We next take conditional expectations of the terms in this expression; we obtain

$$\begin{aligned} \mathbb{P}(V_t^* X V_t | \mathscr{C}_t) = \mathbb{P}(X) &+ \int_0^t \mathbb{P}(V_s^* \mathcal{L}_{L,H}(X) V_s | \mathscr{C}_s) \, ds \\ &+ \int_0^t \mathbb{P}(V_s^* (L^* X + X L) V_s | \mathscr{C}_s) \, d(A_s + A_s^*). \end{aligned}$$

Another application of the quantum Itô rules now yields (6.7). ∎

By applying the Itô rules to the noncommutative Kallianpur–Striebel formula (6.6), we obtain an expression for the normalized conditional state

$$d\pi_t(X) = \pi_t(\mathcal{L}_{L,H}(X)) dt + \Big(\pi_t(L^* X + X L) - \pi_t(L^* + L) \, \pi_t(X) \Big) \Big(dY_t - \pi_t(L^* + L) \, dt \Big). \tag{6.8}$$

This (normalized) *quantum filtering equation* is a quantum analog of the classical Kushner–Stratonovich equation of nonlinear filtering. Note that this is a classical

stochastic differential equation by the spectral theorem: it is a recursive equation that is only driven by the (commutative) observations Y_t. Hence it can be implemented on a classical (digital) signal processor, as depicted in Figure 5.1.

Remark 6.4. Equation (6.8) is expressed in terms of the conditional state $\pi_t(X)$, where $X \in \mathscr{B}$. Now recall from section 2 that any state on a finite-dimensional Hilbert space can be expressed as $\mathrm{Tr}[\rho X]$ for some density matrix ρ. Similarly, if h (and hence $\mathscr{B}$) is finite dimensional, then we can always write $\pi_t(X) = \mathrm{Tr}[\rho_t X]$ where ρ_t, the conditional density matrix, is a (random) density matrix that is a function of the observations up to time t. From (6.8) we obtain explicitly

$$d\rho_t = -i[H,\rho_t]\,dt + (L\rho_t L^* - \tfrac{1}{2}L^*L\rho_t - \tfrac{1}{2}\rho_t L^*L)\,dt + (L\rho_t + \rho_t L^* - \mathrm{Tr}[(L+L^*)\rho_t]\rho_t)\,dW_t,$$

where $dW_t = dY_t - \mathrm{Tr}[(L+L^*)\rho_t]\,dt$. In section 7 we will see that W_t is a Wiener process. It is this representation that is usually found in the physics literature. □

6.2. Photon counting measurements. We now consider the photon counting setup given by (5.8) and (5.10). We would like to follow the same procedure as for homodyne detection. The following lemma, which replaces Lemma 6.1, suggests how to proceed. The proof is identical to that of Lemma 6.1.

LEMMA 6.5. *Let U'_t be the solution of the QSDE*

$$dU'_t = \Big\{L'\,dA_t^* - L'^*\,dA_t - \frac{1}{2}L'^*L'\,dt - iH'\,dt\Big\}U'_t$$

and let V'_t be the solution of

$$dV'_t = \Big\{L'(d\Lambda_t + dA_t^* + dA_t + dt) - \frac{1}{2}L'^*L'\,dt - L'\,dt - iH'\,dt\Big\}V'_t.$$

Then V'_t is affiliated to $\mathrm{vN}(\Lambda_s + A_s^* + A_s + s : s \le t)'$ *and* $\mathbb{P}(U_t'^* X U'_t) = \mathbb{P}(V_t'^* X V'_t)$.

Define $Z_t = \Lambda_t + A_t^* + A_t + t$ and $\mathscr{C}_t = \mathrm{vN}(Z_s : 0 \le s \le t)$. Lemma 6.5 directly provides us with a nondemolition change of measure, provided that we rotate our problem so that $\mathscr{Y}_t = U_t'^*\mathscr{C}_t U'_t$ using a suitable unitary operator U'_t. Then, defining $\sigma_t(X) = U_t'^*\,\mathbb{P}(V_t'^* X V'_t|\mathscr{C}_t)\,U'_t$, the Kallianpur–Striebel formula holds for $\sigma_t(X)$.

Define R_t as the solution of the QSDE

$$dR_t = (dA_t - dA_t^* - \tfrac{1}{2}dt)\,R_t.$$

Recall Example 4.5; evidently R_t is a Weyl operator, and in particular $\Lambda_t = R_t^* Z_t R_t$. But recall that $Y_t = U_t^*\Lambda_t U_t = U_t^* R_t^* Z_t R_t U_t$; thus $U'_t = R_t U_t$ is our rotation of choice. Using the quantum Itô rules we obtain

$$dU'_t = \Big\{(L-I)\,dA_t^* - (L^*-I)\,dA_t - \frac{1}{2}(L^*L + I - 2L + 2iH)\,dt\Big\}U'_t,$$

which corresponds to the nondemolition change of measure

$$dV'_t = \Big\{(L-I)\,dZ_t - \frac{1}{2}(L^*L - I + 2iH)\,dt\Big\}V'_t.$$

For $X \in \mathscr{B}$, using the quantum Itô rules we obtain

$$dV_t'^* X V'_t = V_t'^*\big(\mathcal{L}_{L,H}(X)\big)V'_t\,dt + V_t'^*(L^*XL - X)V'_t\,(dZ_t - dt).$$

Finally, using the definition of σ_t and the quantum Itô rules we obtain

$$d\sigma_t(X) = \sigma_t(\mathcal{L}_{L,H}(X))\,dt + \big(\sigma_t(L^*XL) - \sigma_t(X)\big)\big(dY_t - dt\big),$$

which is the unnormalized quantum filtering equation for counting observations.

Using the Kallianpur–Striebel formula $\pi_t(X) = \sigma_t(X)\,/\,\sigma_t(I)$ we can now obtain an expression for the normalized conditional state

$$d\pi_t(X) = \pi_t(\mathcal{L}_{L,H}(X))\,dt + \left(\frac{\pi_t(L^*XL)}{\pi_t(L^*L)} - \pi_t(X)\right)\big(dY_t - \pi_t(L^*L)\,dt\big),$$

which is the normalized quantum filtering equation for photon counting.

7. The innovations method. In this section we rederive the filtering equation for homodyne detection, (6.8), using martingale methods that are analogous to the classical case [13, 17]. We follow the classical treatment as in [33], [31, Chapter 18], and [65, Chapter 7]. Martingale methods have enjoyed wide and successful application in classical stochastic theory. The procedure is less straightforward than the reference probability method, however, and some familiarity with classical filtering theory would be helpful (see, e.g., [26] for an excellent introduction).

Let $\xi_t, \beta_t, \lambda_t, \mu_t$ be adapted processes affiliated to $\mathscr{Y}_t'$, where

$$\xi_t = \xi_0 + \int_0^t \beta_s\,ds + m_t = \xi_0 + \int_0^t \beta_s\,ds + \int_0^t (\lambda_s\,dA_s + \mu_s\,dA_s^*). \tag{7.1}$$

The measurement process Y_t is given by (5.9), and in what follows we write $h_t = j_t(L + L^*)$ and $Z_t = A_t + A_t^*$. Note that the conditional expectation $\hat{\xi}_t = \mathbb{P}(\xi_t|\mathscr{Y}_t)$ is well defined, and similarly for the coefficients β_t, λ_t, and μ_t.

The main filtering result for a process of the form (7.1) is the following.

THEOREM 7.1 (noncommutative Fujisaki–Kallianpur–Kunita). *Under the above assumptions, the filtered process $\hat{\xi}_t$ satisfies the QSDE*

$$d\hat{\xi}_t = \hat{\beta}_t\,dt + (\hat{\lambda}_t + \widehat{\xi_t h_t} - \hat{\xi}_t\hat{h}_t)\,dW_t, \tag{7.2}$$

where $\hat{r}_t \equiv \mathbb{P}(r_t|\mathscr{Y}_t)$ for any r_t affiliated to $\mathscr{Y}_t'$, and $dW_t = dY_t - \hat{h}_t\,dt$ defines the $\mathscr{Y}_t$-Wiener process (with respect to $\mathbb{P}$) W_t, called the innovations process.

The filtering expression (7.2) is formally identical to the classical case [31, Theorem 18.11] and [65, Proposition 3.2]. Before we prove Theorem 7.1, we will show how to obtain the quantum filtering equation (6.8) using this result.

COROLLARY 7.2. *The conditional state $\pi_t(X)$ is given by* (6.8).

Proof. We set $\lambda_t = -j_t([X, L^*])$, $\mu_t = j_t([X, L])$, $\beta_t = j_t(\mathcal{L}_{L,H}(X))$, and $\xi_t = j_t(X)$. Then $\widehat{\xi_t h_t} = \pi_t(X(L + L^*))$, $\hat{\xi}_t\hat{h}_t = \pi_t(X)\pi_t(L + L^*)$, $\hat{\lambda}_t = -\pi_t([X, L^*])$, and $\hat{\beta}_t = \pi_t(\mathcal{L}_{L,H}(X))$. Hence using (7.2), (6.8) follows. ∎

Proof of Theorem 7.1. *Step* 1. We first show that the process

$$M_t = \hat{\xi}_t - \hat{\xi}_0 - \int_0^t \hat{\beta}_s\,ds$$

is a $\mathscr{Y}_t$-martingale, i.e., $\mathbb{P}(M_t|\mathscr{Y}_s) = M_s$ for all $s \le t$. This property is equivalent to $\mathbb{P}((M_t - M_s)K) = 0$ for all $K \in \mathscr{Y}_s$, or equivalently

$$\mathbb{P}\left[\left(\hat{\xi}_t - \hat{\xi}_s - \int_s^t \hat{\beta}_r\,dr\right)K\right] = \mathbb{P}\left[\left(\xi_t - \xi_s - \int_s^t \beta_r\,dr\right)K\right] = \mathbb{P}[(m_t - m_s)K] = 0$$

for all $K \in \mathscr{Y}_s$, where we have used Definition 3.13 in the first step. But as $K \in \mathscr{Y}_s \subset \mathscr{B} \otimes \mathscr{W}_{s]}$,

$$\mathbb{P}[(m_t - m_s)K] = \mathbb{P}\left[K \int_s^t (\lambda_r \, dA_r + \mu_r \, dA_r^*)\right] = \mathbb{P}\left[\int_s^t (K\lambda_r \, dA_r + K\mu_r \, dA_r^*)\right] = 0,$$

where we have used that the vacuum expectation of quantum Itô integrals vanishes. Thus we have demonstrated that M_t is a $\mathscr{Y}_t$-martingale.

Step 2. We now show that W_t is a Wiener process under $\mathbb{P}$. We begin by verifying that the innovations process

$$W_t = Y_t - \int_0^t \hat{h}_s \, ds \tag{7.3}$$

is a $\mathscr{Y}_t$-martingale. We need to show that $\mathbb{P}[(W_t - W_s)K] = 0$ for any $s \le t$ and $K \in \mathscr{Y}_s$. This is equivalent to

$$\mathbb{P}\left[\left(Y_t - Y_s - \int_s^t \hat{h}_r \, dr\right) K\right] = \mathbb{P}\left[\left(Y_t - Y_s - \int_s^t h_r \, dr\right) K\right] = 0$$

for all $K \in \mathscr{Y}_s$, where the second expression follows from the definition of the conditional expectation. But from (5.9) we obtain

$$\mathbb{P}\left[\left(Y_t - Y_s - \int_s^t h_r \, dr\right) K\right] = \mathbb{P}[(Z_t - Z_s)K] = 0$$

as $K \in \mathscr{Y}_s \subset \mathscr{B} \otimes \mathscr{W}_{s]}$, $(Z_t - Z_s) \in \mathscr{W}_{[s,t]}$, and hence $\mathbb{P}[(Z_t - Z_s)K] = \mathbb{P}(K)\,\mathbb{P}(Z_t - Z_s) = 0$. Thus W_t is a $\mathscr{Y}_t$-martingale.

From (7.3) we read off the Itô rule $dW_t^2 = dt$; classically, a process that obeys this property and is a martingale must be a Wiener process by Lévy's theorem (e.g., [31, Lemma 18.7]). But we can simply apply the classical result, as W_t is a commutative process (note that $\hat{h}_t \in \mathscr{Y}_t$ for $s \le t$ by construction) and is hence equivalent to the corresponding classical process obtained through the spectral theorem.

Now that we have shown that W_t is a Wiener process, we can try to represent the martingale M_t as a stochastic integral with respect to W_t. As usual in filtering theory the ordinary martingale representation theorem does not suffice for this purpose, but the representation theorem of Fujisaki–Kallianpur–Kunita (e.g., [49, Theorem 5.20]) allows us to conclude nonetheless that

$$M_t = \int_0^t \gamma_s \, dW_s \quad \Longrightarrow \quad \hat{\xi}_t = \hat{\xi}_0 + \int_0^t \hat{\beta}_s \, ds + \int_0^t \gamma_s \, dW_s \tag{7.4}$$

for some adapted process $\gamma_t \in \mathscr{Y}_t$.

Step 3. We next obtain a first expression for $\widehat{\xi_t Y_t}$:

$$\widehat{\xi_t Y_t} = \int_0^t [\widehat{\beta_s Y_s} + \widehat{\xi_s h_s} + \hat{\lambda}_s] ds + M_1(t), \tag{7.5}$$

where $M_1(t)$ is a $\mathscr{Y}_t$-martingale. As before, it suffices to show that

$$\mathbb{P}((M_1(t) - M_1(s))K) = \mathbb{P}\left[\left(\xi_t Y_t - \xi_s Y_s - \int_s^t [\beta_s Y_s + \xi_s h_s + \lambda_s] ds\right) K\right] = 0$$

for all $K \in \mathscr{Y}_s$, where we have used the definition of the conditional expectation. But

$$\begin{aligned} d(\xi_t Y_t) &= (d\xi_t)Y_t + \xi_t dY_t + d\xi_t dY_t \\ &= (\beta_t dt + dm_t)Y_t + \xi_t(hdt + dZ_t) + dm_t dZ_t \\ &= (\beta_t Y_t + \xi_t h_t + \lambda_t)dt + (Y_t\lambda_t + \xi_t)dA_t + (Y_t\mu_t + \xi_t)dA_t^*. \end{aligned}$$

Hence exactly as before, it follows that $M_1(t)$ is a $\mathscr{Y}_t$-martingale.

Step 4. Next, we derive a second expression for $\widehat{\xi_t Y_t}$:

$$\widehat{\xi_t Y_t} = \int_0^t [\hat{\beta}_s Y_s + \hat{\xi}_s \hat{h}_s + \gamma_s]ds + M_2(t), \tag{7.6}$$

where $M_2(t)$ is a $\mathscr{Y}_t$-martingale. To show this, note that $\widehat{\xi_t Y_t} = \hat{\xi}_t Y_t$. By Itô's rules,

$$\begin{aligned} d(\hat{\xi}_t Y_t) &= (d\hat{\xi}_t)Y_t + \hat{\xi}_t dY_t + d\hat{\xi}_t dY_t \\ &= (\hat{\beta}_t dt + \gamma_t dW_t)Y_t + \hat{\xi}_t(\hat{h}_t dt + dW_t) + \gamma_t dW_t dW_t \\ &= (\hat{\beta}_t Y_t + \hat{\xi}_t \hat{h}_t + \gamma_t)dt + (\gamma_t Y_t + \hat{\xi}_t)dW_t \end{aligned}$$

which establishes (7.6).

Step 5. We can now identify γ_t. From (7.5) and (7.6) we have two representations for $\widehat{\xi_t Y_t}$. By uniqueness, it follows that the finite variation terms are equal, namely,

$$\widehat{\beta_s Y_s} + \widehat{\xi_s h_s} + \hat{\lambda}_s = \hat{\beta}_s Y_s + \hat{\xi}_s \hat{h}_s + \gamma_s.$$

Therefore $\gamma_s = \widehat{\xi_s h_s} + \hat{\lambda}_s - \hat{\xi}_s \hat{h}_s$ as required. □

8. Conclusion. In this article we have provided an introduction to quantum filtering. Our goal has been to emphasize the mathematical structures of quantum probability and to show their use in system-probe models from quantum optics. We have seen that the techniques employed in quantum filtering theory closely mirror their analogs in the classical theory of nonlinear filtering. As in the classical theory, an important role is played by the conditional expectation as the mean least square estimate of the system given the observations thus far.

The spectral theorem provides a one-to-one correspondence between commutative von Neumann algebras equipped with normal states and classical (Kolmogorov) probability spaces. This enabled us to represent commuting observables (self-adjoint operators acting on a Hilbert space) as random variables on a single classical probability space. For an observable X that commutes with all members of a commutative family of observables $\mathscr{Y}$, we can define the conditional expectation of X onto $\mathscr{Y}$ by pulling back the classical conditional expectation of X onto $\mathscr{Y}$, both represented on a classical probability space via the spectral theorem. For this procedure to work, it is crucial that the family $\mathscr{Y}$ is commutative, the *self-nondemolition* property, and that X commutes with $\mathscr{Y}$, the *nondemolition* property.

As a model for the quantum electromagnetic field we introduced the algebra $\mathscr{W}$ of bounded operators on the Boson Fock space equipped with the vacuum state ϕ. We studied families of commuting operators affiliated to $\mathscr{W}$, and by representing these commuting operators on a classical probability space, we found that $(\mathscr{W}, \phi)$ contains families of Wiener and Poisson processes, all of which can be written as linear combinations of the so-called fundamental noises Λ_t (gauge process), A_t^* (creation process), and A_t (annihilation process). These families of operators do not necessarily

commute with each other, and therefore the different Wiener and Poisson processes in $(\mathscr{W}, \phi)$ cannot be represented on the same classical probability space. Physically, these processes can all be observed with a suitable measurement setup (e.g., homodyne detection or photon counting); however, these observations cannot be made in a single realization of the experiment, as the different families do not commute.

Although the noncommuting processes affiliated to $(\mathscr{W}, \phi)$ cannot be represented on the same probability space, it is still possible to capture them within a single stochastic calculus. Stochastic integrals with respect to the fundamental noises can be defined as operators affiliated with $\mathscr{W}$, and a quantum Itô rule (integration by parts rule) based on a quantum Itô table for the fundamental noises can be shown to hold [42]. The interaction of some fixed system, e.g., a cloud of atoms, with the electromagnetic field in a Markov limit [1, 36] is given by a unitary U_t that satisfies a quantum stochastic differential equation, i.e., a stochastic differential equation given in terms of quantum stochastic integrals with respect to the fundamental noises. Note that the equation for U_t can therefore be driven by different noises that do not necessarily commute with each other.

Given the unitary U_t, the Heisenberg evolution of the observables of the system (e.g., a cloud of atoms) is given by $j_t(X) = U_t^* X U_t$. Instead of directly observing the system at time t, we only had access to field observables up to time t, e.g., $Y_s = U_s^*(A_s + A_s^*)U_s,\ 0 \le s \le t$ (homodyne detection) or $Y_s = U_s^* \Lambda_s U_s,\ 0 \le s \le t$ (photon counting). We showed that these system-observation pairs satisfy the nondemolition requirements that are necessary for the existence of the conditional expectation $\mathbb{P}(j_t(X)|\mathscr{Y}_t)$ of a system observable $j_t(X)$ at time t onto the observations thus far. The quantum filtering equation recursively propagates the conditional expectation, our best estimate of system observables, in time. In close analogy with the classical case, we derived the quantum filtering equation in two ways, once using a change of measure technique and once using martingales and martingale representation.

There are many points we did not touch upon in this introduction to quantum filtering. Some noteworthy omissions are the linear theory [30] and models in discrete time with discrete observables [19]. Another notable omission is how the filtering equations can be used when controlling a quantum system [10, 11, 28, 18]. As in the classical case, a separation theorem can be shown to hold [18]. This means that the optimal controller will depend on the observation history only through the filter. This separates the control problem into an estimation step (filtering) and a control step based on the estimates only.

REFERENCES

[1] L. Accardi, A. Frigerio, and Y. Lu, *The weak coupling limit as a quantum functional central limit*, Comm. Math. Phys., 131 (1990), pp. 537–570.

[2] L. Accardi, J. Gough, and Y. Lu, *On the stochastic limit for quantum theory*, Rep. Math. Phys., 36 (1995), pp. 155–187.

[3] M. A. Armen, J. K. Au, J. K. Stockton, A. C. Doherty, and H. Mabuchi, *Adaptive homodyne measurement of optical phase*, Phys. Rev. Lett., 89 (2002), 133602.

[4] E. Arthurs and J. L. Kelly, *On the simultaneous measurement of a pair of conjugate observables*, Bell Syst. Tech. J., 44 (1965), pp. 725–729.

[5] S. Attal and J. M. Lindsay, *Quantum stochastic calculus with maximal operator domains*, Ann. Probab., 32 (2004), pp. 488–529.

[6] A. Barchielli, *Continual measurements in quantum mechanics and quantum stochastic calculus*, in Open Quantum Systems III: Recent Developments, S. Attal, A. Joye, and C.-A. Pillet, eds., Springer, Berlin, Heidelberg, 2006, pp. 207–292.

[7] A. Barchielli, *Direct and heterodyne detection and other applications of quantum stochastic calculus to quantum optics*, Quantum Opt., 2 (1990), pp. 423–441.
[8] C. Barnett, R. F. Streater, and I. F. Wilde, *Quasi-free quantum stochastic integrals for the CAR and CCR*, J. Funct. Anal., 52 (1983), pp. 19–47.
[9] V. P. Belavkin, *Quantum filtering of Markov signals with white quantum noise*, Radiotechnika i Electronika, 25 (1980), pp. 1445–1453.
[10] V. P. Belavkin, *Theory of the control of observable quantum systems*, Autom. Rem. Control, 44 (1983), pp. 178–188.
[11] V. P. Belavkin, *Nondemolition measurement and control in quantum dynamical systems*, in Information Complexity and Control in Quantum Physics, CISM Courses and Lectures 294, S. Diner and G. Lochak, eds., Springer-Verlag, Vienna, 1987, pp. 331–336.
[12] V. P. Belavkin, *Quantum continual measurements and a posteriori collapse on CCR*, Comm. Math. Phys., 146 (1992), pp. 611–635.
[13] V. P. Belavkin, *Quantum stochastic calculus and quantum nonlinear filtering*, J. Multivariate Anal., 42 (1992), pp. 171–201.
[14] P. Biane, *Calcul stochastique noncommutatif*, in Lectures on Probability Theory (Saint-Flour, 1993), Lecture Notes in Math. 1608, P. Bernard, ed., Springer-Verlag, Berlin, 1995, pp. 1–96.
[15] L. M. Bouten, *Filtering and Control in Quantum Optics*, Ph.D. thesis, Radboud Universiteit Nijmegen, 2004.
[16] L. M. Bouten, S. C. Edwards, and V. P. Belavkin, *Bellman equations for optimal feedback control of qubit states*, J. Phys. B, At. Mol. Opt. Phys., 38 (2005), pp. 151–160.
[17] L. M. Bouten, M. I. Guţă, and H. Maassen, *Stochastic Schrödinger equations*, J. Phys. A, 37 (2004), pp. 3189–3209.
[18] L. M. Bouten and R. Van Handel, *On the separation principle of quantum control,* in Proceedings of the 2006 QPIC Symposium (Nottingham, UK), M. Guta, ed., World Scientific, Singapore, to appear.
[19] L. M. Bouten, R. Van Handel, and M. R. James, *A discrete invitation to quantum filtering and feedback control*, SIAM Rev., to appear.
[20] O. Bratteli and D. Robinson, *Operator Algebras and Quantum Statistical Mechanics* 1, 2nd ed., Springer-Verlag, Berlin, Heidelberg, 1987.
[21] P. Bushev, D. Rotter, A. Wilson, F. Dubin, C. Becher, J. Eschner, R. Blatt, V. Steixner, P. Rabl, and P. Zoller, *Feedback cooling of a single trapped ion*, Phys. Rev. Lett., 96 (2006), p. 043003.
[22] H. J. Carmichael, *An Open Systems Approach to Quantum Optics*, Springer-Verlag, Berlin, Heidelberg, New York, 1993.
[23] C. Cohen Tannoudji, J. Dupont Roc, and G. Grynberg, *Photons and Atoms: Introduction to Quantum Electrodynamics*, Wiley, New York, 1989.
[24] E. B. Davies, *Quantum stochastic processes*, Comm. Math. Phys., 15 (1969), pp. 277–304.
[25] E. B. Davies, *Quantum Theory of Open Systems*, Academic Press, London, New York, San Francisco, 1976.
[26] M. H. A. Davis and S. I. Marcus, *An introduction to nonlinear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., D. Reidel, Dordrecht, 1981, pp. 53–75.
[27] A. C. Doherty, S. Habib, K. Jacobs, H. Mabuchi, and S. M. Tan, *Quantum feedback and classical control theory*, Phys. Rev. A, 62 (2000), 012105.
[28] A. C. Doherty and K. Jacobs, *Feedback-control of quantum systems using continuous state-estimation*, Phys. Rev. A, 60 (1999), pp. 2700–2711.
[29] T. Duncan, *Evaluation of likelihood functions*, Inform. and Control, (1968), pp. 62–74.
[30] S. C. Edwards and V. P. Belavkin, *Optimal Quantum Feedback Control via Quantum Dynamic Programming*, preprint quant-ph/0506018v2, University of Nottingham, 2005.
[31] R. J. Elliott, *Stochastic Calculus and Applications*, Springer-Verlag, New York, 1982.
[32] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*, Springer-Verlag, New York, 1995.
[33] M. Fujisaki, G. Kallianpur, and H. Kunita, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 1 (1972), pp. 19–40.
[34] C. W. Gardiner and M. J. Collett, *Input and output in damped quantum systems: Quantum stochastic differential equations and the master equation*, Phys. Rev. A, 31 (1985), pp. 3761–3774.
[35] J. M. Geremia, J. K. Stockton, A. C. Doherty, and H. Mabuchi, *Quantum Kalman filtering and the Heisenberg limit in atomic magnetometry*, Phys. Rev. Lett., 91 (2003), 250801.

[36] J. Gough, *Quantum flows as Markovian limit of emission, absorption and scattering interactions*, Comm. Math. Phys., 254 (2005), pp. 489–512.
[37] J. Gough, V. P. Belavkin, and O. G. Smolyanov, *Hamilton-Jacobi-Bellman equations for quantum filtering and control*, J. Opt. B Quantum Semiclass. Opt., 7 (2005), pp. S237–S244.
[38] A. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory*, North–Holland, Amsterdam, New York, Oxford, 1982.
[39] A. Holevo, *Quantum stochastic calculus*, J. Soviet Math., 56 (1991), pp. 2609–2624 (in English); Itogi Nauki i Tekhniki, Ser. Sovr. Prob. Mat. 36 (1990), pp. 3–28 (in Russian).
[40] A. Holevo, *Statistical Structure of Quantum Theory*, Springer-Verlag, Berlin, 2001.
[41] R. L. Hudson, *An introduction to quantum stochastic calculus and some of its applications*, in Quantum Probability Communications, vol. XI, S. Attal and J. Lindsay, eds., World Scientific, Singapore, 2003, pp. 221–271.
[42] R. L. Hudson and K. R. Parthasarathy, *Quantum Itô's formula and stochastic evolutions*, Comm. Math. Phys., 93 (1984), pp. 301–323.
[43] M. R. James, *Risk-sensitive optimal control of quantum systems*, Phys. Rev. A, 69 (2004), 032108.
[44] M. R. James, *A quantum Langevin formulation of risk-sensitive optimal control*, J. Opt. B Quantum Semiclass. Opt., 7 (2005), pp. S198–S207.
[45] R. V. Kadison and J. R. Ringrose, *Fundamentals of the Theory of Operator Algebras*, Vol. I, Academic Press, San Diego, 1983.
[46] R. V. Kadison and J. R. Ringrose, *Fundamentals of the Theory of Operator Algebras*, Vol. II, Academic Press, San Diego, 1986.
[47] B. Kümmerer, *Markov dilations on W^*-algebras*, J. Funct. Anal., 63 (1985), pp. 139–177.
[48] G. Lindblad, *On the generators of quantum dynamical semigroups*, Comm. Math. Phys., 48 (1976), pp. 119–130.
[49] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes* I: *General Theory*, Springer-Verlag, Berlin, 2001.
[50] H. Maassen, *Quantum Markov processes on Fock space described by integral kernels*, in QP and Applications II, Lecture Notes in Math. 1136, L. Accardi and W. von Waldenfels, eds., Springer-Verlag, Berlin, 1985, pp. 361–374.
[51] E. Merzbacher, *Quantum Mechanics*, 3rd ed., Wiley, New York, 1998.
[52] P.-A. Meyer, *Quantum Probability for Probabilists*, Springer-Verlag, Berlin, 1993.
[53] R. Mortensen, *Optimal Control of Continuous-Time Stochastic Systems*, Ph.D. thesis, University of California, Berkeley, 1966.
[54] E. Nelson, *Notes on non-commutative integration*, J. Funct. Anal., 15 (1974), pp. 103–116.
[55] K. R. Parthasarathy, *An Introduction to Quantum Stochastic Calculus*, Birkhäuser, Basel, 1992.
[56] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, Vol.* 1: *Functional Analysis,* Academic Press, San Diego, London, 1980.
[57] S. Sakai, *C^*-algebras and W^*-algebras*, Springer-Verlag, Berlin, 1998.
[58] J. K. Stockton, *Continuous Quantum Measurement of Cold Alkali-Atom Spins*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 2006.
[59] M. Takesaki, *Conditional expectations in von Neumann algebras*, J. Funct. Anal., 9 (1971), pp. 306–321.
[60] R. Van Handel and H. Mabuchi, *Quantum projection filter for a highly nonlinear model in cavity QED*, J. Opt. B Quantum Semiclass. Opt., 7 (2005), pp. S226–S236.
[61] R. Van Handel, J. K. Stockton, and H. Mabuchi, *Feedback control of quantum state reduction*, IEEE Trans. Automat. Control, 50 (2005), pp. 768–780.
[62] R. Van Handel, J. K. Stockton, and H. Mabuchi, *Modelling and feedback control design for quantum state preparation*, J. Opt. B Quantum Semiclass. Opt., 7 (2005), pp. S179–S197.
[63] D. Williams, *Probability with Martingales*, Cambridge University Press, Cambridge, UK, 1991.
[64] H. Wiseman, *Quantum theory of continuous feedback*, Phys. Rev. A, 49 (1994), pp. 2133–2150.
[65] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*, Springer-Verlag, New York, 1985.
[66] M. Zakai, *On the optimal filtering of diffusion processes*, Z. Wahrsch. Verw. Geb., 11 (1969), pp. 230–243.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2242–2255

SINGLE OUTPUT-DEPENDENT OBSERVABILITY NORMAL FORM*

GANG ZHENG†, DRISS BOUTAT‡, AND JEAN-PIERRE BARBOT§

Abstract. This paper gives the sufficient and necessary conditions which guarantee the existence of a diffeomorphism in order to transform a nonlinear system without inputs into a canonical normal form that is output dependent. Moreover, we extend our results to a class of systems with inputs.

Key words. observability, normal forms

AMS subject classifications. 15A15, 15A09, 15A23

DOI. 10.1137/050627137

1. Introduction. Since Luenberger's work [9], the design of an observer for observable linear systems with linear outputs has been a well-known concept. In order to use the same observer for nonlinear systems, the so-called observability linearization problem for nonlinear systems was born. The sufficient and necessary conditions which guarantee the existence of a diffeomorphism and of an output injection to transform a single output nonlinear system without inputs into a linear one with an output injection were firstly addressed in [12]. Then, for a multi-output nonlinear system without inputs, the linearization problem was partially solved in [13]. The complete solution to the linearization problem was given in [16]. Another approach was introduced for the analytical systems in [11] by assuming that the spectrum of the linear part must lie in the Poincaré domain, and it was generalized in [14] by assuming that the spectrum of the linear part must lie in the Siegel domain. These assumptions are not generically fulfilled. Other approaches using quadratic normal forms were given in [1] and [3]. All these approaches enable us to design an observer for a larger class of nonlinear systems.

Meanwhile, other researchers worked directly on designing nonlinear observers, such as high-gain observers [6], [4], [7]. Nevertheless, even if the conditions which guarantee the linearization method to design an observer were not generically fulfilled, this method would still remain important for the nonlinear observer design because (1) it works well for nonanalytic systems, and (2) because it could be used not only in adaptive theory but also for the observation of systems with unknown inputs. All these reasons explain why researchers continue to investigate this matter.

In [10], the author gave the sufficient and necessary geometrical conditions to transform a nonlinear system into a so-called output-dependent time scaling linear canonical form, while the author of [5] gave the dual geometrical conditions of [10].

In this paper, as an extension of [17], we will propose a method to deduce the geometrical conditions which are sufficient and necessary to guarantee the existence

*Received by the editors March 18, 2005; accepted for publication (in revised form) June 23, 2007; published electronically December 21, 2007.
http://www.siam.org/journals/sicon/46-6/62713.html
†INRIA Rhône Alpes, 655 Avenue de l'Europe, 38334 St Ismier Cedex, France (gang.zheng@inrialpes.fr).
‡LVR ENSI, 10 Boulevard de Lahitolle, 18020 Bourges, France (driss.boutat@ensi-bourges.fr).
§ECS ENSEA, 6 Avenue du Ponceau, 95014 Cergy-Pontoise, and Project ALIEN, INRIA-Futur, Orsay, France (barbot@ensea.fr).

of a local diffeomorphism $z = \phi(x)$ which transforms the locally observable dynamical system

$$\left\{ \begin{array}{l} \dot{x} = f(x), \\ y = h(x), \end{array} \right. \tag{1.1}$$

where $x \in U \subset \mathbb{R}^n$, $f : U \subset \mathbb{R}^n \to \mathbb{R}^n$ and $h : U \subset \mathbb{R}^n \to \mathbb{R}$ are sufficiently smooth, into the form

$$\left\{ \begin{array}{l} \dot{z} = A(y)z + \beta(y), \\ y = z_n = Cz, \end{array} \right. \tag{1.2}$$

where

$$A(y) = \begin{pmatrix} 0 & \cdots & 0 & 0 & 0 \\ \alpha_1(y) & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \cdots & \vdots \\ 0 & \cdots & \alpha_{n-2}(y) & 0 & 0 \\ 0 & \cdots & 0 & \alpha_{n-1}(y) & 0 \end{pmatrix}, \ \beta(y) = \begin{pmatrix} \beta_1(y) \\ \beta_2(y) \\ \vdots \\ \beta_{n-1}(y) \\ \beta_n(y) \end{pmatrix},$$

and where $\alpha_i(y) \neq 0$ for $y \in]-a, a[$ and $a > 0$. This kind of linearization is called the single output-dependent observability (SODO) normal form.

For dynamical systems in the form of (1.2) we may, for example, apply the following high-gain observer [2]:

$$\left\{ \begin{array}{l} \dot{\hat{z}} = A(y)\hat{z} + \beta(y) - \Gamma^{-1}(y) R_\rho^{-1} C^T (C\hat{z} - y), \\ 0 = -\rho R_\rho - \bar{A}^T R_\rho - R_\rho \bar{A} + C^T C, \end{array} \right. \tag{1.3}$$

where $\Gamma(y)$ is the $n \times n$ diagonal matrix

$$\Gamma(y) = \operatorname{diag}\left[\prod_{i=1}^{n-1} \alpha_i(y), \prod_{i=2}^{n-1} \alpha_i(y), \ldots, \alpha_{n-1}(y), 1 \right],$$

and $\bar{A}$ is the $n \times n$ matrix defined as

$$\bar{A} = \begin{pmatrix} 0 & \cdots & 0 & 0 \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Indeed, here the output of system (1.2) is considered as an input of (1.3). Setting $e = z - \hat{z}$, we see that the observation error can be obtained as follows:

$$\dot{e} = \left(A(y) - \Gamma^{-1}(y) R_\rho^{-1} C^T C \right) e.$$

The convergence of such an observer is proved in [2], and thus in section 4 we simply highlight the design of such an observer for systems in the form (1.2).

Moreover, we generalize our result to a class of systems with inputs. Then we discuss some useful corollaries in order to deal with affine systems and the so-called left invertibility problem.

This paper is organized as follows. The next section addresses notations and technical results which are key to proving our main result. In section 3, we present

our method to deduce the geometrical conditions for a nonlinear system without inputs in order to transform it into a SODO normal form. Section 4 is devoted to the generalization of our results to a class of systems with inputs. Also in section 4, some practical cases are studied, including the state affine systems and the left invertibility problem. Throughout this paper, examples are shown in order to highlight our theoretical results.

2. Notations and technical results. Throughout this article, we denote by $L_f^{i-1}h$ for $1 \leq i \leq n$ the $(i-1)$th Lie derivative of output h in the direction of f, and we set $\theta_i = dL_f^{i-1}h$ as its differential. Assume that system (1.1) is locally observable, and thus that $\theta = (\theta_1, \ldots, \theta_n)^T$ is a basis of the cotangent bundle T^*U of U. Then we also consider the vector field τ_1 defined in [12] as

$$\begin{cases} \theta_i(\tau_1) = 0 \text{ for } 1 \leq i \leq n-1, \\ \theta_n(\tau_1) = 1, \end{cases} \tag{2.1}$$

and by induction we define

$$\tau_k = (-1)^{k-1} ad_f^{k-1}(\tau_1) \text{ for } 2 \leq k \leq n. \tag{2.2}$$

It is clear that $\{\tau_1, \ldots, \tau_n\}$ is a basis of the tangent bundle TU of U.

Let us recall a famous result from [12].

THEOREM 2.1. *The following conditions are equivalent:*

(i) *There exist a diffeomorphism and an output injection which transform system* (1.1) *into normal form* (1.2) *with* $\alpha_k(y) = 1$ *for* $1 \leq k \leq n-1$.

(ii) $[\tau_i, \tau_j] = 0$ *for* $1 \leq i, j \leq n$.

If for some $1 \leq k \leq n-1$ the functions $\alpha_k(y)$ in the form (1.2) are not constant, then (ii) of Theorem 2.1 is not fulfilled. Consequently, the rest of this section is devoted to using $[\tau_i, \tau_n]$ in order to determine all the functions $\alpha_i(y)$ for $1 \leq i \leq n-1$.

LEMMA 2.2. *For a system in the form* (1.2) *we have for* $1 \leq k \leq n-1$,

$$\begin{aligned} \tau_k = {} & \frac{1}{\pi_k}\frac{\partial}{\partial z_k} + \left(A_{k-1}^k(z_n)\, z_{n-1} + \eta_{k-1}^k(z_n)\right)\frac{\partial}{\partial z_{k-1}} \\ & + \sum_{i=1}^{k-2}\left(A_i^k(z_n)\, z_{n-k+i} + \sum_{j=n-k+i+1}^{n-1}\sum_{l=j}^{n-1} T_{j,l}^k(z_n)\, z_j z_l\right)\frac{\partial}{\partial z_i} \\ & + \sum_{i=1}^{k-2}\left(\sum_{j=n-k+i+1}^{n}\eta_i^k(z_n)\, z_j + O_{z_n}^{[3]}(z_{n-k+i+1}, \ldots, z_{n-1})\right)\frac{\partial}{\partial z_i}, \end{aligned} \tag{2.3}$$

where $\pi_n = 1$ *and* $\pi_{k-1} = \pi_k \alpha_{k-1}$ *for* $2 \leq k \leq n$; $\eta_i^k(z_n)$ *and* $T_{j,l}^k(z_n)$ *are some smooth functions of* z_n; $O_{z_n}^{[3]}(z_{n-k+2}, \ldots, z_{n-1})$ *represents the residue higher than order* 2 *with a coefficient which is a function of* z_n; *and*

$$\begin{aligned} A_i^k(z_n) = {} & (-1)^{k-i+1} \\ & \times \left(S_{k-i,1}^k \frac{\pi_i'}{\pi_i^2} + \sum_{m=k-i+1}^{k-1} S_{k-i,m-k+i+1}^k \frac{\pi_{k-m}'}{\pi_{k-m}^2}\left(\prod_{j=k-i+1}^{m}\alpha_{k-j}\right)\right)\pi_{n-k+i}, \end{aligned} \tag{2.4}$$

where $S^k_{k-i,1}$ *and* $S^k_{k-i,m-k+i+1}$ *are defined as*

(2.5) $S^k_{j,1} = 1, S^k_{j,l} = S^{k-1}_{j-1,l} + S^{k-1}_{j,l-1}$ *for* $2 \le k \le n,\ 1 \le j \le k-1,$ *and* $1 \le l \le k-j,$

and $S^k_{0,l} = S^k_{i,0} = 0.$

Proof. For a system in the (1.2) form, (2.1) gives $\tau_1 = \frac{1}{\pi_1}\frac{\partial}{\partial z_1}$. Then we use (2.2) to obtain

$$\tau_2 = \frac{1}{\pi_2}\frac{\partial}{\partial z_2} + \left(\frac{\pi_1'}{\pi_1^2}\pi_{n-1}z_{n-1} + \frac{\pi_1'}{\pi_1^2}\beta_n\right)\frac{\partial}{\partial z_1},$$

and

$$\begin{aligned}\tau_3 = {} & \frac{1}{\pi_3}\frac{\partial}{\partial z_3} + \left(\left(\frac{\pi_1'}{\pi_1^2}\alpha_1 + \frac{\pi_2'}{\pi_2^2}\right)\pi_{n-1}z_{n-1} + \left(\frac{\pi_1'}{\pi_1^2}\alpha_1 + \frac{\pi_2'}{\pi_2^2}\right)\beta_n\right)\frac{\partial}{\partial z_2} \\ & - \left(\frac{\pi_1'}{\pi_1^2}\pi_{n-2}z_{n-2} + \left(\frac{\pi_1'}{\pi_1^2}\pi_{n-1}\right)'\pi_{n-1}z_{n-1}^2\right)\frac{\partial}{\partial z_1} \\ & - \left(\left(\left(\frac{\pi_1'}{\pi_1^2}\pi_{n-1}\right)'\beta_n + \pi_{n-1}\beta_n'\right)z_{n-1} + \frac{\pi_1'}{\pi_1^2}\pi_{n-1}\beta_{n-1} + \beta_n\beta_n'\right)\frac{\partial}{\partial z_1}.\end{aligned}$$

Then by an induction, for $3 < k \le n$, we get

$$\begin{aligned}\tau_k = {} & \frac{1}{\pi_k}\frac{\partial}{\partial z_k} + \left(A^k_{k-1}(z_n)\, z_{n-1} + \eta^k_{k-1}(z_n)\right)\frac{\partial}{\partial z_{k-1}} \\ & + \sum_{i=1}^{k-2}\left(A^k_i(z_n)\, z_{n-k+i} + \sum_{j=n-k+i+1}^{n-1}\sum_{l=j}^{n-1} T^k_{j,l}(z_n)\, z_j z_l\right)\frac{\partial}{\partial z_i} \\ & + \sum_{i=1}^{k-2}\left(\sum_{j=n-k+i+1}^{n}\eta^k_i(z_n)\, z_j + O^{[3]}_{z_n}(z_{n-k+i+1},\ldots,z_{n-1})\right)\frac{\partial}{\partial z_i},\end{aligned}$$

where

$$\begin{aligned}A^k_i(z_n) = {} & (-1)^{k-i+1} \\ & \times \left(S^k_{k-i,1}\frac{\pi_i'}{\pi_i^2} + \sum_{m=k-i+1}^{k-1} S^k_{k-i,m-k+i+1}\frac{\pi_{k-m}'}{\pi_{k-m}^2}\left(\prod_{j=k-i+1}^{m}\alpha_{k-j}\right)\right)\pi_{n-k+i},\end{aligned}$$

with the coefficients S^k_i given by the rule (2.5). ∎

In order to determine the $\alpha_i(y)$ for $1 \le i \le n-1$, we impose that

$$\frac{\partial}{\partial z_i}h \circ \phi^{-1} = \begin{cases} 0 \text{ for } 1 \le i \le n-1, \\ 1 \text{ when } i = n. \end{cases}$$

Now we are ready to state a set of differential equations which enables us to compute functions α_i for $1 \le i \le n-1$.

PROPOSITION 2.3. *If there exists a diffeomorphism which transforms system* (1.1) *into form* (1.2)*, then*

$$[\tau_k, \tau_n] = \lambda_k(y)\tau_k + G^{[1]}_n + R \text{ for } 1 \le i \le n-1,$$

where

$$G_n^{[1]} = \sum_{i=1}^{k-2} \left(\frac{1}{\pi_k} T_{k,n-k+i}^k z_{n-k+i} \right) \frac{\partial}{\partial z_i} + \frac{1}{\pi_k} T_{k,k}^k z_k \frac{\partial}{\partial z_{2k-n}},$$

and

$$R = \sum_{i=1}^{k-2} \left(\sum_{j=n-k+i+1}^{n} \bar{\eta}_i^k (z_n) + O_{z_n}^{[2]} (z_{n-k+i+1}, \ldots, z_{n-1}) \right) \frac{\partial}{\partial z_i}$$

and

$$\lambda_k(y) = \operatorname{diag}\{\delta_1^k(y), \ldots, \delta_i^k(y), \ldots, \delta_k^k(y), 0, \ldots, 0\} \text{ for } 1 \le i \le k-1, \tag{2.6}$$

where $\delta_k^k = A_k^n + \frac{\pi_k'}{\pi_k}$ *and* $\delta_i^k = A_i^n - A_{n-k+i}^n - \frac{(A_i^k)'}{A_i^k}$ *for* $1 \le i \le k-1$, *and* A_i^k *is given as in* (2.4).

Proof. According to (2.3), for $1 \le k \le n-1$ we have

$$\begin{aligned}
[\tau_k, \tau_n] = & \left(A_k^n + \frac{\pi_k'}{\pi_k} \right) \frac{1}{\pi_k} \frac{\partial}{\partial z_k} \\
& + \sum_{i=1}^{k-2} \left(\left(A_i^n - A_{n-k+i}^n - \frac{(A_i^k)'}{A_i^k} \right) A_i^k z_{n-k+i} + \frac{1}{\pi_k} T_{k,n-k+i}^k z_{n-k+i} \right) \frac{\partial}{\partial z_i} \\
& \quad + \frac{1}{\pi_k} T_{k,k}^k z_k \frac{\partial}{\partial z_{2k-n}} \\
& + \sum_{i=1}^{k-2} \left(\sum_{j=n-k+i+1}^{n} \bar{\eta}_i^k (z_n) + O_{z_n}^{[2]} (z_{n-k+i+1}, \ldots, z_{n-1}) \right) \frac{\partial}{\partial z_i}.
\end{aligned}$$

Set $\lambda_k(y) = \operatorname{diag}\{\delta_1^k(y), \ldots, \delta_i^k(y), \ldots, \delta_k^k(y), 0, \ldots, 0\}$, where $\delta_k^k = A_k^n + \frac{\pi_k'}{\pi_k}$ and $\delta_i^k = A_i^n - A_{n-k+i}^n - \frac{(A_i^k)'}{A_i^k}$ for $1 \le i \le k-1$. Then

$$[\tau_k, \tau_n] = \lambda_k(y) \tau_k + G_n^{[1]} + R. \qquad \square \tag{2.7}$$

Remark 1. In (2.7), $\lambda_k(y)$ could be uniquely determined since $G_n^{[1]}$ might be separated according to the coefficients of second-order terms in τ_n.

Finally, the following result enables us to determine all the functions $\alpha_i(y)$ for all $1 \le i \le n-1$.

PROPOSITION 2.4. *If there exists a diffeomorphism which transforms system* (1.1) *into form* (1.2), *then* $\alpha_i = \frac{\pi_i}{\pi_{i+1}}$ *for* $1 \le i \le n-2$, *and* $\alpha_{n-1} = \pi_{n-1}$, *where*

$$\begin{cases}
\pi_i = c_i \exp \left[\int \left(\exp \int \left(\delta_i^i - \delta_i^{n-1} - \delta_{i+1}^{i+1} \right) dy - \bar{B}_i^{n-1} \right) dy \right] \text{ for } 1 \le i \le n-2, \\
\pi_{n-1} = c_{n-1} \exp \left(\int \left(\dfrac{\delta_{n-1}^{n-1} - \bar{A}_{n-1}^n}{2} \right) dy \right),
\end{cases}$$

with $\bar{B}_1^k = 0$, *and for* $1 \le i, k \le n-1$, *and* $1 \le i \le n-1$,

$$\bar{B}_i^k = \sum_{m=k-i+1}^{k-1} S_{k-i,m-k+i+1}^k \frac{\pi'_{k-m}}{\pi_{k-m}}. \tag{2.8}$$

Proof. Define

$$B_i^k = \frac{\pi'_i}{\pi_i} + \bar{B}_i^k. \tag{2.9}$$

According to (2.4), for $1 \le i, k \le n-1$,

$$\frac{\left(A_i^k\right)'}{A_i^k} = \frac{\left(B_i^k\right)'}{B_i^k} - \frac{\pi'_i}{\pi_i} + \frac{\pi'_{n-k+i}}{\pi_{n-k+i}}.$$

As $\delta_k^k = A_k^n + \frac{\pi'_k}{\pi_k}$, hence we have

$$\delta_i^{n-1} = A_i^n - A_{1+i}^n - \frac{\left(B_i^{n-1}\right)'}{B_i^{n-1}} + \frac{\pi'_i}{\pi_i} - \frac{\pi'_{1+i}}{\pi_{1+i}} = \delta_i^i - \delta_{1+i}^{1+i} - \left(\frac{\pi'_i}{\pi_i} + \bar{B}_i^{n-1}\right)' \Big/ \left(\frac{\pi'_i}{\pi_i} + \bar{B}_i^{n-1}\right),$$

which yields

$$\pi_i = c_i \exp\left[\int \left(\exp \int \left(\delta_i^i - \delta_i^{n-1} - \delta_{1+i}^{1+i}\right) dy - \bar{B}_i^{n-1}\right) dy\right] \text{ for } 1 \le i \le n-2,$$

where $\bar{B}_i^{n-1}$ is defined as in (2.8) and $c_i \in R$, $c_i \neq 0$.

As $\delta_{n-1}^{n-1} = 2\frac{\pi'_{n-1}}{\pi_{n-1}} + \bar{A}_{n-1}^n$, where

$$\bar{A}_{n-1}^n = \sum_{m=2}^{n-1} S_{1,m}^n \frac{\pi'_{n-m}}{\pi_{n-m}},$$

then

$$\pi_{n-1} = c_{n-1} \exp\left(\int \left(\frac{\delta_{n-1}^{n-1} - \bar{A}_{n-1}^n}{2}\right) dy\right). \qquad \square$$

Remark 2. For system (1.2), if we set $\alpha_i = s(y)$ for $1 \le i \le n-1$, then

$$\delta_{n-1}^{n-1} = A_{n-1}^n + \frac{\pi'_{n-1}}{\pi_{n-1}} = 2\frac{\pi'_{n-1}}{\pi_{n-1}} + \sum_{i=2}^{n-1} \frac{\pi'_{n-i}}{\pi_{n-i}}.$$

By the definition of π_i for $1 \le i \le n-1$, we have $\pi_k = s^{n-k}$ for $1 \le k \le n-1$, and therefore

$$\delta_{n-1}^{n-1} = 2\frac{s'}{s} + \sum_{i=2}^{n-1} i\frac{s'}{s} = l_n \frac{s'}{s},$$

where $l_n = \frac{n(n-1)}{2} + 1$. In such a way, we obtain the same result as the one stated in [10].

3. Main result. If there exists a diffeomorphism which transforms system (1.1) into form (1.2), then (2.8) of Proposition 2.4 gives all α_i for $1 \leq i \leq n-1$. Therefore, let us consider a new family of vector fields defined as follows:

$$\widetilde{\tau}_1 = \pi_1 \tau_1 \text{ and } \widetilde{\tau}_{i+1} = \frac{1}{\alpha_i}[\widetilde{\tau}_i, f] \text{ for } 1 \leq i \leq n-1. \tag{3.1}$$

Set

$$\theta(\widetilde{\tau}_1, \cdots, \widetilde{\tau}_n) = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & \vdots & \cdots & \pi_{n-1} & \widetilde{l}_{2,n} \\ \vdots & \ldots & \ddots & \ldots & \vdots \\ \vdots & \pi_2 & \cdots & \cdots & \vdots \\ \pi_1 & \widetilde{l}_{n,2} & \cdots & \cdots & \widetilde{l}_{n,n} \end{pmatrix} := \widetilde{\Lambda},$$

where

$$\widetilde{l}_{k,j} = \theta_k(\widetilde{\tau}_j) \text{ for } 2 \leq k \leq n \text{ and } n-k+2 \leq j \leq n.$$

Consider the R^n-valued form ω,

$$\omega = \widetilde{\Lambda}^{-1}\theta := (\omega_1, \omega_2, \ldots, \omega_n)^T, \tag{3.2}$$

where, for $1 \leq s \leq n$, we have

$$\omega_s = \sum_{m=1}^{n} r_{s,m}\theta_m. \tag{3.3}$$

Then the following algorithm gives all the components of ω.

ALGORITHM 1.

For $1 \leq j \leq n$,

$$r_{n,j} = \cdots = r_{n-j+2,j} = 0 \text{ and } r_{n-j+1,j} = 1.$$

For $2 \leq k \leq n-1$ *and* $1 \leq j \leq n$,

$$r_{n-k,j} = -\sum_{i=2}^{k} \widetilde{l}_{k,n-k+i-(j-1)} r_{n-k+i-(j-1),j}.$$

Then (3.3) *becomes*

$$\omega_s = \sum_{m=1}^{n-s+1} r_{s,m}\theta_m.$$

THEOREM 3.1. *The following conditions are equivalent:*

(1) *There exists a diffeomorphism which transforms system* (1.1) *into a SODO normal form* (1.2).

(2) *There exists a family of functions* $\alpha_i(y)$ *for* $1 \leq i \leq n-1$ *such that the family of vector fields* $\widetilde{\tau}_i$ *for* $1 \leq i \leq n$ *defined in* (3.1) *satisfies the following commutativity conditions:*

$$[\widetilde{\tau}_i, \widetilde{\tau}_j] = 0 \text{ for } 1 \leq i, j \leq n. \tag{3.4}$$

(3) *There exists a family of functions* $\alpha_i(y)$ *for* $1 \leq i \leq n-1$ *such that the* R^n*-valued form* ω *defined in* (3.2) *satisfies the following condition:*

$$d\omega = 0. \tag{3.5}$$

Proof. Assume that there exists a diffeomorphism which transforms system (1.1) into form (1.2). Then we compute $\alpha_i(y)$ for $1 \leq i \leq n-1$ from (2.8) in Proposition 2.4. Thus, it is easy to show that $\tau_1 = \frac{1}{\pi_1}\frac{\partial}{\partial z_1}$, which yields that $\widetilde{\tau}_1 = \frac{\partial}{\partial z_1}$, and then by construction we obtain $\widetilde{\tau}_i = \frac{\partial}{\partial z_i}$ for $2 \leq i \leq n$. Consequently, we have $[\widetilde{\tau}_i, \widetilde{\tau}_j] = 0$ for $1 \leq i, j \leq n$.

Reciprocally, assume that there exists $\alpha_i > 0$ for $1 \leq i \leq n-1$ such that $[\widetilde{\tau}_i, \widetilde{\tau}_j] = 0$ for $1 \leq i, j \leq n$. Then we know (see [8], [15]) that we can find a local diffeomorphism $\phi = z$ such that

$$\phi_*(\widetilde{\tau}_i) = \frac{\partial}{\partial z_i}.$$

As $\phi_*(\widetilde{\tau}_i) = \frac{\partial}{\partial z_i}$ is constant, hence we have

$$\frac{\partial}{\partial z_i}\phi_*(f) = \phi_*\left([\widetilde{\tau}_i, f]\right) = \alpha_i \phi_*(\widetilde{\tau}_{i+1}) = \alpha_i \frac{\partial}{\partial z_{i+1}},$$

and thus $\frac{\partial}{\partial z_i}\phi_*(f) = \alpha_i \frac{\partial}{\partial z_{i+1}}$ for $1 \leq i \leq n-1$. Consequently, by integration we obtain $\phi_*(f) = A(y)z + \beta(y)$.

Moreover, as $dh \circ \widetilde{\tau}_i = 0$ for $1 \leq i \leq n-1$ and $dh \circ \widetilde{\tau}_n = 1$, we obtain $h \circ \phi^{-1} = z_n$.

Finally, in order to prove that in Theorem 3.1 condition (2) is equivalent to condition (3), it is sufficient to prove that (3.4) is equivalent to (3.5).

Recall that for any two vector fields X, Y, we have

$$d\omega(X, Y) = L_X\left(\omega(Y)\right) - L_Y\left(\omega(X)\right) - \omega([X, Y]).$$

Setting $X = \widetilde{\tau}_i$ and $Y = \widetilde{\tau}_j$, we obtain

$$d\omega(\widetilde{\tau}_i, \widetilde{\tau}_j) = L_{\widetilde{\tau}_i}\omega(\widetilde{\tau}_j) - L_{\widetilde{\tau}_j}\omega(\widetilde{\tau}_i) - \omega([\widetilde{\tau}_i, \widetilde{\tau}_j]).$$

As $\omega(\widetilde{\tau}_j)$ and $\omega(\widetilde{\tau}_i)$ are constant, then we have

$$d\omega(\widetilde{\tau}_i, \widetilde{\tau}_j) = -\omega([\widetilde{\tau}_i, \widetilde{\tau}_j]).$$

Because ω is an isomorphism and $(\widetilde{\tau}_i)_{1 \leq i \leq n}$ is a basis of TU, then (3.4) is equivalent to (3.5). ∎

Remark 3. (i) The R^n-valued form ω can be viewed as an isomorphism $TU^n \to U \times \mathbb{R}^n$ which brings each $\widetilde{\tau}_i$ to the canonical vector basis $\frac{\partial}{\partial z_I}$. Moreover, $d\omega = 0$ means that there is a local diffeomorphism $\phi : U \to U$ such that ω is the tangent map of ϕ.

(ii) The diffeomorphism $\phi(x) = z$ is determined by $\omega = \phi_*(x)$, which can be given locally as

$$z_i = \phi_i(x) = \int_\gamma \omega_i + \phi_i(0) \text{ for } 1 \leq i \leq n,$$

where γ is a smooth path from 0 to x lying in a neighborhood $V_0 \subseteq U$ of 0.

The following simple example is studied in order to illustrate Theorem 3.1.

Example 1. Let us consider the following system

$$\left\{\begin{array}{l} \dot{x}_1 = \dfrac{\gamma(y)}{1+x_4}x_1x_3, \\ \dot{x}_2 = \dfrac{\beta(y)}{1+x_4}x_1, \\ \dot{x}_3 = \mu(y)x_2, \\ \dot{x}_4 = \gamma(y)x_3, \\ y = x_4, \end{array}\right. \tag{3.6}$$

which gives

$$\left\{\begin{array}{l} \theta_1 = dx_4, \\ \theta_2 = \gamma dx_3 + \gamma' x_3 dx_4, \\ \theta_3 = \gamma\mu dx_2 + 2\gamma'\gamma x_3 dx_3 + \left((\gamma\mu)' x_2 + (\gamma'\gamma)' x_3^2\right) dx_4, \\ \theta_4 = \gamma\mu\dfrac{\beta}{1+x_4}dx_1 + \left(2\gamma'\mu + (\gamma\mu)'\right)\gamma x_3 dx_2 \\ \qquad + \left(2\gamma'\gamma\mu x_2 + \gamma\,(\gamma\mu)' x_2 + 3\gamma\,(\gamma'\gamma)' x_3^2\right) dx_3 + O^{[2]}\,(x_1,\ x_2,\ x_3)\,\theta_1. \end{array}\right.$$

Then we have $\tau_1 = \frac{1+x_4}{\gamma\mu\beta}\frac{\partial}{\partial x_1}$. Consequently, we obtain

$$\left\{\begin{array}{l} \tau_2 = \frac{1}{\gamma\mu}\frac{\partial}{\partial x_2} + (1+x_4)\,\gamma\frac{(\gamma\mu\beta)'}{(\gamma\mu\beta)^2}x_3\frac{\partial}{\partial x_1}, \\ \tau_3 = \frac{1}{\gamma}\frac{\partial}{\partial x_3} - \gamma\mu\,(1+x_4)\,\frac{(\gamma\mu\beta)'}{(\gamma\mu\beta)^2}x_3\frac{\partial}{\partial x_2} + \left(\frac{(\gamma\mu)'}{(\gamma\mu)^2} + \beta\frac{(\gamma\mu\beta)'}{(\gamma\mu\beta)^2}\right)\gamma x_2\frac{\partial}{\partial x_1} + R_{1,3}\tau_1, \\ \tau_4 = \frac{\partial}{\partial x_4} + \left(\frac{\gamma'}{\gamma} + \frac{(\gamma\mu)'}{(\gamma\mu)} + \frac{(\gamma\mu\beta)'}{(\gamma\mu\beta)}\right)x_3\frac{\partial}{\partial x_3} - \left(\frac{(\gamma\mu)'}{(\gamma\mu)} + 2\frac{(\gamma\mu\beta)'}{(\gamma\mu\beta)}\right)x_2\frac{\partial}{\partial x_2} \\ \qquad + \left(\frac{1}{1+x_4} + \frac{(\gamma\mu\beta)'}{\gamma\mu\beta}\right)x_1\frac{\partial}{\partial x_1} + R_{1,4}\,(z_3,z_2)\,\tau_1 + R_{2,3}(z_3^2)\tau_2. \end{array}\right.$$

A straightforward computation gives

$$\begin{aligned} &\delta_1^1 = 2\frac{(\gamma\mu\beta)'}{\gamma\mu\beta},\ \delta_2^2 = -2\frac{(\gamma\mu\beta)'}{\gamma\mu\beta},\ \delta_3^3 = 2\frac{\gamma'}{\gamma} + \frac{(\gamma\mu)'}{\gamma\mu} + \frac{(\gamma\mu\beta)'}{\gamma\mu\beta}, \\ &\delta_1^3 = 4\frac{(\gamma\mu\beta)'}{(\gamma\mu\beta)} - \left[\frac{(\gamma\mu\beta)'}{(\gamma\mu\beta)}\right]' \Big/ \left[\frac{(\gamma\mu\beta)'}{(\gamma\mu\beta)}\right], \\ &\delta_2^3 = -\left(2\frac{\gamma'}{\gamma} + \frac{(\gamma\mu)'}{\gamma\mu} + 3\frac{(\gamma\mu\beta)'}{\gamma\mu\beta}\right) - \left(\frac{(\gamma\mu)'}{\gamma\mu} + \frac{(\gamma\mu\beta)'}{\gamma\mu\beta}\right)' \Big/ \left(\frac{(\gamma\mu)'}{\gamma\mu} + \frac{(\gamma\mu\beta)'}{\gamma\mu\beta}\right). \end{aligned}$$

According to (2.8) in Proposition 2.4, we obtain

$$\left\{\begin{array}{l} \pi_1 = c_1\exp\left[\int\left(\exp\int\left(\delta_1^1 - \delta_1^3 - \delta_2^2\right)dy\right)dy\right] = c_1\gamma\mu\beta, \\ \pi_2 = c_2\exp\left[\int\left(\exp\int\left(\delta_2^2 - \delta_2^3 - \delta_3^3\right)dy - \dfrac{\pi_1'}{\pi_1}\right)dy\right] = c_2\gamma\mu, \\ \pi_3 = c_3\exp\left(\int\left(\dfrac{1}{2}\left(\delta_3^3 - \dfrac{\pi_1'}{\pi_1} - \dfrac{\pi_2'}{\pi_2}\right)\right)dy\right) = c_3\gamma. \end{array}\right.$$

Thus $\alpha_1 = \frac{\pi_1}{\pi_2} = \frac{c_1}{c_2}\beta$, $\alpha_2 = \frac{\pi_2}{\pi_3} = \frac{c_2}{c_3}\mu$, and $\alpha_3 = \frac{\pi_3}{\pi_4} = c_3\gamma$, so the new vector fields are

$$\widetilde{\tau}_1 = c_1\,(1+x_4)\,\frac{\partial}{\partial x_1},\ \widetilde{\tau}_2 = c_2\frac{\partial}{\partial x_2},\ \widetilde{\tau}_3 = c_3\frac{\partial}{\partial x_3},\ \widetilde{\tau}_4 = \frac{\partial}{\partial x_4} + \frac{x_1}{1+x_4}\frac{\partial}{\partial x_1}.$$

It is clear that $[\widetilde{\tau}_i, \widetilde{\tau}_j] = 0$ for all $1 \leq i, j \leq 4$. Therefore, according to Theorem 3.1, system (3.6) can be transformed into SODO normal form (1.2).

Moreover, as

$$\widetilde{\Lambda} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & \gamma & \gamma' x_3 \\ 0 & \gamma\mu & 2\gamma'\gamma x_3 & (\gamma\mu)' x_2 + 2(\gamma'\gamma)' x_3^2 \\ \gamma\mu\beta & (2\gamma'\mu + (\gamma\mu')) \gamma x_3 & 2\gamma'\gamma\mu x_2 + \gamma(\gamma\mu)' x_2 + 6\gamma(\gamma'\gamma)' x_3^2 & \gamma \frac{x_1}{(1+x_4)^2} \mu\beta + R \end{pmatrix},$$

where $R = O^{[2]}_{x_4}(x_1,\ x_2,\ x_3)$, a straightforward computation gives

$$\omega = \widetilde{\Lambda}^{-1}\theta = \begin{pmatrix} d\dfrac{x_1}{c_1(1+x_4)}, & d\left(\dfrac{x_2}{c_2}\right), & d\left(\dfrac{x_3}{c_3}\right), & dx_4 \end{pmatrix}^T.$$

As $\omega = d\phi$, thus the diffeomorphism which transforms system (3.6) into SODO normal form (1.2) is

$$\phi(x) = z = \left(\frac{x_1}{c_1(1+x_4)}, \frac{x_2}{c_2}, \frac{x_3}{c_3}, x_4\right)^T,$$

with which system (3.6) could be transformed into

$$\begin{cases} \dot{z}_1 = 0, \\ \dot{z}_2 = \dfrac{c_1}{c_2}\beta(y) z_1, \\ \dot{z}_3 = \dfrac{c_2}{c_3}\mu(y) z_2, \\ \dot{z}_4 = c_3\gamma(y) z_3. \end{cases}$$

So far in this paper, we have considered only systems without inputs. The next section is devoted to systems that are also driven by an input term.

4. Extension to systems with inputs. Consider a system with inputs in the form

$$\begin{cases} \dot{x} = f(x) + g(x,u), \\ y = h(x), \end{cases} \tag{4.1}$$

where $x \in U \subset \mathbb{R}^n$, $f : U \subset \mathbb{R}^n \to \mathbb{R}^n$, $g : U \times \mathbb{R}^m \to \mathbb{R}^n$, and $h : U \subset \mathbb{R}^n \to \mathbb{R}$ are analytic functions and where $g(x,0) = 0$ for $x \in U$.

For system (4.1), the SODO normal form along its output trajectory $y(t)$ is

$$\begin{cases} \dot{z} = A(y)z + \beta(y) + \eta(y,u), \\ y = z_n = Cz, \end{cases} \tag{4.2}$$

where $A(y)$ and $\beta(y)$ are given as in (1.2), and $\eta(y,u) = [\ \eta_1(y,u),\ \eta_2(y,u), \ldots, \eta_n(y,u)\ ]^T$.

THEOREM 4.1. *System* (4.1) *can be transformed into SODO normal form* (4.2) *by a diffeomorphism if and only if*

(i) *one of the conditions in Theorem* 3.1 *is fulfilled.*

(ii) $[g, \widetilde{\tau}_i] = 0$ *for* $1 \leq i \leq n-1$.

Proof. From Theorem 3.1, we can state that there exists a diffeomorphism ϕ such that

$$\phi_*(f) = A(y)z + \beta(y).$$

For $1 \le i \le n-1$, because $\phi_*(\widetilde{\tau}_i) = \frac{\partial}{\partial z_i}$ is constant, hence we have

$$\frac{\partial}{\partial z_i}\phi_*(g) = \phi_*([g, \widetilde{\tau}_i]) = 0.$$

Therefore $\phi_*(g) = \eta(y,u)$. Thus, we obtain the form (4.2). □

Remark 4. If $g(x,u) = g_1(x)u_1 + \cdots + g_m(x)u_m$, and both conditions (i) and (ii) of Theorem 4.1 are fulfilled, then

$$\eta(y,u) = B_1(y)u_1 + \cdots + B_m(y)u_m.$$

Let us now study some special cases of the term $\eta(y,u)$.

COROLLARY 4.2. *Assume that conditions* (i) *and* (ii) *of Theorem* 4.1 *are fulfilled.*
(a) *If* $[g, \widetilde{\tau}_n] = 0$, *then*

$$\eta(y,u) = \eta(u).$$

(b) *If* $g(x,u) = g_1(x)u_1 + \cdots + g_m(x)u_m$ *and*

$$[g_k, \widetilde{\tau}_i] = 0 \text{ for } 1 \le i \le n \text{ and } 1 \le k \le m,$$

then

$$\eta(y,u) = B_1u_1 + \cdots + B_mu_m,$$

where B_i *are constant vector fields.*

Example 2. Let us consider the system

$$\begin{cases} \dot{x}_1 = \dfrac{\gamma(y)}{1+x_3}x_1x_2 + \dfrac{x_1}{1+x_3}u, \\ \dot{x}_2 = \dfrac{\mu(y)}{1+x_3}x_1, \\ \dot{x}_3 = \gamma(y)x_2 + u, \\ y = x_3. \end{cases} \tag{4.3}$$

A straightforward computation gives

$$\tau_1 = \frac{1+x_3}{\gamma\mu}\frac{\partial}{\partial x_1},\ \tau_2 = \frac{1}{\gamma}\frac{\partial}{\partial x_2} + \left((1+x_3)\frac{(\gamma\mu)'}{\gamma\mu^2}\right)x_2\frac{\partial}{\partial x_1},$$
$$\tau_3 = \frac{\partial}{\partial x_3} + \left(\frac{(\mu\gamma)'}{(\mu\gamma)} + \frac{\gamma'}{\gamma}\right)x_2\frac{\partial}{\partial x_2} + \left(\frac{1}{1+x_3} - \frac{(\gamma\mu)'}{\gamma\mu}\right)x_1\frac{\partial}{\partial x_1}.$$

Then we obtain

$$\delta_1^1 = 0,\ \delta_2^2 = 2\frac{\gamma'}{\gamma} + \frac{(\mu\gamma)'}{(\mu\gamma)},\ \delta_1^2 = -\frac{(\mu\gamma)'}{(\mu\gamma)} - 2\frac{\gamma'}{\gamma} - \left(\frac{(\mu\gamma)'}{(\mu\gamma)}\right)' \Big/ \left(\frac{(\mu\gamma)'}{(\mu\gamma)}\right).$$

Then according to (2.8), we have

$$\begin{cases} \pi_1 = c_1 \exp\left[\int \left(\exp \int \left(\delta_1^1 - \delta_1^2 - \delta_2^2\right) dy\right) dy\right] = c_1\gamma\mu, \\ \pi_2 = c_2 \exp\left(\int \left(\frac{1}{2}\left(\delta_2^2 - \frac{\pi_1'}{\pi_1}\right)\right) dy\right) = c_2\gamma, \end{cases}$$

which yields $\alpha_1(y) = \frac{c_1}{c_2}\mu(y)$ and $\alpha_2(y) = c_2\gamma(y)$. Therefore, we obtain $\widetilde{\tau}_1 = c_1(1+x_3)\frac{\partial}{\partial x_1}$, $\widetilde{\tau}_2 = c_2\frac{\partial}{\partial x_2}$, and $\widetilde{\tau}_3 = \frac{\partial}{\partial x_3} + \frac{x_1}{1+x_3}\frac{\partial}{\partial x_1}$.

As $g = \frac{x_1}{1+x_3}\frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_3} = \widetilde{\tau}_3$, then $[g, \widetilde{\tau}_1] = [g, \widetilde{\tau}_2] = 0$, and system (4.3) is transformed into

$$\begin{cases} \dot{z}_1 = 0, \\ \dot{z}_2 = \dfrac{c_1}{c_2}\mu(y)\, z_1, \\ \dot{z}_3 = c_2\gamma(y)\, z_2 + u, \\ y = z_3 \end{cases} \tag{4.4}$$

by the diffeomorphism

$$\phi(x) = z = \left(\frac{x_1}{c_1(1+x_3)}, \frac{x_2}{c_2}, x_3\right)^T.$$

Following the proposed high-gain observer in the form (1.3), the corresponding observer for system (4.4) can be designed as follows:

$$\begin{cases} \dot{\hat{z}}_1 = -\dfrac{\rho^3}{\gamma\mu}(\hat{z}_3 - z_3), \\ \dot{\hat{z}}_2 = \dfrac{c_1}{c_2}\mu(y)\,\hat{z}_1 - 3\dfrac{\rho^2}{\gamma}(\hat{z}_3 - z_3), \\ \dot{\hat{z}}_3 = c_2\gamma(y)\,\hat{z}_2 - 3\rho(\hat{z}_3 - z_3) + u, \end{cases}$$

where ρ is the tunable gain. For a more specific, yet simple, simulation, choose $c_1 = c_2 = 1$, $u(t) = 1$, $\mu(y) = 1+y^2$, and $\gamma(y) = 2+\cos(y)$. Its simulation results are presented in Figures 4.1, 4.2, and 4.3 which, respectively, present the convergences of the system's states and their estimations.

In addition, in order to solve the left invertibility problem, the observability matching condition (OMC) for system (4.1) with $m = 1$ is

$$\begin{cases} L_g L_f^{i-1} h = 0 \ \forall x \in U,\ 1 \le i \le n-1, \\ L_g L_f^{n-1} h \neq 0. \end{cases}$$

COROLLARY 4.3. *Assume conditions* (i) *and* (ii) *of Theorem* 4.1 *are fulfilled and that the OMC is verified. Then*

$$\eta(y,u) = [\eta_1(y,u),\, 0,\, \ldots,\, 0]^T.$$

Remark 5. The OMC for system (4.1) with $m = 1$ is equivalent to $g \in span\{\widetilde{\tau}_1\}$.

We give another example in order to highlight Corollary 4.3.

Example 3. Consider the system

$$\begin{cases} \dot{x}_1 = u, \\ \dot{x}_2 = \mu(y)x_1 + \mu(y)x_1^2 + \dfrac{x_2}{1+x_1}u, \\ \dot{x}_3 = \gamma(y)\dfrac{x_2}{1+x_1}, \\ y = x_3. \end{cases} \tag{4.5}$$

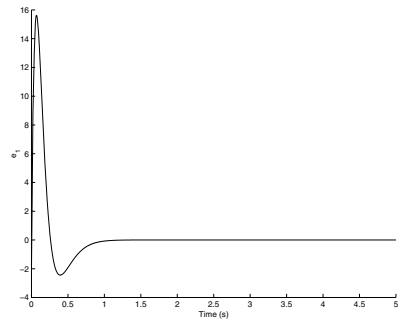


FIG. 4.1. *Observation error between* z_1 *and* $\hat{z}_1$.

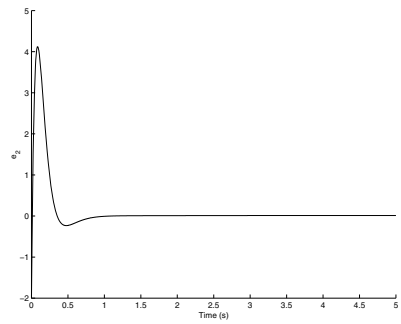


FIG. 4.2. *Observation error between* z_2 *and* $\hat{z}_2$.

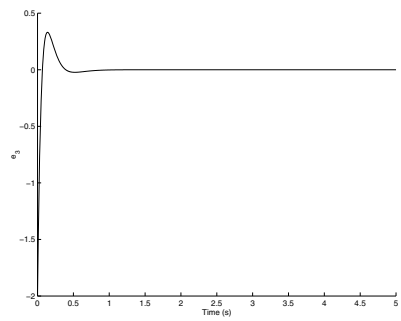


FIG. 4.3. *Observation error between* z_3 *and* $\hat{z}_3$.

A straightforward computation gives $\tau_1 = \frac{1}{\gamma\mu}\frac{\partial}{\partial x_1} + \frac{1}{\gamma\mu}\frac{x_2}{1+x_1}\frac{\partial}{\partial x_2}$. From (2.8) in Proposition 2.4, we can determine $\alpha_1(y) = \frac{c_1}{c_2}\mu(y)$ and $\alpha_2(y) = c_2\gamma(y)$. Thus, we have $\widetilde{\tau}_1 = c_1\frac{\partial}{\partial x_1} + c_1\frac{x_2}{1+x_1}\frac{\partial}{\partial x_2}$, $\widetilde{\tau}_2 = c_2(1+x_1)\frac{\partial}{\partial x_2}$, and $\widetilde{\tau}_3 = \frac{\partial}{\partial x_3}$.

As $g \in span\{\widetilde{\tau}_1\}$, then the OMC condition is fulfilled, therefore system (4.5) could be transformed by the diffeomorphism

$$\phi(x) = z = \left(\frac{x_1}{c_1}, \frac{x_2}{c_2(1+x_1)}, x_3\right)^T$$

into

$$\begin{cases} \dot{z}_1 = \dfrac{u}{c_1}, \\ \dot{z}_2 = \dfrac{c_1}{c_2}\mu(y)\, z_1, \\ \dot{z}_3 = c_2\gamma(y)\, z_2, \\ y = z_3. \end{cases}$$

5. Conclusion. In this paper, we have put forward the geometrical conditions which allow us to determine whether a nonlinear system can be transformed locally into the SODO normal form by means of a diffeomorphism and of an output injection. In our main result we state two equivalent ways to check these conditions. In the first one, we used Lie brackets commutativity, and the second one was based on the one forms. Moreover, an extension of our results is stated for a class of nonlinear systems with inputs.

Acknowledgments. We are deeply grateful to the anonymous reviewers for valuable comments and helpful suggestions that enabled us to improve the presentation of this paper. We are also deeply grateful to Professor Vincent Maki for many corrections.

REFERENCES

[1] L. Boutat-Baddas, D. Boutat, J.-P. Barbot and R. Tauleigne, *Quadratic observability normal form*, in Proceedings of the 40th IEEE Conference on Decision and Control, IEEE, Piscataway, NJ, 2001, pp. 2942–2947.

[2] K. Busawon, M. Farza, and H. Hammouri, *A simple observer for a class of nonlinear systems*, Appl. Math. Lett., 11 (1998), pp. 27–31.

[3] S. Chabraoui, D. Boutat, L. Boutat-Baddas, and J. P. Barbot, *Observability quadratic characteristic numbers*, in Proceedings of the 42nd IEEE Conference on Decision and Control, IEEE, Piscataway, NJ, 2003, pp. 3653–3658.

[4] J. P. Gauthier and I. Kupka, *Deterministic Observation Theory and Applications*, Cambridge University Press, Cambridge, UK, 2001.

[5] M. Guay, *Observer linearization by output-dependent time-scale transformations*, IEEE Trans. Automat. Control, 47 (2002), pp. 1730–1735.

[6] H. Hammouri and J.P. Gauthier , *Bilinearization up to the output injection*, Systems Control Lett., 11 (1988) pp. 139–149.

[7] H. Hammouri and J. Morales, *Observer synthesis for state-affine systems*, in Proceedings of the 29th IEEE Conference on Decision and Control, IEEE, Piscataway, NJ, 1990, pp. 784–785.

[8] A. Isidori, *Nonlinear Control Systems*, 2nd ed., Springer-Verlag, Berlin, 1989.

[9] D. Luenberger, *An introduction to observers*, IEEE Trans. Automat. Control, 16 (1971), pp. 596–602.

[10] W. Respondek, A. Pogromsky, and H. Nijmeijer, *Time scaling for observer design with linearization error dynamics*, IEEE Trans. Automat. Control, 3 (1989), pp. 199–216.

[11] N. Kazantzis and C. Kravaris, *Nonlinear observer design using Lyapunov's auxiliary theorem*, Systems Control Lett., 34 (1998), pp. 241–247.

[12] A. Krener and A. Isidori, *Linearization by output injection and nonlinear observer*, Systems Control Lett., 3 (1983), pp. 47–52.

[13] A. J. Krener and W. Respondek, *Nonlinear observer with linearizable error dynamics*, SIAM J. Control Optim., 30 (1985), pp. 197–216.

[14] A. J. Krener and M. Xiao, *Nonlinear observer design in the Siegel domain*, SIAM J. Control Optim., 41 (2002), pp. 932–953.

[15] H. Nijmeijer and A. J. Van Der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.

[16] X.-H. Xia and W.-B. Gao, *Nonlinear observer design by observer error linearization*, SIAM J. Control Optim., 27 (1989), pp. 199–216.

[17] G. Zheng, D. Boutat, and J. P. Barbot, *Output dependent observability linear normal form*, in Proceedings of the 44th IEEE Conference on Decision and Control, IEEE, Piscataway, NJ, 2005, pp. 7026–7030.

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2256–2279

CONTROLLABILITY OF A CLASS OF NEWTONIAN FILTRATION EQUATIONS WITH CONTROL AND STATE CONSTRAINTS*

XU LIU† AND HANG GAO‡

Abstract. This paper addresses a study of the controllability of a class of Newtonian filtration equations, with nonnegative constraints on the control and state variables. When the control enters the system through the whole domain where the equation evolves, we characterize the set of nonnegative targets which are approximately controllable at any time $T > 0$. The proof combines the Fenchel–Rockafellar duality theory and a fixed point argument. When the control is restricted to be active in a proper open subset of the whole domain, we prove a negative controllability result by means of a localization technique which reflects the underlying obstruction phenomenon in the system.

Key words. Newtonian filtration equation, approximate controllability, Fenchel–Rockafellar duality theory, obstruction phenomenon

AMS subject classifications. 93B05, 35K65

DOI. 10.1137/060649951

1. Introduction and main results.

Consider the following controlled system governed by a Newtonian filtration equation:

$$\begin{cases} y_t(x,t) - (y^m)_{xx}(x,t) = \chi_\omega u(x,t), & (x,t) \in Q, \\ y(-1,t) = y(1,t) = 0, & t \in (0,T), \\ y(x,0) = y_0(x), & x \in (-1,1), \end{cases} \tag{1.1}$$

where $Q = (-1,1) \times (0,T)$, $1 < m < 3$, χ_ω denotes the characteristic function of a domain $\omega \subseteq (-1,1)$, u is a nonnegative control function, and y_0 is a given initial value.

Newtonian filtration equations, as an important class of quasilinear degenerate parabolic equations, come from a variety of diffusion phenomena as a net action of matters. They are suggested as mathematical models of physical problems in many fields such as filtration, phase transition, biochemistry, and dynamics of biological groups. For example, for a homogeneous, isotropic, and rigid porous medium filled with a fluid, the flow is governed by the continuity equation

$$\theta_t + \operatorname{div} \overrightarrow{V} = 0$$

and Darcy's law

$$\overrightarrow{V} = -K(\theta)\nabla\varphi,$$

where θ denotes the volumetric moisture content, $\overrightarrow{V}$ the macroscopic velocity of the fluid, $K(\theta)$ the hydraulic conductivity, and φ the total potential. If absorption and

*Received by the editors January 14, 2006; accepted for publication (in revised form) July 28, 2007; published electronically December 21, 2007. This work was supported by NSF of China under grants 10471021, 10571161, and 10601010.

http://www.siam.org/journals/sicon/46-6/64995.html

†Department of Mathematics, Zhejiang University, Hangzhou, Zhejiang 310027, China; and Institute of Mathematics and Statistics, Northeast Normal University, Changchun, Jilin 130024, China (liux216@nenu.edu.cn).

‡Institute of Mathematics and Statistics, Northeast Normal University, Changchun, Jilin 130024, China (hangg@nenu.edu.cn).

chemical, osmotic, and thermal effects are ignored, then, after a necessary change of variables, for horizontal flow, we have

$$\theta_t = \triangle \theta^m.$$

We shall focus our attention on the simple case—the one-dimensional case. Note that system (1.1) is degenerate whenever $m > 1$. Compared to linear equations and quasilinear equations without degeneracy, such equations, to a certain extent, reflect more exactly the physical reality. Indeed, when $m = 1$, (1.1) becomes the heat equation. In this case the solution possesses the property of infinite speed of propagation of disturbances; i.e., any nontrivial nonnegative initial data implies the positivity of the solution after the initial time. If $m > 1$, (1.1) is degenerate. In this case, the solution possesses the property of finite speed of propagation of disturbances. On the other hand, the appearance of degeneracy makes the problem more difficult. Therefore, in the last four decades the study in this direction has attracted a large number of researchers, and great progress has been made (see [1], [4], [9], [21], and the references cited therein).

The goal of this paper is to study the controllability and noncontrollability of system (1.1). Many papers have been devoted to the controllability of semilinear nondegenerate parabolic systems without control/state constraints. We mention here an incomplete list of related works: [2], [10], [11], [13], [17], [19], [20], and the rich references cited therein. However, as far as we know, only a few papers have been published on the controllability of degenerate parabolic equations. In [5] and [6], the authors study the null controllability of linear degenerate parabolic equations. In [14] and [15], the authors discuss the approximate controllability of quasilinear degenerate p-Laplacian equations. In [8] the authors prove an obstruction phenomenon, which implies that system (1.1) with $0 < m < 1$ is not approximately controllable under a local control at any time. We notice that in the above mentioned papers, the authors do not put any constraint on the control or state variables. Meanwhile, very little is known about the controllability of parabolic equations with control/state constraints. In [18], the constrained controllability problem for the abstract evolution equation is studied. In [23], the author establishes a criterion for approximate controllability of a heat equation under a nonnegative step boundary control function. In [7], the authors prove that the semilinear parabolic equation is approximately controllable with a nonnegative constraint on the control, provided that the nonlinearity is a nondecreasing continuous function. The authors prove this result by means of a cancellation technique, which consists in modifying the control associated to the linear case by means of a perturbation which cancels the nonlinearity appearing in the equation and guarantees the nonnegativity of control function.

In this paper we shall discuss the controllability and noncontrollability with nonnegative constraints on the control and state variables for system (1.1) with $1 < m < 3$, which is a degenerate quasilinear parabolic equation. To the best of our knowledge, no reference has addressed this controllability problem before. In contrast to the case of unconstrained controllability problems, constraint on the control introduces essential difficulties, even if the control acts in the whole domain. Indeed, it is well known that in the case without constraint, for most linear systems, the approximate controllability and null controllability may hold simultaneously. However, in [18] the authors point out that this is not true for the constrained case, even for linear evolution systems.

In [7], by virtue of the monotonicity of the nonlinearity, the key point in deriving the desired controllability is to show the existence of a nonnegative control for

the linearized system. However, it seems that this technique does not work for our quasilinear problem. In fact, since the nonlinearity $(y^m)_{xx}$ of system (1.1) has no monotonicity, we could not construct a nonnegative control function directly as in [7]. Instead, we shall make use of the Fenchel–Rockafellar duality theory to represent the nonnegative control with minimal energy for the linearized system and further solve our quasilinear controllability problem by means of a fixed point argument. Also, since the principal part of system (1.1) is nonlinear, we cannot use the method developed in [11] and [7] to prove the uniform boundedness of control functions, which is critical when employing the fixed point method (see Remark 3.1). Rather, following [12], we shall use two functionals and suitable key estimates for the solution of the linearized system to establish the desired uniform boundedness. On the other hand, in order to show a negative controllability result for system (1.1), we shall prove an obstruction phenomenon for this system by means of a localization technique.

In order to state our main results, we need to introduce some definitions. First, due to the degeneracy of system (1.1), we are interested only in the generalized solution of system (1.1) in the following sense.

DEFINITION 1.1. *A function y is called a nonnegative generalized solution of system* (1.1) *if*

(1) $y \geq 0$ *a.e. in* Q, $y \in C([0,T]; L^2(-1,1))$, *and* $y^m \in L^1(Q)$, *and*

(2) *for any* $\varphi \in C^2(\overline{Q})$ *with* $\varphi(-1,t) = \varphi(1,t) = \varphi(x,T) = 0$, *the following equality holds:*

$$\iint_Q [y(x,t)\varphi_t(x,t) + y^m(x,t)\varphi_{xx}(x,t)]dxdt + \iint_Q \chi_\omega u(x,t)\varphi(x,t)dxdt = -\int_{-1}^{1} y_0(x)\varphi(x,0)dx.$$

It is well known that, for each $y_0 \in L^2(-1,1)$ with $y_0 \geq 0$ a.e. in $(-1,1)$ and $u \in L^2(Q)$ with $u \geq 0$ a.e. in Q, system (1.1) admits one and only one nonnegative generalized solution in the sense of Definition 1.1 (see [1], [4], and [9]). Moreover, the unique generalized solution y satisfies

$$\|y\|_{C([0,T];L^2(-1,1))} \leq \|y_0\|_{L^2(-1,1)} + \int_0^T \|u(\cdot,t)\|_{L^2(-1,1)}dt.$$

Next, set

$$L^2_+(-1,1) = \{\xi \in L^2(-1,1);\ \xi \geq 0 \text{ a.e. in } (-1,1)\},$$
$$L^2_+(Q) = \{\xi \in L^2(Q);\ \xi \geq 0 \text{ a.e. in } Q\},$$
$$L^2_+(-1,1) + Y_d = \{y + Y_d;\ y \in L^2_+(-1,1)\},$$

where $Y_d \in L^2(-1,1)$ is the value at time T of the solution of system (1.1) with $u = 0$. We need the following notions.

DEFINITION 1.2. *Target $y_1 \in L^2_+(-1,1)$ is said to be approximately controllable with a nonnegative constraint on the control if for each $\varepsilon > 0$ and $y_0 \in L^\infty(-1,1)$ with $y_0 \geq 0$ a.e. in $(-1,1)$, there exists a nonnegative control function $u \in L^2_+(Q)$ such that the corresponding solution y of system* (1.1) *satisfies*

$$\|y(\cdot,T;u) - y_1\|_{L^2(-1,1)} < \varepsilon.$$

DEFINITION 1.3. *System* (1.1) *is said to be approximately controllable with a nonnegative constraint on the control if any target* $y_1 \in L^2_+(-1,1) + Y_d$ *is approximately controllable.*

Remark 1.4. Since the solution of system (1.1) reaches Y_d at time T without any control input, by the comparison principle, if we add a nonnegative control to this system, then the value at time T of the solution of system (1.1) belongs to $L^2_+(-1,1) + Y_d$. This is the reason to choose this space as the target set.

Now, the main results in this paper are stated as follows.

THEOREM 1.5. *If* $\omega = (-1,1)$*, then system* (1.1) *is approximately controllable with a nonnegative constraint on the control at any time* $T > 0$*.*

Remark 1.6. If system (1.1) is replaced by

$$\begin{cases} y_t(x,t) - (|y|^{m-1}y)_{xx}(x,t) = u(x,t), & (x,t) \in Q, \\ y(-1,t) = y(1,t) = 0, & t \in (0,T), \\ y(x,0) = y_0(x), & x \in (-1,1), \end{cases} \tag{1.2}$$

we do not need to study the nonnegative solution. The unconstrained approximate controllability of system (1.2) is obvious. Indeed, by the well-known results, for each $\varepsilon > 0$, y_0 sufficiently smooth and $y_1 \in L^2(-1,1)$, there exist two regular functions y and v satisfying

$$\begin{cases} y_t(x,t) - y_{xx}(x,t) = v(x,t), & (x,t) \in Q, \\ y(-1,t) = y(1,t) = 0, & t \in (0,T), \\ y(x,0) = y_0(x), & x \in (-1,1), \end{cases}$$

such that

$$\|y(\cdot,T) - y_1\|_{L^2(-1,1)} < \varepsilon. \tag{1.3}$$

If choosing $u = v + y_{xx} - (|y|^{m-1}y)_{xx} \in L^\infty(Q)$, then y and u satisfy system (1.2) and (1.3). By density, for any $y_0 \in L^2(-1,1)$, system (1.2) is approximately controllable.

Combining Theorem 1.5 and energy estimates for system (1.1), it is easy to show that any nonnegative target $y_1 \in L^2_+(-1,1)$ is approximately controllable with a nonnegative constraint on the control for a long time, i.e.,

COROLLARY 1.7. *Suppose* $\omega = (-1,1)$*. Then for each* $\varepsilon > 0$*,* $y_0 \in L^\infty(-1,1)$ *with* $y_0 \geq 0$ *a.e. in* $(-1,1)$ *and target* $y_1 \in L^2_+(-1,1)$*, there exist a time* $T > 0$ *and a nonnegative control function* $u \in L^2_+(Q)$ *such that the corresponding solution* y *of system* (1.1) *satisfies*

$$\|y(\cdot,T;u) - y_1\|_{L^2(-1,1)} < \varepsilon.$$

Theorem 1.5 tells us that when the control is active in the whole domain ($\omega = (-1,1)$), any target $y_1 \in L^2_+(-1,1) + Y_d$ is approximately controllable at any time. But when the control is restricted to act in a proper subdomain ω of $(-1,1)$, one could not expect the same result. Indeed, we have the following negative result.

THEOREM 1.8. *If* $\omega \subseteq (-1,1)$ *and* $\omega \neq (-1,1)$*, then there exists a time* $T^* > 0$ *such that for any* $0 < T < T^*$*, one can find a target in* $L^2_+(-1,1) + Y_d$ *which is not approximately controllable at time* T*.*

The rest of this paper is organized as follows. In section 2 we study the constrained approximate controllability of linearized system. Section 3 is devoted to the proof of the approximate controllability of system (1.1). In section 4 we discuss the noncontrollability for this system when the control is locally acted.

2. Constrained controllability of the linearized system. In this section, we study the approximate controllability of the linearized system with nonnegative constraints on the control and state.

First, for each $z \in K := \{\xi \in L^2(Q);\ \|\xi\|_{L^\infty(0,T;L^2(-1,1))} \le C_1^*,\ 0 \le \xi \le C_2^*$ a.e. in $Q\}$ (C_1^* and C_2^* are two constants to be determined in section 3), we consider the system

$$\begin{cases} y_t - \left[(\frac{1}{k} + mz^{m-1})y_x\right]_x = u, & (x,t) \in Q, \\ y(-1,t) = y(1,t) = 0, & t \in (0,T), \\ y(x,0) = y_0(x), & x \in (-1,1), \end{cases} \tag{2.1}$$

where k is an arbitrary fixed positive integer and $y_0 \in L^\infty(-1,1)$ with $y_0 \ge 0$ a.e. in $(-1,1)$. Set

$$E = \{y(\cdot,T;u);\ u \in L^2_+(Q)\}.$$

Then we have the following known result (see [7]).

LEMMA 2.1. *E is dense in $L^2_+(-1,1) + y(\cdot,T;0)$, where $y(\cdot,T;0)$ denotes the value at time T of the solution of system (2.1) with $u = 0$.*

Next, we are ready to find a nonnegative control with minimal norm. The method is based on the following Fenchel–Rockafellar duality theory (see [3]).

LEMMA 2.2. *Let X and Y be real Banach spaces. Let $F : X \to (-\infty, +\infty]$ and $G : Y \to (-\infty, +\infty]$ be two proper, convex, and lower semicontinuous functionals. Let $L : X \to Y$ be a linear continuous operator. We suppose that there exists a $u_0 \in X$ such that $F(u_0) < +\infty$ and G is continuous at Lu_0. Set*

$$J(u) = F(u) + G(Lu)$$

and

$$\hat{J}(\varphi_T) = F^*(L^*\varphi_T) + G^*(-\varphi_T).$$

Then we have

$$\inf_{u\in X}\{J(u)\} = -\inf_{\varphi_T \in Y^*}\{\hat{J}(\varphi_T)\}, \tag{2.2}$$

where L^* denotes the adjoint operator of L, $F^*(\varphi) := \sup_{u\in X}\{(\varphi,u)_{X^*,X} - F(u)\}$ $\forall \varphi \in X^*$, and similarly for G^* defined in Y^*. Moreover, if $u^* \in X$ and $\varphi_T^* \in Y^*$ are the minimizers of the functionals J and $\hat{J}$, respectively, then

$$0 \in \partial F(u^*) - L^*\varphi_T^*, \tag{2.3}$$

where ∂F denotes the subdifferential of F.

For each $\varepsilon > 0$ and $y_d \in L^2_+(-1,1)$, consider the following functional defined on $L^2(Q)$:

$$J(u) = \frac{1}{2}\iint_Q u^2(x,t)dxdt + \begin{cases} 0 & \text{if } \|y(\cdot,T;u) - y_1\|_{L^2(-1,1)} \le \varepsilon \text{ and } u \ge 0 \text{ in } Q, \\ +\infty & \text{otherwise,} \end{cases}$$

where $y(\cdot,\cdot;u)$ denotes the solution of system (2.1) associated to u and $y_1 = y_d + y(\cdot,T;0)$.

By Lemma 2.1, $J(u) \not\equiv +\infty$. At the same time, by classical control theory, there exists a unique $u^* \in L^2(Q)$ such that

$$J(u^*) = \inf\{J(u);\, u \in L^2(Q)\}.$$

We rewrite the functional J as

$$J(u) = \frac{1}{2}\iint_Q u^2(x,t)dxdt + \begin{cases} 0 & \text{if } u \geq 0 \text{ in } Q, \\ +\infty & \text{otherwise} \end{cases}$$
$$+ \begin{cases} 0 & \text{if } \|y(\cdot,T;u) - y_1\|_{L^2(-1,1)} \leq \varepsilon, \\ +\infty & \text{otherwise} \end{cases}$$

and apply Lemma 2.2 with $X = L^2(Q)$ and $Y = L^2(-1,1)$.

Taking

$$F(u) = \frac{1}{2}\iint_Q u^2(x,t)dxdt + \begin{cases} 0 & \text{if } u \geq 0 \text{ in } Q, \\ +\infty & \text{otherwise,} \end{cases}$$

$$G(\xi) = \begin{cases} 0 & \text{if } \|\xi - y_d\|_{L^2(-1,1)} \leq \varepsilon, \\ +\infty & \text{otherwise,} \end{cases}$$

and

$$L(u) = \tilde{y}(\cdot,T;u),$$

where $\tilde{y}$ denotes the solution of system (2.1) with $y_0 = 0$, we can easily prove that the conditions of Lemma 2.2 are satisfied.

We consider the system

$$\begin{cases} \varphi_t + \left[(\frac{1}{k} + mz^{m-1})\varphi_x\right]_x = 0, & (x,t) \in Q, \\ \varphi(-1,t) = \varphi(1,t) = 0, & t \in (0,T), \\ \varphi(x,T) = \varphi_T(x), & x \in (-1,1), \end{cases} \tag{2.4}$$

where $\varphi_T \in L^2(-1,1)$.

For any $\varphi_T \in L^2(-1,1)$ and $u \in L^2(Q)$, multiplying the first equation of system (2.4) by $\tilde{y}(\cdot,\cdot;u)$ and integrating on Q, we get

$$\int_{-1}^{1} \varphi_T(x)\tilde{y}(x,T)dx = \iint_Q \varphi(x,t)u(x,t)dxdt,$$

which implies

$$L^*(\varphi_T) = \varphi \quad \forall \varphi_T \in L^2(-1,1),$$

where φ is the corresponding solution of system (2.4). Moreover,

$$G^*(-\varphi_T) = \sup\left\{-\int_{-1}^{1} \varphi_T(x)\xi(x)dx;\, \xi \in L^2(-1,1),\, \|\xi - y_d\|_{L^2(-1,1)} \leq \varepsilon\right\}$$
$$= \varepsilon\|\varphi_T\|_{L^2(-1,1)} - \int_{-1}^{1} \varphi_T(x)y_d(x)dx$$

and

$$\begin{aligned} F^*(L^*\varphi_T) &= \sup\left\{\iint_Q \varphi(x,t)u(x,t)dxdt - \frac{1}{2}\iint_Q u^2(x,t)dxdt;\ u \in L^2_+(Q)\right\} \\ &= \sup\left\{\frac{1}{2}\iint_Q \varphi^2(x,t)dxdt - \frac{1}{2}\iint_Q [u(x,t)-\varphi(x,t)]^2dxdt;\ u \in L^2_+(Q)\right\} \\ &= \frac{1}{2}\iint_Q \varphi_+^2(x,t)dxdt, \end{aligned}$$

where $w_+ := \max\{w, 0\}$. Thus by (2.2), we have

$$\inf_{u\in L^2(Q)} J(u) = -\inf_{\varphi_T\in L^2(-1,1)} \hat{J}(\varphi_T),$$

where

$$\hat{J}(\varphi_T) = \frac{1}{2}\iint_Q \varphi_+^2(x,t)dxdt + \varepsilon\|\varphi_T\|_{L^2(-1,1)} - \int_{-1}^{1}\varphi_T(x)y_d(x)dx.$$

By classical control theory, there exists a unique $\varphi_T^* \in L^2(-1,1)$ such that

$$\hat{J}(\varphi_T^*) = \inf\{\hat{J}(\varphi_T);\ \varphi_T \in L^2(-1,1)\}.$$

At the same time, by (2.3), we get

$$\iint_Q \left(\varphi^* - \frac{u+u^*}{2}\right)(u^*-u)dxdt \geq 0 \quad \forall u \in L^2_+(Q),$$

where φ^* is the solution of system (2.4) with $\varphi_T = \varphi_T^*$. The above inequality implies

$$u^* = \varphi_+^* \quad \text{in } Q. \tag{2.5}$$

Thus the linearized system (2.1) is approximately controllable with nonnegative constraints on the control and state, and the control with minimal norm satisfies (2.5).

3. Constrained controllability of system (1.1). First, we consider the constrained approximate controllability of the quasilinear system

$$\begin{cases} y_t - \left[(\frac{1}{k} + my^{m-1})y_x\right]_x = u, & (x,t) \in Q, \\ y(-1,t) = y(1,t) = 0, & t \in (0,T), \\ y(x,0) = y_0(x), & x \in (-1,1), \end{cases} \tag{3.1}$$

where k is an arbitrary fixed positive integer and $y_0 \in L^\infty(-1,1)$ with $y_0 \geq 0$ a.e. in $(-1,1)$.

For each $z \in K = \{\xi \in L^2(Q);\ \|\xi\|_{L^\infty(0,T;L^2(-1,1))} \leq C_1^*,\ 0 \leq \xi \leq C_2^* \text{ a.e. in } Q\}$ (C_1^* and C_2^* will be specified later) and $\varphi_T \in L^2(-1,1)$, we denote by φ the corresponding solution of system (2.4). For any fixed $y_d \in L^2_+(-1,1)$, denote by φ_T^* the minimizer of the functional defined on $L^2(-1,1)$,

$$\hat{J}(\varphi_T) = \frac{1}{2}\iint_Q \varphi_+^2(x,t)dxdt + \varepsilon\|\varphi_T\|_{L^2(-1,1)} - \int_{-1}^{1}\varphi_T(x)y_d(x)dx,$$

and denote by φ^* the solution of system (2.4) with $\varphi_T = \varphi_T^*$.

Noting that φ_T^* and φ^* depend on $z \in K$, we define the operator

$$L: \ K \to L^2(Q),$$
$$z \mapsto y,$$

where $y = L(z)$ denotes the solution of system (2.1) with $u^* = \varphi_+^*$. From the discussion of section 2, the definition of L is reasonable, and for each $z \in K$, we have

$$\|y(\cdot, T; u^*) - y_d - Y_d^z\|_{L^2(-1,1)} \le \varepsilon,$$

where Y_d^z denotes the value at time T of the solution of system (2.1) with $u = 0$.

In the following, we shall prove that L has a fixed point by Schauder's fixed point theorem, which yields the approximate controllability of system (3.1).

Step 1. We prove that $L(K) \subseteq K$. First, we shall establish the uniform boundedness of $\{\varphi_+^*\}$ for $z \in K$. For any $z \in K$ and $\varphi_T \in L^2(-1,1)$, we denote by φ the corresponding solution of system (2.4). Consider the following two functionals defined on $L^2(-1,1)$:

$$\hat{J}(\varphi_T) = \frac{1}{2}\iint_Q \varphi_+^2(x,t)dxdt + \varepsilon\|\varphi_T\|_{L^2(-1,1)} - \int_{-1}^1 \varphi_T(x)y_d(x)dx,$$

$$\bar{J}(\varphi_T) = \frac{1}{2}\iint_Q \varphi_+^2(x,t)dxdt - \int_{-1}^1 \varphi(x, T-\delta)y_d(x)dx,$$

where $\delta > 0$ is a constant to be determined later. Denote by φ_T^* the minimizer of the functional $\hat{J}$. We shall show that there exists a suitable $\delta > 0$ such that, for any $z \in K$ and $\varphi_T \in L^2(-1,1)$,

$$\varepsilon\|\varphi_T\|_{L^2(-1,1)} - \int_{-1}^1 [\varphi_T(x) - \varphi(x, T-\delta)]\, y_d(x)dx \ge 0. \tag{3.2}$$

This leads to

$$\bar{J}(\varphi_T^*) \le \hat{J}(\varphi_T^*) \le 0.$$

Thus

$$\iint_Q \varphi_+^{*2}(x,t)dxdt \le 2\int_{-1}^1 \varphi^*(x, T-\delta)y_d(x)dx \le 2\int_{-1}^1 \varphi_+^*(x, T-\delta)y_d(x)dx. \tag{3.3}$$

Here we use the fact that $y_d \ge 0$ a.e. in $(-1,1)$. Since φ^* solves

$$\begin{cases} \varphi_t^* + \left[(\frac{1}{k} + mz^{m-1})\varphi_x^*\right]_x = 0, & (x,t) \in Q, \\ \varphi^*(-1,t) = \varphi^*(1,t) = 0, & t \in (0,T), \\ \varphi^*(x,T) = \varphi_T^*(x), & x \in (-1,1), \end{cases} \tag{3.4}$$

multiplying the first equation of system (3.4) by φ_+^* and integrating on $(-1,1) \times (T-\delta, t)$, where $t \in (T-\delta, T)$, we get

$$\int_{-1}^1 \varphi_+^{*2}(x, T-\delta)dx \le \int_{-1}^1 \varphi_+^{*2}(x,t)dx.$$

Integrating the above inequality with respect to t on $(T-\delta, T)$, it follows that

$$\int_{-1}^{1} \varphi_+^{*2}(x, T-\delta)dx \le \frac{1}{\delta} \iint_Q \varphi_+^{*2}(x,t)dxdt. \tag{3.5}$$

Taking into account (3.3) and (3.5), we have

$$\|\varphi_+^*\|_{L^2(Q)} \le \frac{2}{\delta^{1/2}} \|y_d\|_{L^2(-1,1)}, \tag{3.6}$$

which means that $\{\varphi_+^*\}$ is uniformly bounded in $L^2(Q)$ if inequality (3.2) is valid.

Next, we shall specify constants C_1^* and C_2^* in the definition of set K. For any $z \in K$, multiplying by y the first equation of system (2.1) with $u = \varphi_+^*$ and integrating with respect to x on $(-1,1)$, we get

$$\frac{1}{2}\frac{d}{dt}\int_{-1}^{1} y^2(x,t)dx \le \int_{-1}^{1} \varphi_+^*(x,t)y(x,t)dx.$$

By Gronwall's inequality and (3.6), we have

$$\begin{aligned} \|y\|_{L^\infty(0,T;L^2(-1,1))} &\le e^{T/2}(\|y_0\|_{L^2(-1,1)} + \|\varphi_+^*\|_{L^2(Q)}) \\ &\le e^{T/2}(\|y_0\|_{L^2(-1,1)} + \frac{2}{\delta^{1/2}}\|y_d\|_{L^2(-1,1)}) \\ &:= C_1^*. \end{aligned} \tag{3.7}$$

At the same time, using the weak maximum principle (see [16]) for system (2.1) with $u = \varphi_+^*$, by (3.6), we get that there exists a constant $C = C(k) > 0$ independent of z such that

$$\begin{aligned} \|y\|_{L^\infty(Q)} &\le \|y_0\|_{L^\infty(-1,1)} + C\|\varphi_+^*\|_{L^2(Q)} \\ &\le \|y_0\|_{L^\infty(-1,1)} + \frac{2C}{\delta^{1/2}}\|y_d\|_{L^2(-1,1)} \\ &:= C_2^*. \end{aligned}$$

With the choice of C_1^* and C_2^*, we get that $L(K) \subseteq K$.

Now, we show that inequality (3.2) holds for some $\delta > 0$. Without loss of generality, we suppose $y_d \in C_0^1(-1,1)$. Multiplying the first equation of system (2.4) by y_d and integrating on $(-1,1) \times (T-\delta, T)$, we get

$$\int_{-1}^{1} [\varphi_T(x) - \varphi(x, T-\delta)]y_d(x)dx - \int_{T-\delta}^{T}\int_{-1}^{1} \left(\frac{1}{k} + mz^{m-1}\right)\varphi_x(x,t)y_{dx}(x)dxdt = 0.$$

It suffices to show that there exists a $\delta > 0$ such that, for any $z \in K$ and $\varphi_T \in L^2(-1,1)$,

$$\int_{T-\delta}^{T}\int_{-1}^{1} \left(\frac{1}{k} + mz^{m-1}\right)|\varphi_x(x,t)y_{dx}(x)|dxdt \le \varepsilon\|\varphi_T\|_{L^2(-1,1)}. \tag{3.8}$$

By Hölder's inequality and inequality (3.7), we estimate the term on the left-hand

side of inequality (3.8):

$$\begin{aligned}
&\int_{T-\delta}^{T}\int_{-1}^{1}(\tfrac{1}{k}+mz^{m-1})|\varphi_x(x,t)y_{dx}(x)|dxdt\\
&\quad\le \|y_{dx}\|_{L^\infty(-1,1)}\|(\tfrac{1}{k}+mz^{m-1})^{1/2}\varphi_x\|_{L^2(Q)}\|(\tfrac{1}{k}+mz^{m-1})^{1/2}\|_{L^2((-1,1)\times(T-\delta,T))}\\
&\quad\le \|y_{dx}\|_{L^\infty(-1,1)}\|\varphi_T\|_{L^2(-1,1)}\|(1+mz^{m-1})^{1/2}\|_{L^2((-1,1)\times(T-\delta,T))}\\
&\quad\le \|y_{dx}\|_{L^\infty(-1,1)}\|\varphi_T\|_{L^2(-1,1)}\left\{2\delta+2^{(3-m)/2}m\delta\|z\|^{m-1}_{L^\infty(0,T;L^2(-1,1))}\right\}^{1/2}\\
&\quad\le \|y_{dx}\|_{L^\infty(-1,1)}\|\varphi_T\|_{L^2(-1,1)}\cdot\\
&\left\{2\delta+2^{(3-m)/2}m\delta\left[e^{T/2}\left(\|y_0\|_{L^2(-1,1)}+\frac{2}{\delta^{1/2}}\|y_d\|_{L^2(-1,1)}\right)\right]^{m-1}\right\}^{1/2}. \qquad (3.9)
\end{aligned}$$

Since $1<m<3$, we can choose a sufficiently small constant $\delta>0$, such that inequality (3.8) is valid and δ depends only on y_0, y_d, and ε.

Thus inequality (3.2) is valid. At the same time, we also conclude that $\{\|\varphi_+^*\|_{L^2(Q)}\}$ is uniformly bounded and $L(K)\subseteq K$.

Remark 3.1. From the above discussion, a critical point to the proof of Step 1 is the uniform boundedness of $\{\varphi_+^*\}$, i.e., estimate (3.6). Taking into account that the principal part of system (3.1) is nonlinear, we cannot use the same method as that in [11, Proposition 2.3] and in [7, Proposition 15] to prove it. Indeed, for any $z\in K$, we fail to get the uniform boundedness of the minimizer $\{\varphi_T^*\}$ of $\hat{J}(\varphi_T)$. The uniform boundedness of $\{\varphi_T^*\}$ implies that of $\{\varphi_+^*\}$. The latter is enough for our discussion. Here we make use of two functionals $\hat{J}(\varphi_T)$ and $\bar{J}(\varphi_T)$ introduced in [12] and suitable estimates for the linearized system to prove (3.6).

Step 2. We prove the continuity of L. Let $\{z_n\}$ be any convergent sequence in K, say,

$$z_n\to z_0 \quad \text{strongly in } L^2(Q).$$

Denote by $\hat{J}_n$ and $\hat{J}_0$ the functional $\hat{J}$ with $z=z_n$ and $z=z_0$ in (2.4), respectively. Let φ_n be the solution of system (2.4) with $z=z_n$ and $\varphi_T=\varphi_{Tn}$, where φ_{Tn} is the minimizer of $\hat{J}_n$. Let y_n be the solution of system (2.1) with $z=z_n$ and $u=\varphi_{n+}$. Then we have the following result.

LEMMA 3.2. *There exists a constant $C>0$ independent of z_n such that*

$$\|\varphi_{Tn}\|_{L^2(-1,1)}\le C.$$

Proof. Suppose that there exists a subsequence of $\{z_n\}$ (still denoted by $\{z_n\}$) such that $\|\varphi_{Tn}\|_{L^2(-1,1)}\to+\infty$, $n\to+\infty$. Let

$$\hat{\varphi}_n=\frac{\varphi_n}{\|\varphi_{Tn}\|_{L^2(-1,1)}},\quad \hat{\varphi}_{Tn}=\frac{\varphi_{Tn}}{\|\varphi_{Tn}\|_{L^2(-1,1)}}.$$

Then

$$\frac{\hat{J}(\varphi_{Tn})}{\|\varphi_{Tn}\|_{L^2(-1,1)}}=\frac{1}{2}\iint_Q\hat{\varphi}_{n+}^2(x,t)dxdt\|\varphi_{Tn}\|_{L^2(-1,1)}+\varepsilon-\int_{-1}^{1}\hat{\varphi}_{Tn}(x)y_d(x)dx.$$

If $\liminf_{n\to+\infty}\frac{1}{2}\iint_Q\hat{\varphi}_{n+}^2(x,t)dxdt>0$, then

$$\liminf_{n\to+\infty}\frac{\hat{J}(\varphi_{Tn})}{\|\varphi_{Tn}\|_{L^2(-1,1)}}\ge\varepsilon. \qquad (3.10)$$

On the other hand, suppose that

$$\liminf_{n\to+\infty} \frac{1}{2} \iint_Q \hat{\varphi}_{n+}^2(x,t)dxdt = 0. \tag{3.11}$$

Since $\|\hat{\varphi}_{Tn}\|_{L^2(-1,1)} = 1$, we can extract a subsequence (still denoted by $\{\hat{\varphi}_{Tn}\}$) which weakly converges to some element $\hat{\varphi}_T$ in $L^2(-1,1)$. Since $z_n \to z_0$ strongly in $L^2(Q)$ and $0 \le z_n \le C_2^*$ a.e. in Q, we can extract a subsequence (still denoted by $\{z_n\}$) which converges to z_0 in $L^p(Q)$ for any $0 < p < +\infty$. Thus $\hat{\varphi}_n$ converges to some element $\hat{\varphi}$ strongly in $L^2(Q)$ and weakly in $L^2(0,T;H_0^1(-1,1))$. We can easily verify that $\hat{\varphi}$ is the solution of system (2.4) with $z = z_0$ and $\varphi_T = \hat{\varphi}_T$. By (3.11), we get

$$\hat{\varphi}_+ = 0 \quad \text{a.e. in } Q.$$

Thus

$$\hat{\varphi}_T(x) \le 0 \quad \text{a.e. in } (-1,1).$$

Moreover, noticing that $y_d \ge 0$ a.e. in $(-1,1)$, we have

$$\liminf_{n\to+\infty} \frac{\hat{J}(\varphi_{Tn})}{\|\varphi_{Tn}\|_{L^2(-1,1)}} \ge \varepsilon - \int_{-1}^{1} \hat{\varphi}_T(x) y_d(x) dx \ge \varepsilon. \tag{3.12}$$

But $\hat{J}(\varphi_{Tn}) \le \hat{J}(0) = 0$, which is a contradiction to (3.10) and (3.12). This completes the proof. □

By (3.9), there exist a subsequence of $\{\varphi_{Tn}\}$ (still denoted by itself) and $\hat{\varphi}_T \in L^2(-1,1)$ such that

$$\varphi_{Tn} \to \hat{\varphi}_T \text{ weakly in } L^2(-1,1). \tag{3.13}$$

Thus $\{\varphi_n\}$ and $\{y_n\}$ have subsequences (still denoted by themselves) which satisfy, respectively,

$$\begin{aligned} &\varphi_n \to \hat{\varphi} \quad \text{strongly} \quad \text{in } L^2(Q), \\ &y_n \to y \quad \text{weakly} \quad \text{in } L^2(0,T;H_0^1(-1,1)), \\ &y_n \to y \quad \text{strongly} \quad \text{in } L^2(Q), \end{aligned} \tag{3.14}$$

where $\hat{\varphi}$ is the solution of system (2.4) with $z = z_0$ and $\varphi_T = \hat{\varphi}_T$ and y is the solution of system (2.1) with $z = z_0$ and $u = \hat{\varphi}_+$. Since

$$\hat{J}_n(\varphi_{Tn}) \le \hat{J}_n(\varphi_T) \quad \forall \varphi_T \in L^2(-1,1),$$

by (3.13) and (3.14), we have

$$\lim_{n\to+\infty} \hat{J}_n(\varphi_T) = \hat{J}_0(\varphi_T), \quad \hat{J}_0(\hat{\varphi}_T) \le \liminf_{n\to+\infty} \hat{J}_n(\varphi_{Tn}).$$

This implies

$$\hat{J}_0(\hat{\varphi}_T) \le \hat{J}_0(\varphi_T) \quad \forall \varphi_T \in L^2(-1,1);$$

that is, $\hat{\varphi}_T$ is the minimizer of $\hat{J}_0$. Thus L is a continuous operator.

Step 3. We prove the compactness of L. Taking into account the uniform boundedness of $\{\varphi_+^*\}$ in $L^2(Q)$ for $z \in K$, if we denote by y the solution of system (2.1) with $u = \varphi_+^*$, then

$$\|y\|_{L^2(0,T;H_0^1(-1,1))} + \|y_t\|_{L^2(0,T;H^{-1}(-1,1))} \le C$$

with C independent of z. By Aubin's compact theorem, $L(K)$ is a compact subset of $L^2(Q)$. Thus L is a compact operator.

By Schauder's fixed point theorem, the above discussion implies that L has a fixed point. Thus system (3.1) is approximately controllable with nonnegative constraints on the control and state; that is, for each $\varepsilon > 0$ and $y_d \in L_+^2(-1,1)$, there exists a nonnegative control function $u_k \in L_+^2(Q)$ such that the corresponding nonnegative solution y_k of system (3.1) satisfies

$$\|y_k(\cdot,T;u_k) - y_d - y_k(\cdot,T;0)\|_{L^2(-1,1)} < \frac{\varepsilon}{2}, \tag{3.15}$$

where $y_k(\cdot,\cdot;0)$ denotes the solution of system (3.1) with $u = 0$.

Next, we discuss the constrained approximate controllability of system (1.1).

Proof of Theorem 1.5. By (3.6), $\{u_k\}$ can be chosen such that

$$\|u_k\|_{L^2(Q)} \le C,$$

with C independent of k. Thus there exist a subsequence of $\{u_k\}$ (still denoted by $\{u_k\}$) and a nonnegative function $u^* \in L_+^2(Q)$ such that

$$u_k \to u^* \text{ weakly in } L^2(Q).$$

Using Moser's iteration and a similar method used in the proof of Lemma 5 in [1], we have that there exists a constant $C > 0$ independent of k such that

$$\|y_k\|_{L^\infty(Q)\cap C([0,T];L^2(-1,1))} \le C, \quad \|y_k^{(m-1)/2} y_{kx}\|_{L^2(Q)} \le C,$$

$$\|y_k^m(x+h,t) - y_k^m(x,t)\|_{L^2(Q_h)} \le C|h|, \quad \|y_k^m(x,t+h) - y_k^m(x,t)\|_{L^2(Q_h)}^2 \le C|h|^{1/2},$$

where h is a real parameter and $Q_h := \{(x,t) \in Q;\ (x+h,t) \in Q,\ (x,t+h) \in Q\}$.

Thus there exist a subsequence of $\{y_k\}$ (still denoted by itself) and $w \in L^\infty(Q)$ such that for any $p > 1$

$$y_k^m \to w \text{ strongly in } L^p(Q).$$

Define $y := w^{1/m}$. Then there exists a subsequence of $\{y_k\}$ such that

$$y_k \to y \quad \text{a.e. in Q.}$$

By the Lebesgue dominated convergence theorem, this yields that for any $p > 1$

$$y_k \to y \text{ strongly in } L^p(Q).$$

By (3.1), for any test function $\varphi \in C^2(\overline{Q})$ satisfying the conditions of Definition 1.1, y_k satisfies the following integral equality:

$$\iint_Q \left[y_k(x,t)\varphi_t(x,t) + \frac{1}{k} y_k(x,t)\varphi_{xx}(x,t) + y_k^m(x,t)\varphi_{xx}(x,t) \right] dxdt$$
$$= -\iint_Q u_k(x,t)\varphi(x,t)dxdt - \int_{-1}^{1} y_0(x)\varphi(x,0)dx.$$

Letting $k \to +\infty$, we get

$$\begin{aligned}&\iint_Q [y(x,t)\varphi_t(x,t) + y^m(x,t)\varphi_{xx}(x,t)]dxdt\\ &= -\iint_Q u^*(x,t)\varphi(x,t)dxdt - \int_{-1}^{1} y_0(x)\varphi(x,0)dx.\end{aligned}$$

Thus y is the nonnegative generalized solution of system (1.1) with $u = u^*$.

At the same time,

$$y_k(\cdot, T; u_k) \to y(\cdot, T; u^*) \text{ weakly in } L^2(-1,1). \tag{3.16}$$

Similarly, there exists a subsequence of $\{y_k(\cdot,\cdot;0)\}$ (still denoted by itself) such that for any $p > 1$

$$\begin{aligned} y_k(\cdot,\cdot;0) &\to y(\cdot,\cdot;0) \quad \text{strongly} \quad \text{in } L^p(Q),\\ y_k(\cdot,T;0) &\to y(\cdot,T;0) \quad \text{weakly} \quad \text{in } L^2(-1,1),\end{aligned} \tag{3.17}$$

where $y(\cdot,\cdot;0)$ is the generalized solution of system (1.1) with $u = 0$.

By (3.16) and (3.17), for each $\varepsilon > 0$, there exists a $K > 0$ such that

$$\|y(\cdot,T;u^*) - y_d - y(\cdot,T;0)\|_{L^2(-1,1)} \le \|y_K(\cdot,T;u_K) - y_d - y_K(\cdot,T;0)\|_{L^2(-1,1)} + \frac{\varepsilon}{2}.$$

Taking into account (3.15), we have

$$\|y(\cdot,T;u^*) - y_d - y(\cdot,T;0)\|_{L^2(-1,1)} \le \varepsilon,$$

which means that system (1.1) is approximately controllable with nonnegative constraints on the control and state. We complete the proof of Theorem 1.5.

Combining Theorem 1.5 and energy estimates for system (1.1), we also get that any nonnegative target is approximately controllable for a long time.

Proof of Corollary 1.7. Taking into account the proof of Theorem 1.5, we can easily get that for each $\varepsilon > 0$ and $y_1 \in L^2_+(-1,1)$, there exists a constant $\gamma > 0$ such that if $y_0 \in L^\infty(-1,1)$ satisfies $\|y_0\|_{L^2(-1,1)} < \gamma$, then we can find a nonnegative control function $u \in L^2_+(Q)$ satisfying

$$\|y(\cdot,T;u) - y_1\|_{L^2(-1,1)} < \varepsilon,$$

where y denotes the corresponding solution of system (1.1).

For any $y_0 \in L^\infty(-1,1)$ with $y_0 \ge 0$ a.e. in $(-1,1)$, denote by $\bar{y}$ the solution of system (1.1) with $u = 0$. Multiplying the first equation of system (1.1) with $u = 0$ by $\bar{y}^m$ and integrating with respect to x on $(-1,1)$, we have

$$\frac{1}{m+1}\frac{d}{dt}\int_{-1}^{1} \bar{y}^{m+1}(x,t)dx + \int_{-1}^{1} (\bar{y}^m)_x^2(x,t)dx = 0. \tag{3.18}$$

By Poincáre's inequality, there exists a constant $C > 0$ such that

$$\int_{-1}^{1} \bar{y}^{2m}(x,t)dx \le C\int_{-1}^{1} (\bar{y}^m)_x^2(x,t)dx.$$

Set

$$C_1 = \frac{m+1}{C} 2^{\frac{1-m}{1+m}}.$$

By Hölder's inequality and (3.18), we get

$$\begin{aligned}
&\frac{d}{dt}\int_{-1}^{1} \bar{y}^{m+1}(x,t)dx + C_1 \left[\int_{-1}^{1} \bar{y}^{m+1}(x,t)dx\right]^{\frac{2m}{m+1}} \\
&\le \frac{d}{dt}\int_{-1}^{1} \bar{y}^{m+1}(x,t)dx + C_1 2^{\frac{m-1}{m+1}} \int_{-1}^{1} \bar{y}^{2m}(x,t)dx \\
&\le \frac{d}{dt}\int_{-1}^{1} \bar{y}^{m+1}(x,t)dx + (m+1)\int_{-1}^{1} (\bar{y}^m)_x^2(x,t)dx = 0.
\end{aligned}$$

Hence $\int_{-1}^{1} \bar{y}^{m+1}(x,t)dx$ is less than or equal to the solution of the following problem:

$$H'(t) + C_1 H^{2m/(m+1)}(t) = 0,$$

$$H(0) = \int_{-1}^{1} y_0^{m+1}(x)dx.$$

This implies

$$\int_{-1}^{1} \bar{y}^{m+1}(x,t)dx \le H(t) = \left[C_1 \frac{m-1}{m+1} t + \left(\int_{-1}^{1} y_0^{m+1}(x)dx\right)^{\frac{1-m}{1+m}}\right]^{\frac{1+m}{1-m}}.$$

On the other hand,

$$\|\bar{y}(\cdot,t)\|_{L^2(-1,1)} \le 2^{\frac{m-1}{2m+2}} \|\bar{y}(\cdot,t)\|_{L^{m+1}(-1,1)}.$$

Thus

$$\|\bar{y}(\cdot,t)\|_{L^2(-1,1)} \le 2^{\frac{m-1}{2m+2}} \left[C_1 \frac{m-1}{m+1} t + \left(\int_{-1}^{1} y_0^{m+1}(x)dx\right)^{\frac{1-m}{1+m}}\right]^{\frac{1}{1-m}}. \tag{3.19}$$

Taking into account (3.19) and $m > 1$, we can find a sufficiently large $T_1 > 0$ such that

$$\|\bar{y}(\cdot,T_1)\|_{L^2(-1,1)} < \gamma.$$

Therefore, for any $T > T_1$, we can find a nonnegative control function u defined on $(-1,1) \times (T_1,T)$ such that y_1 is approximately controllable at time T.

We choose

$$\tilde{u}(x,t) = \begin{cases} 0, & (-1,1) \times (0,T_1), \\ u(x,t), & (-1,1) \times (T_1,T), \end{cases}$$

and denote by y the solution of system (1.1) with $u = \tilde{u}$; then

$$\|y(\cdot,T;\tilde{u}) - y_1\|_{L^2(-1,1)} < \varepsilon.$$

This implies that for any nonnegative target $y_1 \in L^2_+(-1,1)$, we can find a time $T > 0$ and a nonnegative control function $\tilde{u} \in L^2_+(Q)$ such that y_1 is approximately controllable. We complete the proof of Corollary 1.7.

4. Noncontrollability of system (1.1) under a local control. In this section, we consider the control system

$$\begin{cases} y_t(x,t) - (y^m)_{xx}(x,t) = \chi_\omega u(x,t), & (x,t) \in Q, \\ y(-1,t) = y(1,t) = 0, & t \in (0,T), \\ y(x,0) = y_0(x), & x \in (-1,1), \end{cases} \tag{4.1}$$

where ω is a nonempty strict open subset of $(-1,1)$. First, using a technique of local estimates, we shall prove an L^∞-obstruction phenomenon to the solution of system (4.1). We need the following iteration lemma.

LEMMA 4.1 (see [21]). *Let $\{y_n\}$ $(n = 0,1,2,\cdots)$ be a sequence of positive numbers satisfying*

$$y_{n+1} \le C b^n y_n^{1+\alpha},$$

where $C > 0$, $b > 1$ and $\alpha > 0$ are constants. If

$$y_0 \le C^{-1/\alpha} b^{-1/\alpha^2},$$

then

$$\lim_{n\to+\infty} y_n = 0.$$

In what follows, for a fixed $x_0 \in (-1,1)\backslash\overline{\omega}$ and a positive constant $R > 0$ satisfying $B_{2R}(x_0) = (x_0 - 2R,\, x_0 + 2R) \subseteq (-1,1)\backslash\overline{\omega}$, define

$$\phi(t,\rho) = \sup_{\tau\in(0,t)} \tau^{1/(m+1)} \sup_{\rho\le r\le 2R} \frac{\|y(\cdot,\tau)\|_{L^\infty(B_r(x_0))}}{r^{2/(m-1)}},$$

$$K(t,\rho) = t^{-1} + t^{(1-m)/(1+m)} \phi^{m-1}(t,\rho),$$

$$\psi(t,\rho) = \sup_{\tau\in(0,t)} \sup_{\rho\le r\le 2R} r^{(1+m)/(1-m)} \int_{B_r(x_0)} y(x,\tau)dx,$$

where $0 < t \le T$ and $0 < \rho < R$.

We now prove the following lemma.

LEMMA 4.2. *There exists a constant $C > 0$ such that for any $0 < t \le T$, $0 < \rho < R$, and any control function $u \in L^2_+(\omega \times (0,T))$, the solution of system (4.1) satisfies*

$$\|y(\cdot,t)\|_{L^\infty(B_\rho(x_0))} \le C\,[K(t,\rho)]^{3/(3m+1)} \left(\int_{\frac{t}{4}}^t \int_{B_{2\rho}(x_0)} y^{m+1} dx d\tau \right)^{2/(3m+1)}. \tag{4.2}$$

Proof. For $n = 0,1,2,\ldots$, denote

$$\begin{aligned} &t_n = \frac{t}{2} - \frac{t}{2^{n+2}},\ \rho_n = \rho + \frac{\rho}{2^n},\ \bar{\rho_n} = \frac{1}{2}(\rho_n + \rho_{n+1}), \\ &B_n = B_{\rho_n}(x_0),\ B_n' = B_{\bar{\rho_n}}(x_0), \\ &Q_n = B_n \times (t_n, t),\ Q_n' = B_n' \times (t_{n+1}, t), \\ &k_n = k - \frac{k}{2^n}, \end{aligned}$$

where $k > 0$ will be specified later. Obviously

$$Q'_{n+1} \subseteq Q_{n+1} \subseteq Q'_n \subseteq Q_n.$$

Let ξ_n be a smooth function defined on Q_n, satisfying

$$0 \le \xi_n \le 1 \text{ in } Q_n, \quad \xi_n = 1 \text{ in } Q'_n,$$

$$\xi_n = 0 \text{ on } \{x_0 \pm \rho_n\} \times (t_n, t) \cup B_n \times \{t_n\},$$

$$|\xi_{nx}| \le \frac{2^{n+2}}{\rho}, \quad 0 \le \xi_{nt} \le \frac{2^{n+3}}{t}.$$

Multiplying the first equation of (4.1) by $(y-k_n)_+^m \xi_n^2$ and integrating on Q_n, by Höder's inequality, we get

(4.3)
$$\begin{aligned}
&\frac{1}{m+1}\int_{B_n}(y-k_n)_+^{m+1}(x,t)\xi_n^2(x,t)dx + m^2\iint_{Q_n}(y-k_n)_+^{m-1}y^{m-1}y_x^2\xi_n^2 dxd\tau \\
&\quad = \frac{2}{m+1}\iint_{Q_n}(y-k_n)_+^{m+1}\xi_n\xi_{n\tau}dxd\tau - 2m\iint_{Q_n}(y-k_n)_+^{m}y^{m-1}y_x\xi_n\xi_{nx}dxd\tau \\
&\quad \le \frac{C2^n}{t}\iint_{Q_n}(y-k_n)_+^{m+1}dxd\tau + \frac{1}{2}m^2\iint_{Q_n}(y-k_n)_+^{m-1}y^{m-1}y_x^2\xi_n^2 dxd\tau \\
&\qquad + C\iint_{Q_n}(y-k_n)_+^{m+1}y^{m-1}\xi_{nx}^2 dxd\tau \\
&\quad \le \frac{C2^n}{t}\iint_{Q_n}(y-k_n)_+^{m+1}dxd\tau + \frac{1}{2}m^2\iint_{Q_n}(y-k_n)_+^{m-1}y^{m-1}y_x^2\xi_n^2 dxd\tau \\
&\qquad + \frac{C4^n}{\rho^2}\sup_{t_n\le\tau\le t}\|y(\cdot,\tau)\|_{L^\infty(B_n)}^{m-1}\iint_{Q_n}(y-k_n)_+^{m+1}dxd\tau;
\end{aligned}$$

here and hereafter C denotes a different constant. Moreover,

(4.4)
$$\begin{aligned}
\iint_{Q_n}(y-k_n)_+^{m-1}y^{m-1}y_x^2\xi_n^2 dxd\tau &\ge \iint_{Q_n}(y-k_n)_+^{2m-2}y_x^2\xi_n^2 dxd\tau \\
&= \frac{1}{m^2}\iint_{Q_n}\left[(y-k_n)_+^m\right]_x^2\xi_n^2 dxd\tau.
\end{aligned}$$

Hence if we denote

$$w_n = (y-k_n)_+^m,$$

then by (4.3) and (4.4) we have

$$\sup_{t_{n+1}\le\tau\le t}\int_{B'_n} w_n^{\frac{m+1}{m}}(x,\tau)dx + \iint_{Q'_n} w_{nx}^2 dxd\tau \le C4^n K(t,\rho)\iint_{Q_n} w_n^{\frac{m+1}{m}}dxd\tau. \tag{4.5}$$

Let $\bar{\xi_n}$ be another smooth function defined on B'_n, satisfying

$$0 \le \bar{\xi_n} \le 1, \quad \bar{\xi_n} = 1 \text{ in } B_{n+1}, \quad |\bar{\xi}_{nx}| \le \frac{2^{n+2}}{\rho}.$$

By the Gagliardo–Nirenberg embedding inequality, we get

$$\begin{aligned}\iint_{Q_{n+1}} w_{n+1}^{(4m+2)/m}dxd\tau &\le \iint_{Q'_n} (w_{n+1}\bar{\xi_n})^{(4m+2)/m}dxd\tau\\ &\le \iint_{Q'_n} (w_n\bar{\xi_n})^{(4m+2)/m}dxd\tau\\ &\le C\iint_{Q'_n} (w_{nx}^2 + w_n^2\bar{\xi_{nx}}^2)dxd\tau\\ &\quad\cdot\left[\sup_{t_{n+1}\le\tau\le t}\int_{B'_n} w_n^{(m+1)/m}(x,\tau)dx\right]^2.\end{aligned} \tag{4.6}$$

We notice that the definition of w_n and $\phi(t,\rho)$ implies

$$\begin{aligned}\iint_{Q'_n} w_n^2\bar{\xi_{nx}}^2dxd\tau &\le \frac{C4^n}{\rho^2}\iint_{Q'_n}(y-k_n)_+^{2m}dxd\tau\\ &\le \frac{C4^n}{\rho^2}\sup_{t_{n+1}\le\tau\le t}\|(y-k_n)_+\|_{L^\infty(B'_n)}^{m-1}\iint_{Q'_n}(y-k_n)_+^{m+1}dxd\tau\\ &\le C4^nK(t,\rho)\iint_{Q'_n}(y-k_n)_+^{m+1}dxd\tau.\end{aligned} \tag{4.7}$$

Substituting (4.5) and (4.7) into (4.6) and estimating its right-hand side, we obtain

$$\iint_{Q_{n+1}} w_{n+1}^{(4m+2)/m}dxd\tau \le C4^{3n}K^3(t,\rho)\left[\iint_{Q_n}(y-k_n)_+^{m+1}dxd\tau\right]^3. \tag{4.8}$$

Set

$$A_n = \{(x,\tau)\in Q_n;\, y(x,\tau) > k_n\},\quad |A_n| = \operatorname{meas} A_n.$$

By Hölder's inequality and (4.8), we get

$$\begin{aligned}\iint_{Q_{n+1}}(y-k_{n+1})_+^{m+1}dxd\tau &\le \left[\iint_{Q_{n+1}}(y-k_{n+1})_+^{4m+2}dxd\tau\right]^{\frac{m+1}{4m+2}}|A_{n+1}|^{\frac{3m+1}{4m+2}}\\ &\le 4^{\frac{3m+3}{4m+2}n}[K(t,\rho)]^{\frac{3m+3}{4m+2}}|A_{n+1}|^{\frac{3m+1}{4m+2}}\\ &\quad\cdot\left[\iint_{Q_n}(y-k_n)_+^{m+1}dxd\tau\right]^{\frac{3m+3}{4m+2}}.\end{aligned} \tag{4.9}$$

On the other hand, we notice that

$$\begin{aligned}\iint_{Q_n}(y-k_n)_+^{m+1}dxd\tau &\ge \iint_{Q_{n+1}}(k_{n+1}-k_n)^{m+1}dxd\tau\\ &= \frac{k^{m+1}}{2^{(n+1)(m+1)}}|A_{n+1}|.\end{aligned} \tag{4.10}$$

Thus we derive from (4.9) and (4.10) that

$$\iint_{Q_{n+1}}(y-k_{n+1})_+^{m+1}dxd\tau \le C4^{n\beta_1}2^{n\beta_2}k^{\beta_3}K^{\beta_4}(t,\rho)\left[\iint_{Q_n}(y-k_n)_+^{m+1}dxd\tau\right]^{\frac{3m+2}{2m+1}},$$

where $\beta_1 = (3m+3)/(4m+2)$, $\beta_2 = (m+1)(3m+1)/(4m+2)$, $\beta_3 = -(m+1)(3m+1)/(4m+2)$, and $\beta_4 = (3m+3)/(4m+2)$. Applying Lemma 4.1, we conclude that if

$$\iint_{Q_0} y^{m+1} dx d\tau \le C k^{(3m+1)/2} K^{-3/2}(t,\rho),$$

then

$$\int_{\frac{t}{2}}^{t} \int_{B_\rho(x_0)} (y-k)_+^{m+1} dx d\tau = 0.$$

Therefore, if we take

$$k = C[K(t,\rho)]^{3/(3m+1)} \left(\iint_{Q_0} y^{m+1} dx d\tau \right)^{2/(3m+1)},$$

then

$$\|y(\cdot,t)\|_{L^\infty(B_\rho(x_0))} \le k;$$

that is,

$$\|y(\cdot,t)\|_{L^\infty(B_\rho(x_0))} \le C[K(t,\rho)]^{3/(3m+1)} \left(\int_{\frac{t}{4}}^{t} \int_{B_{2\rho}(x_0)} y^{m+1} dx d\tau \right)^{2/(3m+1)}.$$

We get the conclusion of Lemma 4.2. □

LEMMA 4.3. *There exist two positive constants C_1 and C_2 such that for any $t > 0$,*

$$\phi(t,\rho) \le C_1 \int_0^t \tau^{(1-m)/(1+m)} \phi^m(\tau,\rho) d\tau + C_2 [\psi(t,\rho)]^{2/(m+1)}. \tag{4.11}$$

Proof. Denote

$$\Phi(t,\rho) = t^{1/(m+1)} \frac{\|y(\cdot,t)\|_{L^\infty(B_\rho(x_0))}}{\rho^{2/(m-1)}}.$$

Multiplying (4.2) by $t^{1/(m+1)} \rho^{2/(1-m)}$, we get

(4.12)

$$\begin{aligned}
\phi(t,\rho) &\le C t^{1/(m+1)} \rho^{2/(1-m)} [K(t,\rho)]^{3/(3m+1)} \left(\int_{\frac{t}{4}}^{t} \int_{B_{2\rho}(x_0)} y^{m+1} dx d\tau \right)^{2/(3m+1)} \\
&\le C t^{1/(m+1)} \rho^{2/(1-m)} \left[t^{-1} + t^{(1-m)/(1+m)} \phi^{m-1}(t,\rho) \right]^{3/(3m+1)} \\
&\quad \cdot \left(\int_{\frac{t}{4}}^{t} \int_{B_{2\rho}(x_0)} y^{m+1} dx d\tau \right)^{2/(3m+1)} \\
&\le C t^{-2/[(m+1)(3m+1)]} \rho^{2/(1-m)} \left(\int_{\frac{t}{4}}^{t} \int_{B_{2\rho}(x_0)} y^{m+1} dx d\tau \right)^{2/(3m+1)} \\
&\quad + C t^{4/[(m+1)(3m+1)]} \rho^{2/(1-m)} [\phi(t,\rho)]^{(3m-3)/(3m+1)} \\
&\quad \cdot \left(\int_{\frac{t}{4}}^{t} \int_{B_{2\rho}(x_0)} y^{m+1} dx d\tau \right)^{2/(3m+1)}.
\end{aligned}$$

Denote

$$H^1 = t^{-2/[(m+1)(3m+1)]}\rho^{2/(1-m)}\left(\int_{\frac{t}{4}}^{t}\int_{B_{2\rho}(x_0)} y^{m+1}dxd\tau\right)^{2/(3m+1)},$$

$$H^2 = t^{4/[(m+1)(3m+1)]}\rho^{2/(1-m)}[\phi(t,\rho)]^{(3m-3)/(3m+1)}\left(\int_{\frac{t}{4}}^{t}\int_{B_{2\rho}(x_0)} y^{m+1}dxd\tau\right)^{2/(3m+1)}.$$

By the definition of $\phi(t,\rho)$ and $\psi(t,\rho)$, we have

$$\begin{aligned} H^1 &= \left[\int_{\frac{t}{4}}^{t} t^{-1/(m+1)}\int_{B_{2\rho(x_0)}} \rho^{(3m+1)/(1-m)}y^{m+1}dxd\tau\right]^{2/(3m+1)} \\ &\le C\left[t^{-1}\int_{\frac{t}{4}}^{t}\tau^{\frac{m}{m+1}}\frac{\|y(\cdot,\tau)\|^m_{L^\infty(B_{2\rho}(x_0))}}{\rho^{2m/(m-1)}}\rho^{\frac{1+m}{1-m}}\int_{B_{2\rho}(x_0)} ydxd\tau\right]^{2/(3m+1)} \\ &\le C[\phi(t,\rho)]^{2m/(3m+1)}[\psi(t,\rho)]^{2/(3m+1)} \\ &\le \frac{1}{4}\phi(t,\rho) + C[\psi(t,\rho)]^{2/(m+1)} \end{aligned} \tag{4.13}$$

and

$$\begin{aligned} H^2 &= [\phi(t,\rho)]^{(3m-3)/(3m+1)}\left(\int_{\frac{t}{4}}^{t}\int_{B_{2\rho}(x_0)} t^{\frac{2}{m+1}}\rho^{\frac{3m+1}{1-m}}y^{m+1}dxd\tau\right)^{2/(3m+1)} \\ &\le C[\phi(t,\rho)]^{(3m-3)/(3m+1)}\left[\int_{\frac{t}{4}}^{t}\tau^{(1-m)/(1+m)}\frac{1}{\rho}\right. \\ &\quad \left.\cdot\int_{B_{2\rho(x_0)}}\tau\frac{\|y(\cdot,\tau)\|^{m+1}_{L^\infty(B_{2\rho}(x_0))}}{\rho^{2(m+1)/(m-1)}}dxd\tau\right]^{2/(3m+1)} \\ &\le C[\phi(t,\rho)]^{(3m-3)/(3m+1)}\left[\int_{\frac{t}{4}}^{t}\tau^{(1-m)/(1+m)}\Phi^{m+1}(\tau,2\rho)d\tau\right]^{2/(3m+1)} \\ &\le C[\phi(t,\rho)]^{(3m-1)/(3m+1)}\left[\int_{\frac{t}{4}}^{t}\tau^{(1-m)/(1+m)}\phi^m(\tau,\rho)d\tau\right]^{2/(3m+1)} \\ &\le \frac{1}{4}\phi(t,\rho) + C\int_0^t\tau^{(1-m)/(1+m)}\phi^m(\tau,\rho)d\tau. \end{aligned} \tag{4.14}$$

Now the conclusion of Lemma 4.3 follows from (4.12), (4.13), and (4.14). □

LEMMA 4.4. *$\psi(t,\rho)$ and $\phi(t,\rho)$ satisfy*

$$\begin{aligned} \psi(t,\rho) \le C\Bigg\{&\int_0^t\tau^{-m/(1+m)}[\phi(\tau,\rho)]^{(m-1)/2}\psi(\tau,\rho)d\tau \\ &+\int_0^t\tau^{(2-m)/(1+m)}[\phi(\tau,\rho)]^{(3m-3)/2}\psi(\tau,\rho)d\tau\Bigg\} \\ &+|||y_0|||_\rho, \end{aligned} \tag{4.15}$$

where $|||y_0|||_\rho = \sup_{\rho\le r\le 2R} r^{(1+m)/(1-m)}\int_{B_r(x_0)} y_0(x)dx$.

Proof. We choose a smooth function ξ defined on $B_{2\rho}(x_0)$ such that

$$0 \le \xi \le 1, \quad \xi = 1 \text{ in } B_\rho(x_0), \quad |\xi_x| \le \frac{1}{\rho}.$$

Multiplying the first equation of (4.1) by $\tau^{1/2} y^{(m-1)/2}\xi^2$ and integrating on $B_{2\rho}(x_0)\times(0,t)$, we get

$$\begin{aligned}
&\frac{2}{m+1}\int_{B_{2\rho}(x_0)} t^{1/2}[y(x,t)]^{(m+1)/2}\xi^2(x)dx \\
&\quad + \frac{m(m-1)}{2}\int_0^t\int_{B_{2\rho}(x_0)} \tau^{1/2} y^{(3m-5)/2} y_x^2\xi^2 dxd\tau \\
&= \frac{1}{m+1}\int_0^t\int_{B_{2\rho}(x_0)} \tau^{-1/2} y^{(m+1)/2}\xi^2 dxd\tau - 2m\int_0^t\int_{B_{2\rho}(x_0)} \tau^{1/2} y^{(3m-3)/2} y_x\xi\xi_x dxd\tau \\
&\le \frac{1}{m+1}\int_0^t\int_{B_{2\rho}(x_0)} \tau^{-1/2} y^{(m+1)/2}\xi^2 dxd\tau + C\int_0^t\int_{B_{2\rho}(x_0)} \tau^{1/2} y^{(3m-1)/2}\xi_x^2 dxd\tau \\
&\quad + \frac{m(m-1)}{4}\int_0^t\int_{B_{2\rho}(x_0)} \tau^{1/2} y^{(3m-5)/2} y_x^2\xi^2 dxd\tau.
\end{aligned}$$

Thus we have

$$\begin{aligned}
\int_0^t\int_{B_{2\rho}(x_0)} \tau^{1/2} y^{(3m-5)/2} y_x^2\xi^2 dxd\tau &\le C\int_0^t\int_{B_{2\rho}(x_0)} \tau^{-1/2} y^{(m+1)/2} dxd\tau \\
&\quad + \frac{C}{\rho^2}\int_0^t\int_{B_{2\rho}(x_0)} \tau^{1/2} y^{(3m-1)/2} dxd\tau.
\end{aligned} \tag{4.16}$$

Denote

$$L(t) = \frac{1}{\rho^2}\int_0^t\int_{B_{2\rho}(x_0)} \tau^{1/2} y^{(3m-1)/2} dxd\tau,$$

$$J_2(t) = \int_0^t\int_{B_{2\rho}(x_0)} \tau^{-1/2} y^{(m+1)/2} dxd\tau.$$

Then we can derive

$$\begin{aligned}
J_2(t) &\le \int_0^t \tau^{-1/2}\rho^{\frac{1+m}{m-1}} \|y(\cdot,\tau)\|^{(m-1)/2}_{L^\infty(B_{2\rho}(x_0))}\rho^{\frac{1+m}{1-m}}\int_{B_{2\rho}(x_0)} y(x,\tau)dxd\tau \\
&\le C\int_0^t \tau^{\frac{-m}{m+1}}\rho^{\frac{2m}{m-1}}\left[\tau^{1/(m+1)}\frac{\|y(\cdot,\tau)\|_{L^\infty(B_{2\rho}(x_0))}}{\rho^{2/(m-1)}}\right]^{\frac{m-1}{2}}\psi(\tau,\rho)d\tau \\
&\le C\rho^{2m/(m-1)}\int_0^t \tau^{-m/(m+1)}[\phi(\tau,\rho)]^{(m-1)/2}\psi(\tau,\rho)d\tau
\end{aligned} \tag{4.17}$$

and

$$\begin{aligned}
L(t) &\le \frac{1}{\rho^2}\int_0^t \tau^{1/2}\rho^{\frac{1+m}{m-1}}\|y(\cdot,\tau)\|^{\frac{3m-3}{2}}_{L^\infty(B_{2\rho}(x_0))}\int_{B_{2\rho}(x_0)}\rho^{\frac{1+m}{1-m}} y(x,\tau)dxd\tau \\
&\le \int_0^t \rho^{\frac{2m}{m-1}}\tau^{\frac{2-m}{m+1}}\left[\tau^{1/(m+1)}\frac{\|y(\cdot,\tau)\|_{L^\infty(B_{2\rho}(x_0))}}{\rho^{2/(m-1)}}\right]^{(3m-3)/2}\psi(\tau,\rho)d\tau \\
&\le C\rho^{2m/(m-1)}\int_0^t \tau^{(2-m)/(m+1)}[\phi(\tau,\rho)]^{(3m-3)/2}\psi(\tau,\rho)d\tau.
\end{aligned} \tag{4.18}$$

Substituting (4.17) and (4.18) into (4.16), we get

$$\begin{aligned}&\int_0^t\int_{B_{2\rho}(x_0)}\tau^{1/2}y^{(3m-5)/2}y_x^2\xi^2dxd\tau\\ &\le C\rho^{2m/(m-1)}\int_0^t\tau^{-m/(m+1)}[\phi(\tau,\rho)]^{(m-1)/2}\psi(\tau,\rho)d\tau\\ &\quad+C\rho^{2m/(m-1)}\int_0^t\tau^{(2-m)/(m+1)}[\phi(\tau,\rho)]^{(3m-3)/2}\psi(\tau,\rho)d\tau.\end{aligned}\tag{4.19}$$

Multiplying the first equation of (4.1) by ξ^2 and integrating on $B_{2\rho}(x_0)\times(0,t)$, we have

$$\begin{aligned}\int_{B_\rho(x_0)}y(x,t)dx&\le\int_{B_{2\rho}(x_0)}y_0(x)dx+2m\int_0^t\int_{B_{2\rho}(x_0)}y^{m-1}y_x\xi\xi_xdxd\tau\\ &\le\int_{B_{2\rho}(x_0)}y_0(x)dx\\ &\quad+\frac{C}{\rho}\left(\int_0^t\int_{B_{2\rho}(x_0)}\tau^{1/2}y^{\frac{3m-5}{2}}y_x^2\xi^2dxd\tau\right)^{1/2}\\ &\quad\cdot\left(\int_0^t\int_{B_{2\rho}(x_0)}\tau^{-1/2}y^{\frac{m+1}{2}}dxd\tau\right)^{1/2}.\end{aligned}\tag{4.20}$$

Taking into account (4.17) and (4.19), in (4.20), we get

$$\begin{aligned}\int_{B_\rho(x_0)}y(x,t)dx&\le\int_{B_{2\rho}(x_0)}y_0(x)dx\\ &\quad+C\rho^{(1+m)/(m-1)}\int_0^t\tau^{-m/(m+1)}[\phi(\tau,\rho)]^{(m-1)/2}\psi(\tau,\rho)d\tau\\ &\quad+C\rho^{(1+m)/(m-1)}\int_0^t\tau^{(2-m)/(m+1)}[\phi(\tau,\rho)]^{(3m-3)/2}\psi(\tau,\rho)d\tau.\end{aligned}$$

Multiplying both sides of the above inequality by $\rho^{(1+m)/(1-m)}$, we get (4.15). ∎

LEMMA 4.5. *There exist constants γ_1, γ_2, γ_3, and $t^*>0$ such that*

$$\phi(t,\rho)\le\gamma_1|||y_0|||_\rho^{2/(m+1)},\quad \psi(t,\rho)\le\gamma_2|||y_0|||_\rho,\tag{4.21}$$

where $t<\min\{t^*,T,\gamma_3|||y_0|||_\rho^{1-m}\}$.

Proof. Since $\psi(\cdot,\rho)$ is increasing, by (4.11), it follows that for any fixed $t^*>0$,

$$\phi(t,\rho)\le C_1\int_0^t\tau^{(1-m)/(1+m)}\phi^m(\tau,\rho)d\tau+C_2[\psi(t^*,\rho)]^{2/(m+1)},\ t<t^*.$$

Hence $\phi(t,\rho)$ is less than or equal to the solution of the following problem:

$$H'(t)=C_1t^{(1-m)/(1+m)}H^m(t),$$

$$H(0)=C_2[\psi(t^*,\rho)]^{2/(m+1)}.$$

This implies

$$\phi(t,\rho) \le H(t) = C_2[\psi(t^*,\rho)]^{\frac{2}{m+1}} \left[1 - \frac{C_1(1+m)(m-1)}{2C_2^{1-m}}(t\psi^{m-1}(t^*,\rho))^{\frac{2}{1+m}}\right]^{1/(1-m)},$$

provided that the value in the bracket is positive.

If we take t^* such that

$$1 - \frac{C_1(1+m)(m-1)}{2C_2^{1-m}}(t^*\psi^{m-1}(t^*,\rho))^{2/(1+m)} > 0,$$

then

$$\phi(t,\rho) \le C_2[\psi(t,\rho)]^{\frac{2}{m+1}} \left[1 - \frac{C_1(1+m)(m-1)}{2C_2^{1-m}}(t\psi^{m-1}(t,\rho))^{\frac{2}{1+m}}\right]^{1/(1-m)}$$

for $t < t^*$.

Thus there exist positive constants δ_1 and δ_2 such that

$$(t\psi^{m-1}(t,\rho))^{2/(1+m)} \le \delta_1, \tag{4.22}$$

$$\phi(t,\rho) \le \delta_2[\psi(t,\rho)]^{2/(1+m)} \tag{4.23}$$

for $t < t^*$.

Substituting (4.22) and (4.23) into (4.15), we obtain

$$\begin{aligned}
\psi(t,\rho) &\le C\int_0^t \tau^{-m/(1+m)}[\psi(\tau,\rho)]^{2m/(1+m)}d\tau \\
&\quad + C\int_0^t \tau^{(2-m)/(1+m)}[\psi(\tau,\rho)]^{(4m-2)/(1+m)}d\tau + |||y_0|||_\rho \\
&\le C\int_0^t \tau^{-m/(1+m)}[\psi(\tau,\rho)]^{2m/(1+m)}\left(1 + \tau^{\frac{2}{1+m}}[\psi(\tau,\rho)]^{2(m-1)/(1+m)}\right)d\tau \\
&\quad + |||y_0|||_\rho \\
&\le C\int_0^t \tau^{-m/(1+m)}[\psi(\tau,\rho)]^{2m/(1+m)}d\tau + |||y_0|||_\rho
\end{aligned}$$

for $t < t^*$.

Hence $\psi(t,\rho)$ is less than or equal to the solution of the following problem:

$$M'(t) = Ct^{-m/(1+m)}[M(t)]^{2m/(1+m)},$$

$$M(0) = |||y_0|||_\rho.$$

This implies

$$\psi(t,\rho) \le M(t) = |||y_0|||_\rho \left[1 - \frac{C(m-1)t^{1/(1+m)}}{|||y_0|||_\rho^{(1-m)/(1+m)}}\right]^{(1+m)/(1-m)},$$

provided that

$$1 - \frac{C(m-1)t^{1/(1+m)}}{|||y_0|||_\rho^{(1-m)/(1+m)}} > 0.$$

Thus there exist two constants γ_2 and γ_3 such that if $t < \min\left\{t^*, \gamma_3|||y_0|||_\rho^{1-m}\right\}$, then

$$\psi(t,\rho) \le \gamma_2|||y_0|||_\rho.$$

Furthermore, by (4.23), there exists a constant $\gamma_1 > 0$ such that

$$\phi(t,\rho) \le \gamma_1|||y_0|||_\rho^{2/(1+m)}.$$

Thus we get the conclusion of Lemma 4.5. □

The above inequality implies that

$$\|y(\cdot,t)\|_{L^\infty(B_\rho(x_0))} \le \gamma_1 t^{-1/(1+m)}\rho^{2/(m-1)}|||y_0|||_\rho^{2/(1+m)} \tag{4.24}$$

for $t < \min\left\{t^*, T, \gamma_3|||y_0|||_\rho^{1-m}\right\} := T^*$.

By (4.24), for any time $T < T^*$ and initial value $y_0 \in L^\infty(-1,1)$ with $y_0 \ge 0$ a.e. in $(-1,1)$, we can always find a target $y_1 \in L^2_+(-1,1) + Y_d$ satisfying $\|y_1\|_{L^\infty(B_\rho(x_0))} > \gamma_1 T^{-1/(1+m)}\rho^{2/(m-1)}|||y_0|||_\rho^{2/(1+m)} + 1$ such that for any $u \in L^2_+(\omega \times (0,T))$, the corresponding solution y of system (4.1) satisfies

$$\|y(\cdot,T;u) - y_1\|_{L^2(-1,1)} > \frac{1}{2},$$

which shows that y_1 is not approximately controllable at time T. The above result implies the conclusion of Theorem 1.8.

Remark 4.6. Using the same method as that in [22], we can prove that the solution of system (4.1) has the property of finite speed of propagation for each nonnegative control. However, by this result, we cannot explain why system (1.1) is not approximately controllable at small time, unlike the case of a hyperbolic system. Indeed, the method used in [22] determines that the speed of propagation of solution depends on the control function closely. Here we use a technique of local estimates to prove an obstruction phenomenon, which leads to a negative controllability result. In fact, inequality (4.24) implies the finite speed of propagation of solution for system (4.1).

Acknowledgment. The authors thank all reviewers for their constructive comments.

REFERENCES

[1] S. N. Antontsev and S. I. Shmarev, *A model porous medium equation with variable exponent of nonlinearity: Existence, uniqueness and localization properties of solutions*, Nonlinear Anal., 60 (2005), pp. 515–545.

[2] V. Barbu, *Controllability of parabolic and Navier-Stokes equations*, Sci. Math. Jpn., 56 (2002), pp. 143–211.

[3] V. Barbu and Th. Precupanu, *Convexity and Optimization in Banach Spaces*, D. Reidel, Dordrecht, The Netherlands, 1986.

[4] Ph. Benilan, *Equations d'evolution dans un espace de Banach quelconque et applications*, Thése, Orsay, 1972.

[5] P. Cannarsa, P. Martinez, and J. Vancostenoble, *Persistent regional null controllability for a class of degenerate parabolic equations*, Commun. Pure Appl. Anal., 3 (2004), pp. 607–635.

[6] P. Cannarsa, P. Martinez, and J. Vancostenoble, *Null controllability of degenerate heat equations*, Adv. Differential Equations, 10 (2005), pp. 153–190.

[7] J. I. Diaz, J. Henry, and A. M. Ramos, *On the approximate controllability of some semilinear parabolic boundary value problems*, Appl. Math. Optim., 37 (1998), pp. 71–97.

[8] J. I. Diaz and A. M. Ramos, *Some results about the approximate controllability property for quasilinear diffusion equations*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 1243–1248.

[9] J. I. Diaz and I. I. Vrabie, *Existence for reaction diffusion systems: A compactness method approach*, J. Math. Anal. Appl., 188 (1994), pp. 521–540.

[10] A. Doubova, E. Fernández-Cara, M. González-Burgos, and E. Zuazua, *On the controllability of parabolic systems with a nonlinear term involving the state and the gradient*, SIAM J. Control Optim., 41 (2002), pp. 798–819.

[11] C. Fabre, J. P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[12] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.

[13] A. V. Fursikov and O. Y. Imanvilov, *Controllability of Evolution Equations*, Lecture Notes Series 34, Seoul National University, Seoul, Korea, 1996.

[14] H. Gao, X. Hou, and N. Pavel, *Optimal control and controllability problems for a class of nonlinear degenerate diffusion equations*, Panamer. Math. J., 13 (2003), pp. 103–126.

[15] H. Gao, P. Lei, and B. Zhang, *A class of nonlinear degenerate integrodifferential control systems*, SIAM J. Control Optim., 43 (2004), pp. 986–1010.

[16] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1967 (in Russian).

[17] A. Lopéz, X. Zhang, and E. Zuazua, *Null controllability of the heat equation as singular limit of the exact controllability of dissipative wave equations*, J. Math. Pures Appl. (9), 79 (2000), pp. 741–808.

[18] K. Phung, G. Wang, and X. Zhang, *On the existence of time optimal control of some linear evolution equations*, Discrete Contin. Dyn. Syst. Ser. B, to appear.

[19] S. Tang and X. Zhang, *Carleman inequality for backward stochastic parabolic equations with general coefficients*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 775–780.

[20] G. Wang and L. Wang, *Local internal controllability of the Boussinesq system*, Nonlinear Anal., 53 (2003), pp. 637–652.

[21] Z. Wu, J. Zhao, J. Yin, and H. Li, *Nonlinear Diffusion Equations*, World Scientific, River Edge, NJ, 2001.

[22] J. Yin and H. Gao, *Behavior of solutions to a degenerate diffusion problem*, Acta Math. Appl. Sinica, 13 (1997), pp. 188–195.

[23] H. Zhou, *The controllability problem for a parabolic system with nonnegative boundary control*, Kexue Tongbao, 42 (1997), pp. 246–249 (in Chinese).

SIAM J. CONTROL OPTIM.
Vol. 46, No. 6, pp. 2280–2302

ROBUST CONTROL APPROACH TO OPTION PRICING: A REPRESENTATION THEOREM AND FAST ALGORITHM*

PIERRE BERNHARD†, NAÏMA EL FAROUQ‡, AND STÉPHANE THIERY†

Abstract. The so-called interval model for security prices, together with a robust control approach, allows one to construct a consistent theory of option pricing, including discrete time trading and arbitrary transaction costs. In this context, a new representation theorem for the viscosity solution of the relevant Isaacs (differential) quasi-variational inequality leads to simple formulas and fast numerical algorithms to compute a hedging portfolio. We argue that in spite of a less satisfactory market model, the overall theory is not much less realistic than the classical Black and Scholes theory but rather only that it shifts from the portfolio model to the market model the place where the model is violated when sudden large price changes occur on the market. As such, and subject to a more detailed validation, the new theory might be the basis of a possible alternative as a normative theory whenever transaction costs or discrete time trading are the main concerns.

Key words. option pricing, hedging, transaction costs, robust control, impulse control

AMS subject classifications. 49L25, 49N25, 49N70, 91B28

DOI. 10.1137/050626016

1. Introduction.

1.1. The robust control approach to option pricing. In [7, 8, 11], we introduced a robust control approach to option pricing, and more specifically to the design of a hedging portfolio and management strategy, using the so-called interval model for the market and a robust control approach to hedging.

The main claims of that new approach are, on the one hand, the possibility of constructing a consistent theory of hedging portfolios with either continuous or discrete time trading paradigms, the former being the limit of the latter for vanishing time steps, with one and the same (continuous time) market model, and, on the other hand, to accommodate transaction costs and closing costs in a natural way, with a nontrivial hedging portfolio.

We postpone until the last section the discussion of the drawbacks of the "interval model" as compared to the classical Samuelson geometric diffusion. But we dispel at once one criticism, that it does not make use of probabilistic knowledge on the price trajectories. Indeed, it is now known that Black and Scholes formula can be recovered via a purely deterministic scheme. Cox, Ross, and Rubinstein [13] and Kolokol'tsov [16] derive it in passing to the limit as the step size vanishes in a discrete time model. In their model, though, the underlying market model changes with the step size, in a way so as to generate a random walk in the limit. In [7] we obtain that same formula directly with a continuous time model of price evolution and continuous trading but still without endowing the set of possible price histories with a probability law.

Here, after summarizing some previous results, we show a new representation of the solution of the problem at hand—and thus of the pricing function—in terms of

*Received by the editors March 4, 2005; accepted for publication (in revised form) July 30, 2007; published electronically December 21, 2007.
http://www.siam.org/journals/sicon/46-6/62601.html

†I3S, University of Nice-Sophia Antipolis and CNRS, P.O. Box 145, 06903 Sophia Antipolis Cedex, France (Pierre.Bernhard@essi.fr, Thiery@i3s.unice.fr).

‡University Blaise Pascal, B.P. 185, 63006 Clermont Ferrand Cedex, France (ElFarouq@i3s.unice.fr).

the solution of a pair of simple coupled first order linear PDEs in two variables. This yields a fast algorithm to compute both the seller's price and the hedging strategy, thus alleviating the computational complexity that could heretofore be considered a drawback of that approach. Also, we provide a more detailed comparison of the new pricing paradigm with the classical Black and Scholes formula than in the previous papers on that topic and some numerical evaluations of the robustness of the new theory to violations of its market model.

At this point, we wish to stress two things: on the one hand, we do not claim a kind of overall superiority of the new theory over that of Black and Scholes. We claim only that this is a possible approach that may, concerning certain problems, be interesting to use in comparison with the more classical ones.

On the other hand, if we can convince the reader of this fact, he will accept that, concerning a new theory, it is not as yet as well documented as the long-standing Black and Scholes theory, nor as well understood in all its implications.

Among other shortcomings at this time, we do not yet have detailed and comparative validation data. And, in contrast with [16, 21, 22] and most of the literature on superreplication, we have investigated only the *seller's price* of our model, i.e., the price that the seller must charge to make sure to always hedge his costs, that is, as long as the market does not violate the hypotheses of the model.

1.2. Related contributions. Among previous uses of this type of model, let us quote the following.

McEneaney [18] may have been the first to replace the stochastic framework with a robust control approach. He derives the so-called stop loss strategy for bounded variation trajectories. He also recovers the Black and Scholes theory, but this is done at the price of artificially modifying the portfolio model with no other justification than recovering the Itô calculus and the Black and Scholes PDE.

As already noted, Cox, Ross, and Rubinstein [13] introduced a nonstochastic approach to the theory of option pricing in a discrete time setting. Their market model is related to ours in that where we allow for an interval of possible future stock prices, they allow only the end points of such an interval. This model clearly involves no claim of being realistic for any finite time step. Its only objective is to converge, as the time step vanishes, to a continuous random walk, to recover either the Black and Scholes theory or another one with possible price jumps, depending on how the market model behaves in that limiting process.

Kolokol'tsov [16, 17] generalizes the approach of Cox, Ross, and Rubinstein by allowing the same "interval model" as we use here. He notices, as we do in [6], the coincidence with Cox, Ross, and Rubinstein's theory when the final payment is convex. He expands that theory to "rainbow" options, based on several common stocks, and performs a limiting operation similar to that of Cox, Ross, and Rubinstein to derive an equivalent of the Black and Scholes formula for these options. In the case of the interval model, the author points out that the strategy proposed leads to superreplication, so that the trader might have a positive result. This leads him to define a range of "reasonable" or "fair" prices and a mean price. (In this setting, we consider here only his upper bound, i.e., a seller's price.)

In [7], we recover both the stop-loss strategy and the complete Black and Scholes theory without any probability in the model, without having to artificially modify the portfolio model, simply by choosing carefully the set of admissible underlying stock price trajectories and using a weak version of a lemma of Föllmer [15]. The model used there is *not* an interval model. To the contrary, it uses a set of trajectories

in which the trajectories of Samuelson's model almost surely lie. This explains why we recover the Black and Scholes theory while in the present setting we arrive at a different set of formulas.

A special quotation must be made for the amazing book by Shafer and Vovk [24]. They start from the same analysis of hedging as we do in terms of a game against "nature." But where we conclude that we can do without probability theory, they claim that this *is* probability theory. Or rather they claim that this can be an alternative to Kolmogorov's measure-theoretical foundation of probability theory, and they proceed to recover many results, such as asymptotic and ergodic theorems, from this new viewpoint. Our seller's price is their "upper expectation" (of which they claim that it is the price likely to be found on the market). Yet, they are more interested in recovering classical probabilities and elaborating on the classical Black and Scholes theory than in providing alternative models and tackling the problems of transaction costs and discrete trading that we consider. We did not find hints to something like the interval model in this fascinating book, but the relationship of that theory with our model deserves further thinking.

Pujal [20] and Aubin, Pujal, and Saint-Pierre [2] have also adopted the robust control approach (they call it "tychastic" approach), with a market model which is a more general version of our model, since it applies a priori to rainbow options as well. Yet, they do not allow jumps in the content of the portfolio, limiting its rate of change. Saint-Pierre [23] has done efficient implementations of that theory with exactly the interval market model that we use below. He has a fast algorithm which bears a strong resemblance to ours, although precisely asserting their relationship is made difficult by the fact that since their portfolio model is more restrictive than ours, the value function cannot coincide. And our fast algorithm is based upon a representation theorem for our value function.

Similar thoughts are developed by Olsder [19], although this is only a preliminary analysis as stated by the author. And very similar ideas have been developed by Dupire [14] in the context of what he calls "dominance" theory. Barles and Soner [3] developed a different approach to option pricing that let them deal with transaction costs, but then, the price they arrive at depends on the rest of the trader's portfolio.

We took the phrase "interval model" from Roorda, Engwerda, and Schumacher [21, 22], where the authors adopt a viewpoint close to that of robust control. In particular their definition of a market model is the same as ours: a set of possible price trajectories. Their analysis is somewhat different from ours, as they do endow the set of possible trajectories with a probability law. Yet, because they have an interval model, they also run into a superreplication problem and, in a fashion similar to Kolokol'tsov, define a range of "fair" prices.

1.3. Paper outline. In the next section, we present the interval model, both in the continuous trading formalism and in its discretized form, and the portfolio model we adopt, which includes transaction costs and closing costs at will. The consideration of closing costs obliges us to distinguish the cases where the closing is "in kind" or "in cash," because the closing costs born by the trader are not the same.

Section 3 is devoted to the continuous trading problem. We recall the main results we have obtained so far, stressing the case of simple call and put. Next we show a new representation theorem of the solution of the pricing problem. The complete proof of this theorem, and its main use, relies on results of the next section. We also investigate the optimal trading strategies, which have a simple form.

In section 4, we investigate the discrete trading theory. We provide a discrete

time version of the representation theorem and derive from it a new fast algorithm to compute the seller's price. And as we have a convergence theorem of the discrete trading seller's price toward the continuous trading one as the time step vanishes [11], this is also a discretization algorithm for the continuous problem.

Finally, having displayed what can be achieved with this new model, we discuss in the final section its relative strengths and weaknesses compared with the classical Black and Scholes theory. As the main weakness of the new theory is in the unrealistic assumption of the market model, this discussion is mostly based upon an investigation of the robustness of our hedging strategies to violations of that hypothesis. In that discussion, we take a normative view of our theory, i.e., as a decision aid, rather than the positive view of a predictive theory which is more common in economical thinking.

2. Interval model.

2.1. Riskless interest rate. We assume a fixed, known, riskless interest rate ρ characteristic of that economy. In a classical fashion, all monetary values will be assumed to be expressed in end-time value computed at that fixed riskless rate, so that, without loss of generality, the riskless rate can be taken as (seemingly) zero. (It reappears in the theory of American options, but we have not covered it here for lack of space.)

2.2. Market. We share with Roorda, Engwerda, and Schumacher [21, 22] the view that a market model is a set Ω of possible price trajectories. Our model is defined by two real numbers $\tau^- < 0$ and $\tau^+ > 0$, and Ω is the set of all absolutely continuous functions $u(\cdot)$ such that for any two time instants t_1 and t_2,

$$\mathrm{e}^{\tau^-(t_2-t_1)} \le \frac{u(t_2)}{u(t_1)} \le \mathrm{e}^{\tau^+(t_2-t_1)}\,. \tag{1}$$

The notation τ^ε will be used to handle both τ^+ and τ^- at a time. Hence, in that notation, it is understood that $\varepsilon \in \{-,+\}$, sometimes identified to $\{-1,+1\}$.

In the continuous trading theory, we shall use the equivalent characterization

$$\dot{u} = \tau u\,, \quad \tau \in [\tau^-, \tau^+]\,. \tag{2}$$

In that formulation, $\tau(\cdot)$ is a measurable function, which plays the role of the "control" of the market. We shall let Ψ denote the set of measurable functions from $[0,T]$ into $[\tau^-,\tau^+]$. It is equivalent to specify a $u(\cdot)\in\Omega$ or a $(u(0),\tau(\cdot))\in\mathbb{R}^+\times\Psi$. This is an a priori unknown time function. The concept of nonanticipative strategies embodies that fact.

In the discrete trading theory, we shall call h our time step with $T = \mathcal{K}h$, $\mathcal{K}$ being an integer. The hypothesis (1) translates into[1]

$$u(t+h) \in [\mathrm{e}^{\tau^- h}u(t), \mathrm{e}^{\tau^+ h}u(t)]\,.$$

For convenience, we let

$$u(t+h) = (1+\tau(t))u(t)\,, \quad \tau(t) \in [\tau_h^-, \tau_h^+] \tag{3}$$

with

$$\tau_h^\varepsilon = \mathrm{e}^{\tau^\varepsilon h} - 1\,, \quad \varepsilon = \pm\,. \tag{4}$$

[1]It does *not* translate into $u(t+h) \in \{\mathrm{e}^{\tau^- h}u(t), \mathrm{e}^{\tau^+ h}u(t)\}$ as in the Cox–Ross–Rubinstein theory.

Alternatively, we shall write, for any integer k, $u(kh) = u_k$, so that (3) also reads

$$u_{k+1} = (1 + \tau_k)u_k\,, \quad \tau_k \in [\tau_h^-, \tau_h^+], \tag{5}$$

and we let Ψ denote the set of such sequences $\{\tau_k\}$.

The case where h goes to zero will be of interest also. But, contrary to the classical limit process in the Cox–Ross–Rubinstein theory, we keep the underlying continuous time model, hence here τ^+ and τ^-, fixed. Then τ_h^ε behaves as $h\tau^\varepsilon$.

2.3. Portfolio. A (hedging) portfolio will be composed of an amount v (in end-time value) of underlying stock, and an amount y of riskless *bonds*, for a total worth of $w = v + y$. In the normalized (or end-value) representation, the bonds are seemingly with zero interest.

2.3.1. Buying and selling. We let $\xi(t)$ be the buying rate (a sale if $\xi(t) < 0$), which is the trader's control. Therefore we have, in continuous time,

$$\dot v = \tau v + \xi\,. \tag{6}$$

However, there is no reason to restrict the buying/selling rate, so that there is no bound on ξ. To avoid mathematical ill-posedness, we explicitly admit "infinite" buying/selling rate in the form of instantaneous block buy or sale of a finite amount of stock at time instants chosen by the trader together with the amount. Thus the control of the trader also involves the choice of finitely many time instants t_k and trading amounts ξ_k, and the model must be augmented with

$$v(t_k^+) = v(t_k) + \xi_k\,, \tag{7}$$

meaning that $v(\cdot)$ has a jump discontinuity of size ξ_k at time t_k. Equivalently, we may keep formula (6) but allow for impulses $\xi_k\delta(t - t_k)$ in $\xi(\cdot)$.

We shall therefore let $\xi(\cdot) \in \Xi$, the set of real time functions (or rather distributions) defined over $[0, T]$ which are the sum of a measurable function $\xi_c(\cdot)$ and a finite number of weighted translated Dirac impulses $\xi_k\delta(t - t_k)$.

2.3.2. Transaction costs. We assume that there are transaction costs. In this paper, we assume that they are proportional to the transaction amount. But we allow for different proportionality coefficients for a buy or a sale of underlying stock. Hence let C^+ be the cost coefficient for a buy, and $-C^-$ for a sale, so that the cost of a transaction of amount ξ is $C^\varepsilon\xi$ with $\varepsilon = \operatorname{sign}(\xi)$. We have chosen C^- negative, so that, as it should, that formula always gives a positive cost.

We shall use the convention that when we write $C^\varepsilon(\textit{expression})$, and except if otherwise specified, the symbol ε in C^ε stands for the sign of the *expression*.

Our portfolio will always be assumed *self-financed*; i.e., the sale of one of the commodities, underlying stock or riskless bonds, must exactly pay for the buy of the other one *and* the transaction costs. It is a simple matter to see that the worth w of the portfolio then obeys

$$\dot w = \tau v - C^\varepsilon\xi\,, \tag{8}$$

and at jump instants,

$$w(t_k^+) = w(t_k^-) - C^{\varepsilon_k}\xi_k. \tag{9}$$

This is equivalent to

$$w(t) = w(0) + \int_0^t (\tau(s)v(s) - C^\varepsilon\xi(s))\,\mathrm{d}s - \sum_{k|t_k<t} C^{\varepsilon_k}\xi_k\,. \tag{10}$$

2.3.3. Discrete trading. The discrete trading case can be seen as a sequence of jumps at prescribed time instants $t_k = kh$, $k \in \mathbb{K} = \{0, 1, \dots \mathcal{K}\text{–}1\} \subset \mathbb{N}$, and $h \in \mathbb{R}^+$ a prescribed time step, such that $\mathcal{K}h = T$. Writing u_k, v_k, w_k for $u(kh), v(kh), w(kh)$, it leads to

$$v_{k+1} = (1 + \tau_k)(v_k + \xi_k)\,, \tag{11}$$

$$w_{k+1} = w_k + \tau_k(v_k + \xi_k) - C^{\varepsilon_k}\xi_k\,. \tag{12}$$

We shall use the explicit form

$$w_n = w_0 + \sum_{k=0}^{n-1}(\tau_k(v_k + \xi_k) - C^{\varepsilon_k}\xi_k)\,. \tag{13}$$

A *dynamic portfolio* will be a pair of time functions $(v(\cdot), w(\cdot))$, whether time is continuous or discrete, also denoted $(\{v_k\}, \{w_k\})_{k\in\mathbb{K}}$ in the latter case.

2.4. Hedging.

2.4.1. Strategies. The initial portfolio is to be created at step 0. As a consequence the seller's price is obtained taking $v(0) = 0$. Then, formally, admissible hedging strategies will be functions $\varphi : \Omega \to \Xi$ which enjoy the property of being nonanticipative:

$$\forall (u_1(\cdot), u_2(\cdot)) \in \Omega \times \Omega, \quad [u_1|_{(0,t)} = u_2|_{(0,t)}] \Rightarrow [\varphi(u_1(\cdot))|_{[0,t]} = \varphi(u_2(\cdot))|_{[0,t]}].$$

(It is understood here that the restriction of $\delta(t-t_k)$ to a closed interval not containing t_k is 0, and its restriction to a closed interval containing t_k is an impulse.)

In practice, we shall find optimal hedging strategies made of a jump at initial time, followed by a state feedback law $\xi(t) = \phi(t, u(t), v(t))$.

In discrete time, the situation is much simpler. We need only a nonanticipative strategy $\varphi : \Omega \to \mathbb{R}^T$ giving $\xi_k = \varphi_k(u_0, u_1, \dots, u_k)$. Again, we shall find it in the form of a state feedback $\xi_k = \phi_k(u_k, v_k)$.

Yet, these are only nonanticipative laws, the equivalent of stochastic adapted strategies. We have shown in [11] how to handle *strictly* nonanticipative strategies, the equivalent of the stochastic predictable strategies.

We shall call Φ the set of admissible trading strategies.

2.4.2. Closing costs. The idea of a hedging portfolio is that at exercise time, the writer is going to close off its position after abiding by its contract, buying or selling some of the underlying stock according to the necessity. We assume that it sustains proportional costs on this final transaction. We allow for the case where these costs would be different from the running transaction costs because compensation effects might lower them and also allow for the case without closing costs just by making their rate 0. Therefore let $c^+ \le C^+$ and $-c^- \le -C^-$ be these rates.

It is a simple matter to see that, in order to cover both cases where the buyer does or does not exercise its option, the portfolio worth at final time should be $N(u, v)$, given for a call and a closure in kind by

$$N(u, v) = \max\{c^\varepsilon(-v)\,,\ u - K + c^\varepsilon(u - v)\}\,,$$

where the notation convention for $c^\varepsilon(expression)$ holds. We expect that on a typical optimum hedging portfolio for a call, $0 \le v(T) \le u(T)$. Hence

$$N(u, v) = \max\{-c^- v\,,\ u - K + c^+(u - v)\}\ . \tag{14}$$

In the case of a put, where $-u(T) \le v(T) \le 0$, we need to replace the above expression by

$$N(u,v) = \max\{-c^+ v\,,\ K - u - c^-(u+v)\}. \tag{15}$$

The case of a closure in cash is similar but leads to less appealing mathematical formulas in later developments. The details can be found in [10].

2.4.3. Hedging portfolio. An initial portfolio $(v(0), w(0))$ and an admissible trading strategy φ, together with a price history $u(\cdot)$, generate a dynamic portfolio. We set the following.

DEFINITION 2.1. *An initial portfolio* $(v(0) = 0, w(0) = w_0)$ *and a trading strategy* φ *constitute a* hedge *at* u_0 *if for any* $u(\cdot) \in \Omega$ *such that* $u(0) = u_0$ *(equivalently, for any admissible* $\tau(\cdot)$*), the dynamic portfolio thus generated satisfies*

$$w(T) \ge N(u(T), v(T))\,. \tag{16}$$

Now, we may use (10) at time T to rewrite this:

$$\forall \tau(\cdot) \in \Psi\,, \quad N(u(T), v(T)) + \int_0^T \Big(-\tau(t)v(t) + C^\varepsilon \xi(t)\Big)\,\mathrm{d}t + \sum_k C^{\varepsilon_k}\xi_k - w_0 \le 0\,.$$

This in turn is clearly equivalent to

$$w_0 \ge \sup_{\tau(\cdot)\in\Psi} \left[N(u(T), v(T)) + \int_0^T \Big(-\tau(t)v(t) + C^\varepsilon \xi(t)\Big)\,\mathrm{d}t + \sum_k C^{\varepsilon_k}\xi_k \right].$$

We further set the following.

DEFINITION 2.2. *The* seller's price *of the option at* u_0 *is the worth of the cheapest hedging portfolio at* u_0*.*

The seller's price at u_0 is therefore

$$P(u_0) = \inf_{\varphi\in\Phi} \sup_{\tau(\cdot)\in\Psi} \left[N(u(T), v(T)) + \int_0^T \Big(-\tau(t)v(t) + C^\varepsilon \xi(t)\Big)\,\mathrm{d}t + \sum_k C^{\varepsilon_k}\xi_k \right], \tag{17}$$

where it is understood that $v(0) = 0$ and that $\xi(\cdot) = \varphi(u_0, \tau(\cdot))$.

In the case of discrete trading, we get similarly as the seller's price at u_0

$$P(u_0) = \min_{\varphi\in\Phi} \sup_{\{\tau_k\}\in\Psi} \left[N(u_K, v_K) + \sum_{k=0}^{K-1} \Big(-\tau_k(v_k + \xi_k) + C^{\varepsilon_k}\xi_k\Big) \right]. \tag{18}$$

3. Continuous trading.

3.1. The differential game. We are therefore led to the investigation of the impulse control differential game whose dynamics are given by (2), (6), and (7) and the criterion by (17). In a classical fashion we introduce its Isaacs value function:

$$W(t,u,v) = \inf_{\varphi\in\Phi} \sup_{\tau(\cdot)\in\Psi} \left[N(u(T), v(T)) + \int_t^T \Big(-\tau(s)v(s) + C^\varepsilon \xi(s)\Big)\,\mathrm{d}s + \sum_{k|t_k\ge t} C^{\varepsilon_k}\xi_k \right], \tag{19}$$

where the dynamics are integrated from $u(t) = u$, $v(t) = v$. Hence the seller's price is $P(u_0) = W(0, u_0, 0)$.

There are new features in that game, in that, on the one hand, impulse controls are allowed, and hence an Isaacs quasi-variational inequality (or QVI; see Bensoussan and Lions [4]) should be at work, but, on the other hand, impulse costs have a zero infimum. As a consequence, that QVI is degenerate, and no general result is available. In Bernhard, El Farouq, and Thiery [11], we introduce the so-called Joshua transformation that lets us show the following fact.

THEOREM 3.1. *The function W defined by* (19) *is a continuous viscosity solution of the following "differential QVI":*

$$
\begin{aligned}
0 = \min\Bigg\{ & \frac{\partial W}{\partial t} + \max_{\tau\in[\tau^-,\tau^+]} \tau \left[\frac{\partial W}{\partial u} u + \left(\frac{\partial W}{\partial v} - 1 \right) v \right], \\
& \frac{\partial W}{\partial v} + C^+, \quad -\left(\frac{\partial W}{\partial v} + C^- \right) \Bigg\}, \\
W(T, u, v) = {} & N(u, v)\,.
\end{aligned}
\tag{20}
$$

This PDE in turn lends itself to an analysis, either along the lines of the Isaacs–Breakwell theory through the construction of a field of characteristics for a transformed game (see [11]) or using the theory of viscosity solutions and the representation theorem as outlined hereafter. The solution we seek is further characterized by its behavior at infinity. Yet its uniqueness does not derive from the classical results on viscosity solutions. We have to take this into account in our proof of Theorem 3.2 below.

3.2. Simple call or put.

3.2.1. Representation formula. We give here a new theory of (20). We introduce two functions $\check{v}(t, u)$, a representation of the singular manifold, and $\check{w}(t, u)$, the restriction of W to that manifold, handled jointly as

$$
\mathcal{V}(t, u) = \begin{pmatrix} \check{v}(t, u) \\ \check{w}(t, u) \end{pmatrix}.
$$

That pair of functions is entirely defined by a linear PDE that involves the following two matrices (q^- and q^+ are defined hereafter in (22)):

$$
\mathcal{S} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \mathcal{T} = \frac{1}{q^+ - q^-} \begin{pmatrix} \tau^+ q^+ - \tau^- q^- & \tau^+ - \tau^- \\ -(\tau^+ - \tau^-) q^+ q^- & \tau^- q^+ - \tau^+ q^- \end{pmatrix},
$$

and it seems to play a very important role in the overall theory. Namely,

$$
\mathcal{V}_t + \mathcal{T}(\mathcal{V}_u u - \mathcal{S}\mathcal{V}) = 0\,. \tag{21}
$$

The definitions of q^+ and q^-, as well as the terminal conditions at T for (21), depend on the type of option considered. For a simple call or put, and a closure in kind, we have

$$
\begin{aligned}
q^-(t) &= \max\{(1 + c^-) \exp[\tau^-(T - t)] - 1\,,\ C^-\}\,, \\
q^+(t) &= \min\{(1 + c^+) \exp[\tau^+(T - t)] - 1\,,\ C^+\}\,.
\end{aligned}
\tag{22}
$$

Notice that $q^\varepsilon = C^\varepsilon$ for $t \le t_\varepsilon$, with

$$T - t_\varepsilon = \frac{1}{\tau^\varepsilon} \ln \frac{1 + C^\varepsilon}{1 + c^\varepsilon} \ . \tag{23}$$

The terminal conditions are given, for a call, by

$$\mathcal{V}^t(T,u) = \begin{cases} (\,0 \quad 0\,) & \text{if } u < \dfrac{K}{1+c^+}\,, \\ \dfrac{(1+c^+)u - K}{c^+ - c^-}(\,1 \quad -c^-\,) & \text{if } \dfrac{K}{1+c^+} \le u < \dfrac{K}{1+c^-}\,, \\ (\,u \quad u - K\,) & \text{if } u \ge \dfrac{K}{1+c^-} \end{cases} \tag{24}$$

and symmetric formulas for a put. (All combinations call/put, closure in kind/in cash, are detailed in [10]).

We claim the following fact.

THEOREM 3.2. *The function W defined by* (19) *is given by*

$$W(t,u,v) = \check{w}(t,u) + q^\varepsilon(\check{v}(t,u) - v)\,, \quad \varepsilon = \operatorname{sign}(\check{v} - v)\,, \tag{25}$$

where q^ε is given by formula (22) *(for a simple call or put), and $(\check{v}\ \check{w}) = \mathcal{V}^t$ is given by* (21) *and the terminal conditions* (24) *for a call (and symmetrical formulas for a put).*

Proof. The proof is done by checking first that the function (25) is indeed a viscosity solution of (20). That complete check is rather lengthy, as it involves checking the viscosity condition on many manifolds where ∇W may be discontinuous. It is given in the appendix.

Then we notice that the function W thus constructed has (in the Joshua transformed form of [11]) the regularity required[2] by the classical verification theorem in its detailed form of [5] (replacing "lower value" by "upper value"). The viscosity conditions imply the satisfaction of the old corner conditions, as developed in that paper. Thus the verification theorem applies, and that function is indeed the value function of the original differential game. (Let us add that numerical integration supports that claim with great accuracy.) □

It can also be shown that the solution of (21) is nontrivial only in the region where the option may end either in the money or out of the money, i.e., the region

$$\frac{K}{1+c^+} \mathrm{e}^{-\tau^+(T-t)} \le u \le \frac{K}{1+c^-} \mathrm{e}^{-\tau^-(T-t)}\,. \tag{26}$$

Outside of this region, it keeps the form of the terminal condition.

COROLLARY 3.3. *The seller's price of a call is $\check{w}(0,u_0) + q^+(0)\check{v}(0,u_0)$, with $\check{v}$ and $\check{w}$ initialized as in* (24) *(and symmetrically for a put).*

3.2.2. Optimal trading strategy. A hedging strategy is $\xi = 0$ (does no trading) as long as $w \ge W(t,u,v)$. When $w = W(t,u,v)$, it is defined in terms of $\varepsilon = \operatorname{sign}(\check{v}(t,u) - v)$ and is $\xi = 0$ if $t \ge t_\varepsilon$, a positive jump towards $\check{v}$ if $\varepsilon = +1$ and $t < t_+$, and a negative jump towards $\check{v}$ if $\varepsilon = -1$ and $t < t_-$. On the manifold $v = \check{v}$,

[2] It is continuous, which is more than required by the theorem, piecewise C^2 in domains defined by C^2 manifolds on which it has simple gradient discontinuities.

we have shown that there is a control, depending on τ, that keeps $w(t)$ on or "above" the graph of W.

The dependence of the control $\xi(t)$ on the instantaneous rate $\tau(t)$ is undesirable. It is not implementable as such and is not an admissible causal strategy. (Accepting such strategies would create arbitrage opportunities.) However, the convergence theorem of [11], recalled in the next section, provides a practical solution: use the discrete time theory with whatever time step is feasible. It gives an exact (within our model) admissible hedging strategy for a portfolio value close to the optimum one.

4. Discrete trading.

4.1. The multistage game. In the case of discrete trading, we have to investigate the game whose dynamics are given by (5) and (11), and the criterion by (18). This is a completely classical dynamic game. Let $W^h(kh,u,v) = W^h_k(u,v)$ be its Isaacs value function. We immediately obtain its Isaacs equation and the following theorem.

THEOREM 4.1. *The value function W^h is given by the recursion*

$$
\begin{aligned}
&\forall k < \mathcal{K}, \forall (u,v)\,, \quad W^h_k(u,v)\\
&= \min_{\xi} \max_{\tau\in[\tau_h^-,\tau_h^+]} [W^h_{k+1}((1+\tau)u,(1+\tau)(v+\xi)) - \tau(v+\xi) + C^\varepsilon \xi],\\
&\forall (u,v)\,, \quad W^h_{\mathcal{K}}(u,v) = N(u,v)\,.
\end{aligned}
\tag{27}
$$

Finally, the main theorem of [11], and a central result in that theory, is the following convergence theorem. Let $W^h(t,u,v)$ be the function obtained by linear interpolation in time between $W^h_k(u,v)$ and $W^h_{k+1}(u,v)$ with $t \in [kh,(k+1)h]$.

THEOREM 4.2. *The functions W^h converge uniformly on every compact towards the function W (of the continuous trading theory) when the step h goes to zero (in a dyadic fashion: $h = T/(2^n)$, $n \to \infty$).*

Optimal hedging strategy. An important consequence of this theorem is that, even if we are almost in a "continuous trading" situation, the optimal portfolio and trading strategy can be approached by a discrete trading strategy. However, the optimal discrete trading strategy does *not* make use of τ_k to compute ξ_k. Thus it alleviates the problem of the dependence of the optimal strategy on τ in the continuous time theory.

As a matter of fact, one computes a sequence of $\check{v}^h_k(u)$ (see the next paragraph), and let $\varepsilon = \operatorname{sign}(\check{v}^h_k(u_k) - v_k)$. The optimal discrete time hedging strategy is just to do nothing if $t_k \geq t_\varepsilon$ (see (23))—but for most realistic value of the parameters, this is immaterial because $T - t_\varepsilon < h$—and for all other discrete time instants to jump to $v = \check{v}^h_k(u_k)$, which therefore plays the role of an optimum portfolio composition.

4.2. A fast algorithm. We propose here a new fast algorithm to compute the solution of (27), which, in view of Theorem 4.2, also yields a fast algorithm to approximate a solution of the continuous trading problem. It can be viewed as a particular difference scheme for (21).

Define the following recursion:

$$
\begin{aligned}
q^\varepsilon_K &= c^\varepsilon\,,\\
q^\varepsilon_{k+\frac{1}{2}} &= (1+\tau^\varepsilon_h) q^\varepsilon_{k+1} + \tau^\varepsilon_h\,,\\
q^\varepsilon_{k+1} &= \varepsilon \min\{\varepsilon q^\varepsilon_{k+\frac{1}{2}}\,,\, \varepsilon C^\varepsilon\}\,,
\end{aligned}
\tag{28}
$$

and let, for every integer ℓ,

$$Q_\ell^\varepsilon = \begin{pmatrix} q_\ell^\varepsilon & 1 \end{pmatrix} \quad \text{and} \quad \mathcal{V}_\ell^h(u) = \begin{pmatrix} \check{v}_\ell^h(u) \\ \check{w}_\ell^h(u) \end{pmatrix}. \tag{29}$$

Take $\check{v}_{\mathcal{K}}^h(u) = \check{v}(T,u)$, $\check{w}_{\mathcal{K}}^h(u) = \check{w}(T,u)$ as given by (24) for a call (symmetrically for a put) and

$$\mathcal{V}_k^h(u) = \frac{1}{q_{k+\frac{1}{2}}^+ - q_{k+\frac{1}{2}}^-} \begin{pmatrix} 1 & -1 \\ -q_{k+\frac{1}{2}}^- & q_{k+\frac{1}{2}}^+ \end{pmatrix} \begin{pmatrix} Q_{k+1}^+ \mathcal{V}_{k+1}^h((1+\tau^+)u) \\ Q_{k+1}^- \mathcal{V}_{k+1}^h((1+\tau^-)u) \end{pmatrix}. \tag{30}$$

We leave to the reader the tedious, but straightforward, task to check that this is indeed a consistent finite difference scheme for (21).

We claim the following.

THEOREM 4.3. *The solution of* (27) *is given by* (28), (29), (30), *and* (24) *for a call, as*

$$W_k^h(u,v) = \check{w}_k^h(u) + q_k^\varepsilon(\check{v}_k^h(u) - v) = Q_k^\varepsilon \mathcal{V}_k^h(u) - q_k^\varepsilon v, \quad \varepsilon = \operatorname{sign}(\check{v}_k^h(u) - v).$$

The proof is given in appendix, together with that of the equivalent "continuous" theorem, Theorem 3.2.

COROLLARY 4.4. *The seller's price of a call is* $Q_0^+ \mathcal{V}_0^h(u_0)$ *with* $\mathcal{V}_{\mathcal{K}}^h$ *initialized as in* (24) *(a symmetric form holds for a put.)*

The important fact, of course, is that we now have two sequences of functions of one variable to compute, $\{\check{v}_k^h(\cdot)\}$ and $\{\check{w}_k^h(\cdot)\}$, instead of one sequence of functions of two variables $\{W_k^h(\cdot,\cdot)\}$. This is a major saving in computer time and memory. We have typically discretized u and v with 300 to 3000 points each. Therefore the saving is in a ratio of 1:100 to 1:1000. This algorithm has been programmed.[3] The results were indeed identical to those obtained with the straightforward discretization of the Isaacs equation but much faster and with the above reduction in memory space.

5. Discussion. We wish to discuss here the strengths and weaknesses of this new theory as compared to the classical Black and Scholes theory [12] and related work. While we have no pretense to a global superiority of the new theory over the classical one, we wish to show that it might for some purposes be a possible alternative, given more experience and validation work.

Let us first notice that for a put or call option, and for small transaction costs, the general appearance of our seller's price as a function of $u(0)$ is very similar to that of the classical Black and Scholes theory. (Some curves are published in [7].) This can be understood as a consequence of the convergence theorem, Theorem 4.2, and of the fact that for such a convex terminal payoff, our discrete transaction price with zero transaction costs coincides with that of Cox, Ross, and Rubinstein, which itself is close to a Black and Scholes price curve for small time steps. The difference between the two theories is likely to show up more for digital options, which we are currently investigating. Indeed, our PDE is first order, and as a consequence the discontinuity in the boundary value—the terminal payment—propagates backward in time, while the second order Black and Scholes PDE would yield a smooth solution for any time less than exercise time. (But in the absence of transaction costs and in continuous

[3]This was done by Laurent Fischer and Loïc Maitrehut, students at ESSI whose contribution we acknowledge.

time, the optimal hedging portfolio is bound to go to infinity in the neighborhood of the discontinuity. Our theory avoids that difficulty.)

To go further in a comparison, one must distinguish what the mathematics strictly say, and what is practically possible beyond the mathematically grounded facts, thanks to some robustness in the theory.

5.1. Strict mathematical properties. Clearly, a major weakness of our model is that it rules out from the start very fast price variations in the market. If we try to take τ^- and τ^+ so large that the model is (essentially) always satisfied, then we will end up with too large a price. This is a classical fact that because our market model is incomplete we have to resort to superreplication, potentially ending up with an unrealistically large price. Hence we shall get a reasonable premium only by tolerating some violations of the market model.

Now, the Black and Scholes theory has its own theoretical shortcomings. On the one hand, it fundamentally assumes that trading is continuous and with no delay. It is impossible, within Samuelson's model, to achieve hedging if the trading is not done continuously, except with the trivial—and too expensive—portfolio $v = u$. On the other hand, within Samuelson's model, *there is no nontrivial hedging portfolio for option pricing with transaction costs* [25]. The first problem arises from the fact that Samuelson's model may display arbitrarily large variations in any finite time, the second one from the closely related fact that it has almost surely trajectories of unbounded total variation. (This in itself could be considered as not very realistic.)

Let us concentrate on the continuous versus discrete trading issue. Real trading has to be discrete, forcing a discrepancy between real trading and the Black and Scholes theory. This is of little consequence as long as the price of the underlying stock does not change too quickly. But when it does, that discrepancy becomes potentially fatal.

Hence both theories fail under the same circumstances: when there are unusually fast variations of the price of the underlying stock on the market. In our theory the market model is violated; in that of Black and Scholes, it is the portfolio model which fails.

Mathematically, it is impossible to reconcile a model that allows for arbitrarily large stock price variations within one time step with discrete time hedging. Hence a mathematical theory has to give up one of the two features. The Black and Scholes theory gives up the (theoretical) ability to do discrete trading. We wanted to develop a theory of discrete trading, the discrete time market model being consistent with (i.e., the time sampling of) a continuous time underlying market model, kept fixed as the time step is decreased. Thus we had to give up a model that would allow for arbitrarily large price variations in one step of time. Yet we wanted a model less idealized than that of Cox, Ross, and Rubinstein—and not dependent on the time step. Thus we were forced to invent the interval model, at the price of giving up market completeness. And it is no surprise that other authors came up with the same model.

Turning now to the transaction costs issue, they are a natural ingredient of our theory. Indeed we were forced to introduce them to avoid the naive "stop loss" strategy, which is the only solution of the hedging problem in the absence of transaction costs. While we view our ability to deal with transaction costs, even large ones, as a strength of the new theory, the fact that in their absence the only solution is the naive one may be viewed as a limitation of any model with bounded variation trajectories. In contrast, it takes the difficult theory of diffusion limits to deal approximately with

small transaction costs in the Black and Scholes theory.

Finally, once transaction costs are introduced, it is only natural to assume that there are closing costs as well. The introduction of those costs creates a difference between closure in kind or closure in cash. Yet, if this difference is deemed annoying, closing costs may be removed, just by setting c^- and c^+ to zero, this time with no detrimental effect on the theory.

5.2. Robustness. Now, it is well known that in practice, the Black and Scholes theory, and the derived hedging strategy, can be used with discrete transactions, provided that they are frequent enough. Also, small transaction costs can be tolerated, and furthermore, an approximation of perfect hedging can be developed with the concept of diffusion limits (see [1]), although this does not contradict the claim of [25]. These are features of robustness of that theory to small violations of the hypotheses used to derive it.

The new theory also seems to exhibit a fair degree of robustness, as displayed by the following numerical experiments. In all these experiments, the exercise price K is taken as the unit of monetary value.

5.2.1. Theoretical average payments. In a first series of experiments, we assumed that the ratio $\tau = (u_{k+1} - u_k)/u_k$ obeys a probability law with compact support $[\sigma^-, \sigma^+]$ but that this compact is strictly larger than the range $[\tau^-, \tau^+]$ used to compute the hedging strategy of our theory. And we computed the expected overall cost to a trader using that hedging strategy under this hypothesis. This was done with the help of the discrete Kolmogorov equation which gives us an "exact" (up to the precision of the numerical computation) expected value, in contrast with a Monte Carlo simulation.[4]

For the law of τ, we used either a uniform law over $[\sigma^-, \sigma^+]$, a rather pessimistic case, or a "hat" law, with a piecewise affine density, null at the end points and maximal at $\tau = 0$. And we used $\sigma^\varepsilon = (1 + \Delta)\tau^\varepsilon$, with $\Delta > 0$. In all the simulations, we used the formulas for a closure in cash (deemed more realistic) and the following set of parameters:

τ^-	τ^+	C^-	C^+	c^-	c^+	$\mathcal{K}$
-5%	3%	$-.7\%$	$.7\%$	$-.35\%$	$.35\%$	44

In Figures 1 and 2, we have plotted the premium $P(u_0)$ computed with the hypothesis $\tau \in [\tau^-, \tau^+]$ and, on the same graphics, the total expected expense $Q_\Delta(u_0)$ computed with the same data as P, except that τ is distributed over $[\sigma^-, \sigma^+]$, for various values of Δ. Figure 1 corresponds to a uniformly distributed τ and Figure 2 to a "hat" law. The curve $Q_\Delta(\cdot)$ with $\Delta = .85$ in the first case ($\Delta = 162$ in the second) can hardly be distinguished from the curve $P(\cdot)$.

In Figure 3, we plot the difference $P(K) - Q_\Delta$ as a function of Δ for the two probability laws of σ: uniform or "hat." The conclusion is that for a "spillover" up to 85% in the case of the uniform law, and 162% in the case of the "hat" law, the expected result is still positive for the trader.

5.2.2. Simulations with price series. In a second set of experiments, we started from stock price time series. We computed sets of intervals $[\tau^-, \tau^+]$ aimed to contain a variable fraction, say p, of the realizations of τ on this sample and to be centered, i.e., chosen such that the number of occurrences of τ less than τ^- is the

[4]These computations were done by M'hamed Oumouhou during an internship.

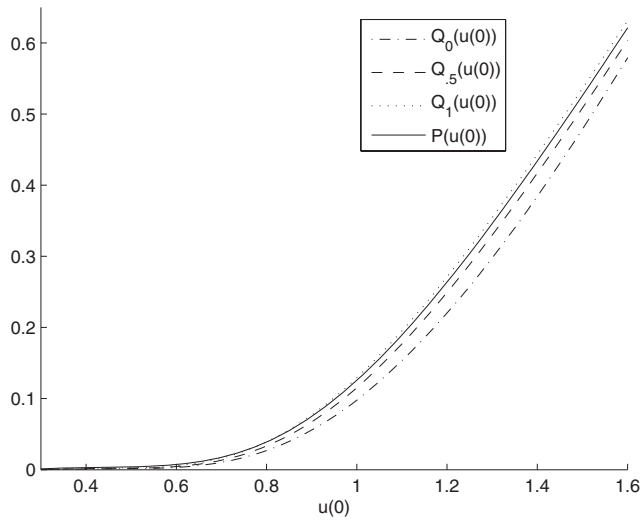


FIG. 1. *Total expense compared with computed premium as a function of u_0 for various values of the spillover ratio Δ. Case τ uniformly distributed.*

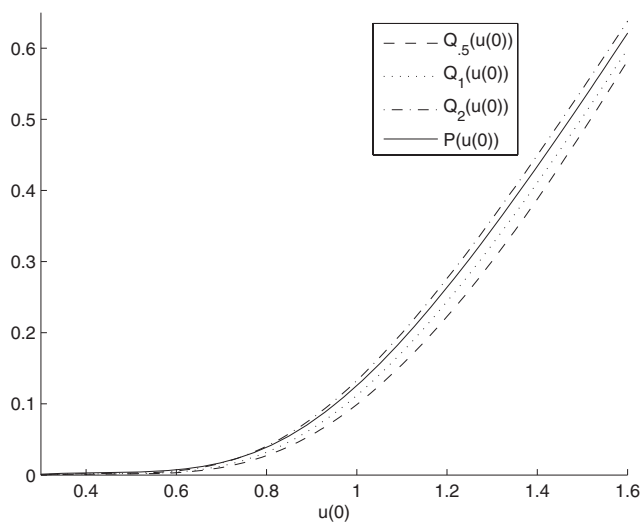


FIG. 2. *Total expense compared with computed premium as a function of u_0 for various values of the spillover ratio Δ. Case of the "hat" law.*

same as the number of occurrences above τ^+. With each of these intervals $[\tau^-, \tau^+]$, we computed the premium and the hedging strategy advocated by our theory. Then we simulated the effect of that hedging strategy if it had been used by the trader to hedge an option and computed the total cost to the trader. When this cost ends up higher than the premium, it means that the trader using this interval $[\tau^-, \tau^+]$ and acting according to our theory would have lost money.

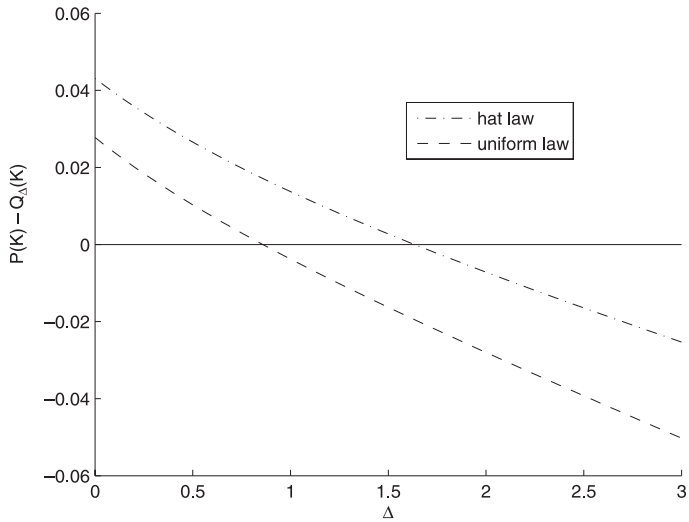


FIG. 3. *The difference* $P(K) - Q_\Delta(K)$ *as a function of* Δ.

As a complement to that test, we compute an empirical (a posteriori) historical volatility of the underlying price, compute the Black and Scholes theoretical premium for that volatility, and see for which $[\tau^-, \tau^+]$, if any, our theory gives approximately the same premium. Remember, though, that our theory always assumes non zero transaction costs, contrary to that of Black and Scholes. Therefore to make a more significant comparison, in the incurred cost of the simulations, we separated the transaction costs from those due to price variations.

We ran that experiment both with real stock price series and for simulated "log-normal" price series, a situation ideally favorable to the Black and Scholes theory since it is the hypothesis it is based upon.

Figures 4–7 show the result of four of these experiments, typical of the many we performed. We plot our premium and the cost incurred by the trader excluding transaction costs against the fraction p of points lying outside of the interval $[\tau^-, \tau^+]$ used. We also show on the same graph the Black and Scholes premium for $K = u_0$. The conclusion is that, for transaction costs less than 1%, for many simulations, taking an interval $[\tau^-, \tau^+]$ that excludes up to 30% of the observed τ_k's yields a premium larger than or equal to the total cost, but excluding transaction costs, the coincidence happens at a higher exclusion ratio and close to the Black and Scholes premium. Some simulations, mainly at low volatilities, show a Black and Scholes premium significantly less than ours. More investigations are needed to completely understand these cases.

5.3. Conclusion. A careful analysis shows that it is rather natural to resort to such "interval models," and this explains why several authors developed that idea independently. To this remark, we add that for the strict problem of hedging a contingent claim, the robust control approach, also used by several of these authors, whether explicitly or implicitly, allows us to proceed without endowing the set of admissible stock price trajectories with a probability law. This is so since what is sought is a hedge for *every* possible trajectory. (And this remark carries over to the Black and Scholes theory if one carefully picks the set of admissible trajectories, as

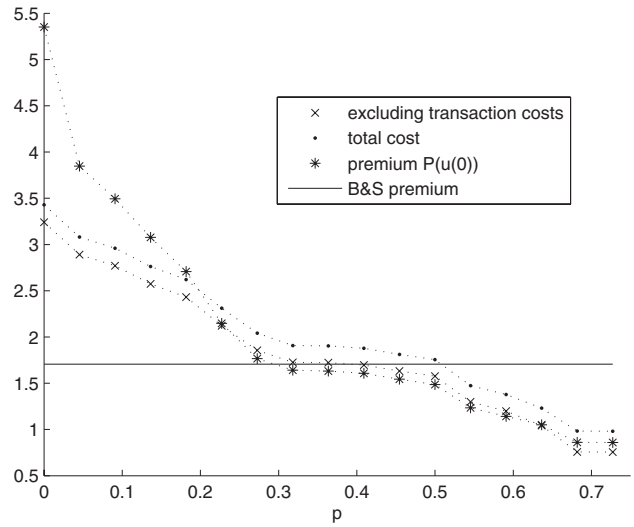


FIG. 4. *Premium and costs incurred as a function of the fraction p; AirFrance series, August* 18, 1998 *to October,* 20, 1998 *(volatility* $\simeq .03$*).*

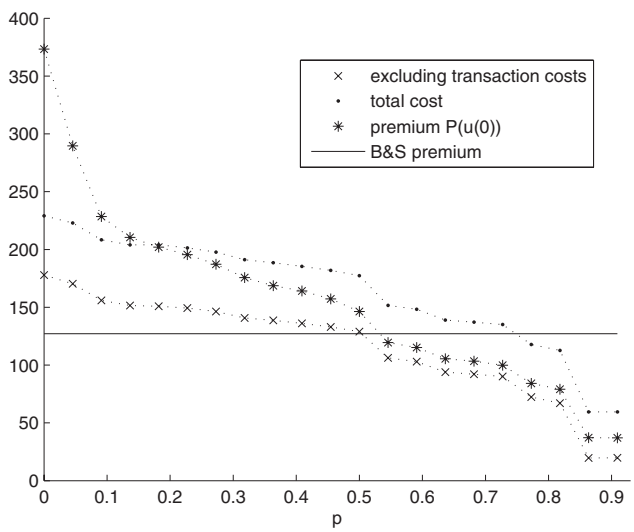


FIG. 5. *Premium and total cost as a function of p; CAC*40 *series, March* 6, 1998 *to May* 8, 1998 (volatility $\simeq .01$).

shown in [7].)

The resulting theory exhibits a strong mathematical structure, that can be exploited to get semiexplicit formulas via a fast algorithm transaction costs, whether in discrete trading or continuous trading. The latter is the limit of the former with vanishing step size; this, we stress, keeps *the same* continuous time model for the underlying price trajectories. Thus the discrete trading strategy, which is very simple

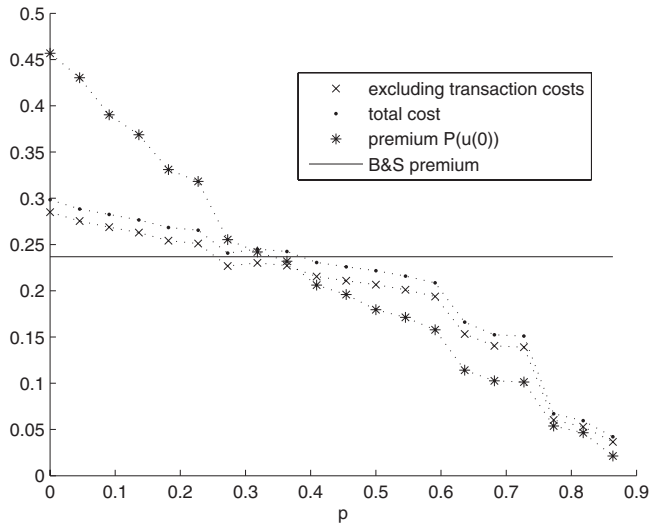


FIG. 6. *Premium and total cost as a function of p; log-normal series,* $\sigma = .1$.

to implement, is a good approximation of the theoretical continuous strategy.

The seller's prices computed qualitatively and quantitatively resemble the Black and Scholes prices, although the presence of transaction costs makes them larger.

The hedging strategies derived from the theory exhibit a fair degree of robustness to violations of the market model. Our simulations show that picking an interval containing 70% of the actual relative price variations often leads to a premium comparable to the Black and Scholes premium and an effective hedging. Yet, these simulations were ran using a posteriori "statistical" information on the price series. If an approach based upon the interval model is to be routinely used, one must also develop new statistical tools to inform it.

This robustness analysis is carried out in a *normative* perspective, to show that this theory can be used on the actual market as a *decision aid*. We have at this point no claim to a *positive* theory that would *explain* premiums actually used by the operators, less so that the current market is overwhelmingly dominated by the Black and Scholes theory, which is, in that respect, self-enforcing.

We feel that the results provided so far point to the conclusion that this theory might be useful in some situations, for instance, when transaction costs are too high to be neglected or when time discretization is critical.

In any extent, this is by all means a young theory. There remains a large work of sensitivity analysis, simulations, and validation to perform. We hope to have proved that it is worthwhile pursuing.

Appendix A. Proof of Theorems 3.2 and 4.3.

A.1. Theorem 4.3. We make the proof in the case of a call. The argument for a put is completely similar.

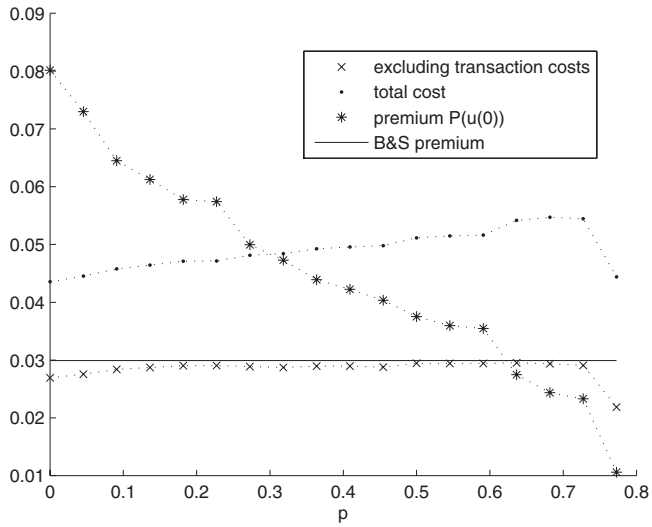


FIG. 7. *Premium and total cost as a function of p; log-normal series, $\sigma = .01$.*

It is useful to notice an alternate, "two-stage" form of the recursion (27):

$$W^h_{k+\frac{1}{2}}(u,v) = \max_{\tau\in[\tau_h^-,\tau_h^+]} [W^h_{k+1}((1+\tau)u,(1+\tau)v) - \tau v]\,, \tag{31}$$

$$W^h_k(u,v) = \min_{\xi} [W^h_{k+\frac{1}{2}}(u,v+\xi) + C^\varepsilon \xi]\,. \tag{32}$$

This form shows that the convexity of N is preserved, and the W^h_k are convex.[5]

Note that the formula of the theorem is correct at final time, $k = K$. Assume it is correct at time $k+1$. Consider the step (31). Because W^h_{k+1} is convex, the maximum is reached either at τ_h^- or at τ_h^+. For each u, the function to be maximized in τ is piecewise affine in v, and its graph as a function of v can be represented as two wedges with one branch sloping downwards (see Figure 8), one for each τ^ε. These can be written as

$$W^+_{k+\frac{1}{2}} := \check{w}^+_{k+\frac{1}{2}} + q^\varepsilon(\check{v}^+_{k+\frac{1}{2}} - v)\,,$$
$$W^-_{k+\frac{1}{2}} := \check{w}^-_{k+\frac{1}{2}} + q^\varepsilon(\check{v}^-_{k+\frac{1}{2}} - v)\,,$$

where $\check{v}^+_{k+\frac{1}{2}}$, $\check{v}^-_{k+\frac{1}{2}}$, $\check{w}^+_{k+\frac{1}{2}}$, and $\check{w}^-_{k+\frac{1}{2}}$ are easily written in terms of $\check{v}^h_{k+1}$ and $\check{w}^h_{k+1}$ evaluated at $(1+\tau^+)u$ and $(1+\tau^-)u$.

As a result, $\check{v}_k$ is obtained as the abscissa of the intersection of the two wedges in this graph. (In the figure, $\check{v}^\varepsilon$ stands for $\check{v}^\varepsilon_{k+\frac{1}{2}} = \check{v}_{k+1}((1+\tau^\varepsilon)u)/(1+\tau^\varepsilon)$, $\varepsilon = \pm$.)

Now, we claim the following fact.

PROPOSITION A.1. *We have for all (k,u)*

$$\frac{1}{1+\tau_h^-}\check{v}^h_{k+1}((1+\tau_h^-)u) \le \check{v}^h_k(u) \le \frac{1}{1+\tau_h^+}\check{v}^h_{k+1}((1+\tau_h^+)u)\,.$$

[5]Hence, from the convergence theorem, so is $W(t,\cdot,\cdot)$.

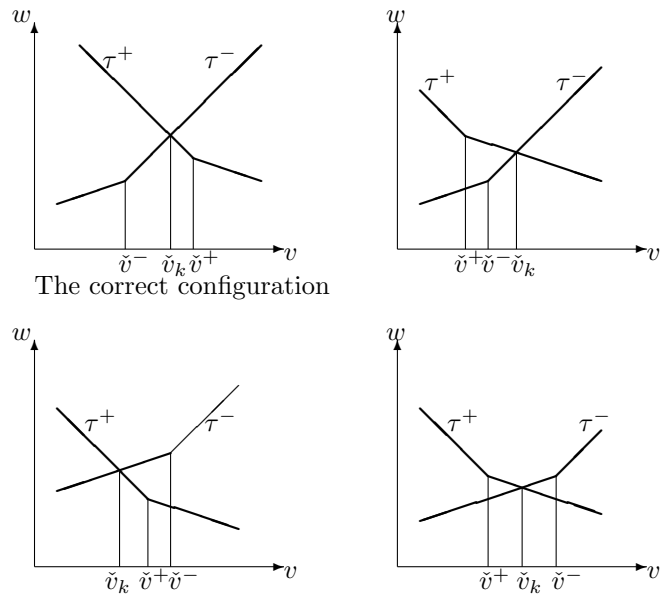


FIG. 8. *Four possible configurations.*

Proof. Assume that the left-hand inequality does not hold. Then a *decrease* of the price of the underlying stock (by a factor $1+\tau^-$) would result in the cheapest hedging portfolio having a *larger* content (in number of shares) in this stock than the previous one, a contradiction for a call (and for any option with an increasing payment function). □

Only the first possibility in the figure is consistent with the proposition, and it results in the max being again a simple wedge. Its minimum is achieved at the intersection of the right branch of the graph with τ^- and the left branch of the graph with τ^+. This gives the formulas (30). (One needs to notice that the q^ε_k as given by (28) coincide with $q^\varepsilon(kh)$ as defined by (22).)

Their remains to carry out (32). It is an inf convolution with a wedge function acting on the v variable only. It leaves unchanged branches with a slope between $-C^+$ and $-C^-$ (and the min is then reached at $\xi = 0$) and replaces steeper slopes by these two limit ones, hence the min or max operations in (22).

A.2. Theorem 3.2. We have to show that formula (25), where $\varepsilon = \operatorname{sign}(\check{v} - v)$, q^ε is given by (22), and $\mathcal{V}(t,u)$ is the solution of the PDE (21), is the (regular) viscosity solution of (20). Let

$$
\begin{aligned}
&H(t,u,v,\mathrm{D}W,\tau) := W_t + \tau[W_u u + (W_v - 1)v]\,,\\
&\bar{H}(t,u,v,\mathrm{D}W) := \max\nolimits_{\tau\in[\tau^-,\tau^+]} H(t,u,v,\mathrm{D}W,\tau)\,.
\end{aligned}
\tag{33}
$$

Then (20) reads

$$
\min\{\bar{H}(t,u,v,\mathrm{D}W)\,, W_v + C^+\,, -W_v - C^-\} = 0\,. \tag{34}
$$

Define $Q^\varepsilon = (q^\varepsilon \;\; 1)$ and $\mathbb{1} = \left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)$.

A.2.1. Preliminary propositions. The proof of the theorem is by checking that formula (25) indeed provides a (sufficiently regular) viscosity solution of (34). We

therefore review the manifolds where our formula allows for a gradient discontinuity of $\mathcal{V}$.

We first stress a simple fact, as a consequence of the definition (22).

PROPOSITION A.2. *For $\varepsilon = 1$ and $\varepsilon = -1$,*

- *if $t \le t_\varepsilon$, $q^\varepsilon = C^\varepsilon$;*
- *if $t > t_\varepsilon$, $q^\varepsilon \in [c^\varepsilon, C^\varepsilon]$ and*

$$\dot{q^\varepsilon} = -\tau^\varepsilon (q^\varepsilon + 1). \tag{35}$$

We also claim the following important fact.

PROPOSITION A.3. *For all $(t,u,v) \in [0,T] \times \mathbb{R}^+ \times \mathbb{R}$,*

$$Q^\varepsilon \mathcal{V}_t \le 0\,, \quad \textit{or equivalently} \quad \operatorname{sign}[Q^\varepsilon(\mathcal{V}_u u - \mathbb{1}\check{v})] = \varepsilon. \tag{36}$$

Proof. The equivalence of the two forms of the claim comes from the fundamental PDE (21) and the the fact that

$$Q^\varepsilon \mathcal{T} = \tau^\varepsilon Q^\varepsilon\,. \tag{37}$$

Simple geometry shows that Proposition A.1 implies

$$\check{w}^+_{k+\frac{1}{2}} + q^-_{k+\frac{1}{2}}(\check{v}^+_{k+\frac{1}{2}} - \check{v}^-_{k+\frac{1}{2}}) \le \check{w}^-_{k+\frac{1}{2}} \le \check{w}^+_{k+\frac{1}{2}} + q^+_{k+\frac{1}{2}}(\check{v}^+_{k+\frac{1}{2}} - \check{v}^-_{k+\frac{1}{2}})\,. \tag{38}$$

In the limit as $h \to 0$, $W^h \to W$, but also $\mathcal{V}^h \to \mathcal{V}$ that satisfies the PDE (21). And since the defining recursion (30) is a consistent discretization scheme for (21), the differentials converge, and, as a tedious but simple calculation shows, (38) converges to (36). □

For a given (t,u,v), let $\varepsilon = \operatorname{sign}(\check{v}(t,u) - v)$. As a consequence of (36), and keeping in mind that $q^\varepsilon + 1 > 0$,

$$\operatorname{sign}[(\check{w}_u + q^\varepsilon \check{v}_u)u - (q^\varepsilon + 1)v] = \varepsilon\,, \tag{39}$$

so that the max in $\bar{H}$ is reached at $\tau = \tau^\varepsilon$.

A.2.2. Differentiable case. We first investigate regions of (t,u,v) space where our formula (25) gives a differentiable function. The partials of (25) are given by

$$W_t = Q^\varepsilon \mathcal{V}_t + \dot{q^\varepsilon}(\check{v} - v)\,, \quad W_u = Q^\varepsilon \mathcal{V}_u\,, \quad W_v = -q^\varepsilon\,,$$

where $Q^\varepsilon \mathcal{V}_t = -\tau^\varepsilon Q^\varepsilon V_u u + \tau^\varepsilon(1+q^\varepsilon)\check{v}$ using (21) with (37). Replacing these partials in (33) with property (39), it follows that

$$\bar{H}(t,u,v,\mathrm{D}W) = \left(\dot{q^\varepsilon} + \tau^\varepsilon(1+q^\varepsilon)\right)(\check{v} - v)\,.$$

If $t > t_\varepsilon$, Proposition A.2 leads to $\bar{H} = 0$, while the other two terms in (34) are positive because of Proposition A.2.

If $t < t_\varepsilon$, $\bar{H}(t,u,\check{v},\mathrm{D}W) = 0$ and

$$\bar{H}(t,u,v,\mathrm{D}W) = \tau^\varepsilon(1+C^\varepsilon)(\check{v}(t,u) - v) \ge 0$$

since $\tau^+ > 0$ and $\tau^- < 0$, so that $\varepsilon = \operatorname{sign}(\check{v} - v)$ is also the sign of τ^ε and $1 + C^\varepsilon > 0$ by hypothesis. Moreover, according to Proposition A.2, one of the other two terms in (34) is zero and the other one positive.

A.2.3. The singular manifold $v = \check{v}$. On the manifold $v = \check{v}(t,u)$, formula (25) for W is nondifferentiable. It has a nonvoid subdifferential, obtained by replacing q^ε by $q = \lambda q^+ + (1-\lambda)q^-$ in the formulas for the partial derivatives in either of the regions $\varepsilon = -1$ or $\varepsilon = 1$. This is so because these partials are affine in q^ε.

Now, for each ϵ, the maximum in τ in $\bar{H}$, reached at τ^ε, is 0. Therefore, for $\tau^{-\varepsilon}$, $H \leq 0$. Hence, as an affine function of q (for fixed τ) which ranges from 0 to a negative number, H is nonpositive for all possible q's. Hence so is its max in τ, $\bar{H}$. The other two terms in (34) are trivially nonpositive for all λ. Therefore the minimum of the three terms is nonpositive, and this is the viscosity condition.

A.2.4. Boundaries of the nontrivial region. We adapt and complete here a result of [11] (Proposition 8.1).

PROPOSITION A.4. *Along the manifolds $u = K\exp(-\tau^\varepsilon(T-t))$, the gradients of $\mathcal{V}$ may be discontinuous, with discontinuities $\delta\mathcal{V}_t$ and $\delta\mathcal{V}_u$ satisfying $\delta\mathcal{V}_t = -\tau^\varepsilon\delta\mathcal{V}_u$ and $(q^{-\varepsilon}\ 1)\delta\mathcal{V}_u = 0$.*

Proof. Let $(\hat{t},\hat{u})$ be a tangent to a manifold bearing a gradient discontinuity of a continuous solution of (21). Let $(\delta\mathcal{V}_t, \delta\mathcal{V}_u)$ be that discontinuity. Continuity of $\mathcal{V}$ imposes that

$$\hat{t}\delta\mathcal{V}_t + \hat{u}\delta\mathcal{V}_u = 0\,.$$

Now, because $\mathcal{V}$ satisfies (21) in both open half spaces on each side of the manifold, it follows that

$$\delta\mathcal{V}_t + \mathcal{T}\delta\mathcal{V}_u u = 0\,.$$

Hence, combining the two equations, we get

$$\left(\frac{1}{u}\frac{\hat{u}}{\hat{t}}I - \mathcal{T}\right)\delta\mathcal{V}_u = 0\,,$$

which is possible with a nonzero $\delta\mathcal{V}_u$ if and only if $\hat{u}/u\hat{t}$ is an eigenvalue of $\mathcal{T}$, hence either τ^- or τ^+. Therefore, the gradient discontinuity has to be born by a curve of the form $u = u_T\exp(-\tau^\varepsilon(T-t))$, and in this formula, we have to choose $u_T = K$ to embed the gradient discontinuity of $N(u,v)$ in that curve.

Then $\delta\mathcal{V}_u$ has to be the eigenvector of $\mathcal{T}$ associated with the corresponding eigenvalue, i.e., $(1\ -q^{-\varepsilon})$, as is easily checked ∎

Assume we are hedging a call, thus with $0 \leq v \leq u$. On the left boundary $u = K\exp(-\tau^+(T-t))$, we have $\varepsilon = -1$, and the discontinuities of the gradient of our function W are given by $(\delta W_t, \delta W_u, \delta W_v) = (Q^-\delta\mathcal{V}_t, Q^-\delta\mathcal{V}_u, 0) = 0$. Therefore the function (25) is smooth. A similar argument applies along the boundary $u = K\exp(-\tau^-(T-t))$. And symmetric arguments hold for a put.

A.2.5. Boundaries of the jump regions. Finally, one has to check the two manifolds $t = t_\varepsilon$, where $\mathcal{V}_t$ is discontinuous, because q_t^ε is. It can be seen that the superdifferential of W is nonempty there, and is made of all the vectors $(Q^\varepsilon\mathcal{V}_t + \delta, W_u, -C^\varepsilon)$ with $\delta \in [-\tau^\varepsilon(1+C^\varepsilon)(\check{v}-v), 0]$, and notice that $-\tau^\varepsilon(1+C^\varepsilon)(\check{v}-v) < 0$ (which shows that it is the superdifferential which is nonempty). As a consequence, the viscosity condition reads

$$\forall\delta \in [-\tau^\varepsilon(1+C^\varepsilon)(\check{v}-v), 0]\,, \quad Q^\varepsilon\mathcal{V}_t + \delta + \tau^\varepsilon[Q^\varepsilon\mathcal{V}_u u - (C^\varepsilon+1)v] \geq 0\,.$$

However, we have already seen that this quantity is zero for δ at the lower end of the interval. And thus the inequality does hold, ending the proof.

Remark A.1.

1. It can be noted that this direct check confirms only the fact that the field of extremal trajectories constructed in [11] satisfies the relevant corner conditions as developed in the Isaacs–Breakwell theory. Yet, here we have an explicit formula that guarantees that there are no other singular surfaces in the state space. This is difficult to ascertain with the previous theory.
2. We have shown in [9] further relationships between (20) on the one hand and (21) and (25) on the other hand.

Acknowledgments. We wish to acknowledge the help of many intern students from École Polytech'Nice-Sophia and from the department of mathematics, both at the University of Nice-Sophia Antipolis, and notably M'hamed Oumouhou, whose many numerical computations helped build confidence in the theory. We also wish to thank three anonymous reviewers for providing much help in improving this paper and one of them for pointing out the book by Shafer and Vovk.

REFERENCES

[1] H. Ahn, M. Dayal, E. Grannan, and G. Swindle, *Option replication with transaction costs: general diffusion limits*, Ann. Appl. Probab., 8 (1998), pp. 676–707.

[2] J.-P. Aubin, D. Pujal, and P. Saint-Pierre, *Dynamic management of portfolios with transaction costs under tychastic uncertainty*, in Numerical Methods in Finance, H. Ben Hameur and M. Breton, eds., Springer, New York, 2005, pp. 59–89.

[3] G. Barles and H. M. Soner, *Option pricing with transaction costs and a nonlinear Black and Scholes equation*, Finance Stoch., 2 (1998), pp. 369–397.

[4] A. Bensoussan and J.-L. Lions, *Contrôle impulsionnel et inéquations quasi variationnelles*, Dunod, Paris, 1982.

[5] P. Bernhard, *Singular surfaces in differential games, an introduction*, in Differential Games and Applications, Lecture Notes in Control Inform. Sci. 3, P. Haggedorn, G.-J. Olsder, and H. Knoboloch, eds., Springer, Berlin, 1977, pp. 1–33.

[6] P. Bernhard, *Une approche déterministe de l'évaluation d'options*, in Optimal Control and Partial Differential Equations, J.-L. Menaldi, E. Rofman, and A. Sulem, eds., IOS Press, Amsterdam, 2001, pp. 511–520.

[7] P. Bernhard, *Robust control approach to option pricing*, in Applications of Robust Decision Theory and Ambiguity in Finance, M. Salmon, ed., City University Press, London, 2003.

[8] P. Bernhard, *A Robust Control Approach to Option Pricing Including Transaction Costs*, Ann. Internat. Soc. Dynam. Games 7, A. Nowak ed., Birkhäuser, Boston, 2005, pp. 391–416.

[9] P. Bernhard, *On the singularities of an impulsive differential game arising in mathematical finance*, Int. Game Theory Rev., 8 (2005), pp. 219–229.

[10] P. Bernhard, *The robust control approach to option pricing and interval models: An overview*, in Numerical Methods in Finance, M. Breton and H. Ben-Ameur, eds., Springer, New York, 2005, pp. 91–108.

[11] P. Bernhard, N. El Farouq, and S. Thiery, *An impulsive differential game arising in finance with interesting singularities*, in Advances in Dynamic Games, Ann. Internat. Soc. Dynam. Games 8, A. Haurie, S. Muto, L. A. Petrosjan, and T. E. S. Raghavan, eds., Birkhäuser, Boston, 2006, pp. 335–363.

[12] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Political Economy, 81 (1973), pp. 637–659.

[13] J. C. Cox, S. A. Ross, and M. Rubinstein, *Option pricing: A simplified approach*, J. Financial Economics, 7 (1979), pp. 229–263.

[14] B. Dupire, *Résultats de dominance*, Seminar at University of Nice, Sophia Antipolis Cedex, France, 2003.

[15] H. Föllmer, *Calcul d'Ito sans Probabilités*, in Séminaire de probabilités XV, Lecture Notes in Math. 850, J. Azéma and M. Yor, eds., Springer, Berlin, 1981, pp. 143–151.

[16] V. Kolokol'tsov, *Nonexpansive maps and option pricing theory*, Kybernetica, 34 (1998), pp. 713–724.

[17] V. Kolokol'tsov, *Idempotent structures in optimization*, J. Math. Sci., 104 (2001), pp. 847–880.
[18] W. M. McEneaney, *A robust control framework for option pricing*, Math. Oper. Res., 22 (1997), pp. 22–221.
[19] G-J. Olsder, *Control-theoretic thoughts on option pricing*, Int. Game Theory Rev., 2 (2000), pp. 209–228.
[20] D. Pujal, *Évaluation et gestion dynamiques de portefeuilles*, Thesis, Paris Dauphine University, Paris Cedex, France, 2000.
[21] B. Roorda, J. Engwerda, and H. Schumacher, *Performance of hedging strategies in interval models*, Kybernetica, 41 (2005), pp. 575–592.
[22] B. Roorda, J. Engwerda, and H. Schumacher, *Coherent acceptability measures in multiperiod models*, Math. Finance, 15 (2005), pp. 589–612.
[23] P. Saint-Pierre, *Viable capture basin for studying differential and hybrid games*, Int. Game Theory Rev., 6 (2004), pp. 109–136.
[24] G. Shafer and V. Vovk, *Probability and Finance: It's Only a Game*, Wiley, New York, 2001.
[25] H. M. Soner, S. E. Shreve, and J. Cvitanic, *There is no non-trivial hedging portfolio for option pricing with transaction costs*, Ann. Appl. Probab., 5 (1995), pp. 327–355.